

CUNY 2021
34th Annual CUNY Conference on Human Sentence Processing
March 4-6, 2021
University of Pennsylvania

Program and Abstracts

The 34th CUNY Conference on Human Sentence Processing will take place Thursday, March 4 – Saturday, March 6, 2021. It will be hosted by the University of Pennsylvania.

The theme of the Special Session is Language Acquisition and Language Processing: Finding New Connections, funded by the National Science Foundation.

Invited Speakers

Anne Christophe, École Normale Supérieure

Cynthia L. Fisher, University of Illinois

Jeffery Lidz, University of Maryland

Christopher Manning, Stanford University

Elissa Newport, Georgetown University

Linda Smith, Indiana University

Questions? Contact: cuny2021@easychair.org

The CUNY 2021 Organizing Committee

John Trueswell (Chair), Delphine Dahan, Anna Papafragou, Gareth Roberts, Kathryn Schuler, Florian Schwarz, and Charles Yang.

Table of Contents

Main Session Talks	Page
Thursday March 4, 2021	3
Friday March 5, 2021	22
Saturday March 6, 2021	42
Parallel Session Talks	
Thursday Afternoon, March 4, 2021	61
Friday Morning, March 5, 2021	213
Friday Evening, March 5, 2021page.....	360
Saturday Afternoon, March 6, 2021.....	441

34th Annual CUNY Conference on Human Sentence Processing

Thursday March 4, 2021

Session	Time	Type	Title	Authors
1	8:45	Talk	Opening Remarks	John Trueswell
1	9:00	Talk	Evaluating "each"- (but not "every"-) sentences encourages encoding individual properties	Tyler Knowlton, Justin Halberda, Paul Pietroski and Jeffrey Lidz
1	9:30	Talk	Are logical representations quantifier-specific? Evidence from priming for a non-quantifier-specific representation of scope	Mieke Sarah Slim, Peter Lauwers and Robert J. Hartsuiker
1	10:00	Talk	What primes what - An experimental framework to explore alternatives for Scalar Implicatures	Paul Marty, Jacopo Romoli, Yasutada Sudo, Richard Breheny and Joe Cowan
2	10:30	Break		
3	11:00	Invited Talk	Developmental plasticity and lateralization of function for language	Elissa Newport
3	11:45	Talk	Acquiring recursive structures through distributional learning	Daoxin Li and Kathryn Schuler
4	12:15	Break		
5	12:30	Parallel Session	Link to Thursday Parallel Session	
6	14:30	Break		
7	15:00	Invited Talk	Second-year syntax: Discovering Dependencies	Jeffrey Lidz
7	15:45	Talk	Evidence of accurate logical reasoning in online sentence comprehension	Maksymillian Dabkowski and Roman Feiman
8	16:15	Break		
9	16:45	Talk	Syntactic and semantic parallelism guides filler-gap processing in coordination	Stephanie Rich and Matt Wagers
9	17:15	Talk	The laboratory discovered: Place-for-institution metonyms appearing in subject position are processed as agents	Matthew Lowder, Adrian Zhou and Peter Gordon
9	17:45	Talk	Social and communicative biases jointly influence grammatical choices in learning	Gareth Roberts and Masha Fedzechkina

Evaluating *each*- (but not *every*-) sentences encourages encoding individual properties

Tyler Knowlton (Maryland), Justin Halberda (Johns Hopkins), Paul Pietroski (Rutgers), Jeffrey Lidz (Maryland)

The meaning of a universally quantified sentence like *each/every circle is green* is standardly thought to express a relation between two independent sets [1], as in (1). But the same content could be represented in speakers' minds in terms of individuals and their properties, as in (2).

- (1) $\text{TheX:Circle}(X) \subseteq \text{TheY:Green}(Y) \approx \text{the circles}_X \text{ are a subset of the green-things}_Y$
(2) $\forall x:\text{Circle}(x)[\text{Green}(x)] \approx \text{each individual circle}_x \text{ is such that it}_x \text{ is green}$

There is some evidence that speakers have a group-implicating meaning for *every*, in line with (1), but have a purely individual-based meaning for *each*, as in (2). For example, participants have been found to offer better estimates of the number of circles when asked whether *every circle is green* compared to when they were shown similar images but asked whether *each circle is green* [2]. This difference might reflect the distinction between (1-2), as only (1) calls for treating the circles as a group whose cardinality can then be estimated [3].

Here, we test another prediction of the (1-2) distinction: evaluating sentences with *each*, represented as in (2), will lead to encoding the circles' individual properties. In contrast, evaluating sentences with *every*, represented more like (1), is predicted to call for mentally grouping the circles in a way that abstracts away from the particular details of each individual.

In this novel task, we consider the individual property color. On each trial, participants were shown three circles that were different shades of blue, green, or orange (e.g., Fig. A) and asked to evaluate sentences like *each circle is green* or *every circle is green*. Colors were selected from an independently-normed set [4] so that half of the trials were "true" according to a majority of adults' empirically-determined color category boundaries. After participants responded to the first question, the circles were briefly masked (300ms). On half of the trials one circle's hue was then changed and participants were asked to evaluate whether *one circle changed its color*.

If *each* is understood as in (2) and *every* is understood more like (1), then participants who evaluated *each*-sentences should be more likely to notice when an individual circle changes its color compared to participants who saw the same pictures but evaluated *every*-sentences. This prediction of superior performance following *each* also controls for a potential confound in the results from [2]: *each*-sentences may have led to inferior cardinality estimation performance simply because *each* is less frequent than *every*, and thus requires extra cognitive effort that could have otherwise been devoted to encoding cardinality. Here though, the less frequent quantifier is predicted to result in better performance on a follow-up memory task.

In Experiment 1 ($n=36$), we find that this prediction is borne out. Participants were more accurate at the change-detection question if they first evaluated an *each*-statement than if they first evaluated an *every*-statement (Fig. B; $t_{33.97}=2.33$, $p<.05$). The two groups showed no significant difference in their reaction times for sentence evaluation ($t_{31.67}=0.71$, $p=.49$) or change detection ($t_{33.83}=0.08$, $p=.94$). Experiment 2 ($n=36$) replicated this effect using a staircased design in which the change-detection task got easier when participants failed to detect the change and harder when they correctly detected it, maintaining an average accuracy of around 70% for both conditions. Specifically, the new hue on a trial with a change was drawn from a normal distribution centered on the original color. If a participant correctly detected the change, the standard deviation of this distribution decreased, making subsequent trials harder; if they failed to detect the change, the standard deviation increased, making subsequent trials easier. We find that participants in the *each* condition had a smaller average standard deviation than those in the *every* condition (Fig. C; $t_{3022}=11.65$, $p<.001$). In both experiments, participants with an average reaction time exceeding 3 standard deviations above the mean were excluded.

These results support the hypothesis that *every* calls for abstracting away from individuals whereas *each* calls for their explicit representation. More generally, they offer a new tool for probing a specific dimension (group- vs. individual-highlighting) of meaning representations.

Fig. A: Example trial

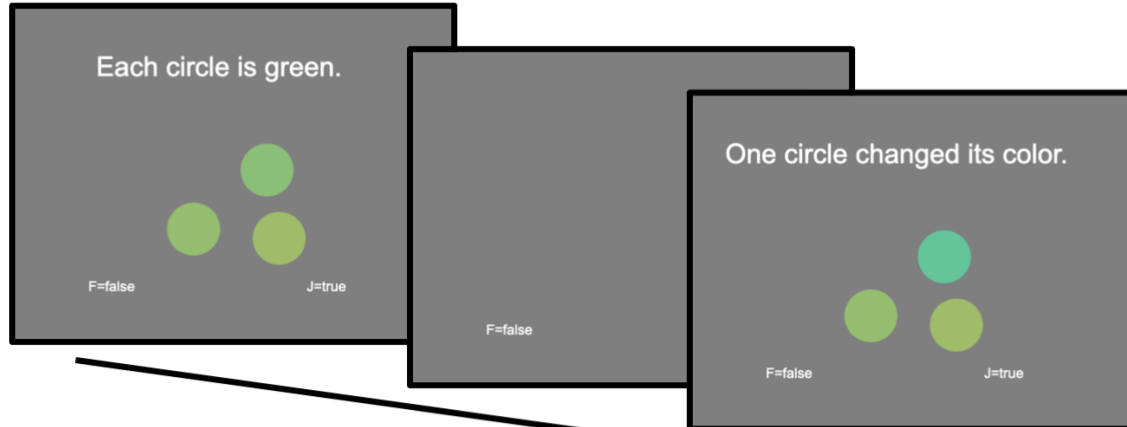


Fig. B: Exp 1 - Change detection accuracy

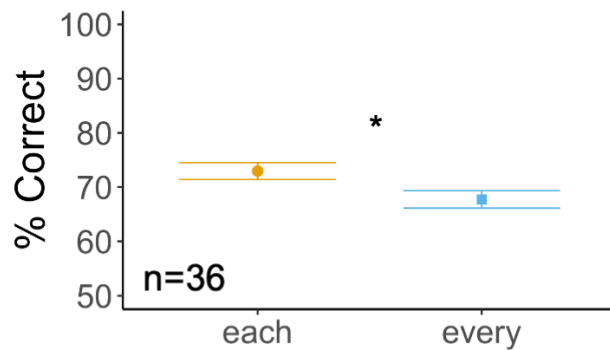
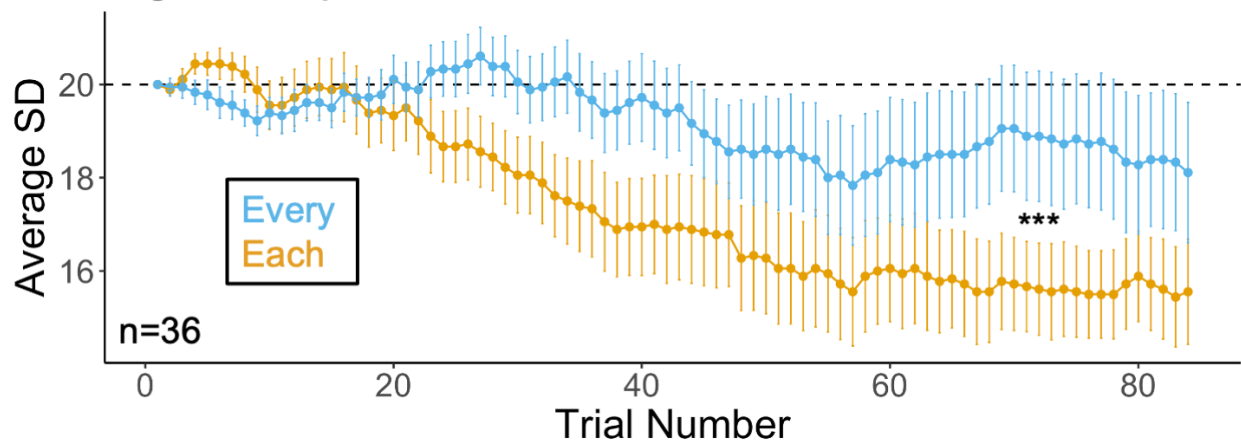


Fig. C: Exp2 - SD of new color distribution



References

- [1] Barwise & Cooper (1981) *Generalized quantifiers and natural language*
- [2] Knowlton, Pietroski, Halberda, & Lidz (2019) *The mental representation of universal quantifiers: evidence from verification*
- [3] Ariely (2001) *Seeing sets: representation by statistical properties*
- [4] Bae, Olkkonen, Allred, & Flombaum (2015) *Why some colors appear more memorable than others: a model combining categories and particulars in color working memory*

Are logical representations quantifier-specific?
Evidence from priming for a non-quantifier-specific representation of scope
Mieke Sarah Slim, Peter Lauwers and Robert J. Hartsuiker
Ghent University, Belgium

Scopally ambiguous sentences (e.g., *Every bear approached a tent*) allow two scopal configurations: a *universal-wide* (wide scope *every*: every bear approached a different tent) and an *existential-wide* configuration (wide scope *a*: every bear approached the same tent). The assignment of scope is mentally represented as logical representations. A key question about logical representations is whether scope is represented following quantifier-specific scope-taking operations or following more general scope operations. This question is relevant, because quantifiers differ from each other in their scope-taking biases (e.g., *each* is more likely to take wide scope than *all*, loup, 1975). Feiman and Snedeker (2016; henceforth F&S) previously tested this question using the structural priming paradigm in comprehension. They observed that logical representations are only susceptible to priming if prime and target contained the same quantifiers. This finding indicates that logical representations are differentiated according to quantifier-specific scope-taking mechanisms. We replicated F&S's study in Dutch. Dutch quantifier words are slightly different than English quantifier words. More specifically, the Dutch distributive quantifiers *iedere* and *elke* are closer in meaning than their rough English translation equivalents *each* and *every*. Our original aim was therefore to test whether priming emerged between *elke* and *iedere*. However, the outcome of Exp. 1 led us to re-examine F&S's hypothesis that logical representations are quantifier-specific.

We used sentence-picture matching tasks to elicit priming of logical representations in language comprehension (similar to F&S; Fig. 1). Prime sentences either had the form *elke...een* ('every...a'), *iedere...een* ('every...a') or *alle...een* ('all...a'). Target sentences were always *elke...een*. In Exp. 1 ($n = 188$), we manipulated Prime Quantifier (*elke*, *iedere*, *alle*) between participants (following F&S). The results of Exp. 1 revealed priming from *elke* to *elke*, but also between the different quantifiers *alle* and *elke*. There was no priming between *iedere* and *elke* (Fig. 2). Given these inconclusive results, we ran a replication (Exp. 2; $n = 180$) in which Prime Quantifier was manipulated within participants. Exp. 2 showed priming in all conditions (with no differences in the magnitude of the effect; Fig. 3). This finding contrasts with F&S's hypothesis. Rather, people seem to generalise in scope assignment across different quantifier words if they are exposed to similar interpretations of different quantifier words. Note that the contrasts between F&S's findings and our findings is likely not due to the difference in language tested in both studies (Dutch vs English). Like English quantifiers, Dutch quantifiers differ from each other in scope-taking behaviour (*elke* and *iedere* are more likely to take wide scope than *alle*; e.g., Dik, 1975). Therefore, it is more likely that these differences are due to differences in experimental design.

Some structural priming studies in language production showed that abstract priming sometimes requires presence of a lexical overlap condition in the same experiment (Muylle, Bernolet, & Hartsuiker, in press). This may also explain our results: In Exp. 1, within-quantifier and between-quantifiers were never both presented to the participants, whereas this was the case in Exp. 2. We tested this hypothesis in Exp. 3 ($n = 260$), in which the presence of the within-quantifier condition (*elke-elke*) was manipulated between blocks. Exp. 3 showed that priming emerged between quantifiers in the absence of a within-quantifier condition (Fig. 4). This suggests that people generalise across different quantifier words as long as they are exposed to both possible readings of multiple quantifier words (i.e., also if they are exposed to multiple between-quantifier conditions). Altogether, our results therefore suggest that the absence of between-quantifier priming does not denote a quantifier-specific representation of scope assignment. Rather, people seem to generalise across the scope-taking behaviour of different quantifiers if they are exposed to the scope-taking behaviour of multiple quantifiers. Therefore, we conclude that logical representations do not involve a quantifier-specific representation of scope assignment: Quantifiers bias us towards the construction of a particular logical representation, but logical representations themselves do not specify quantifier-specific scope-taking mechanisms.

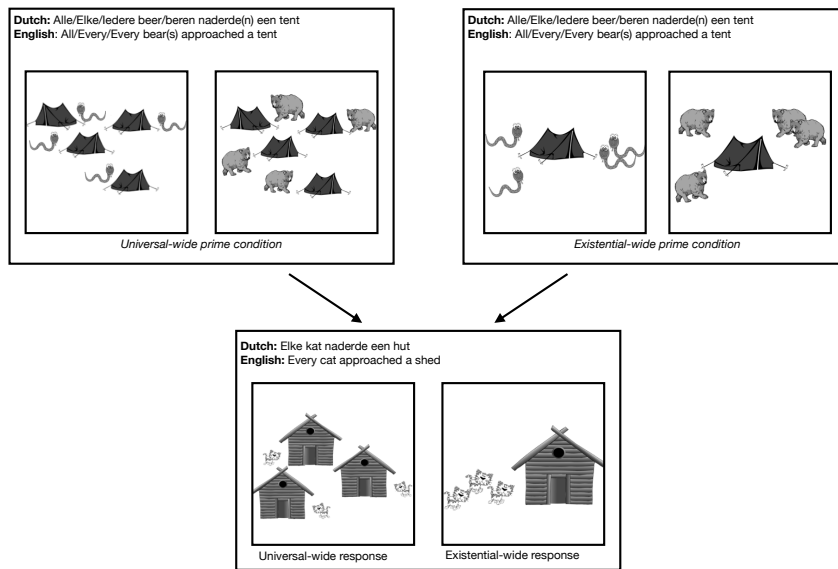


Fig. 1. Example of a prime-target trial of the sentence-picture matching tasks used in Experiments 1-3. Participants matched the sentence with one out of the two pictures. In the primes, they were forced to select one interpretation, in the targets, they could choose between both interpretations.

Prime sentences always involved one universal quantifier (*elke, iedere* or *alle*). The labels *Universal-wide prime*, *Existential-wide prime*, *Universal-wide response* and *Existential-wide response* and the English translations are added to this figure for ease of illustration.

Experiment 1

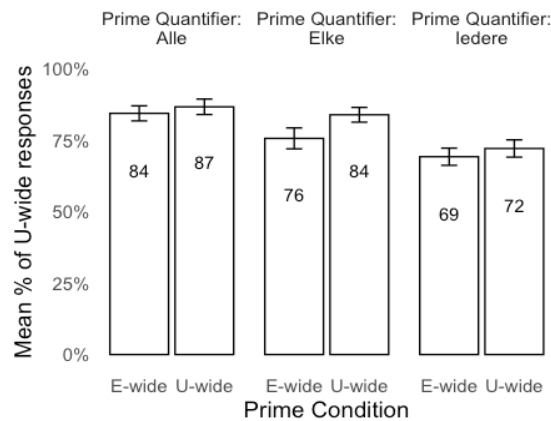


Fig. 2. Percentage of u-wide target choices per Prime Quantifier and Prime Condition configuration in Exp. 1. Logit mixed-effect models comparisons revealed a main effect of Prime Condition ($p < 0.001$), which was modulated by Prime Quantifier ($p = 0.013$; post-hoc comparisons: priming was stronger in *elke* compared to *iedere* ($p = 0.011$), but not compared to *alle* ($p = 0.127$)).

Experiment 2

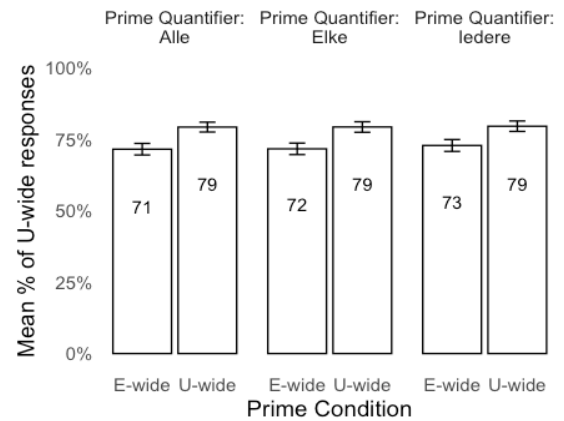


Fig. 3. Percentage of u-wide target choices per Prime Quantifier and Prime Condition configuration in Exp. 2. The statistical analyses revealed a main effect of Prime Condition ($p < 0.001$), which was not modulated by Prime Quantifier ($p = 0.935$).

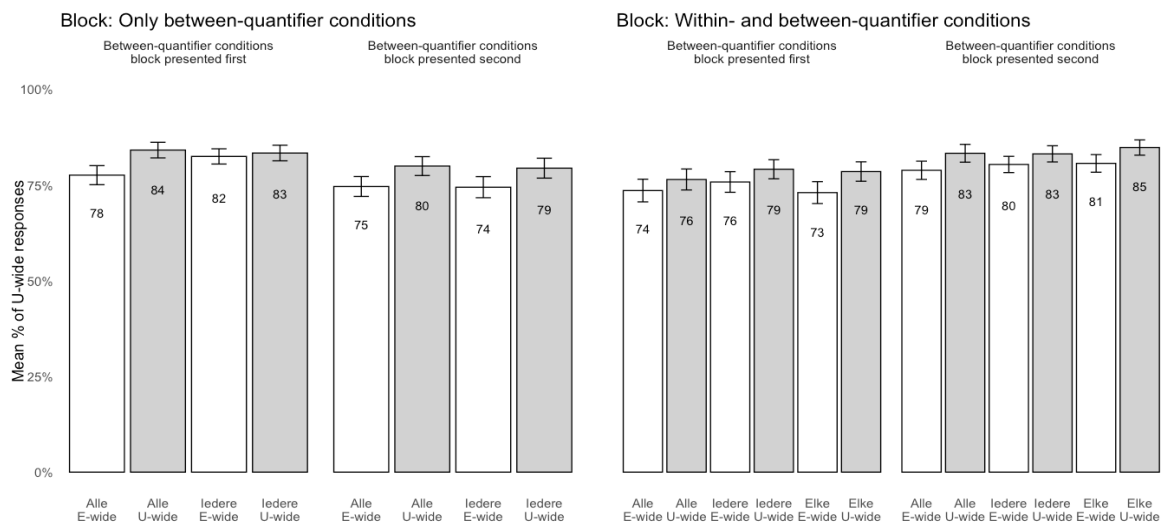


Fig. 4. Percentage of u-wide target choices per Prime Quantifier, Prime Condition, Block, and Block Order configuration in Exp. 3. The statistical analyses revealed a main effect of Prime Condition ($p < 0.001$) which was not modulated by the Prime Quantifier, Block, or Block Order conditions.

References: Feiman, R., & Snedeker, J. (2016). The logic in language: How all quantifiers are alike, but each quantifier is different. *Cognitive psychology*, 87, 29-52.; Ioup, G. (1975). Some universals for quantifier scope. In *Syntax and Semantics volume 4* (pp. 37-58); Muylle, M., Bernolet, S., & Hartsuiker, R.J. (in press). On the limits of shared syntactic representations: When word order variation blocks priming between an artificial language and Dutch. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

What primes what – an experimental framework to explore alternatives for SIs

P. Marty (UCL), J. Cowan (UCL), J. Romoli (U. of Bergen), Y.Sudo (UCL) & R. Breheny (UCL)

Background. Recent studies have shown that Scalar Implicatures (SIs) can be primed. Bott & Chemla (B&C, 2016) demonstrate that people make more pragmatic response in TARGET trials after STRONG priming trials, where they make the equivalent response, than after WEAK priming trials, where the pragmatic response is unavailable – see rows 1, 2 and 4 in Fig.1. Rees & Bott (R&B, 2018) further show that, compared to the WEAK priming trials, rates of pragmatic responses are also higher after ALT priming trials, where stronger alternatives to the expression involved in the TARGET trials are evaluated, see rows 1, 3a and 4 in Fig.1. These results, however, leave open the question whether STRONG and ALT primes boost the rate of pragmatic response, or WEAK primes lower it, or both. Here we address these open questions by adding novel baseline conditions to the set of comparisons. We also use Exp.3 to test different theories of alternatives.

Experiments. All three experiments used the same priming method and stimuli, and tested three types of expressions, AD-HOC, SOME and NUMBER (see Fig. 1). TARGET trials followed two prime trials and involved a forced choice between a card consistent only with the literal meaning and the option to select ‘Better Picture?’. The latter choice indicates an enriched meaning is accessed. Our BASELINE trials were the same as the TARGET trials but did not follow any prime trials. Importantly, these trials were presented in a separate block, prior to the main block of prime and target trials, ensuring that prime trials could not have cross-item effects on the BASELINE trials. Responses were analyzed using mixed effects logistic regression (maximal random effects structure).

Exp.1 ($n = 56$) was a partial replication of B&C’s Exp.2 comparing WEAK vs. STRONG with the addition of BASELINE. Results replicate the contrasts in B&C between WEAK and STRONG (see Fig.2, left). Crucially, BASELINE conditions reveal that STRONG primes increase pragmatic responses for AD-HOC, but not for SOME and NUMBER, suggesting that priming effects for the latter are actually due to below-baseline rates after WEAK primes. We also note that, for NUMBER, the presence of WEAK primes in the main block of trials seems to have lowered the rate of pragmatic responses after STRONG trials, suggesting that priming a lower frequency interpretation in a block of trials can have wider-ranging effects on decisions beyond the immediate triplet of interest.

Exp.2 ($n = 50$) was a control study designed to test whether the visual similarity between the cards used in the priming and target trials could explain (part of) the effects found in **Exp.1** (for similar concerns, see B&C and R&B). The materials and design were the same as in **Exp.1** except that we removed the linguistic stimuli from the WEAK and STRONG priming trials: participants were presented with a single card, either Weak or Strong, examined it and then clicked on it to proceed. Rates of pragmatic responses were about the same after these novel WEAK and STRONG primes and no different from the BASELINE condition for each type of sentence (see Fig.2, right).

Exp.3 ($n = 179$) follows up **Exp.1** by probing for any difference between Stronger (STR-ALT, as in R&B) and Non-Weaker (NW-ALT) alternatives for AD-HOC, see rows 3a and 3b in Fig.1. NW-ALT are predicted to have same effect as STR-ALT by structural theories of alternatives (e.g., Fox & Katzir, 2011, F&K). The ALT conditions for SOME and NUMBER always involved stronger alternatives. To minimize the risk that other primes affect choices to TARGET trials outside of the immediate triplet of interest, the STRONG, STR-ALT and NW-ALT conditions were split between groups: following the block of BASELINE trials, participants were tested on the WEAK priming condition plus one of the three priming conditions just mentioned. Focusing on the case of AD-HOC, the results from STRONG (Fig.3, leftmost panel) essentially replicate the results **Exp.1**; the results from STR-ALT (Fig.3, middle panel) reveal a priming effect above WEAK and BASELINE, showing that the salience of conjunctive sentences increased pragmatic responses for AD-HOC; by contrast, no such priming effects were found in NW-ALT for AD-HOC (Fig.3, rightmost panel).

Discussion. Our findings clarify the extent to which alternatives can prime SIs, by singling out for which type of SI and alternative it can happen. In particular, they provide evidence for priming by alternatives, but only for *ad-hoc* SIs. This refines the results of R&B, but contrasts with those of Waldon & Degen (W&D, 2020) who did not find above-baseline priming effects for *ad-hoc* SIs. We attribute the different outcomes to the fact that baselines in W&D were located in the same block as the prime and target trials, and so the baseline rates may have been increased in the presence of STRONG primes elsewhere in the testing block. In addition, our results reveal a contrast between STR and NW alternatives in their ability to prime *ad-hoc* SIs. These findings pose a challenge for theories that include both types of alternatives without distinguishing between them (e.g., F&K).

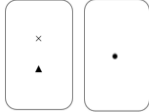
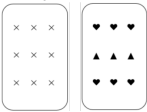
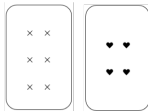
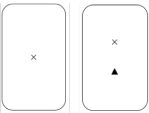
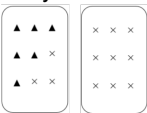
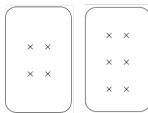
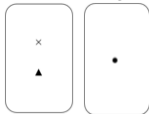
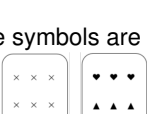
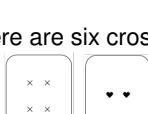
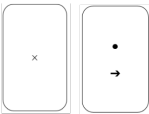

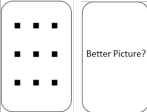
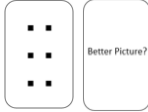
		Ad-hoc	Some	Number
PRIMES	1. WEAK	There is a cross. 	Some of the symbols are crosses. 	There are four crosses. 
	2. STRONG	There is a cross. 	Some of the symbols are crosses. 	There are four crosses. 
	a. STR	There is a cross and a triangle. 	All of the symbols are crosses. 	There are six crosses. 
	b. NW	There is a cross. 		
TARGET		There is a square. 	Some of the symbols are squares. 	There are four squares. 

Figure 1: Example prime and target trials for each priming condition and expression type. In the prime trials, participants are asked to choose the card that best fits the sentence (the expected choice corresponds to the left card).

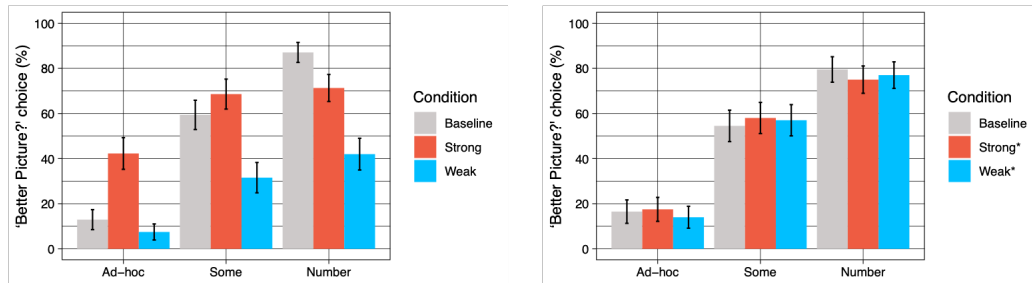


Figure 2: Results from Exp.1 (left) and Exp.2 (right). Error bars denote 95% confidence intervals.

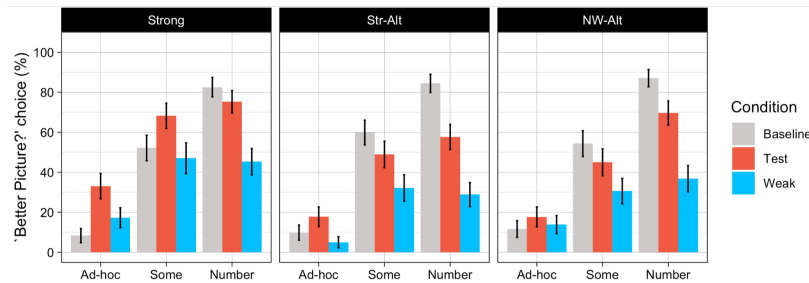


Figure 3: Results from Exp.3. Error bars denote 95% confidence intervals.

Selected references: Bott & Chemla, 2016, *Shared and distinct mechanisms in deriving linguistic enrichment* • Breheny, Klinedinst, Romoli & Sudo, 2018, *The symmetry problem: current theories and prospects* • Fox & Katzir, 2011, *On the characterization of alternatives* • Katzir, 2007, *Structurally-defined alternatives* • Rees & Bott, 2018, *The role of alternative salience in the derivation of scalar implicatures* • Trinh & Haida, 2015, *Constraining the derivation of alternatives* • Waldon & Degen, 2020, *Symmetric alternatives and semantic uncertainty modulate scalar inference*

Developmental plasticity and lateralization of function for language

Elissa L. Newport (Georgetown University Medical Center)

Adults show striking lateralization of function; for example, sentence processing is typically lateralized to the left hemisphere, whereas processing vocal emotion or intonation is typically lateralized to the right hemisphere. A number of investigators have hypothesized inherent differences between the hemispheres that might underlie these processing differences, which produce lateralization differences in over 90% of healthy adults. However, our own recent research finds that both hemispheres are surprisingly capable of conducting either of these processes perfectly well when one hemisphere is damaged at birth. I will show the evidence for this claim and then try to clarify what I think these findings tell us about lateralization of function and what kinds of differences there may be between the hemispheres in early development that set the stage for the strong laterality we see in healthy adult language processing.

Acquiring Recursive Structures through Distributional Learning

Daoxin Li, Kathryn Schuler (University of Pennsylvania)

This study investigates the learning mechanism that enables the acquisition of recursive structures. Languages differ regarding the depth, structure, and syntactic domains of recursive structures (Pérez-Leroux et al. 2018). For example, English allows infinite free embedding of genitive -s, (1), whereas German restricts this structure to only one level and to a limited set of items, (2), (Weiss 2008). Thus, while the ability for recursion may be innate and universal (e.g. Hauser et al. 2002), speakers need to learn in which syntactic domains this ability can be applied.

It has been proposed that explicit evidence for deep embedding in the input is necessary for the acquisition of recursive structures (e.g. Roeper 2011), but experiments have reported early acquisition of recursive structures even though evidence for deep embedding is rarely attested in young children's input (e.g. Giblin et al. 2019). This present study tests an alternative proposal that does not require evidence for multiple-level embeddings. Instead, learners acquire recursion through distributional learning (e.g. Braine 1987; Maratsos & Chalkley 1980). Specifically, the proposal (Grohe et al. 2020; Li et al. 2020) suggests that productivity, defined as structural interchangeability, is the prerequisite for recursion; so the recursion of a structure (e.g. **X's-Y**) is licensed if a sufficiently large proportion of nouns attested in the **X** position in the input are also attested in the **Y** position in the input.

We used two artificial language learning experiments to test the proposal. In each experiment, 25 adults were exposed to 88 **X-ka-Y** phrases, where 12 different words were attested in the **X** position. In Experiment 1, only some of the words were also attested in **Y** position (6 out of the 12); in Experiment 2, nearly all were (10 out of the 12 words). The frequency of 12 words followed a Zipfian distribution, and the total frequency of each word was the same across two experiments. At test, we asked the participants to rate on a scale of 1 to 5 the acceptability of one-level (**X-ka-Y**) and two-level (**X-ka-Y-ka-Z**) attested phrases (i.e. phrases or combinations of two phrases heard during exposure), unattested phrases (i.e. phrases or combinations of two phrases whose post-**ka** position (**Y** or **Z**) was occupied by a word that never appeared in **Y** position the input), and ungrammatical phrases with wrong word order (e.g. **ka-X-Y**, **ka-X-Y-Z-ka**). The distributional learning proposal predicts participants from Exp2 would learn the **X-ka-Y** structure was productive and thus recursive, so they would rate unattested phrases higher than participants from Exp1 at both one- and two-level, even though two-level phrases were never attested in the input.

Results are shown in Figure 1 (one-level) and Figure 2 (two-level). We analyzed the results using ordinal regression. There was a main effect of sentence type (attested, unattested, or ungrammatical) for both one- ($\chi^2(2)=253.00$, $p<0.001$) and two-levels ($\chi^2(2)=323.82$, $p<0.001$); in particular, as predicted, unattested recursive phrases were rated significantly higher than ungrammatical phrases in Exp2 ($p<0.001$) but not in Exp1 ($p=0.47$). There was also a significant interaction between sentence type and experiment (Exp1, Exp2) for both one-level ($\chi^2(2)=8.67$, $p=0.01$) and two-level ($\chi^2(2)=52.74$, $p<0.001$). Comparison between experiments showed that unattested phrases were rated marginally lower in Exp1 than in Exp2 at one-level ($p=0.08$) and significantly lower at two-level ($p<0.01$). Overall, our results suggest that speakers can use distributional information at one level to learn whether a structure can be recursive.

- (1) a. the man's neighbor's book
- (2) a. Vaters Buch ('father's book') vs. *Manns Buch ('man's book')
 - b. *das Manns Nachbars Buch ('the man's neighbor's book')

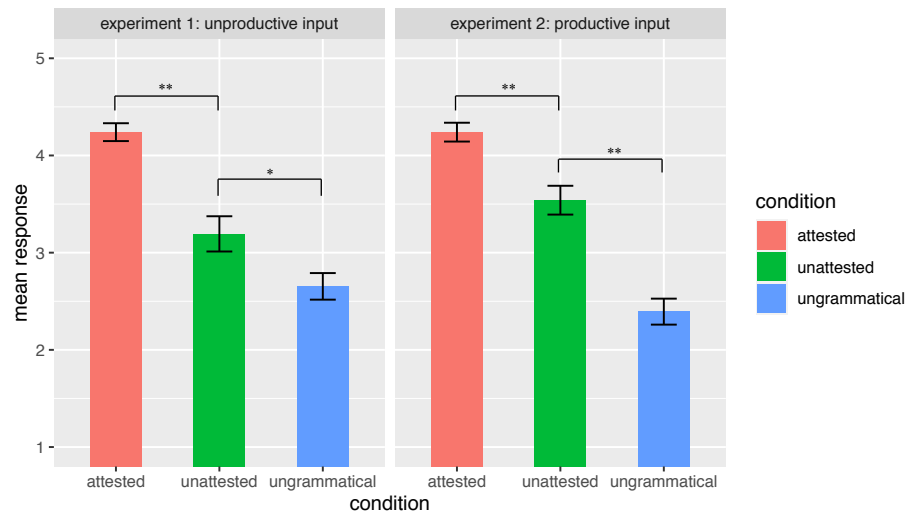


Figure 1. Mean rating response to one-level test phrases in each condition by participants in experiment 1 and experiment 2. Error bars indicate standard errors of the mean.

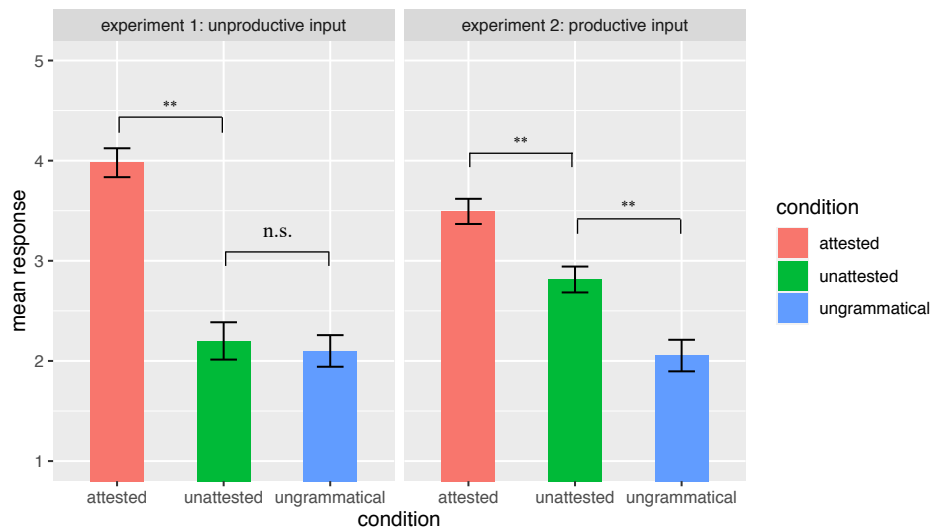


Figure 2. Mean rating response to two-level (recursive) test phrases in each condition by participants in experiment 1 and experiment 2. Error bars indicate standard errors of the mean.

Selected References

Daoxin Li, Lydia Grohe, Petra Schulz, & Charles Yang. (2020). *The distributional learning of recursive structures*. Paper presented at BUCLD-45.

Second-year syntax: Discovering Dependencies

Jeffrey Lidz (University of Maryland)

We examine the interplay between the acquisition of argument structure and wh-movement during the second year of life. We show that argument structure knowledge is acquired prior to wh-movement and that this knowledge makes discovery of the form of wh-movement possible. We further identify the role that a predictive parser plays in shaping early sentence understanding and in filtering the input that contributes to learning.

Evidence of accurate logical reasoning in online sentence comprehension

Maksymilian Dąbkowski (University of California, Berkeley), Roman Feiman (Brown University)

From Wason's (1968) selection task to dual-process theories of cognition (Evans & Stanovich, 2013; Kahneman, 2011), a rich psychological literature has argued that fast and automatic reasoning is not normatively accurate. On the other hand, linguistic theories that seek to explain reliable patterns of judgments attribute a high degree of logical sophistication to all linguistic humans. For example, most accounts of the distribution of negative polarity items (NPIs) invoke *entailment directionality* (e.g. Ladusaw, 1983), presupposing that this logical property can be computed automatically and accurately without logical training. However, outside of acceptability judgements, which have alternative interpretations (c.f. Hoeksema, 2012), there is little evidence that speakers compute this logical property during sentence comprehension (see Agmon et al., 2019).

Two novel self-paced reading experiments tested for signatures of accurate inferences made during sentence comprehension. Experiment 1 (N = 400) tested whether speakers detect logical contradictions. Participants read 12 target items displayed line by line, with line breaks at clausal boundaries. They pressed [SPACE] to advance the next line. Each item contained a "premise" in line 4 and a "conclusion" in line 5, which began with *now that they knew that ...*, presupposing that what comes next appeared earlier in the discourse. (see Figure 1). Otherwise, the two lines differed only in the quantifiers they used (*some*, *all*, *none*, *not all*). There were two conditions where the premise with *QUANT1* was identical to the conclusion with *QUANT2*, two conditions where it differed from but entailed the conclusion, and two conditions where it contradicted it (Figure 2). Participants took significantly longer to advance the conclusion line when it contradicted the premise than when it was entailed by the premise (Figure 5, LMER effect of condition: $\chi^2 = 161.31$, $p < 0.001$), consistent with rapid, normatively accurate sensitivity to the logical relations between these clauses.

Experiment 2 (N = 400) used the same paradigm to test for the capacity to detect subtler unlicensed inferences, even in the absence of strict contradictions. We manipulated the quantifiers (*QUANT*) in both the premise and the conclusion as well as the noun phrase (NP) in the premise (Figure 3). The quantifier was kept constant between the premise and the conclusion. The premise NP appeared with two modifiers (e.g. *male spotted rats*), one modifier (e.g. *spotted rats*), or no modifiers (e.g. *rats*). The conclusion NP always appeared with one modifier. Thus, the premise NP was a subset (*male spotted rats* \subset *spotted rats*), identical to (*spotted rats* = *spotted rats*), or a superset (*rats* \supset *spotted rats*) of the conclusion NP. Four quantifiers and three containment relations (*IDENTITY*, *SUBSET*, *SUPERSET*) yielded $4 \times 3 = 12$ experimental conditions in total. Depending on the combination of the quantifier and containment, there were four conditions where the premise was identical to the conclusion, four conditions where it differed from but entailed the conclusion, and four where it did not entail the conclusion (Figure 4). A significant interaction of containment by direction of entailment (Figure 6, $\chi^2 = 10.9$, $p < 0.001$) revealed that participants took longer to advance the conclusion line when it was not entailed by the premise, again consistent with rapid sensitivity to logical relations between clauses.

Our findings suggest that language processing involves automatic, accurate, and spontaneous logical computations, even in the absence of a question that requires making these inferences to verify text comprehension (Tiemann, 2014). We discuss our findings in relation to decades of psychological research on dual-process theories which argues the opposite, as well as to more sympathetic accounts of 'natural logic' in reasoning (e.g. Braine & O'Brien, 1998) and in grammar (e.g. Gajewski, 2002). We argue that logical competence is inherent in language comprehension, which can reveal the human capacity for reasoning more reliably than puzzle-solving tasks.

- (1) A group of scientists wanted to know whether spotted rats,
- (2) who are pickier eaters than other rats, liked a new kind of food.
- (3) They tested white, black, and spotted rats of both sexes.
- (4) The scientists discovered that QUANT1 of the rats loved the food.
- (5) Now that they knew that QUANT2 of the rats loved the food,
- (6) they decided to issue a recommendation based on their findings.

Figure 1: An example item in Experiment 1. The conclusion line is boxed.

	QUANT1	QUANT2
CONTR	none	some
CONTR	all	not all
ENTAIL	all	some
ENTAIL	none	not all
IDENT	some	some
IDENT	not all	not all

Figure 2: Exp 1 conditions.

- (1) A group of scientists wanted to know whether spotted rats,
- (2) who are pickier eaters than other rats, liked a new kind of food.
- (3) They tested white, black, and spotted rats of both sexes.
- (4) The scientists discovered that QUANT of the ((male) spotted) rats loved the food.
- (5) Now that they knew that QUANT of the spotted rats loved the food,
- (6) they decided to issue a recommendation based on their findings.

Figure 3: An example item in Experiment 2. The conclusion line is boxed.

	ID	SUB	SUP
all	ID	¬ENT	ENT
none	ID	¬ENT	ENT
not all	ID	ENT	¬ENT
some	ID	ENT	¬ENT

Figure 4: Exp 2 conditions.

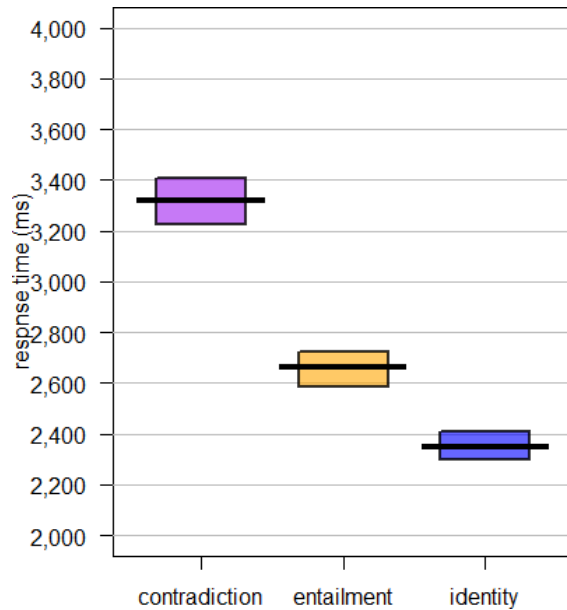


Figure 5: Experiment 1 results.

REFERENCES

- Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2019). Measuring the cognitive cost of downward monotonicity by controlling for negative polarity. *Glossa: A Journal of General Linguistics*, 4(1).
- Braine, M. D. S., & O'Brien, D. P. (1998). *Mental logic*. Psychology Press.
- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3).
- Gajewski, J. (2002). *L-analyticity and natural language* (Manuscript). MIT. Cambridge, MA.
- Hoeksema, J. (2012). On the natural history of negative polarity items. *Linguistic Analysis*, 38(1/2).
- Kahneman, D. (2011). *Thinking, Fast and Slow*.
- Ladusaw, W. A. (1983). Logical form and conditions on grammaticality. *Linguistics and Philosophy*, 6(3).
- Tiemann, S. (2014). *The processing of wieder ('again') and other presupposition triggers* (Doctoral dissertation). Eberhard Karls Universität Tübingen.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3).

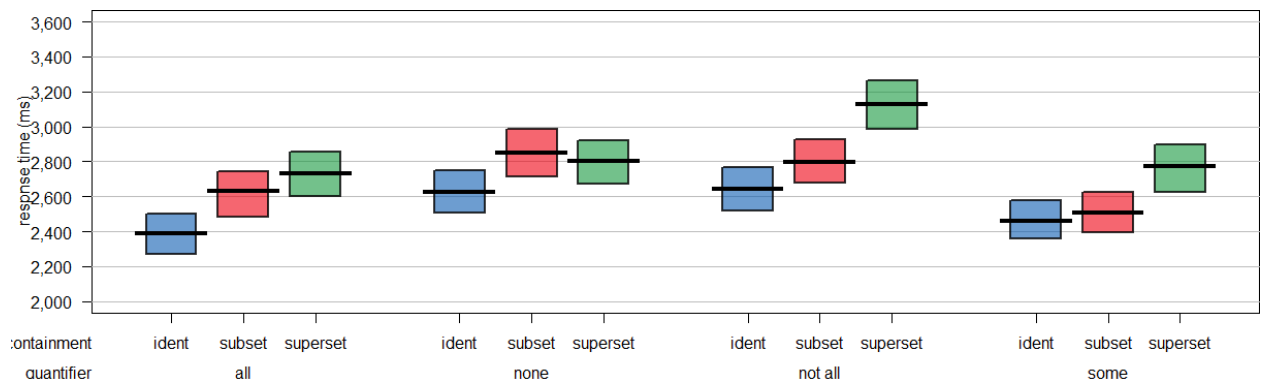


Figure 6: Experiment 2 results. The y-axis shows RT for the conclusion line in each condition.

Syntactic and semantic parallelism guides filler-gap processing in coordination

Stephanie Rich & Matt Wagers (UC Santa Cruz)

skrich@ucsc.edu

Background. The processing of filler-gap dependencies is active and eager, with the language processor postulating a gap at the first grammatically accessible position [1-5]. Therefore, in sentences like (1), comprehenders would likely postulate a direct object (DO) gap early, after *ate*. This would be correct in (1a) but incorrect in (1b), which instead contains a prepositional object (PO) gap, resulting in a filled-gap effect [4]. Prior exposure to PO gaps has been shown to lessen the filled-gap effect [6], suggesting that the parser may alter the default gap-filling strategies given sufficient evidence, or that recovery following a filled-gap is facilitated if similar structures had been recently encountered. Given the general preference for parallel conjuncts in coordination [9, 10], we hypothesize that parallelism will influence the accessibility of later potential gap positions. We test the idea that a suitably parallel first conjunction (without a gap) can facilitate the recovery of a PO gap in the second conjunct. In Experiment 1, we manipulate the presence of the prepositional phrase in the first conjunct, and thus the syntactic parallelism of the two conjuncts. In Experiment 2, we introduce an instrument in a syntactically non-parallel first conjunct, to test whether highlighting particular argument roles could lead to the anticipation of a specified instrument in the second clause as well. We find that an instrument in the first clause reduces, but does not eliminate, the filled-gap effect in the following clause, with and without parallel structure.

Method. We constructed 24 sentences containing two conjuncts (Table 1), manipulating whether the first clause mentioned an instrument (**Instrument**; +instr, -instr) and whether the second clause contained a PO gap (**Gap**; +gap, -gap). In Experiment 1, the +instr condition contained a PP, resulting in parallelism in both syntactic structure and argument role structure. In Experiment 2, the +instr condition highlighted the instrument role periphrastically, using a different syntactic structure and therefore removing the possibility of syntactic parallelism. Both experiments were presented in word-by-word self-paced reading. The critical region in both experiments was the DO in the second clause.

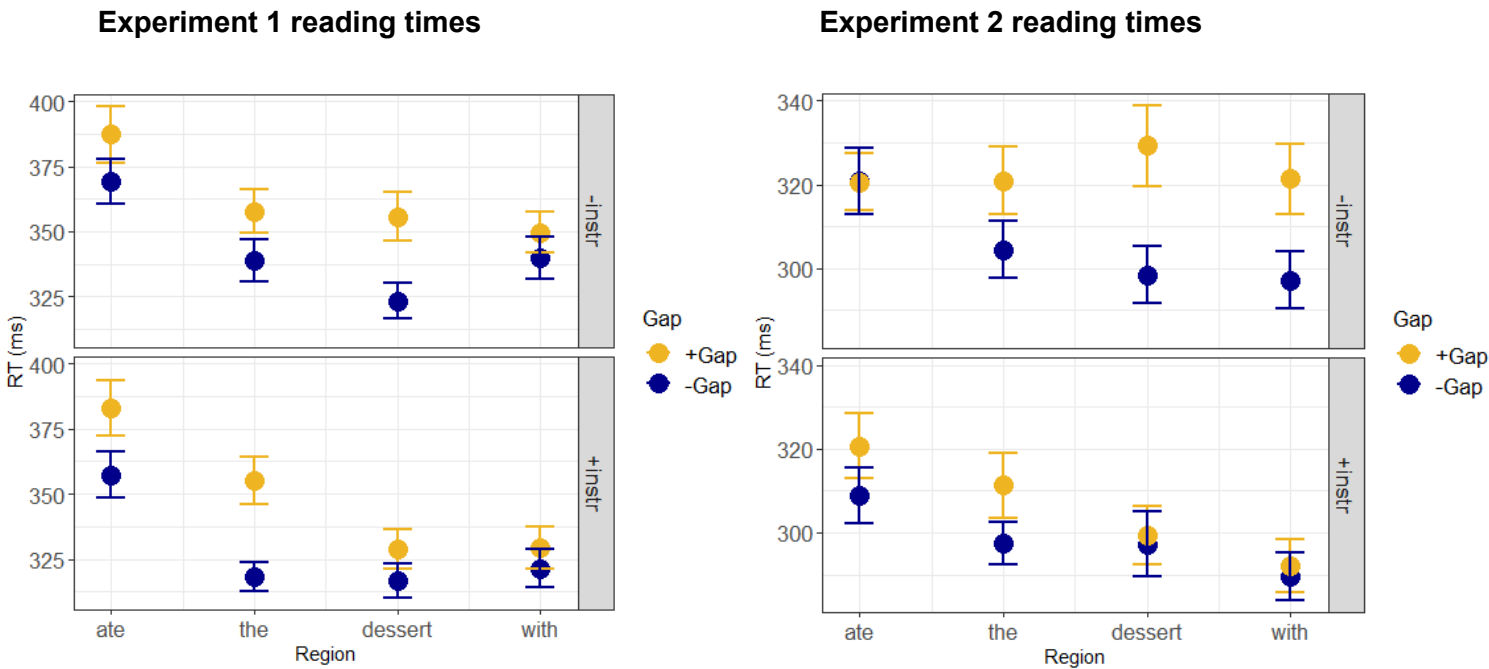
In **Experiment 1 (N = 84)**, the filled-gap effect, or penalty for the +gap conditions, was found at the DO article and DO noun ($p < .05$). The DO noun (*dessert*) showed an interaction ($p < .05$) in which there was a filled-gap effect in the -instr condition, but not in the +instr condition. In **Experiment 2 (N = 88)**, a filled-gap effect was found at the DO article ($p = .06$), and at the DO noun and at the preposition ($p < .05$). There was an advantage for the +instr conditions on the DO noun ($p < .05$) and preposition ($p < .01$) that appears to be driven by a smaller filled-gap effect in the +instr condition compared to the -instr condition, but the interaction was not significant.

Discussion. We add to previous studies of filler-gap processing by showing that the disruption caused by a filled gap in DO position can be lessened by encountering an earlier PP inside a VP in coordination (Exp. 1). However, the results of Exp. 2 suggest that parallel syntactic structure isn't strictly necessary and that processing later gap sites is facilitated if there are other argument roles comprehenders can expect to encounter (cf. [11]), an effect that depends on semantic as well as syntactic information. A third experiment testing parallelism in information structure is underway.

- (1) a. Ben wondered what Carla ate ____.
 b. Ben wondered what Carla ate the dessert with ____.

Experiment 1 (n = 84)	Experiment 2 (n = 88)
Ben saw that... +instr Carla ate the dessert with a spoon... -instr Carla ate the dessert...	Ben saw that... +instr Carla used a spoon to eat the dessert... -instr Carla ate the dessert...
+gap ...but he wondered what Dan ate critical the dessert with spillover at the party on Sunday. -gap ...but he wondered if Dan ate critical the dessert with a fork spillover at the party on Sunday.	

Table 1. Same item across Experiments 1 and 2.



Figures 1 & 2. Reading times on the critical region in the second clause in Experiments 1 & 2.

References

[1] Frazier, 1987; [2] Frazier & Clifton, 1989; [3] Traxler & Pickering, 1996; [4] Stowe, 1986; [5] Omaki et al., 2015; [6] Atkinson & Omaki, 2016; [7] Wagers & Phillips, 2009; [8] Parker, 2017; [9] Frazier et al., 2000; [10] Sturt et al., 2010; [11] Boland et al., 1995

The laboratory discovered: Place-for-institution metonyms appearing in subject position are processed as agents

Matthew W. Lowder, Adrian Zhou (University of Richmond), & Peter C. Gordon (University of North Carolina at Chapel Hill)

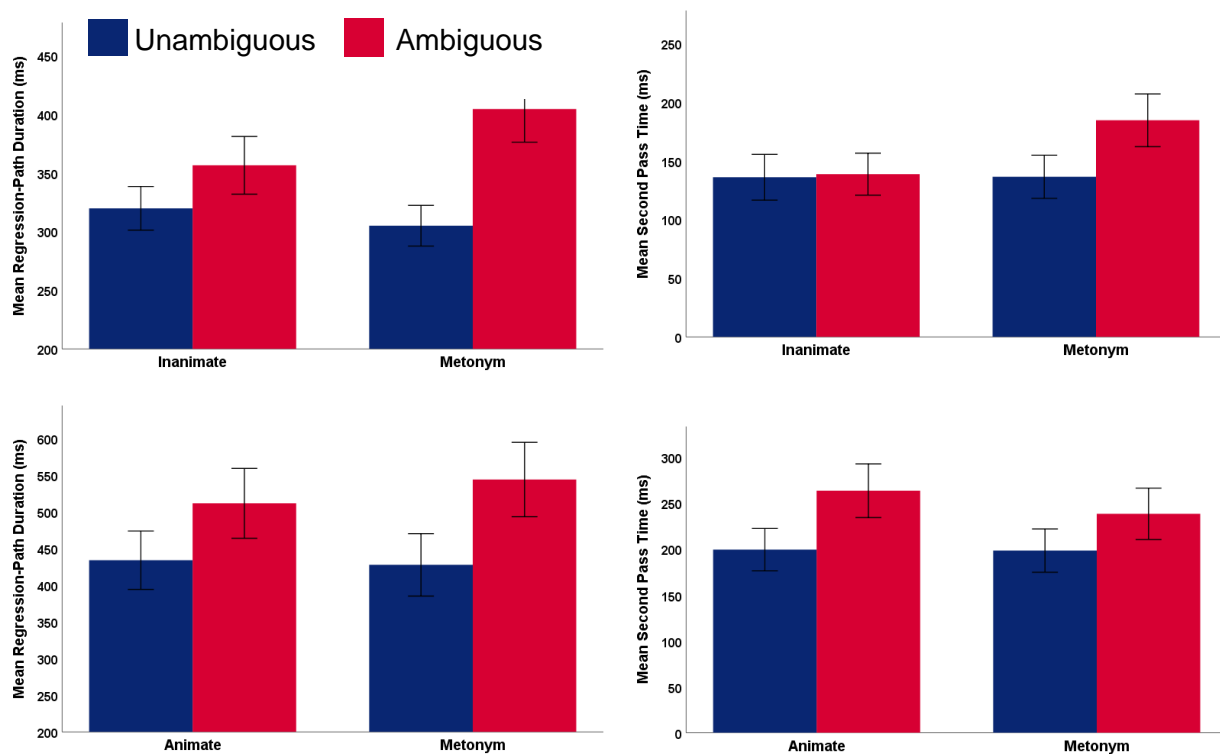
Metonymy is a type of figurative language in which an entity is referred to by a salient characteristic of the entity. For example, in place-for-institution metonymy, as in (1), *college* can refer to a literal physical place (1a) or can be used figuratively to refer to the people associated with the college (1b). Although metonymy is extremely common in everyday language, it is relatively understudied in the psycholinguistic literature, and the research to date presents an unclear picture about how metonyms are processed. Whereas early work suggested that familiar metonyms are processed just as quickly as literal expressions (e.g., Frisson & Pickering, 1999), more recent research suggests that the grammatical role of the metonym can have a large effect on the relative ease or difficulty of processing. For example, Lowder and Gordon (2013) showed that readers had greater difficulty processing familiar metonyms in their figurative sense (1b) versus their literal sense (1a), but only when the metonym was in a focused syntactic position (i.e., object of the verb). In contrast, when the metonym appeared in a defocused position (i.e., part of an adjunct phrase), the processing difference was eliminated.

Metonyms can also appear in subject position in sentences where there is no preceding context to point the comprehender toward a literal or figurative interpretation. Fishbein and Harris (2014) examined the processing of these structures using producer-for-product metonyms as a test case, as in (2). Readers experienced greater difficulty when the metonym was used in its figurative sense (2b) than its literal sense (2a). Fishbein and Harris interpreted this pattern as supporting a “Subject as Agent Principle,” according to which the comprehender immediately assigns sentence subjects the thematic role of agent. In the case of producer-for-product metonyms, this leads to immediate selection of the literal, animate sense of the metonym, as opposed to its figurative, inanimate sense. In the current study, we conducted two eyetracking-while-reading experiments that examined whether similar effects would emerge for place-for-institution metonyms. In contrast to producer-for-product metonyms, place-for-institution metonyms are inanimate in their literal sense but animate in their figurative sense. Thus, if comprehenders have a bias to interpret place-for-institution metonyms that appear in subject position as agents, they should experience difficulty if the structure of the sentence later indicates that the metonym should be assigned the role of patient (i.e., a garden-path effect).

In Experiment 1, participants ($n = 44$) read sentences like those in (3), in which we systematically manipulated whether the sentence subject was a familiar metonym or an inanimate noun without a figurative sense, as well as whether the structure of the sentence was temporarily ambiguous or not. Analyses of regression-path duration and second pass time on the disambiguating by-phrase revealed significant interactions such that there was a large garden-path effect in the metonym condition but not in the inanimate condition. This pattern suggests that readers had a strong tendency to initially select the figurative sense of the metonym and assign it the role of agent. In contrast, there was no available agentive sense for the inanimate subjects. In Experiment 2, participants ($n = 40$) read sentences like those in (4), in which the inanimate condition from Experiment 1 was replaced by an animate condition. Analyses of regression-path duration and second pass time at the disambiguating by-phrase revealed a robust main effect of sentence structure indicating garden-path effects for both the metonym and animate condition. There was no hint of an interaction in any measure, suggesting that the magnitude of this effect was equivalent regardless of whether the sentence subject was animate or was a metonym.

The results provide further support for a Subject as Agent Principle in the processing of metonymy. In the case of place-for-institution metonyms, this heuristic prompts the comprehender to immediately access the figurative sense of the metonym and later revise this interpretation if necessary.

- (1a) Sometime in August, the journalist photographed the college after he had... (Literal)
 (1b) Sometime in August, the journalist offended the college after he had... (Figurative)
 (2a) As planned, Kafka was contacted by the publisher shortly after the... (Literal)
 (2b) As planned, Kafka was printed by the publisher shortly after the... (Figurative)
 (3a) The hospital requested by the doctor was not... (Metonym, Ambiguous)
 (3b) The hospital that was requested by the doctor was not... (Metonym, Unambiguous)
 (3c) The equipment requested by the doctor was not... (Inanimate, Ambiguous)
 (3d) The equipment that was requested by the doctor was not... (Inanimate, Unambiguous)
 (4a) The hospital requested by the doctor was not... (Metonym, Ambiguous)
 (4b) The hospital that was requested by the doctor was not... (Metonym, Unambiguous)
 (4c) The specialist requested by the doctor was not... (Animate, Ambiguous)
 (4d) The specialist that was requested by the doctor was not... (Animate, Unambiguous)



Mean regression-path duration and second pass time for Experiment 1 (top row) and Experiment 2 (bottom row) on the disambiguating by-phrase a function of subject type and sentence structure. Error bars represent 95% confidence intervals.

References

- Fishbein, J., & Harris, J. A. (2014). Making sense of Kafka: Structural biases induce early sense commitment for metonyms. *Journal of Memory and Language*, 76, 94-112.
- Frisson, S., & Pickering, M. J., (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1366-1383.
- Lowder, M. W., & Gordon, P. C. (2013). It's hard to offend the college: Effects of sentence structure on figurative-language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 993-1011.

Social and communicative biases jointly influence grammatical choices in learning

Gareth Roberts (University of Pennsylvania), Masha Fedzechkina (University of Arizona)

Language users are faced with a variety of choices at different levels of organization. These choices are guided by various biases, some stemming from communicative or processing constraints [1], others from social factors [2]. While the role of these biases in isolation is well-documented, less is known about their *joint* influences on speakers' choices. We employed an artificial language learning paradigm to investigate this, using two well-established biases as a test case: a bias for communicative efficiency [3, 4] and a social bias towards identifying with particular speaker groups [2]. We hypothesized that the social bias would interact with the communicative bias, such that learners would converge on languages that satisfied both pressures.

Exp. 1. 60 English speakers were recruited online via Prolific to learn an artificial language. Participants first learned the nouns referring to characters (e.g., 'barsa' for 'mountie') and then learned the grammar by watching transitive scenes ('mountie kicks chef') accompanied by sentences in the language. At the end of the experiment, participants produced sentences in the artificial language to describe novel scenes. They were informed that there are two dialects in the language (each associated with a different alien color during training, Fig. 1). Both dialects had uninformative word order (SOV/OSV 50/50%). The *case dialect* used consistent case marking on the object (100%) and thus left no uncertainty about sentence meaning; the *no-case dialect* used no case and thus permitted maximal uncertainty (subjects were not case marked in either dialect). The language overall had 50% case-marking (conditioned on the dialect) and 50% SOV order (not conditioned on the dialect). Earlier work has shown that, in the absence of social pressures, learners favor robust communication and reduce uncertainty by maintaining case in languages with flexible word order [5, 6]. We modeled a social bias by manipulating which aliens were cast as 'potential trading partners' in the instructions, encouraging participants to feel positive towards one dialect (bias-for-case or bias-for-no-case conditions) or no specific dialect (no-bias condition). **Results.** We assessed the amount of case used in production using a mixed logit model (max RE). Learners in the bias-for-case condition used the same amount of case as learners in the no-bias condition ($\beta = -0.3$, $z = -0.61$, $p = 0.54$, Fig. 2a). Learners in the bias-for-no-case condition produced significantly less case compared to the no-bias condition ($\beta = -1.65$, $z = -3.12$, $p < 0.001$). We further tested whether learners develop strategies to mitigate the increased uncertainty due to dropped case (e.g., fixing word order) by comparing the conditional entropy (i.e., the amount of uncertainty) in the linguistic systems produced across conditions. Learners in the bias-for-no-case condition had significantly higher uncertainty compared to the no-bias condition ($\beta = 0.11$, $z = 4.23$, $p < 0.001$; Fig. 2b), while the no-bias and bias-for-case conditions did not differ from each other ($\beta = -0.02$, $z = -0.79$, $p = 0.42$), suggesting that learners of the bias-for-no-case condition did not adopt strategies to mitigate the increased uncertainty due to dropping case.

Exp. 2 asked whether learners develop such strategies with increased training. We recruited 20 participants in the bias-for-no-case condition with tripled exposure compared to Exp. 1 (administered over 3 consecutive days). **Results.** On the final day of Exp. 2, learners used the same amount of case as learners in the bias-for-no-case condition in Exp. 1 ($\beta = -0.08$, $z = -0.19$, $p = 0.84$) but their linguistic systems had significantly less uncertainty compared to the bias-for-no-case learners in Exp. 1 ($\beta = -0.05$, $z = -2.43$, $p < 0.05$). Thus, increased exposure led learners to change linguistic systems to mitigate uncertainty while still expressing the social bias by dropping case.

Conclusion. Our findings suggest that communicative and social biases jointly shape speakers' linguistic choices and pathways for language change. Interestingly, while participants responded to social biases even in relatively early stages of learning, communicative strategies to mitigate uncertainty played a role only after more substantial exposure.

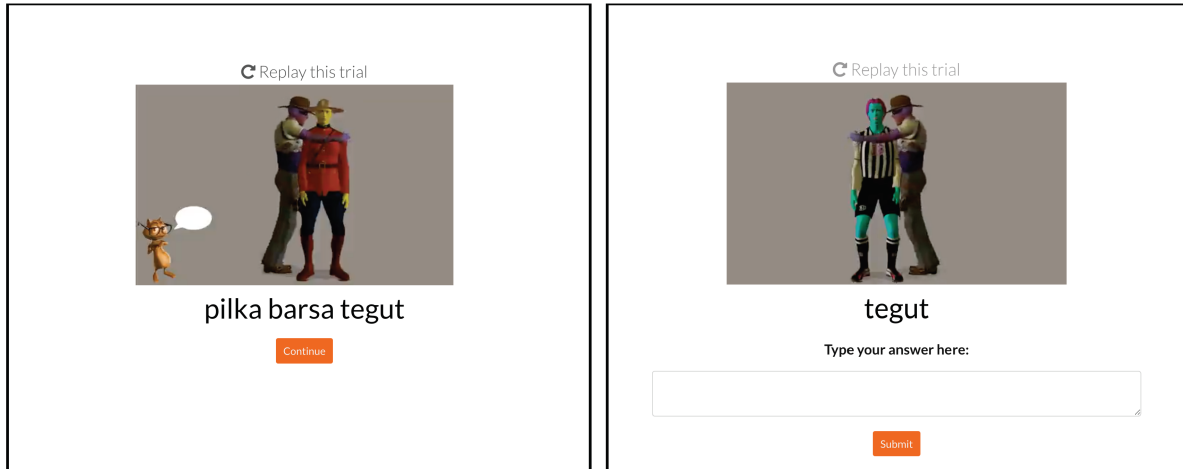


Figure 1: Examples of sentence exposure (left) and sentence production (right) trials. Pictures represent still images of the videos participants saw. The alien informant was present in each sentence-exposure video but absent during sentence-production trials.

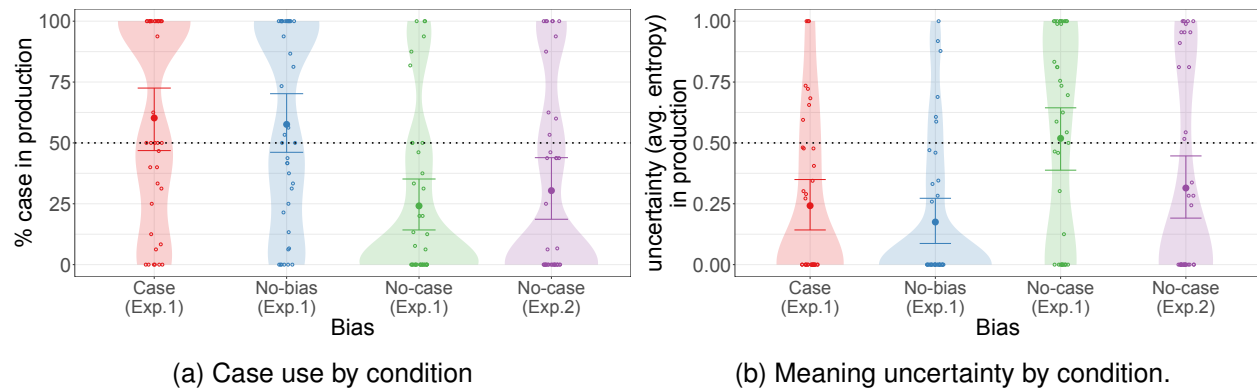


Figure 2: Case use and meaning uncertainty in production. Dashed line represents input value. Large dots represent condition means. Small dots represent individual participant means. Error bars represent bootstrapped 95% confidence intervals.

References

- [1] Harry Tily, Michael Frank, and T Florian Jaeger. The learnability of constructed languages reflects typological patterns. In *Proceedings of the Annual Meeting of the 33rd Cognitive Science Society*, pages 1364–1369, 2011.
- [2] William Labov. *Principles of Linguistic Change Volume 2: Social Factors*. Wiley, Hoboken, NJ, 2001.
- [3] Chigusa Kurumada and T Florian Jaeger. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, 83:152–178, 2015.
- [4] Gerhard Jäger. Evolutionary game theory and typology: A case study. *Language*, pages 74–109, 2007.
- [5] Maryia Fedzechkina, Elissa L Newport, and T Florian Jaeger. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive science*, 41(2):416–446, 2017.
- [6] Masha Fedzechkina and T Florian Jaeger. Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, 196:104115, 2020.

34th Annual CUNY Conference on Human Sentence Processing

Friday March 5, 2021

Session	Time	Type	Title	Authors
1	9:00	Parallel Session	Link to Friday Morning Parallel Session	
2	11:00	Break		
3	11:30	Invited Talk	Bootstrapping the syntactic bootstrapper	Anne Christophe
3	12:15	Talk	English-learning preschoolers use negative sentences to constrain novel word meanings	Alex de Carvalho, Victor Gomes and John Trueswell
4	12:45	Break		
5	13:00	Talk	Speakers extrapolate community-level knowledge from individual linguistic encounters	Anita Tobar, Hugh Rabagliati and Holly Branigan
5	13:30	Talk	Interference in the comprehension of filler-gap and filler-resumptive dependencies	Niki Koesterich, Maayan Keshev, Daria Shamai and Aya Meltzer-Asscher
5	14:00	Talk	Manipulating difficulty at different levels of language production elicits distinct patterns of disfluency	Aurélien Pistono and Robert Hartsuiker
6	14:30	Break		
7	15:00	Invited Talk	Discourse with few words: How infants form durable and expressible memories of objects and their names	Linda Smith and Hadar Raz
7	15:45	Talk	English-learning children's processing of salient phonetic distinctions varying in phonological relevance for word identity	Carolyn Quam and Daniel Swingley
8	16:15	Break		
9	16:45	Talk	A multifactorial approach to constituent orderings	Zoey Liu
9	17:15	Talk	The dual nature of subjecthood: Unifying subject islands and that-trace effects	Rebecca Tollan and Bilge Palaz
9	17:45	Talk	Differential impacts of linguistic alignment across caregiver-child dyads and levels of linguistic structure	Ruthe Foushee, Dan Byrne, Marisa Casillas and Susan Goldin-Meadow
10	18:15	Break		
11	18:30	Parallel Session	Link to Friday Evening Parallel Session	

Bootstrapping the syntactic bootstrapper

Anne Christophe (Ecole Normale Supérieure / PSL University Paris)

For a long time, children were thought to acquire first the sounds of their native language (phonology), then its words (lexicon), then the way in which words are organized into sentences (syntax). This corresponds to what young children produce: first they babble (between 6 and 12 months), then they speak in isolated words (1-2 years), and then they start combining words together. Accordingly, researchers have looked for ways in which children may acquire the sound system of their language before they know words, words before they know syntax, and so on. In many cases however, computational studies have shown that some learning problems are intractable unless one postulates access to at least partial information from other domains, and experimental studies have shown that children have managed to learn some of this partial information.

I will present experimental work on the acquisition of the lexicon, focussing on how children could gather and use syntactic information to facilitate their learning of word meanings – the *syntactic bootstrapping* hypothesis (Gleitman, 1990). Although many experiments show that infants are able to use the syntactic contexts in which unknown words appear to infer something about their potential meanings, what remains unclear is *how* children learn which syntactic contexts correspond to which conceptual features – for instance, how do they figure out that words occurring in noun contexts usually refer to objects, and how do they learn the characteristics of noun contexts in their language? I will present the hypothesis that children might learn these by generalizing from a handful of words for which they already have a meaning, a *semantic seed*. I will back up this hypothesis with computational work (showing that this learning mechanism is feasible), and experimental work (showing that toddlers do indeed learn syntactic contexts in this way).

English-learning preschoolers use negative sentences to constrain novel word meanings

Alex de Carvalho (Université de Paris), Victor Gomes & John Trueswell (University of Pennsylvania)

A central topic in language acquisition is how children use linguistic context to learn the meaning of words [e.g., syntactic bootstrapping, 1]. This application of sentence meaning to word meaning requires toddlers to parse and interpret utterances, perhaps in real-time. It is thus surprising that children's understanding of a common combinatory element, negation ("not") has been found to be delayed, with English-learning infants incorrectly interpreting negative sentences as affirmatives [2] and even 2-to-4-year-olds showing difficulty understanding negative sentences in certain tasks [e.g., 2-3]. This is surprising because parents commonly use negation in labeling events ("That's not a stone!") presumably in an effort to restrict/correct generalizations. If children treat negative labeling as affirmative, parents' attempts would be thwarted ("That's a stone!"). Here we show that children ages 2-4 years do correctly parse and interpret negative labeling events and even use such labeling to restrict the meaning of novel words. We argue this occurs because negation requires contrastive support [see e.g., 4-6].

In our experiment, we tested how affirmative and negative labeling influence children's categorization of objects that vary along a perceptual continuum (from 0 to 100%, Fig1). 2-to-4-year-olds ($n=20$; $\text{Mage}=39.5\text{mo}$, from 26 to 46.7 months) were presented with a continuum of novel creatures embedded into two videos labeled with a novel word (e.g., *blicket*). Each video was played on different televisions within a single video and introduced by a speaker (see Fig1). In the first video (TV1, common to all participants), participants saw two objects from one end of the continuum (e.g., yellowish objects – exemplars 10% and 30%) labeled several times in the affirmative: "Oh look! These are *blickets*!". In the second video (presented in TV2), participants were assigned to either the negative ($n=11$) or the affirmative condition ($n=9$) and saw two other creatures from the other end of the continuum (e.g., pinkish objects – 70% and 90%). Participants in the negative condition heard sentences like "Oh look! These are **not** *blickets*," (from which they should think that *blicket* only applies to yellowish creatures, not pinkish). Participants in the affirmative condition heard sentences like "Oh look! These are **also** *blickets*!" (from which they should think that *blicket* applies to all creatures). Participants were then tested in a selection task with images side-by-side (a new yellowish object (20%) versus a new pinkish object (80%)) and were asked to find the *blicket* (Test Trial 1). After responding and performing on two filler trials with known animals (e.g., *where is the cow?*), participants were asked to find another exemplar of the novel word "Show me the *blicket*!" (Test Trial 2) while seeing a new exemplar similar to a creature labeled as "not a *blicket*" in the teaching phase (e.g., a 85%) vs a novel completely unrelated creature.

The results showed that participants in the negative condition correctly used negative sentences to narrow down the possible referents for the novel words. In Test 1 trials, they selected the exemplar from the bottom of the continuum (i.e., 20% - a new yellowish object) more often than participants in the affirmative condition ($\beta = -1.59$, $\text{SE}=0.64$, $z=-2.49$, $p=.013$). In Test 2 trials, participants in the negative condition chose the unrelated picture more often than participants in the affirmative condition ($\beta=-1.99$, $\text{SE}=0.98$, $z=-2.02$, $p=.043$).

Our results show for the first time that English-learning preschoolers can use negative sentences as a tool to understand the boundaries of a word's meaning. They were even able to remember the restrictive information provided by negative sentences to apply a mutual exclusivity strategy when faced with a novel object (member of the not-blickets family) vs. an unrelated object. The contrasting information provided by negative sentences seem therefore to have helped children to discard the possibility that *blickets* refers to all creature-like objects while without such information, participants in the affirmative condition interpreted both yellowish and pinkish creatures as possibly being "blickets". This study provides direct evidence that preschoolers can take advantage of negative sentences to constrain the extension of a word's meaning.

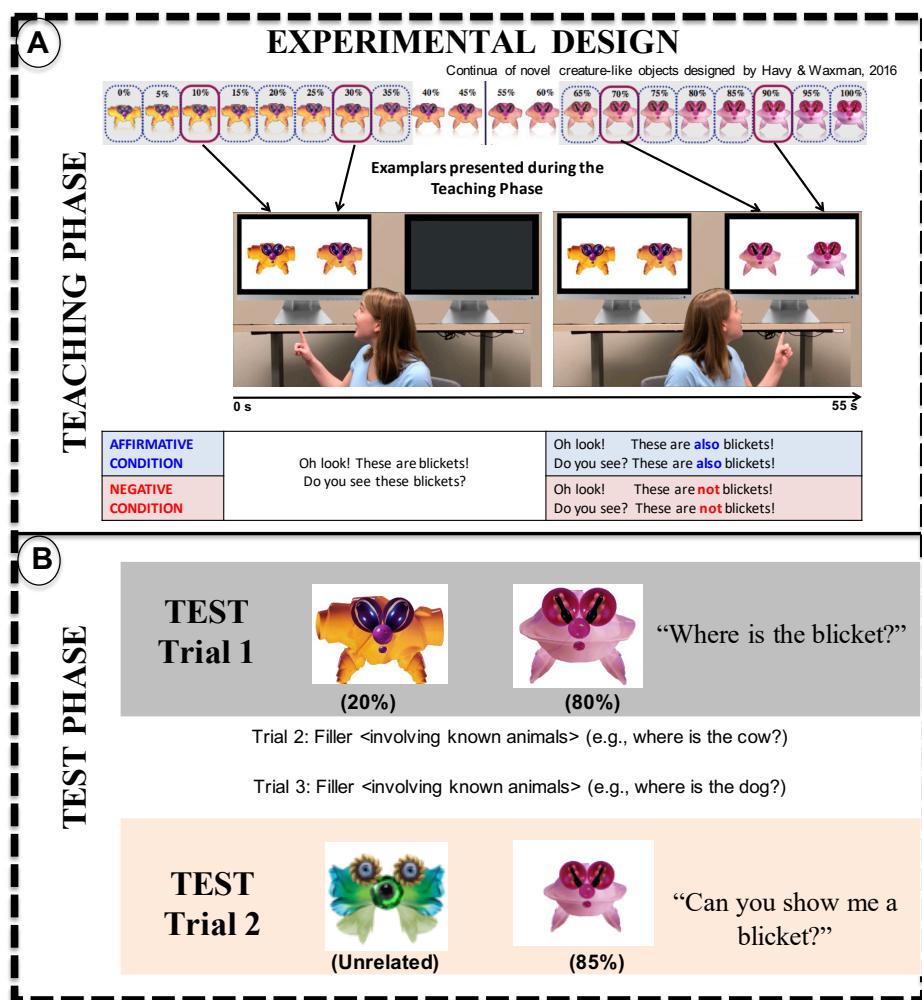


Figure 1: Experimental design - All participants watched the videos in Television 1 and 2 during the teaching phase. The video in Television 1 was the same for all participants. Depending on their condition, in television 2, they heard either negative or affirmative sentences. Finally, they all went through the same test phase with novel exemplars of the continua of novel creature-like objects (e.g., 20% vs 80% in T1 trials or 85% vs unrelated object in T2 trials) and were asked to find the novel word referent (e.g., Where is the blicket?). The experiment contained four novel words in total (2 trials, T1 vs T2, for each novel word). Participants were taught and tested on 4 novel words in this manner, using 4 different perceptual continuums of novel creatures.

Results

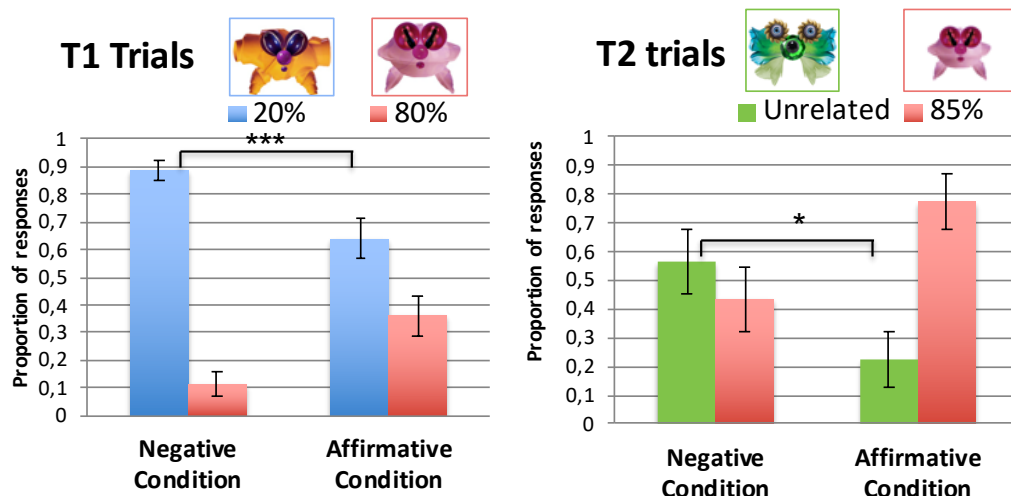


Figure 2: Proportion of picture selection in each type of test trials. T1 trials on the left and T2 trials on the right.

References

- [1] Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, 1, 3–55.
- [2] Feiman, Mody, Sanborn, & Carey (2017). What Do You Mean, No? Toddlers' Comprehension of Logical "No" and "Not." *LLD*
- [3] Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*
- [4] Reuter, Feiman, & Snedeker (2018). Pragmatic and Semantic Factors in Two- and Three-Year-Olds' Understanding of Negation. *Child Dev.*
- [5] Waxman & Klibanoff (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*.
- [6] Gelman, S. A., Wilcox, S. A., & Clark, E. V. (1989). Conceptual and lexical hierarchies in young children. *Cognitive Development*
- [7] Havy & Waxman (2016). Naming influences 9-month-olds' identification of discrete categories along a perceptual continuum. *Cognition*

Speakers extrapolate community-level knowledge from individual linguistic encounters

Anita Tobar Henríquez, Hugh Rabagliati, Holly Branigan, University of Edinburgh

Speakers vary their lexical choices depending on recent lexical processing, e.g., they tend to reuse the same words as their interlocutors (Brennan & Clark, 1996; Branigan et al., 2011). However, it is unclear how speakers' lexical choices are affected by community-level factors, e.g., whether the interlocutor is from their own speech community (*in-community partner*) or not (*out-community partner*). Indeed, we know very little about how speakers learn community-level linguistic knowledge. In three experiments, we examined (i) how speakers' referential choices varied depending on their partner's choices and speech community, and (ii) how speakers' extrapolation of their own choices to a subsequent partner was modulated by their partners' speech communities.

In Experiment 1, 160 Spanish participants completed two sessions of a picture-naming task, where they took turns with a confederate to select and name a target. They encountered different confederates in each session. Experimental items comprised targets with both a high-frequency and a low-frequency label in participants' linguistic community (e.g., *patata* [potato] vs *papa* [spud]). In Session 1, the confederate named targets before the participant, using only low-frequency labels, and we measured participants' tendency to reuse such labels (**Lexical Entrainment**). In Session 2, only participants named the targets, and we measured participants' tendency to reuse the entrained terms they had used in Session 1 (**Maintenance of Entrained Terms**). As shown in Figure 1, we manipulated participants' beliefs about their confederates' linguistic community using a 2x2 design: In Session 1, the confederate was either an *in-community partner* from Spain or an *out-community partner* from Latin America (i.e., **First Partner's Community**); in Session 2, the confederate was either from the same community as the first partner or not (**Second Partner's Community**). Experiment 2 reproduced Experiment 1 in Mexican population (N=160). In Session 1, the confederate was either an *in-community partner* (Mexico) or an *out-community partner* (Argentina), and we measured **Lexical Entrainment**; in Session 2, the confederate was either from the same community as the first partner or not, and we measured **Maintenance of Entrained Terms**. In addition, Experiment 3 tested the effects of perceived linguistic status on entrainment and maintenance in 80 Mexican participants. In Session 1, the confederate was either a high-status out-community partner (Spain) or a low-status out-community partner (Argentina), and we measured **Lexical Entrainment**; in Session 2, all participants interacted with a middle-status in-community (Mexican) partner, and we measured **Maintenance of Entrained Terms**.

In Experiment 1 (Figure 2), disfavoured terms were used significantly more in Session 1 (50%[30%]) than in a spontaneous naming task (4%[6%]; $V=0$, $p<.0001$), suggesting a **Lexical Entrainment Effect**. But lexical entrainment was not affected by **First Partner Community** ($\beta=.038$, $SE=.15$, $z=.25$, $p>.05$): Participants entrained to similar rates with a partner from another speech community (52%[29%]) or from their own speech community (49%[32%]). In Session 2, however, participants generalised their expressions from Session 1 based on their confederates' communities. There was a significant interaction between **First Partner's Community** and **Second Partner's Community** ($\beta=-.3$, $SE=.13$, $z=-2.3$, $p=.02$): Participants who first entrained to an out-community partner maintained those entrained terms less often with an in-community partner in Session 2 (57% [SD=32%]) than with an out-community partner (71%[21%]; $\beta=-.38$, $SE=.18$, $z=-2.2$, $p=.027$); in contrast, participants who entrained to an in-community partner maintained terms to similar extents with an in-community partner as an out-community partner (72%[30%] vs. 79%[20%]; $\beta=0.21$, $SE=.19$, $z=1.1$, $p>.05$). Experiment 2 (Figure 3) replicated this pattern of results in Mexican participants and Experiment 3 (Figure 4) showed that linguistic status had no effect on either lexical entrainment ($\beta=.09$, $SE=.24$, $z=.38$, $p>.05$) or maintenance ($\beta=.24$, $SE=.23$, $z=1.02$, $p>.05$), suggesting that our results were driven by differences in confederates' communities, rather than linguistic status.

These results suggest that speakers encode speech community information during language processing and store that information to inform future contexts of language use, even when such community information has not affected speakers' language use during that particular linguistic encounter. Critically, they show that speakers learn community-level knowledge by extrapolating linguistic information from individual-level experiences.

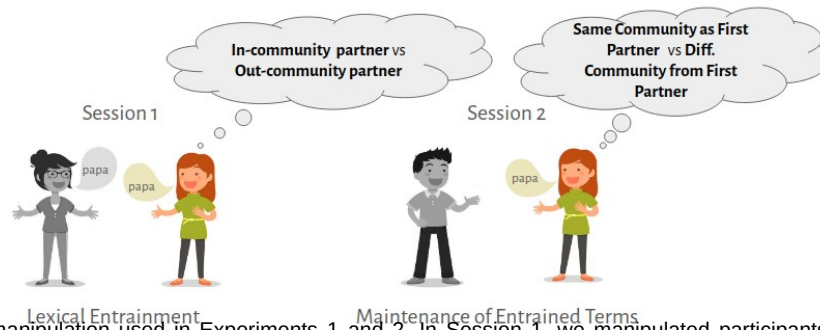


Figure 1. Experimental manipulation used in Experiments 1 and 2. In Session 1, we manipulated participants' beliefs about whether the confederate was either an in-community partner or an out-community partner (First Partner's Community: In-community partner vs out-community partner). In Session 2, we manipulated participants' beliefs about whether the confederate was either from the same community as the first partner or not (Second Partner's Community: Same Community as First Partner vs Different Community from First Partner).

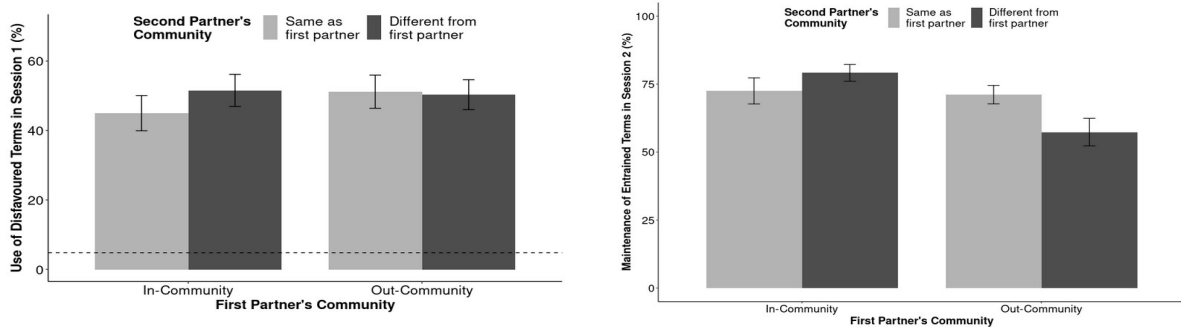


Figure 2. Experiment 1. **Left:** Mean and standard error of the percentage of use of disfavoured terms in **Session 1** (y-axis) by First Partner's Community (x-axis) and Second Partner's Community (colour-coded). The horizontal dashed line represents the baseline, i.e., the mean of percentage of use of disfavoured terms used in Session 1 during **Session 2** (y-axis), by First Partner's Community (x-axis) and Second Partner's Community (colour-coded).

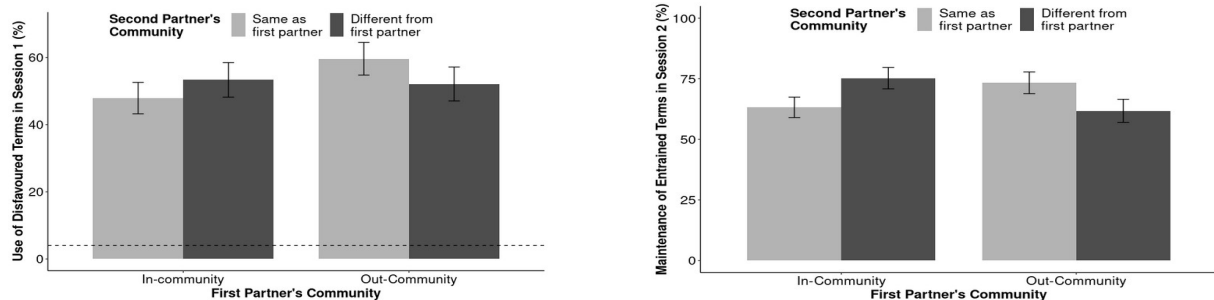


Figure 3. Experiment 2. **Left:** Mean and standard error of the percentage of use of disfavoured terms in **Session 1** (y-axis) by First Partner's Community (x-axis) and Second Partner's Community (colour-coded). The horizontal dashed line represents the baseline, i.e., the mean of percentage of use of disfavoured terms on a spontaneous naming task. **Right:** Mean and standard error of percentage of maintenance of disfavoured terms used in Session 1 during **Session 2** (y-axis), by First Partner's Community (x-axis) and Second Partner's Community (colour-coded).



Figure 4. Experiment 3. **Left:** Mean and standard error of the percentage of use of disfavoured terms in **Session 1** (y-axis) across First Partner's Community (x-axis). The dashed line represents the mean of percentage of use of disfavoured terms on the pretest. **Right:** Mean and standard error of percentage of maintenance of disfavoured terms used in Session 1 during **Session 2** (y-axis) across First Partner's Community (x-axis).

References

- Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9), 2355-2368.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482-1493.

Interference in the comprehension of filler-gap and filler-resumptive dependencies

Niki Koesterich, Maayan Keshev, Daria Shamai, Aya Meltzer-Asscher (Tel Aviv University)

Background. Language processing is subject to interference in various dependency types. Extensive research attributes interference effects to *Retrieval Interference* (RI), namely failure to integrate the correct item, or slow integration, arising when a retrieval cue matches the features of two or more items in memory. This mechanism entails that only cues available at the retrieval site can create interference [1-3]. Recent research, however, has argued that interference effects, at least in part, must be attributed to *Encoding Interference* (EI), i.e. degradation of memory representations when features are shared by items co-activated in memory. In contrast to RI, EI can occur even when the overlap is in features not relevant for retrieval [4-8].

In two comprehension experiments we show that, in Hebrew object relative clauses, a gender matching distractor reduces accuracy both when gender is a retrieval cue (in filler-resumptive dependencies, where gender is marked on the resumptive pronoun [RP]) and when it is not (in filler-gap dependencies). We used right branching grammatical object relatives, such that the main clause subject was the distractor, matching or mismatching the filler and the RP in gender. Participants read the sentences in rapid serial presentation and had to answer yes/no comprehension questions (with confidence ratings) directed at the correct (target) and incorrect (distractor) interpretations. A translated sample set is provided in Table 1.

Experiment 1: obligatory RPs (64 participants, 32 sets). We used verbs that take an *Indirect Object* (IO) complement, where relativization is obligatorily realized by an RP in Hebrew. In addition to the manipulation of distractor match and question type we also manipulated dependency length by increasing the distance between the filler and the verb using a temporal adverb and an adjective phrase. We observed main effects of distractor match ($p < .001$), and question type ($p < .001$), and a significant interaction of the two such that participants were less accurate at distractor questions (i.e. answered 'yes' at a higher rate) when the distractor matched the filler ($p < .001$, Fig. 1A). No main effects or interaction were detected for the length manipulation. In addition, we generated ROC curves separately for Match and Mismatch conditions (see Fig. 2A). A bootstrap test comparing the two curves revealed that participants had significantly lower sensitivity when the distractor matched the filler ($p < .001$).

Experiment 2: optional RPs and gaps (65 participants, 32 sets). We used verbs that take a *Direct Object* (DO) complement, where relativization can be realized either by a RP or a gap. This allowed us to manipulate the retrieval site (gap vs. RP), such that only in RP conditions gender is a retrieval cue. A pre-test ensured that both the fillers and the distractors were similarly likely complements of the RC verb, and that this likelihood was not different for DO verbs and IO verbs from Exp. 1. The results revealed the same main effects and interaction as in Exp. 1 (all $p < .001$, Fig. 1B). Resumption did not produce significant effects apart from a two-way interaction with question type ($p = .02$), suggesting that RPs increased accuracy on filler questions but not on distractor questions (regardless of distractor match). A bootstrap test comparing ROC curves of Match and Mismatch conditions revealed significantly lower sensitivity for Match cases, in both RP ($p < .001$) and gap conditions ($p < .001$, Fig. 2B-C).

Discussion. The current study provides evidence for EI effects in comprehension of relatives. As we detected interference in gap conditions, where gender is not a retrieval cue, the results cannot be attributed to RI. In addition, RI is sometimes argued not to predict interference in grammatical sentences [2], in contrast to our results. The results also cannot be attributed to simple recency of the distractor, as it precedes the target. The results are in line with previous evidence for the effect of NP type on processing of relative clauses [7-8]. We show that EI leads not only to slower RTs [4-8], but also to misinterpretation (i.e. low accuracy). Interestingly, EI is thought to arise when two NPs with overlapping features are co-activated, while in our experiments the distractor (the main clause subject) integrates with its verb before the filler (the target) is encountered. This raises questions as to the type of co-activation which leads to EI.

Exp. 1, <i>obligatory RP</i>	<i>distractor</i> The manager {M/F} knew the cashier.F that {yesterday morning} the {demanding and opinionated} customers listened to her during the busy shift
Exp. 2, gap <i>optional RP</i>	The manager {M/F} knew the cashier.F that the demanding and opinionated customers interested { __ her } during the busy shift
FillerQ: DistractorQ:	Did the customers {listen to/interest} the cashier? (correct: Yes) Did the customers {listen to/interest} the manager? (correct: No)

Table 1. Translation of an example set from the materials of Exp. 1-2.

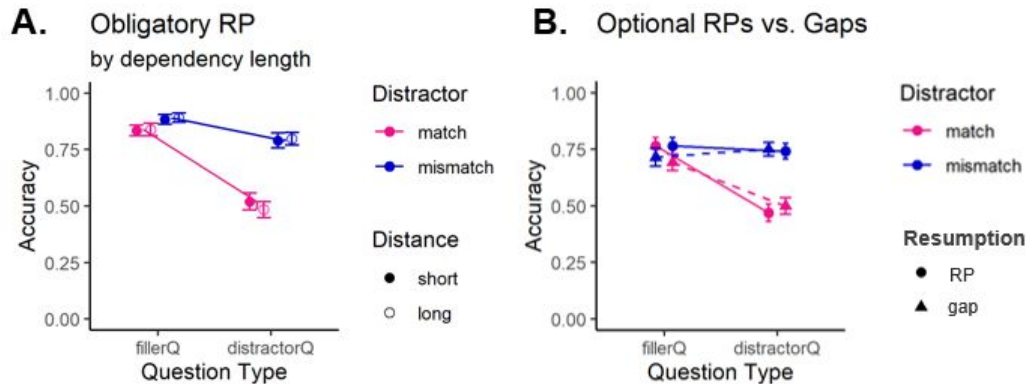


Figure 1. Percent correct responses across experimental conditions in Exp. 1 (left) and 2 (right).

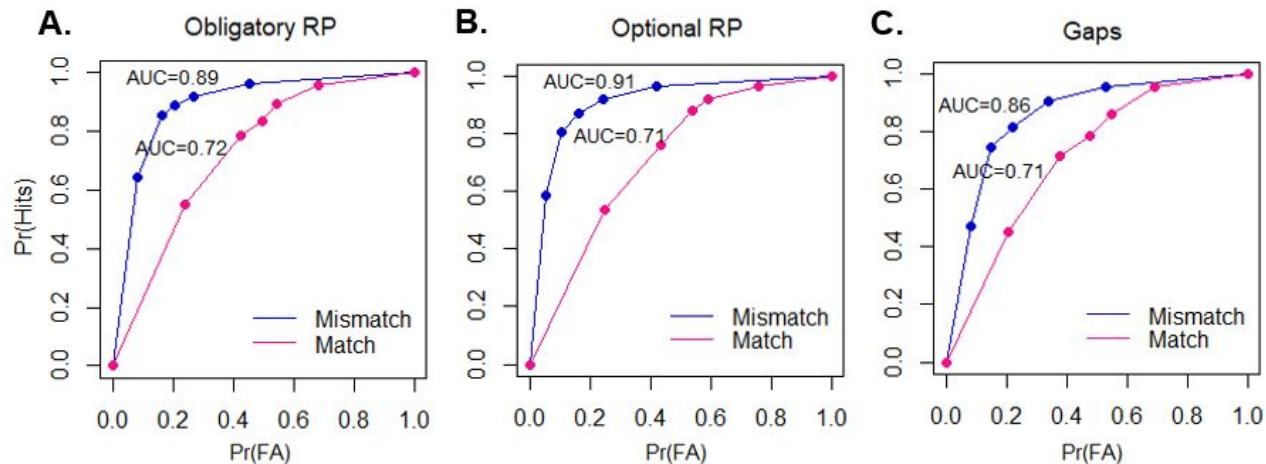


Figure 2. ROC curves for obligatory RPs (Experiment 1, collapsed across dependency length), optional RPs (Experiment 2) and gaps (Experiment 2).

References: [1] Lewis & Vasishth (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cog Sci*. [2] Wagers, Lau, & Phillips (2009). Agreement attraction in comprehension: Representations and processes. *JML*. [3] Jaeger, Engelmann, & Vasishth (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *JML*. [4] Villata, Tabor, & Frank (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psych*. [5] Parker & Konrad (2020). Teasing apart encoding and retrieval interference in sentence comprehension: Evidence from agreement attraction. *CogSci annual meeting*. [6] Smith, Franck, & Tabor (2021). Encoding interference effects support self-organized sentence processing. *Cog Psy*. [7] Gordon, Hendrick, & Johnson (2001). Memory interference during language processing. *JEP: LMC*. [8] Gordon, Hendrick, & Johnson (2004). Effects of noun phrase type on sentence complexity, *JML*.

Manipulating difficulty at different levels of language production elicits distinct patterns of disfluency

Aurélie Pistono and Robert J. Hartsuiker

To reveal the underlying cause of disfluency, several authors attempted to relate the pattern of disfluencies to difficulties at specific levels of production, using a Network Task (e.g. Oomen & Postma, 2002). In this task, participants describe a route through a network of pictures (Fig. 1). This allows for the manipulation of the items to create difficulties at specific stages (e.g. conceptual generation) while holding others constant (e.g. lexical selection). We conducted two experiments to examine the pattern of disfluency related to lexical selection difficulty (i.e. low name agreement), grammatical selection difficulty (i.e. neuter gender, which occurs less frequently than common gender in Dutch), and conceptual difficulty (i.e. blurriness). We also examined whether, by contrast, the manipulated difficulty could be predicted based on the pattern of disfluency associated with it, using multivariate pattern analyses (MVPA, Haynes & Rees, 2006).

In Experiment 1, 20 native Dutch speakers performed 20 network tasks. To examine the initial stage of lexical access we manipulated name agreement; to examine grammatical selection we manipulated grammatical gender. Linear-mixed effects models were performed with name agreement (low/high), gender (neuter/common), and their interaction as fixed effects. In Experiment 2, we examined the conceptual generation of the message, by manipulating the visual identification of some items. Twenty further native Dutch speakers performed 20 network tasks. We ran linear-mixed effects models with conceptual difficulty (blurred/non-blurred items) as a fixed effect. In both experiments, we analyzed: self-corrections, silent pauses, filled pauses, and prolongations. We then used MVPA, training classifiers on disfluency features for each participant, to predict whether s/he was about to mention a low or high name agreement item, a common gender or neuter gender item, or a blurred or non-blurred item.

In Experiment 1, low name agreement items induced more self-corrections and silent pauses than high name agreement items, while common gender items elicited more prolongations than neuter gender items. MVPA demonstrated that lexical selection difficulty is predictable from disfluency patterns, and that silent pauses are the most reliable feature across participants (Fig. 2). Classification accuracies were also above chance when classifying items' gender and only prolongations were consistent across participants. In Experiment 2, contrary to what was expected, blurriness did not induce more disfluency. MVPA yielded complementary findings. They revealed that the classifier could predict whether each participant was about to name blurred or control pictures, but that none of the features was affected in a consistent way across participants. In other words, impeding the conceptual generation of a message affected the pattern of disfluencies of each participant, but this pattern differed from one participant to another.

We replicated the finding that lexical access difficulties elicit self-corrections and pauses (Hartsuiker & Notebaert, 2010). However, contrary to what was expected, neuter gender did not elicit more disfluency than common gender. This effect could be related to the phonological form of the common gender determiner ('de' in opposition to the neuter one 'het'), which is more likely to encourage prolongations. MVPA reinforced these findings, by showing their consistency across participants. On the contrary, these analyses showed that conceptual difficulty manifests itself differently from one participant to another. They therefore point to a need for current models of language production to capture inter-individual variability.

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology*, 57, 169–177.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523–534.

Oomen, C.C.E., Postma, A., (2002). Limitations in processing resources and speech monitoring. *Lang. Cogn. Process.* 17, 163–184.

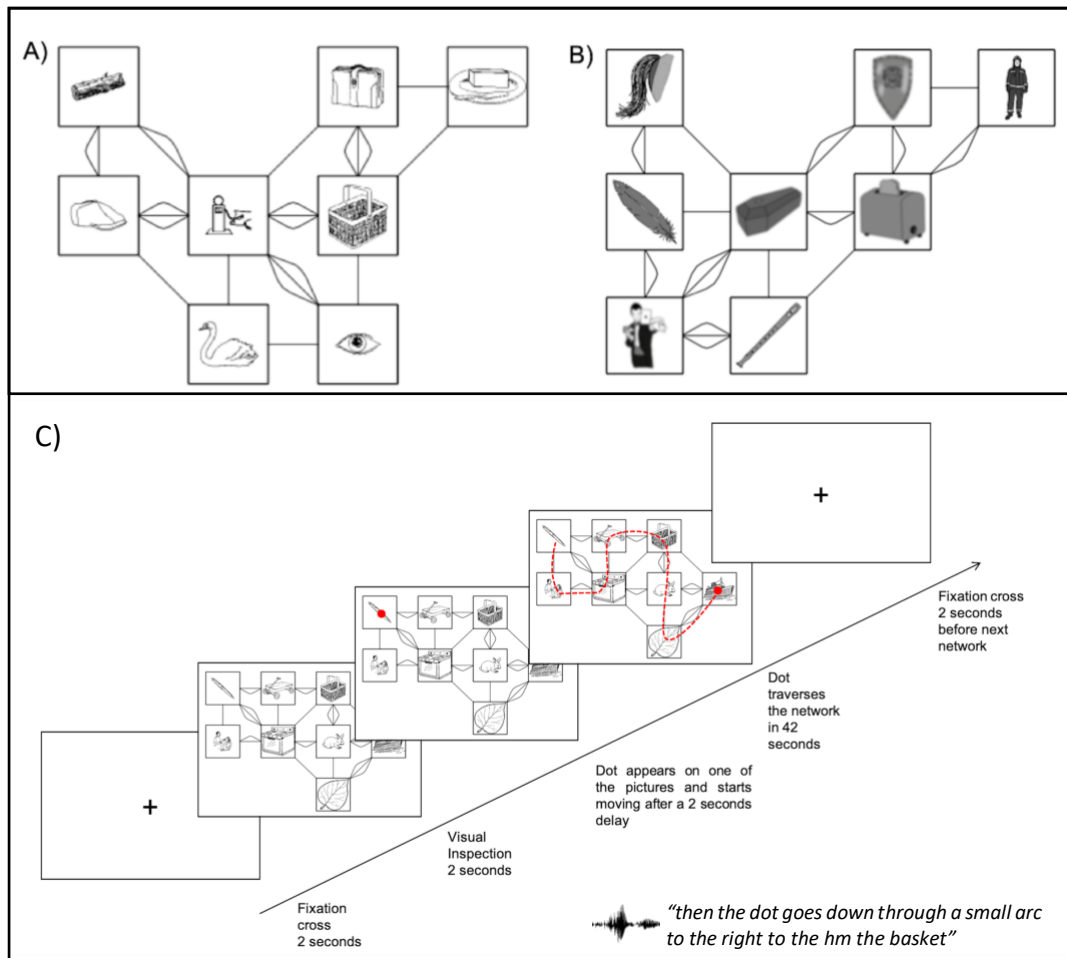


Figure 1. Example of a network for A) Experiment 1 and B) Experiment 2. Panel C) represents the procedure of each experiment. The arrow represents the time course of the experiment. Instructions were given to provide an accurate description of the network using complete sentences and to synchronize the description with the dot that moved through the network.

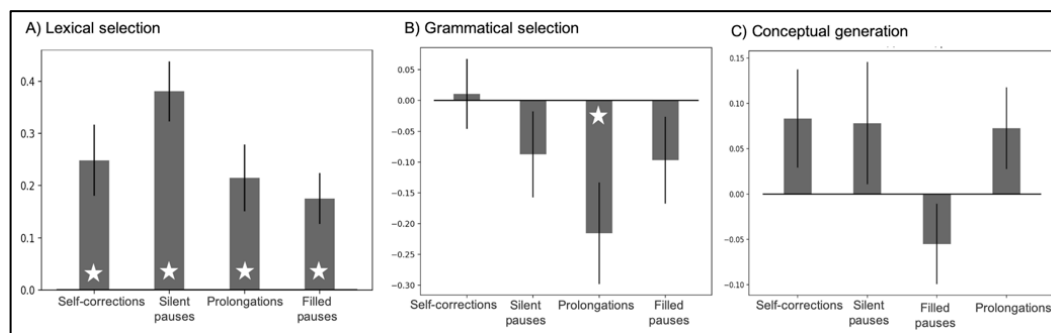


Figure 2. Contribution of each feature when classifying the pattern of disfluency related to each manipulation. White stars indicate significance. On the y-axis, positive values indicate the predicted difficulty (i.e. A) low name agreement; B) neuter gender; C) blurriness).

A) Lexical selection: self-corrections ($t(19)=3.6$, $p<.01$); silent pauses ($t(19)=6.5$, $p<.0001$); prolongations ($t(19)=3.2$, $p<.01$); filled pauses: $t(19)=3.5$, $p<.01$).

B) Grammatical selection: prolongations ($t(19)=-2.5$, $p<.05$).

Discourse with few words: How infants form durable and expressible memories of objects and their names

Linda Smith and Hadar Raz (Indiana University Bloomington)

Learning depends on both the internal processes that do the learning and on the experiences that engage those mechanisms. We know infants learn common object names well before they speak those names because infants 12 months old and younger look reliably to referent upon hearing the name. I will propose a new solution and present new work (with Hadar Raz) including a formal model based on the alignment between the dynamics of early memory formation and the temporal structure of the parent-infant interactions. The talk has three parts. In Part 1 we consider the frequency and temporal structure of the multimodal stream of parent and infant behaviors that surround highly infrequent parent naming and do so at an extended temporal scale characteristic of interactions between parents and their 12-month-old infants. In Part 2, we note how the observed properties of parent-infant interactions in Part 1 align with recent evidence on how durable and expressible cortical memories can rapidly form without hippocampal involvement. We instantiate these ideas in a mathematical model and show how the dynamic properties of the entire stream of events, not just naming, can create an internal environment of persistent activations on which the formation of durable memory depends. In Part 3, the conclusion, we argue for a reconceptualization of the environment for learning object names, one that is less about name-referent co-occurrences and transparency and more about the dynamic structure of the extended social and multimodal experiences and the internal memory processes. Human development evolved to take place in active social contexts. The time scales, temporal properties and multimodal nature of human behavior likely shaped the dynamic properties of infant memory systems. Thus we should not be surprised that the statistical properties of everyday social experiences fit the learning mechanisms available in infancy. This line of reasoning suggests the general importance of studying the natural statistics of everyday human behavior and experience.

English-learning children's processing of salient phonetic distinctions varying in phonological relevance for word identity

Carolyn Quam (Portland State University), Daniel Swingley (University of Pennsylvania)

Processing spoken language requires attributing only some types of phonetic variation to lexical distinctions. Children learning an intonation language like English must rule out pitch contour as lexically contrastive, attributing pitch variation to other sources, like stress or phrasal intonation. The developmental time-course of this learning is unclear. Quam and Swingley (2010), using a language-guided looking method, found that English-speaking 30-month-olds and adults disregarded pitch-contour changes, but did attend to vowel changes, in newly learned words. Two further studies indicated that children learning English rule out pitch as lexically contrastive prior to 24 months, but different methods have led to different developmental timelines. Singh et al. (2014), using a similar method, found 18-month-olds responded to both tone and vowel mispronunciations (MPs), whereas 24-month-olds responded only to vowel MPs. Hay et al. (2015), using the Switch habituation procedure, found 14-month-olds detected mismatches of tone-object pairings, whereas 17- and 19-month-olds did not; however, no segmental baseline was tested. Understanding how children learn to correctly interpret readily perceptible phonetic variation is important, with implications for development of the lexicon and acquisition of prosody.

Here, we compared children's interpretation of the same stimuli across ages and methods. Using the pitch and vowel contrasts of Quam and Swingley (2010), we tested English learning 3- to 5-year-olds, 24-month-olds, and 18-month-olds. Three- to five-year-olds ($N=35$) and 24-month-olds ($N=37$) were tested in the language-guided looking procedure from Quam and Swingley (2010). Children were taught a label ("deebo") for a novel toy with a consistent, exaggerated pitch contour in a story and then via ostensive labeling. At test, the toy was accompanied on the screen by a previously unlabeled (but equally familiar) distracter. Children's fixation of the target image (**Fig. 1**) was measured in response to the correct pronunciation, a vowel MP ("dahbo"), or a pitch MP (rise-fall to low fall, or vice-versa). Three- to five-year-olds were tested with both MPs; 24-month-olds were each tested with either pitch or vowel MPs. Preschoolers showed phonologically constrained responses, attending to vowel but not tone changes, replicating Quam and Swingley's (2010) finding with 30-month-olds. In an ANOVA, an effect of Pronunciation, $F(2,108) = 16.7$, $p < .001$, reflected lower target fixation in response to the vowel MP than the correct pronunciation, $t(36) = 5.53$, $p < .001$ —but there was no looking decrement for tonal MPs. Surprisingly, 24-month-olds ignored changes to *both* pitch and vowel, an effect that conflicts with prior findings of phonological constraint at 24 months; if anything, 24-month-olds showed numerically stronger effects of pitch changes than vowel changes, in contradiction to English phonology. Perhaps the rich teaching context increased the task difficulty relative to Singh et al. (2014), impairing learning.

Other work in our lab indicates 18-month-olds do not always learn words robustly in the procedure used by Quam & Swingley (2010). Thus, here we tested 18-month-olds ($N=64$) in the Switch habituation method, with two word-object pairs presented during habituation. Half of children were habituated to vowel-contrasted words ("veedo" and "vahdo"), the other half to pitch-contrasted words (rise-fall and low-falling contours). Within each cue condition (pitch or vowel), half of children were habituated with stimuli in which variability was added on the non-criterial dimension (children learned vowel-contrasted words in the presence of pitch variability, or pitch-contrasted words in the presence of vowel variability, e.g., "veedo," "vahdo," "viddo"). The results (**Fig. 2**) show that 18-month-olds could be induced to learn word pairs whether the words contrasted in pitch contour or vowel identity. This learning effect was significant for children who habituated ($N=53$), $F(1,49) = 4.46$, $p = .04$, and across all children, $F(1,60) = 4.05$, $p = .049$. Our results suggest that 18-month-olds can flexibly learn lexical distinctions inconsistent with English phonology; 24-month-olds are still in transition, apparently accepting vowel MPs in novel words; and from 30 months (Quam & Swingley, 2010) through preschool and onward, children detect arbitrary vowel changes, while accepting ("listening through") salient pitch variation.

References

- Hay, J. F., Graf Estes, K., Wang, T., & Saffran, J. R. (2015). From flexibility to constraint: The contrastive use of lexical tone in early word learning. *Child Development*, 86, 10–22.
- Quam, C., & Swingley, D. (2010). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *JML*, 62, 135–150.
- Singh, L., Hui, T. J., Chan, C., & Golinkoff, R. M. (2014). Influences of vowel and tone variation on emergent word knowledge: A cross-linguistic investigation. *Devel. Science*, 17, 94–109.

Figures

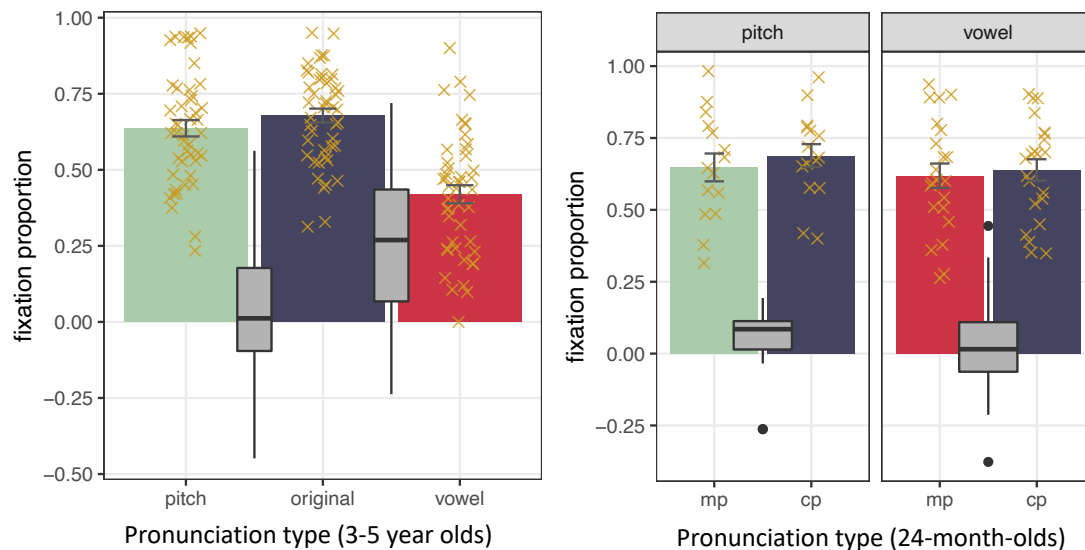


Figure 1. Looking patterns in the language-guided looking method. *Left:* like adults tested in prior work, 3- to 5-year-olds ($N=35$ tested in all 3 conditions) fixated the target picture less only in vowel-mispronunciation (“vowel”) trials vs. correct-pronunciation (“original”) trials. *Right:* 24-month-olds did *not* fixate the target picture significantly less in mispronunciation (“mp”) trials vs. correct-pronunciation (“cp”) trials, for either pitch ($n=15$; trend *ns*) or vowel changes ($n=22$). Box plots indicate within-subject difference scores between correct-pronunciation and mispronunciation trials.

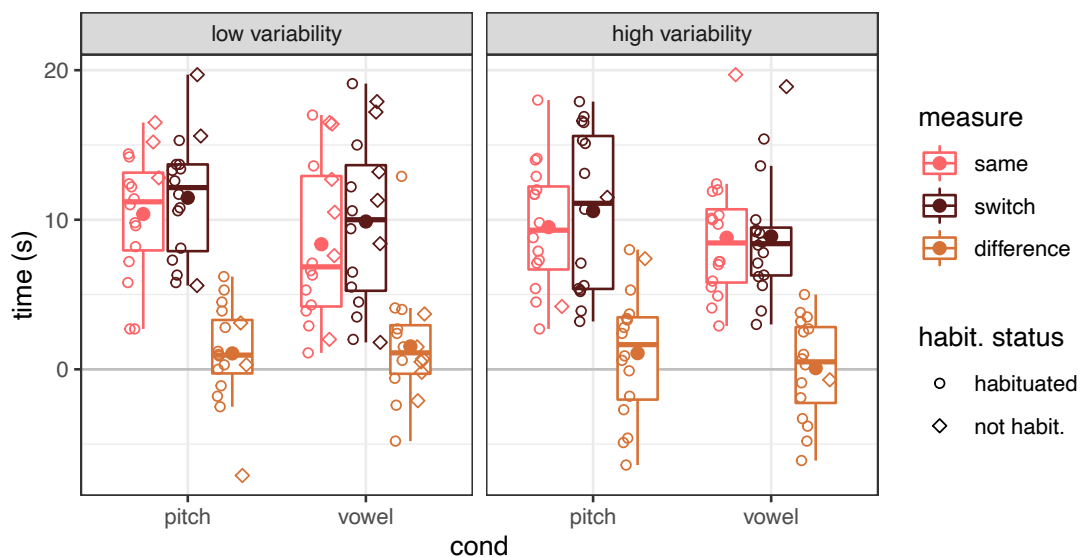


Figure 2. 18-month-olds in the Switch procedure. Switch > Same looking times mark recovery from habituation, indicating label-object learning. Learning was not significantly different across pitch-contour vs. vowel differentiated word-contrast conditions (“cond”), though learning appears informally to not be evident for vowel distinctions (/i/-a/) amid high contour variability.

A multifactorial approach to crosslinguistic constituent orderings

Zoey Liu (Boston College)

Motivation Hawkins (2014) proposed crosslinguistic syntactic variation is a multifactorial process shaped by processing efficiency, that ordering preferences are driven by several competing and cooperating factors simultaneously. Nevertheless, this proposal still lacks proper quantitative support, as most previous studies have focused on a limited set of factors and languages. Their findings are not directly comparable as most experiments have examined syntactic constructions that do not necessarily allow flexible orderings. These limitations mean that it is not currently clear what the best typological determinants are for syntactic orders across languages.

Current study We aim to bridge this gap with the double adpositional phrase (PP) construction as the test case (Liu, 2020), using multilingual corpora from Universal Dependencies (Zeman et al., 2020). We searched for verb phrases (VP) in which the head verb has two PP dependents occurring on the same side (*She danced* [PP1 *with the band*] [PP2 *at the dinner party*]), the order of which allows flexibility in at least some contexts. Preprocessing yielded an initial dataset of 40 languages (33 ended up being Indo-European (IE)). The PP orderings for these languages fall into three different patterns: (1) one for languages with only preverbal PP orders (e.g. Hindi); (2) one for languages with only postverbal PP orders (e.g. Greek); (3) one for languages with both preverbal and postverbal orders (e.g. Czech). A subset of 20 languages was then selected based on data availability, language family and genus coded following The World Atlas of Language Structures (Dryer and Haspelmath 2013), as well as their observed PP ordering pattern in corpora, which included fifteen IE, one Sino-Tibetan (Chinese), one Japanese (Japanese), one Austronesian (Indonesian), and two Afro-Asiatic (Arabic and Hebrew).

Measures We investigated the roles of four theoretically motivated constraints that have been shown to affect syntactic alternations or reflect processing complexity: (1) dependency length, measured as the linear distance between the head verb and the adposition of each PP; (2) semantic closeness, calculated as the semantic similarity between the verb and the lexical head of the PP, using fastText word embeddings (Bojanowski et al., 2017) and cosine similarity; (3) lexical frequency, which was the product of the probability of the adposition and just the lexical head to separate the contribution of phrasal length and frequency; frequency counts were taken from the Python package wordfreq; (4) contextual predictability, which was the product of the conditional probability of the adposition and the lexical head given preceding sentential context; conditional probability was estimated with neural long-short term memory models trained for each language using large-scale texts from Ginters et al. (2017). Specifically, we examined whether there is a typological tendency for the PP that is shorter, or semantically closer to be closer to the verb, and for the more frequent or the more predictable PP to appear first. To better handle issues of missing data, we eventually fit the same model architecture to every language: the order of the two PPs in each VP as the dependent variable, the four factors along with pronominality of each PP as fixed effects and the head verb as a random effect. The predictive power of each factor was evaluated with Bayesian mixed-effects models.

Results Overall, dependency length is the strongest predictor and it is more effective in postverbal than preverbal domains. In certain preverbal cases where dependency length is not effective, semantic closeness and lexical frequency play a weak role. By contrast, contextual predictability does not seem to have a consistent effect across languages.

In each figure, for better representation, statistical significance is indicated by colors: red triangle represents the factor in question has a significant positive effect; green square indicates the factor has a significant negative effect; blue dot means the factor has no effect. 95% confidence intervals for each factor were derived from their respective posterior distributions in the Bayesian regression.

Figure 1: Coefficients for the four factors in languages with only preverbal PP orderings. We included Hindi due to its typologically distinct features, yet without calculating its effect of contextual predictability due to limited training data.

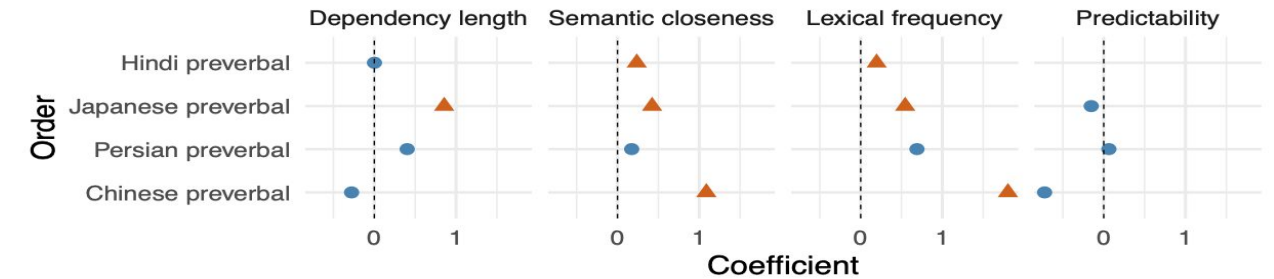


Figure 2: Coefficients for the four factors in languages with only postverbal PP orderings.

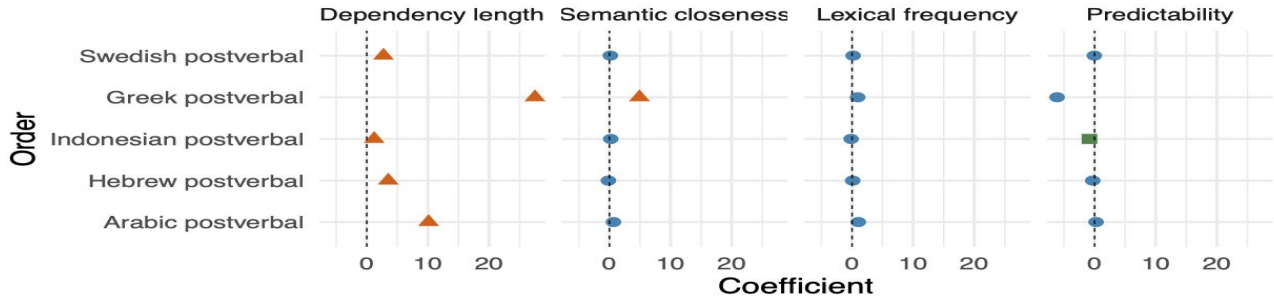
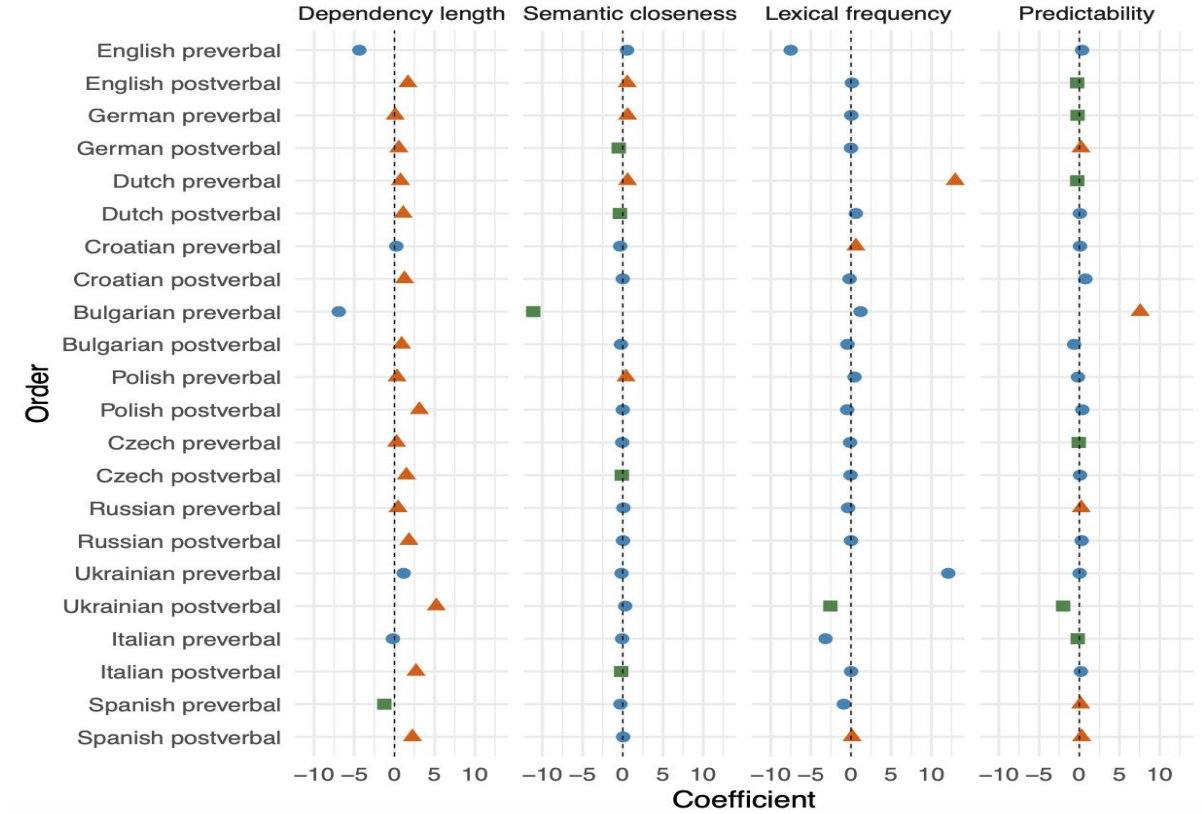


Figure 3: Coefficients for the four factors in languages with both preverbal and postverbal PP orderings.



The dual nature of subjecthood: Unifying subject islands and *that*-trace effects

Rebecca Tollan & Bilge Palaz (University of Delaware)

Overview: Filler-gap dependencies have long been shown to be processed more easily when the gap is associated with the subject position than with object position (e.g., Holmes & O'Regan, 1981). Yet there are (at least) two environments where this 'subject advantage' flips: (i) a dependency *within* a subject NP ('subject island'; Ross, 1967), and (ii) dependency of a subject NP in an embedded clause with a complementizer ('that-trace effect'; Perlmutter, 1968): both cause ill-formedness as in (1) and (2) respectively (yet their object counterparts don't).

(1) *Which car did [the color of __] please Jo? (2) *Which car did Mary think [that __ pleased Jo]?
We propose a unified account of (1) and (2), which we test with a series of rating experiments.

Background: Abeillé et al. (2020) argue for (1) that, as subjects are default topics, *wh*-focusing a sub-constituent of a subject as in (1) creates a topic-focus clash, giving rise to ill-formedness. Note, however, that this account overgeneralizes: when a *wh* filler is "*in situ*" (but still focused, as in "[The color of *which car*] pleased John?"), the result is grammatical despite the discourse clash. Therefore, ill-formedness of (1) must be specific to filler-gap dependency formation.

Proposal: We propose the COMBINATORIAL DISCOURSE ROLE HYPOTHESIS in (3):

(3) A dependency chain bears maximally one discourse role: a combinatoric of filler and gap.

In both (1) and (2), the parser cannot associate filler and gap with a combinatoric discourse role, because the gap position (i.e., subject NP in 1, *that*-clause in 2) falls within a constituent already bearing a unique discourse role itself ("topic" for subject NPs; see Rizzi, 1990 on discourse status of *that*-clauses). Thus, (3) cannot hold, and the non-resolved clash between topic status of the (subject) gap and focus status of the filler causes ill-formedness, as per Abeillé et al.

Experiment 1 offers evidence for our unified account of (1) and (2). We suggest that the well-known alleviation of that-trace effect by adverbials (Culicover, 1992) occurs because the presence of an adverb weakens the topic status of the embedded subject, so the penalty for violating (3) is milder. This predicts that adverbials should ameliorate *subject islands* as well. We ran a 2x2 grammaticality rating study (crossing presence of adverbial with gap location, see A1) via Amazon's *Mechanical Turk* (32 subjects). Mean z scores are shown in Figure 1. A linear mixed-effects model revealed significant main effects and interaction (all $ps < .001$). Importantly, planned comparisons showed that presence of an adverbial significantly *increased* ratings for subject NPs ($t = 2.6$; $p = .01$), in line with our hypothesis, but *decreased* them for object NPs ($t = -4.2$; $p < .0001$), indicating that an adverbial does not increase grammaticality *generally*.

Experiment 2 tests a further prediction; since an embedded *that*-clause, unlike a null clause, already bears a unique discourse function, an object gap should also violate (3) and incur a penalty (albeit mild, since objects are default foci already). We ran a 2x3 study, crossing complementizer (*that*, null) with Q type (subject *wh*Q, object *wh*Q, Yes-NoQ; see A2). Figure 2 shows mean ratings. As expected, subject *wh that* Qs were rated worse than subject *wh* null Qs ($p < .0001$) and crucially, ratings were worse for object *wh that* Qs compared with object *wh* null Qs ($p = .0017$). This indicates a mild object *that*-trace effect, which is unpredicted by current generative syntactic accounts (Anti-Locality; Erlewine, 2020) or prosodic accounts of *that*-trace (Sato & Dobashi, 2016), but correctly predicted under the discourse approach we propose.

Experiment 3 follows Abeillé et al., testing whether *that*-t effects are weaker inside relative clauses compared to *wh* questions (A3). We conducted three further ratings studies, comparing subject *that*-t RCs (Exp 3a: restrictive RCs; Exp 3b and 3c: non-restrictive RCs) with subject *that*-t *wh*Qs (Exp 3a and 3b: matrix *wh*Qs; Exp 3c: embedded *wh*Qs), crossing dependency type with complementizer type in a 2x2 design. Results (Fig. 3) showed that subject *that*-t violations are rated better in RCs vs. in *wh*Qs (planned comparison $t = -2.1$, $p = .03$) when the ratings for null *wh* Qs vs. null RCs are equal ($t = .035$, $p = .97$), consistent with an account in which the function of the construction (i.e., the filler is focused as in *wh*Qs but not in RCs) impacts ratings.

Conclusion: Whereas the long-attested "subject advantage" may arise from syntactic-semantic factors, we propose that all "anti-subject" effects (as in 1 and 2) arise from discourse factors.

Supplemental Materials

A1. Sample stimuli (Experiment 1)

- No Adverb, Subject NP gap:** Which car did [the color of _] delight Jo?
- Adverb, Subject NP gap:** Which car, according to rumor, did [the color of_] delight Jo?
- No Adverb, Object NP gap:** Which car did Jo adore [the color of_]?
- Adverb, Object NP gap:** Which car, according to rumor, did Jo adore [the color of_]?

A2. Sample stimuli (Experiment 2)

- Subject whQ:** Which family member did Lucy think {that/Ø} could drive grandad home?
- Object whQ:** Which family member did Lucy think {that/Ø} Kate could drive home?
- Yes-No Q (baseline):** Did Lucy think {that/Ø} Kate could drive grandad home?

Figures

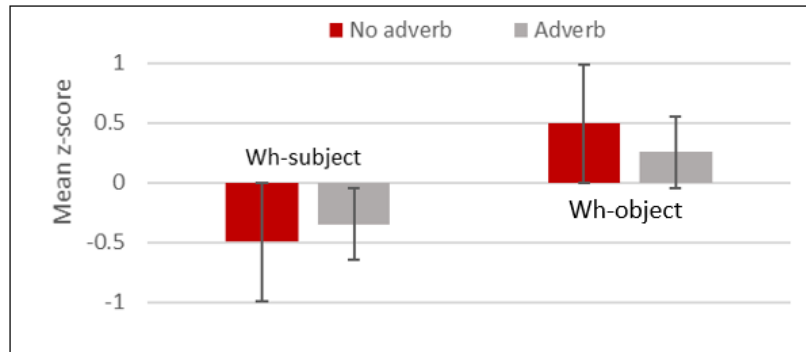


Figure 1. Results of Experiment 1

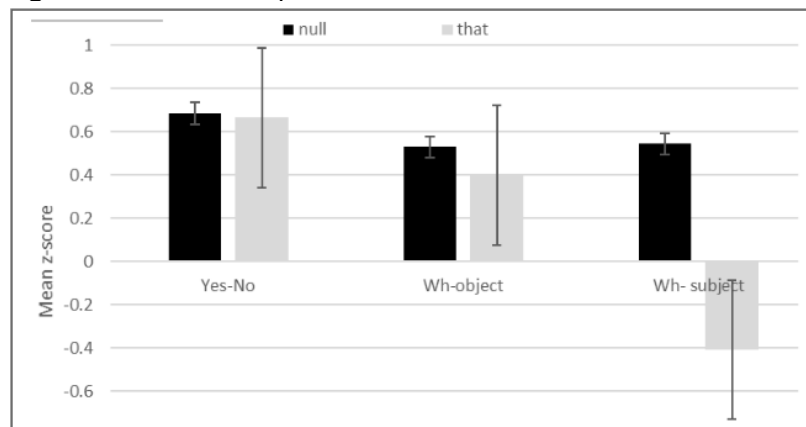


Figure 2. Results of Experiment 2

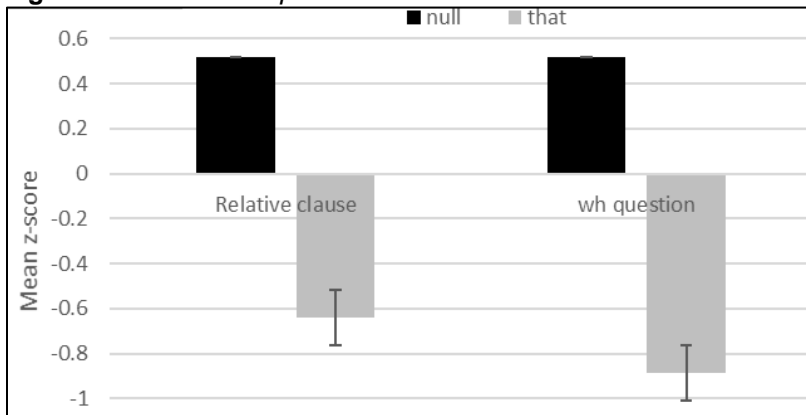


Figure 3. Experiment 3: Results for participants with equal null whQ-null RC baseline ratings

A3. Sample stimuli (Experiment 3)

- That whQ:** Which family member did Lucy think that could drive grandad home?
- Null whQ:** Which family member did Lucy think could drive grandad home?
- That RC:** The family member, who Lucy thought that could drive grandad home, knew Pat.
- Null RC:** The family member, who Lucy thought could drive grandad home, knew Pat.

References

Abeillé et al. (2020), *Cognition*;
 Culicover (1992). *NELS*
 proceedings; Erlewine (2020),
Glossa; Holmes & O'Regan
 (1981), *J. of Verbal Learning*
and Verbal Behavior;
 Perlmutter (1968), PhD thesis,
 MIT; Rizzi (1990), *Linguistic*
Inquiry monographs; Ross
 (1967), PhD thesis, MIT; Sato
 & Dobashi (2016), *Linguistic*
Inquiry.

Differential impacts of linguistic alignment across caregiver-child dyads and levels of linguistic structure

Ruthe Foushee¹ (foushee), Dan Byrne¹ (djbyrne), Marisa Casillas^{1,2} (mcasillas), & Susan Goldin-Meadow¹ (sgsg)

¹ University of Chicago, Chicago, IL, USA (@uchicago.edu)

² Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Keywords: *linguistic alignment* | *caregiver-child interaction* | *language development*
individual differences | *computational linguistics* | *dialogue*

In conversation, we tend to re-use each others' words, phrases, and structures, and become increasingly similar in our pronunciation and rate of speech. This process of *linguistic alignment* has been proposed to play a role in communicative success (Pickering and Garrod, 2004), and, recently, language acquisition: caregivers' alignment to their young interlocutors might reflect their 'tuning' of their child-directed speech (Denby and Yurovsky, 2019; Yurovsky et al., 2016), and even directly promote language learning by facilitating children's real-time speech processing and production. Alignment has typically been studied in adult conversation; however, recent analyses have used (largely cross-sectional) child language corpora to show that children align less than their caregivers, but more with age (Misiek et al., 2020), and to provide early evidence that syntactic alignment predicts vocabulary development (Denby and Yurovsky, 2019). Here, we capitalize on longitudinal data ideally suited to test (1) the robustness of these trends within individual dyads, and (2) the claim that alignment is broadly supportive of language development, by examining the relation between directional caregiver-child alignment at multiple levels of linguistic structure, and child vocabulary outcomes.

Our data represent 90-minute transcripts recorded every four months in the homes of 65 caregiver-child dyads, between the ages of 14 and 58 months (12 transcripts/child; 780 total), along with three administrations of the Peabody Picture Vocabulary Test (PPVT; Dunn and Dunn, 1981) at 30, 42, and 54 months. We quantified linguistic alignment at three levels of structure: LEXICAL, SYNTACTIC, and SEMANTIC. LEXICAL alignment reflected the proportion of shared words between speaker turns, while SYNTACTIC alignment measured the ratio of shared part-of-speech tags. SEMANTIC alignment was calculated by computing the similarity between adjacent utterances, represented in a high dimensional vector space (spacy2; Honnibal and Montani, 2017).

We first fit linear mixed effects models (Bates et al., 2015) to the alignment data for each level separately, including child age, speaker (CHILD/PARENT), the interaction of age and speaker, child sex, and maternal education as predictors. As expected, children aligned less than adults at the LEXICAL ($\beta = -0.02$,

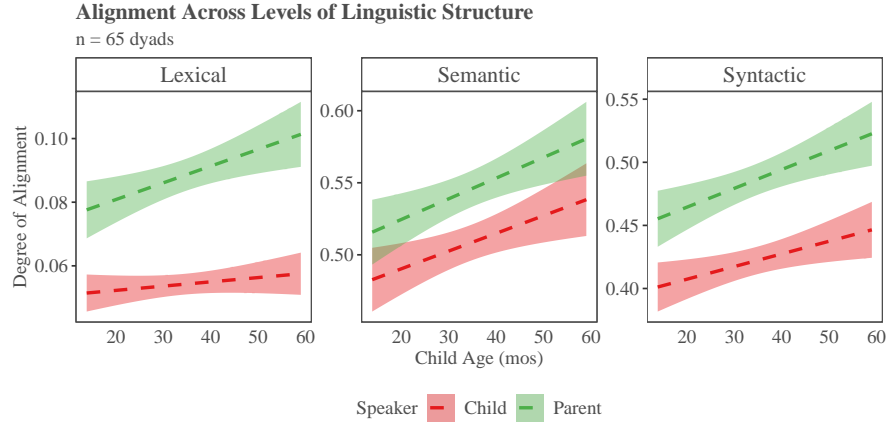


Figure 1: Developmental trajectories of child- and parent alignment at three levels of linguistic structure, across 65 dyads in the Language Development Project dataset (Goldin-Meadow et al., 2014).

$\chi^2(1) = 139.92, p < .001$) and SYNTACTIC levels ($\beta = -0.05, \chi^2(1) = 72.01, p < .001$), and overall LEXICAL ($\beta = 0.001, \chi^2(1) = 4.33, p < .05$) and SYNTACTIC ($\beta = 0.002, \chi^2(1) = 4.51, p < .05$) alignment within dyads increased reliably with age (see Figure 1). To evaluate the hypothesis that increased alignment might itself promote language development, we predicted children’s PPVT scores from caregiver and child alignment, controlling for demographic variables known to correlate with vocabulary outcomes, including maternal education and child sex. Remarkably, caregivers’ levels of LEXICAL ($\beta = 261.00, \chi^2(1) = 22.90, p < .001$) and SYNTACTIC ($\beta = 129.00, \chi^2(1) = 14.46, p < .001$) alignment were significant predictors of children’s PPVT scores, while neither caregiver SEMANTIC alignment ($\beta = 69.50, \chi^2(1) = 1.24, p = .26$), nor *children’s* tendency to align to their parents at any level were significantly related to their vocabulary scores.

Together, our results are consistent with proposals that alignment plays a causal role in advancing language development, but that its impact may differ across levels of linguistic structure — a question left open by previous research. Specifically, our results suggest that *lexical* and *syntactic* alignment, which reflect caregiver’s re-use of children’s immediately preceding words and sentence structures, may promote learning.

References

- Bates, D., Mächler, M., B., B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

- Denby, J., & Yurovsky, D. (2019). Parents' linguistic alignment predicts children's language development. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.
- Dunn, L. M., & Dunn, L. M. (1981). Peabody Picture Vocabulary Test-Revised.
- Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, 69(6), 588–599.
- Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- Misiek, T., Favre, B., & Fourtassi, A. (2020). Development of multi-level linguistic alignment in child-adult conversations. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 54–58.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02), 169–190.
- Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 2093–2098.

34th Annual CUNY Conference on Human Sentence Processing

Saturday March 6, 2021

Session	Time	Type	Title	Authors
1	9:00	Talk	How do people interpret implausible sentences?	Zhenguang Cai, Nan Zhao and Martin Pickering
1	9:30	Talk	New neighbours make bad fences: Form-based semantic shifts in word learning	David A. Haslett and Zhenguang G. Cai
1	10:00	Talk	A random walk down the garden path: A new implementation of self-organized parsing	Garrett Smith
2	10:30	Break		
3	11:00	Invited Talk	Learning verb argument-structure: Syntax and statistics	Cynthia Fisher
3	11:45	Talk	Three-year-olds' comprehension of contrastive and descriptive adjectives: Evidence for contrastive inference	Catherine Davies, Jamie Lingwood, Bissera Ivanova and Sudha Arunachalam
4	12:15	Break		
5	12:30	Parallel Session	Link to Saturday Parallel Session	
6	14:30	Break		
7	15:00	Invited Talk	Natural Language Processing has been overrun by large neural language models! What should we make of that?	Christopher Manning
7	15:45	Talk	Anaphoric dependencies in the digital age: On the relation between emoji and text	Elsi Kaiser and Patrick Georg Grosz
8	16:15	Break		
9	16:45	Talk	Language Production Under Uncertainty: Advance Planning and Incrementality	Arella Gussow and Maryellen MacDonald
9	17:15	Talk	Number attraction in pronoun production: evidence for antecedent feature retrieval	Cassidy Wyatt, Margaret Kandel and Colin Phillips
9	17:45	Talk	A computational model of reference production based on listener visual-search costs	Julian Jara-Ettinger and Paula Rubio-Fernandez

How do people interpret implausible sentences?

Zhenguang G. Cai (Chinese University of Hong Kong), Nan Zhao (Baptist University of Hong Kong), Martin J. Pickering (University of Edinburgh)

People may literally interpret an implausible sentence (e.g., treating *the candle* as the recipient of *the daughter* in *The mother gave the candle the daughter*) or re-interpret it (e.g., treating *the daughter* of the recipient) [1]. To arrive at a plausible re-interpretation, they might resort to *structural reanalysis* by revising their representation of its syntax and using that representation to derive its new interpretation (e.g., revising the sentence into *The mother gave the candle to the daughter*) [1-4]. Alternatively, they might resort to *semantic reanalysis* and revise its semantic representation directly (e.g., swapping the thematic roles of *the candle* and *the daughter*) [5,6]. We report two structural priming experiments to distinguish the two accounts. The structural reanalysis account predicts that participants represent re-interpreted POs as having DO syntax and re-interpreted DOs as having PO syntax; therefore, priming should be reduced following implausible than plausible primes. In contrast, the semantic reanalysis account does not have such a prediction.

In E1 (96 participants, 20 target items, 60 fillers), participants heard double-object (DO) or prepositional-object (PO) sentences that were plausible or implausible and answered a comprehension question (so that it was clear whether they reinterpreted the sentences or not; see Fig 1; cf. [4]).

Plausible DO/PO: *The mother gave the daughter the candle / the candle to the daughter.*

Implausible DO/PO: *The mother gave the candle the daughter / the daughter to the candle.*

Then they described a dative event (e.g., a pirate handing a boxer a cake). Question answering showed that participants re-interpreted plausible DO and PO 10% and 4% and implausible DO and PO 48% and 23% of the time, replicating earlier results [1]. LME modelling of picture descriptions (Table 1) shows that the structural priming was modulated by plausibility, with reduced priming following implausible than plausible primes, suggesting that implausible primes were somehow structurally reanalysed. In addition, priming was also reduced following a re-interpreted than literally-interpreted implausible primes, suggesting a greater extent of structural reanalysis when people re-interpreted than literally interpreted an implausible sentence. Indeed, a re-interpreted implausible prime led to reversed priming (e.g., numerically more PO descriptions following a re-interpreted than literally interpreted implausible DO prime).

Is it possible that participants are triggered to reinterpret by the comprehension question itself? To investigate this issue, E2 (96 participants, 20 target items, 60 fillers) had participants describe the picture before answering the comprehension question. Again, there was reduced priming following implausible than plausible primes, though here priming following implausible primes was comparable following (later) literally-interpreted and re-interpreted implausible primes. A between-experiment analysis showed some marginal evidence that structural priming was reduced following re-interpreted than literally-interpreted implausible primes in E1 but not E2.

The findings suggest that people consider a revised structure when interpreting an implausible sentence, resulting in reduced priming following implausible than plausible primes in both experiments. Note that such a result would not be expected if people only swapped the semantic roles of the two nouns in re-interpreting implausible sentences. There is also some evidence that people also further commit to a revised structure when they explicitly re-interpret an implausible sentence.

REFERENCES

1. Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). *PNAS*, 110(20), 8051-8056.
2. Kim, A., & Osterhout, L. (2005). *J Mem Lang*, 52(2), 205-225.
3. Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). *J Mem Lang*, 33, 285-318.
4. Christianson, K., Luke, S. G., & Ferreira, F. (2010). *JEP: LMC*, 36, 538.

5. Kuperberg, G. R. (2007). *Brain research*, 1146, 23-49.
6. Bornkessel, I., & Schlesewsky, M. (2006). *Psychol Rev*, 113, 787-821.

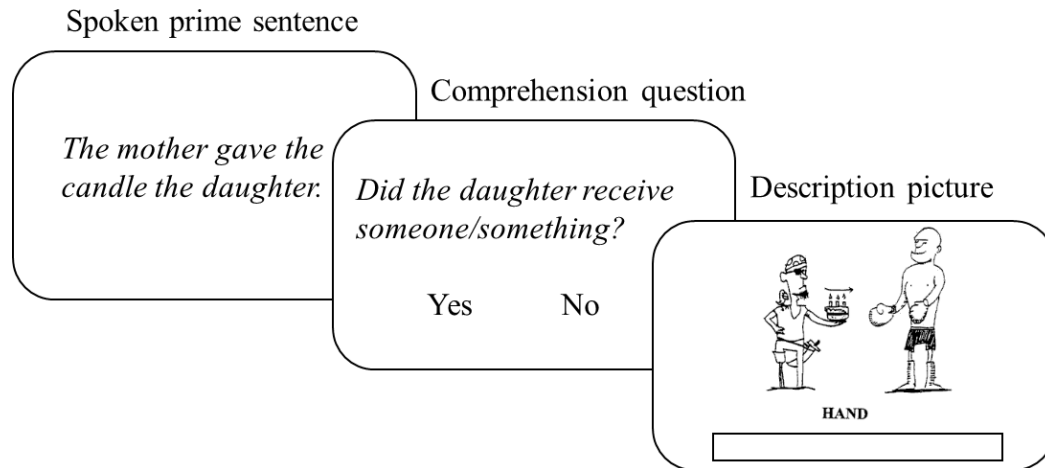


Fig 1. Trial structure in Experiment 1.

Table 1. DO, PO and “other” responses as a function of plausibility, interpretation, and structure in Experiment 1. Priming refers to difference in the proportion of DO responses between DO and PO primes.

			DO	PO	Other	Prop DO	Priming
Plausible	Literally interpreted	DO	122	247	27	0.33	0.16
		PO	65	326	30	0.17	
	Re-interpreted	DO	6	29	9	0.17	0.03
		PO	2	12	5	0.14	
Implausible	Literally interpreted	DO	56	155	20	0.27	0.04
		PO	72	238	31	0.23	
	Re-interpreted	DO	41	146	22	0.22	-0.01
		PO	21	69	9	0.23	

Table 2. DO, PO and “other” responses as a function of plausibility, interpretation, and structure in Experiment 2.

			DO	PO	Other	Prop DO	Priming
Plausible	Literally interpreted	DO	119	206	32	0.37	0.12
		PO	91	273	23	0.25	
	Re-interpreted	DO	15	39	4	0.28	0.11
		PO	4	19	5	0.17	
Implausible	Literally interpreted	DO	79	130	15	0.38	0.10
		PO	79	199	25	0.28	
	Re-interpreted	DO	44	133	14	0.25	0.05
		PO	20	79	13	0.20	

New neighbours make bad fences: Form-based semantic shifts in word learning

David A. Haslett & Zhenguang G. Cai (The Chinese University of Hong Kong)

Words sometimes shift in meaning towards other words that are similar in form. For example, *expunge* is etymologically related to *puncture* but now tends to refer to wiping away, and according to the *Oxford English Dictionary*, this shift “is probably influenced by phonetic association with *sponge*”. The OED has identified over 70 likely cases of such form-based semantic shifts while overlooking or leaving unacknowledged many more, and the *Oxford Guide to Etymology* recognizes similarity of form as a motivation for semantic change (Durkin, 2009), although these changes have also been dismissed as irregular (e.g., Traugott & Dasher, 2001). In recent years, corpus studies have found that words sound similar to words that are similar in form at above-chance levels in 100 languages (Dautriche et al., 2017), and in English this cannot be attributed to etymological relationships and is true throughout the lexicon, not only in pockets of sound symbolism (Monaghan et al., 2014). This subtle correspondence between form and meaning might shape the lexicon to facilitate learning (Kirby et al., 2015), and iterated learning experiments have indeed demonstrated that word forms can converge due to similarity of meaning (Silvey, Kirby & Smith, 2015). However, there is as of yet no experimental evidence for the inverse: that word meanings can converge due to similarity of form.

We therefore conducted two novel word learning experiments, implemented on Qualtrics.com, with 30 items and 60 participants each (native English speakers recruited from Prolific.co), manipulated within subject and within item. Each novel word is either similar in form to an existing “attractor” word or not and is initially presented in a sentence context that implies a meaning that conflicts with the attractor word’s meaning. For example, participants inferred the meaning of either *tormest* or *plonch* from the sentence *The firefighters tormested / plonched the child from the burning building*. The sentence implies the meaning of *rescue*, as confirmed by a cloze test pretest, and the novel target *tormest* is an orthographic neighbour of the attractor word *torment*, whereas the novel control *plonch* has no orthographic neighbours and was generated by the ARC Nonword Database (Rastle, Harrington & Coltheart, 2002). Participants then read an ambiguous (low-cloze) sentence containing the same novel word (e.g., *Chen was tormested / plonched*) and answered a comprehension question by giving a rating on a 7-point scale (e.g., *How thankful was Chen? 1 - Not at all; 7 - Very*). The implied word (*rescue*) elicited low ratings, and the attractor word (*torment*) elicited high ratings (or vice versa in half the items, inverted for analysis). Participants gave low ratings for both novel words, like for the implied word, but as predicted, the novel target (*tormest*) elicited slightly higher ratings than the novel control (*plonch*), indicating that the inferred meaning of the novel target shifted towards the meaning of the attractor word. Experiment 2 required participants to recall and spell the novel word, demonstrating that they had not confused it for the attractor word. Linear mixed effects modelling shows that this difference is significant in both experiments.

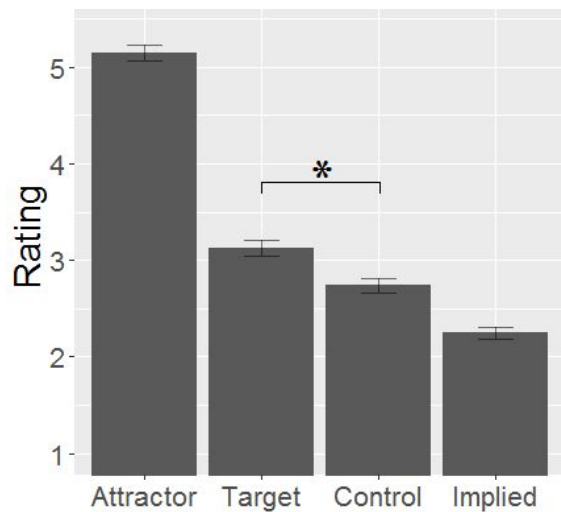
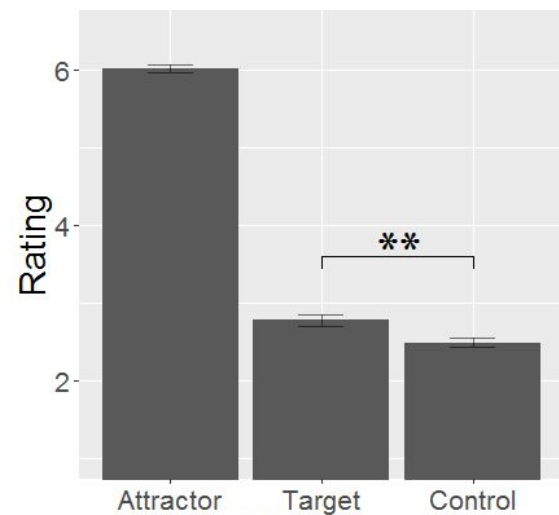
These experiments support the corpus finding that words sometimes shift in meaning towards words that are similar in form, providing evidence that this type of semantic change is regular. The results also suggest that the clustering of form and meaning in the lexicon arises partially as a consequence of how words are learned, which is consistent with the theory that language evolves via learning to constrain arbitrariness and thereby facilitate transmission (e.g., Kirby et al., 2015). Form-based semantic shifts are explicable in terms of the complementary learning systems account of word learning, in which novel words continue to phonologically prime existing words until overnight consolidation, when lexical competition emerges (Davis & Gaskell, 2009). The meanings of newly learned words could in this way be influenced by similar-sounding words following initial exposure, prior to sleep. However, form exerts only a small influence on meaning in these experiments (and across the lexicon), which is to be expected, given that words must be learned primarily according to context (lest communication break down) and that language also evolves to preserve arbitrariness (Kirby et al., 2015).

Table 1. Comparison among word types in Experiment 1

Comparison	β	SE	z	p
Attractor - Implied	2.92	0.24	12.01	< .001
Attractor - Target	-2.01	0.23	-8.67	< .001
Attractor - Control	2.39	0.22	10.74	< .001
Implied - Target	0.91	0.18	5.03	< .001
Implied - Control	-0.53	0.17	-4.93	< .001
Target - Control	0.38	0.15	2.51	.012

Table 2. Comparison among word types in Experiment 2

Comparison	β	SE	z	p
Attractor - Target	-3.25	0.22	-14.79	< .001
Attractor - Control	3.52	0.21	16.43	< .001
Target - Control	0.26	0.09	3.05	.002

**Fig. 1:** Ratings by word type in Experiment 1**Fig. 2:** Ratings by word type in Experiment 2

Works Cited

- Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41, 2149–2169.
- Davis, M. H. & Gaskell M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B*, 364, 3773–3800.
- Durkin, P. (2009). *Oxford guide to etymology*. Oxford UP.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102.
- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B*, 369, 1–12.
- Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology Section A*, 55(4), 1339–1362.
- Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39, 212–226.
- Traugott, E. & Dasher, R. (2001). *Regularity in semantic change*. Cambridge UP.

A random walk down the garden path: A new implementation of self-organized parsing

Garrett Smith (Universität Potsdam)

gasmith@uni-potsdam.de

Models of human sentence comprehension typically assume that the parses people build during word-by-word language understanding are globally consistent with the grammar of their language: Only structures that follow all the rules of the grammar are considered as (partial) parses of a string of words. These models have been widely successful in explaining how people parse sentences (Levy, 2008; Lewis & Vasishth, 2005). However, local coherence effects, where locally viable but globally ungrammatical structures seem to compete with grammatical ones, present a challenge for traditional theories of human sentence comprehension (Tabor, Galantucci, & Richardson, 2004). An alternative theory, based on principles of self-organization, can explain these effects in a natural way. Under self-organization, words assemble themselves into larger syntactic structures according to violable constraints via purely local interactions. There is no global consistency checking; nevertheless, grammatical parses typically emerge on their own. Despite the theoretical innovation of previous self-organizing models, their implementations have suffered from opaque mathematical formalisms and limited coverage of empirical phenomena (e.g., Kempen & Vosse, 1989; Smith & Tabor, 2018). We present a new implementation, called *mparse*, that shows promise for overcoming these issues and making self-organization a more broadly and easily testable theory.

Mparse applies the master equation—used in chemistry and physics to describe continuous-time, discrete-state random walks (Oppenheim, Shuler, & Weiss, 1977)—to model human sentence comprehension. This is how it works: At each word in a sentence, mparse enumerates all possible partial and complete parses that are possible given the words so far and a grammar of binary dependency relations between words. These parse states include both merely locally viable structures and globally grammatically ones. The model jumps stochastically between parse states that differ only by a single dependency link (nearest neighbors), with jumps to more well-formed states being more probable than jumps to less well-formed states. A noise parameter controls how much the model prefers well-formed states over ill-formed ones (low noise = strong preference for well-formed states, high-noise = well- and ill-formed states treated equally). Well-formedness is penalized if a state has too few dependency links, includes longer dependencies, and/or includes word order violations. Reading times are modeled as the amount of time it takes mparse to reach a state with the maximum possible dependency links for the number of words so far (up to $w - 1$ links for w words). Once mparse reaches such a state, it stops processing the current word, inputs the next word, adds new states based on the syntactic affordances of the new word, and resumes the random walk among the states. The master equation formalism offers powerful tools for understanding incremental sentence parsing and making detailed, quantitative predictions. Importantly, mean reading times for each word in a sentence can be calculated easily.

We tested mparse on local coherence effects (1) and the contrast between two types of garden path effects (2) in English (Sturt, Pickering, & Crocker, 1999). As shown in Fig. 1 (left), mparse correctly produces disproportionately slow mean reading times for ... *at the player tossed*... from Tabor et al. (2004). It also correctly produces larger garden path effects (ambiguous - unambiguous) for NP/Z materials than NP/S materials (Fig. 1, right, Sturt et al., 1999).

These results demonstrate that this implementation of self-organization produces reading time predictions in line with existing experimental results. The proof-of-concept results presented here, though, barely scratch the surface of the information that can be gleaned from mparse's mathematical formalism. Future work will explore how the mathematical theory behind mparse can drive new empirical work. Work is also underway for extracting mparse's grammar from large, parsed corpora, opening the door to truly broad-coverage comparisons with competing models like surprisal (Levy, 2008) and cue-based retrieval (Lewis & Vasishth, 2005).

- (1) The coach smiled at the player [who was] [tossed/thrown] the frisbee.
- (2)
 - a. NP/S: The woman saw [that] the doctor had been drinking.
 - b. NP/Z: Before the woman visited[,], the doctor had been drinking.

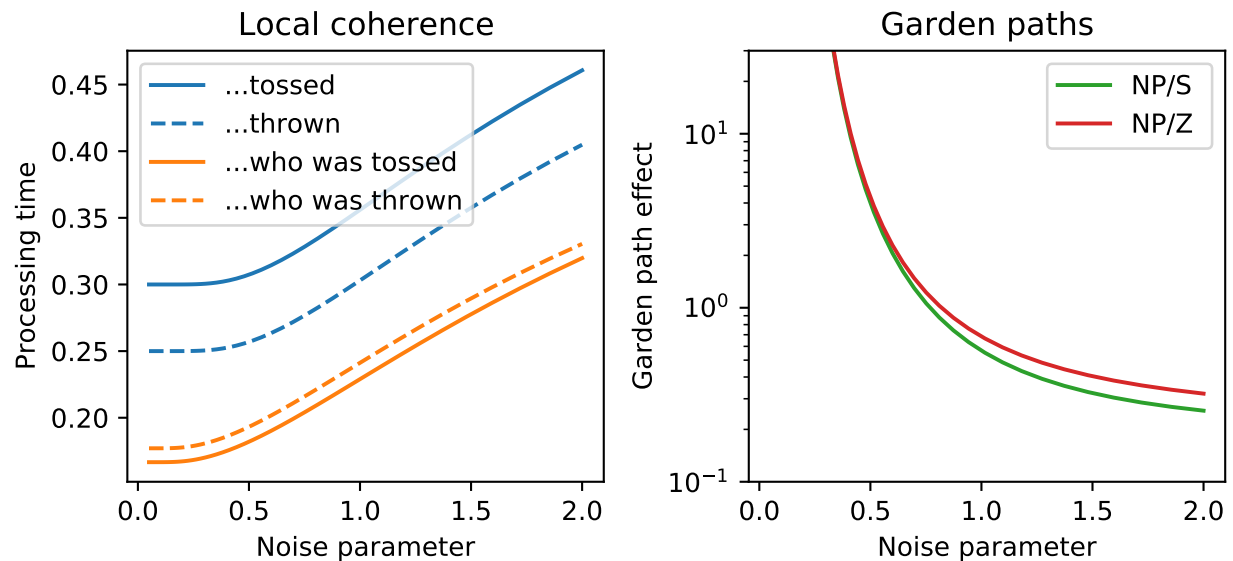


Figure 1: Mean processing times (arbitrary units) at *tossed/thrown* in (1) (left) and mean garden path effects at *had* in (2) (right). The garden path effects are the difference between the ambiguous and unambiguous conditions in (2). Note the logarithmic y-axis in the right panel. As the noise level decreases, the size of both garden path effects explodes because the probability of jumping from a relatively well-formed garden-path state to an ill-formed state intermediate between the garden path and the correct parse decreases rapidly, making repairing the garden path nearly impossible.

References

- Kempen, G., & Vosse, T. (1989). Incremental syntactic tree formation in human sentence processing: A cognitive architecture based on activation decay and simulated annealing. *Connection Science*, 1(3), 273–290.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 375–419.
- Oppenheim, I., Shuler, K. E., & Weiss, G. (1977). *Stochastic processes in chemical physics: The master equation*. MIT Press.
- Smith, G., & Tabor, W. (2018). Toward a theory of timing effects in self-organized sentence processing. In I. Juvina, J. Hout, & C. Myers (Eds.), *Proceedings of the 16th International Conference on Cognitive Modeling* (pp. 138–143). Madison, Wisconsin: University of Wisconsin.
- Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40, 136–150.
- Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4), 355–370.

Learning verb argument-structure: Syntax and statistics

Cynthia Fisher (University of Illinois at Urbana-Champaign)

Learning a language requires balancing lexical and abstract knowledge, to learn patterns ranging from the idiosyncrasies of individual words to structures that can be generalized to almost any word. In this talk I will discuss the role of syntactic-distributional learning about verbs in creating and maintaining this balance. I will present evidence that, both early in acquisition and in the ongoing adaptation of the linguistic system, distributional learning creates probabilistic syntactic-semantic combinatorial knowledge about verbs and verb classes. This knowledge plays two roles: (a) it permits syntactic bootstrapping, as children use each verb's combinatorial behavior in sentences to help identify its meaning; and (b) it supports sentence processing (this is known as *verb bias*), by reducing ambiguity in online comprehension, and guiding sentence production.

Three-year-olds' comprehension of contrastive and descriptive adjectives: Evidence for contrastive inference.

Catherine Davies (University of Leeds), Jamie Lingwood (Liverpool Hope University), Bissera Ivanova (Aix-Marseille University), Sudha Arunachalam (New York University).

Combining information from adjectives with the nouns they modify is essential for comprehension. Previous research suggests that preschoolers do not always integrate adjectives and nouns, and may instead over-rely on noun information when processing referring expressions [1; 2]. This disjointed processing has implications for pragmatics, apparently preventing under-fives from making contrastive inferences.

Two visual world experiments investigated how English-speaking three-year-olds ($N=73$, $M_{age}=44$ months) process size adjectives across syntactic (prenominal; postnominal) and pragmatic (descriptive; contrastive) contexts (Fig. 1). The first experiment used an established paradigm [3] and the second used a novel experimental design that allowed children time to demonstrate their abilities in adjective-noun integration and in contrastive inference. We asked:

1. Do 3-year-olds integrate adjectives and nouns to resolve reference by utterance end?
2. Do 3-year-olds show contrastive inference?
3. Do 3-year-olds process modified noun phrases more quickly when adjectives occur pre- or post-nominally?
4. Is there an association between 3-year-olds' contrastive inferencing ability and their language ability or speed of processing?

Using growth curve analysis [4] (and replicated with logistic regression), we show that preschoolers are able to integrate adjectives and nouns to resolve reference accurately by the end of the referring expression in a variety of pragmatic and syntactic contexts and in the presence of multiple distractors (RQ1). Crucially, by modelling the effect of pragmatic function (contrastive - where the prenominal adjective was informative, vs. descriptive - where it was not) on visual preference for the target object during the unfolding utterance, we reveal for the first time that when task demands are reduced (exp. 2), 3-year-olds show a stronger target preference during the adjective in the contrastive condition and greater distraction from the property competitor in the descriptive condition (Fig. 2; upper panel). Using both manifestations of contrastive inference, we conclude that young children can contrastively infer, given a slowed-down speed of presentation and visually enhanced size contrasts (RQ2; exp. 2). Against our hypothesis that participants would resolve reference more quickly when adjectives appear postnominally [5], we find no effect of syntactic frame (RQ3). Finally, correlational analyses reveal no association between preschoolers' contrastive inferencing ability and their semantic and syntactic abilities, or their speed of processing (RQ4).

Our findings provide novel evidence for a continuity in young children's pragmatic development. By analysing high-resolution online data in response to stimuli that require integration of an adjective with a noun, in younger children than have been tested before, we show that children can coordinate lexical, referential, and pragmatic information to interpret language in real time. We discuss mechanisms driving this coordination, and their relationship to task demands.

References

1. Fernald, A., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjective–noun phrases. *Cognitive Psychology*, 60(3), 190-217.
2. Thorpe, K., Baumgartner, H., and Fernald, A. (2006). Children's developing ability to interpret adjective–noun combinations. In *Proceedings of the 30th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadia Press

3. Huang, Y., & Snedeker, J. (2013). The use of referential context in children's on-line interpretation of scalar adjectives. *Developmental Psychology*, 49, 1090-1102.
4. Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R*. Boca Raton, FL: Chapman and Hall / CRC Press.
5. Ninio, A. (2004). Young children's difficulty with adjectives modifying nouns. *J Child Language*, 31, 255–285.

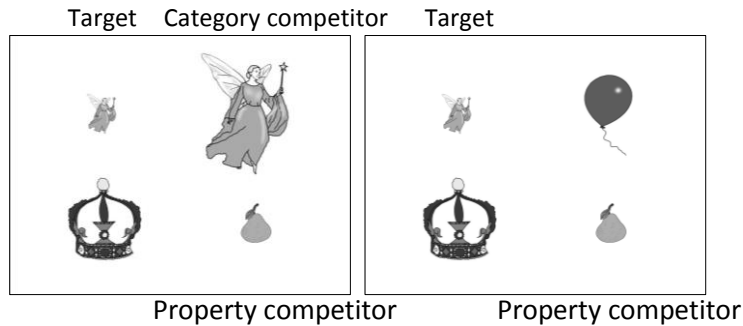


Figure 1. Pragmatic context was manipulated using contrastive (left panel) and descriptive (right panel) visual arrays, crossed with prenominal and postnominal syntactic frames presented auditorily, e.g., *Where's the little fairy?* / *Where's the fairy that's little?*

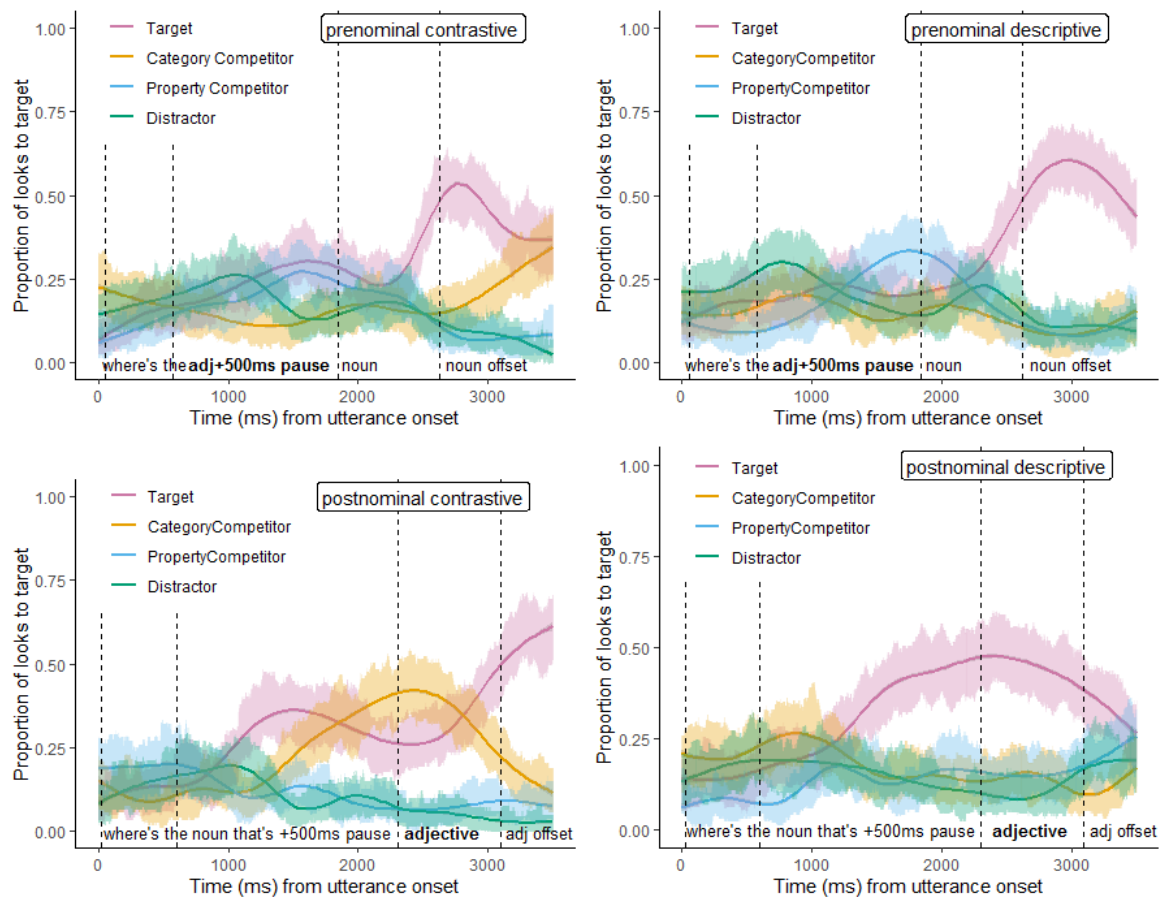


Figure 2. Proportion of looks to each interest area across syntactic and pragmatic conditions. Vertical dashed lines represent mean onset times. Bold text indicates disambiguation points.

Natural Language Processing has been overrun by large neural language models! What should we make of that?

Christopher Manning (Stanford University)

In Natural Language Processing, the long dominant way of using the structure of human languages in systems for various downstream tasks was by building context-free grammar or richer parsers from hand-annotated morphosyntactic resources that display linguistic structure, that is, treebanks. However, recent deep learning language models are simply large artificial neural networks trained in a self-supervised fashion to predict a masked word in a given context. Nevertheless, once fine-tuned, these models yield much better task performance seemingly without any structural knowledge. What is a right-thinking (psycho)linguist meant to think of this? I first consider recurrent neural network models and introduce the notion of bounded hierarchical languages, showing that RNNs can generate such languages with optimal memory. I then examine how deep contextual language models like BERT learn knowledge of linguistic structure because it helps them in word prediction. Using a new method for identifying linguistic hierarchical structure emergent in artificial neural networks, I show how components in these models focus on syntactic grammatical relationships and anaphoric coreference, and, moreover, there seems to be significant shared cross-linguistic structure, or a kind of Universal Grammar. These results both help explain why recent neural models have brought such large improvements across many language-understanding tasks and provide intriguing hints about the possibility of learning language from observed evidence alone, as human children appear to do.

Anaphoric dependencies in the digital age: On the relation between emoji and text

Elsi Kaiser (University of Southern California) and Patrick Georg Grosz (University of Oslo)

Emoji are widely used [2], but have received relatively little attention in psycholinguistics [4,6]. Regardless of one's views about the linguistic status of emoji, readers presumably construct some link between emoji and text. Thus, emoji offer a new window into dependency formation. Based on two studies on emoji-text relations, we argue for (at least) two types of emoji-text dependencies, and explore initial steps to integrate emoji into language processing theories.

Referential dependencies in language include (i) the dependency between a pronoun (or another form) and the individual that it refers to, and (ii) the dependency between an expressive (e.g. *damn*, *f*cking*) and the individual whose opinion it expresses [1,7,10,11]. We extend discussion of dependencies to emoji: We investigate **face emoji** which convey affective information (e.g. 😊, 😐, 😞) and non-face object-related/action-related emoji (e.g. 🍪, 🍷, 🍰); we call these **action emoji**. *We hypothesize both face and action emoji involve anaphoric dependencies (i.e. can be linked to linguistic content), but in different ways:*

We propose **face emoji** resemble expressives (e.g. *damn*), in that they tend to be interpreted as expressing the opinion of a salient **experiencer** (the person experiencing the emotion expressed by the face emoji or the expressive word). This experiencer is typically, but not always, the 1st-person speaker [1,7,10]. In contrast, we propose **action emoji** are interpreted based on principles of discourse coherence (e.g. relations like Explanation [9]), potentially akin to coherence-based accounts of pronoun resolution (see [9], Tables 1-2).

Exp1-2 presented participants (56 L1 English speakers/exp) with text messages with emoji (32 targets, 20 fillers). In Exp1, people indicated who the emoji provides information about (Fig.1). Exp2 was identical but the question for *face emoji* was reworded to ensure an opinion-based response (Fig.2). The three relevant referents/individuals are the message sender (i.e. 1st-person) and the people mentioned in the message (subject and object, see Table 1).

Verbs. To test whether we see discourse coherence effects (similar to those seen on pronoun resolution) on the interpretation of **action emoji**, we tested *transfer verbs* and *two kinds of implicit causality verbs* [3,5,8]: Stimulus-Experiencer (SE) (exp=obj) and Exp-Stim (ES, exp=sub, Table 1). Using both transfer and SE/ES verbs also allows us to test if **face emoji** are akin to expressives, i.e. sensitive to the presence of experiencers in subject/object position.

Emoji. Messages ended in a face or action emoji (Table 1). Faces were compatible with all 3 candidates (sender/sub/obj; results confirm this). Action emoji with transfer verbs depicted transferred objects. Action emoji with IC verbs provided an explanation of the event (Table 2).

Results are in Figs.3-4. **Face emoji** with transfer verbs disprefer objects and prefer senders (Exp1: $p=.078$, Exp2: $p<.001$). The (1st-p) sender preference fits with our hypothesis that face emoji resemble expressives and tend to be interpreted as expressing the opinion of a salient experiencer, often the 1st-person. What about face emoji with IC verbs? Here, the *linguistically-expressed experiencer argument competes with the sender for the role of attitude-holder*. With SE verbs, presence of an experiencer object wipes out the sender preference and boosts the object. With ES verbs, the face emoji strongly prefer the subject (*experiencer*).

Action emoji with transfer-verbs prefer the subject, disprefer the sender and object in both Exp1-2: A depicted object-of-transfer is interpreted as associated with the subject. This fits with the observation that (agentive) subjects are prominent in discourse. Action-emoji with IC verbs in both Exp1-2 show exactly the patterns we expect if action emoji are interpreted based on discourse coherence, perhaps akin to the domain of reference resolution: the explanation-providing emoji is interpreted as linked to the subject with SE, object with ES. (Note that other interpretations are in principle possible, (4c), as with pronouns, but people disprefer them.)

Our results point to two kinds of emoji-text relations, reflected by action vs. face emoji (maybe affective emoji generally; 🍪, 🍷). We suggest these two relations resemble existing linguistic dependencies, suggesting a need for more work on emoji in sentence comprehension.

Examples			
Verb type		Action emoji	Face emoji
Transfer verbs		(1a) abigail brought dessert to emily 🍰	(1b) abigail brought dessert to emily 😊
Implicit causality verbs	Stimulus-experiencer (SE) verbs	(2a) richie annoyed adrian 🥁	(2b) richie annoyed adrian 😞
	Experiencer-stimulus (ES) verbs	(3a) daniel admires aaron 🏆	(3b) daniel admires aaron 😊

Table 1. (Both positive and negative face emoji and negative and positive IC verbs were used)

Implicit causality verbs	Stimulus-experiencer (SE) verbs	(4a) richie _{stim} annoyed adrian _{exp} 🥁 [possible linguistic paraphrase of emoji, not shown in experiment: because he _{ritchie} played the drums]
	Experiencer-stimulus (ES) verbs	(4b) daniel _{exp} admires aaron _{stim} 🏆 [because he _{aaron} won first prize]
	Other readings are also possible in principle:	(4c) richie _{stim} annoyed adrian _{exp} 🥁 [because he _{adrian} hates drums]

Table 2. Illustration of how emoji in IC verb conditions were chosen to provide explanations in line with verb bias (ES/SE verbs are known to elicit explanations about what the *stimulus* did)

daniel admires aaron 😊

The emoji provides information about ____.

☐ daniel

☐ aaron

☐ the sender of the message

Fig.1 Exp.1 sample item

tanya hates amy 😞

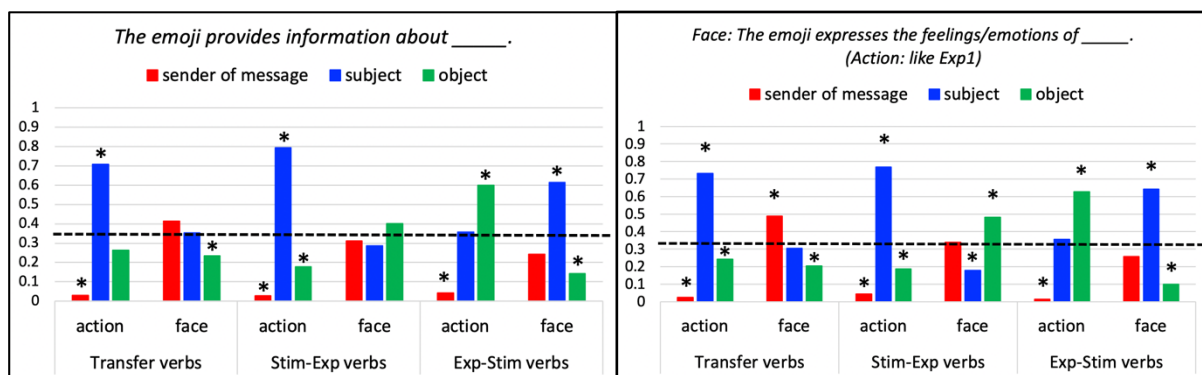
The emoji expresses the feelings/emotions of ____.

☐ tanya

☐ amy

☐ the sender of the message

Fig.2 Exp.2 sample item illustrating question used on face emoji trials (action trials were as in Exp.1)



References: [1] Amaral et al 2007 [2] Bai et al. 2019 [3] Bott & Solstad 2014 [4] Cohn et al. 2018 [5] Garvey & Caramazza 1974 [6] Gawne & McCulloch 2019 [7] Harris & Potts 2009 [8] Hartshorne & Snedeker 2013 [9] Kehler 2002 [10] Lasersohn 2007 [11] Potts 2007

Language Production Under Uncertainty: Advance Planning and Incrementality

Arella Gussow & Maryellen MacDonald (University of Wisconsin-Madison)

gussow@wisc.edu

Language production researchers typically investigate the process of utterance planning in situations where producers know their message. Less is known about a common occurrence in conversation, where A's message will depend on B's ongoing utterances. Although A may not yet know how to reply, some advance planning might be possible, in order to manage turn taking efficiently [1,2]. Prior studies suggest that incrementality, the degree to which planning precedes execution, is under some strategic control [e.g., 3]. Here we investigate the degree of advance planning under message uncertainty in two picture naming studies, permitting precise control over the timing of when the message becomes certain.

In Experiment 1, 64 native English speakers viewed displays showing two pairs of objects (see Figure 1). To avoid screen position effects on naming, the two images in each pair rotated around each other throughout a trial. Displays appeared in one of two conditions: 1) Overlap (Figure 1A), where one image appeared in both the left pair (e.g., vest, stool) and the right pair (e.g., vest, pear); or 2) Different (Figure 1B), where the left pair (e.g., wig, stool) had no overlap with the right pair (e.g., vest, pear). After 2.2 seconds of exposure, a gray background appeared behind one side of the screen, indicating the target pair (vest, pear in Figure 1). Participants' task was to answer the question "Which are the target images?" in a conjoined noun phrase (e.g., "the vest and the pear"), and they were free to name the two images in either order. Participants were told to respond as soon as possible, and that their recordings would later be used for another participant who would have to identify the targets. Dependent measures were the order of images named and the initiation latencies of all words in the noun phrase (automatically extracted by FAVE [4]). If speakers plan ahead while uncertain of the targets and thus their message, they should prioritize planning of elements common to either message when possible (Figure 1A). Such planning should yield tendencies to name the overlapping image in the Overlap condition first, with shorter initiation latencies in this situation compared to other outcomes.

Results: Figure 2 shows that in the Overlap condition, participants were more likely to place the overlapping target first in their response, suggesting they had planned the overlapping target in advance. Moreover, Figure 3 shows that overlap-first utterances in the Overlap condition had shorter initiation latencies than when the overlapping image was uttered last and all utterances in the Different condition, for which advanced planning was not possible. Exp. 2 replicated these results in an online experiment using typed responses (N= 84), indicating similar planning strategies in both spoken and typed productions (Figures 4-5).

These results show evidence of early planning and utterance initiation in the face of message uncertainty. Specifically, producers who are uncertain of their message tend to plan and produce portions of their utterance that are guaranteed to be useful, and they continue planning the rest incrementally. Initiation latencies in both studies (Figures 3 & 5) show that advance planning (overlap-first utterances) yields an initiation latency advantage throughout the entire utterance, emphasizing the benefits of early planning. More generally, these results suggest that including situations of production under uncertainty not only addresses a common conversational situation that is under-studied in the lab, but it could also inform theories of incremental planning during language production.

References

- [1] Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes*, 55(2), 230–240.
- [2] Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific reports*, 5, 12881.
- [3] Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, 46(1), 57-84.
- [4] Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) Program Suite v1. 2.2.

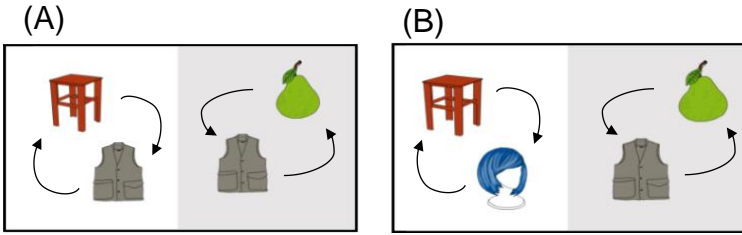


Fig. 1. Examples of visual displays in the (A) Overlap condition, (B) Different condition. Every two images rotated around each other as illustrated by the arrows (arrows did not appear during the experiment). The gray background appeared after 2.2 seconds of exposure, indicating the target images.

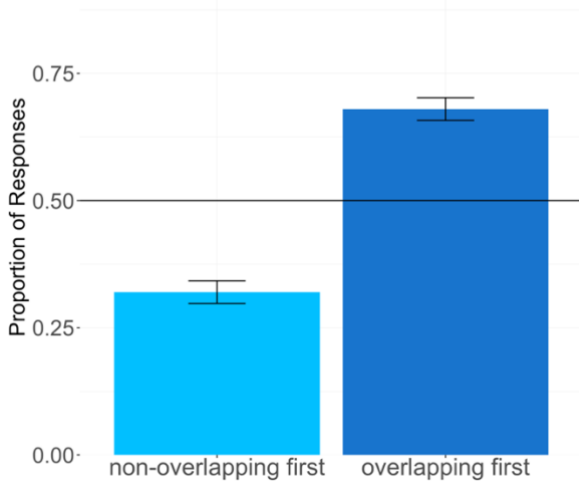


Fig. 2. Order choice in the Overlap condition in Exp 1.

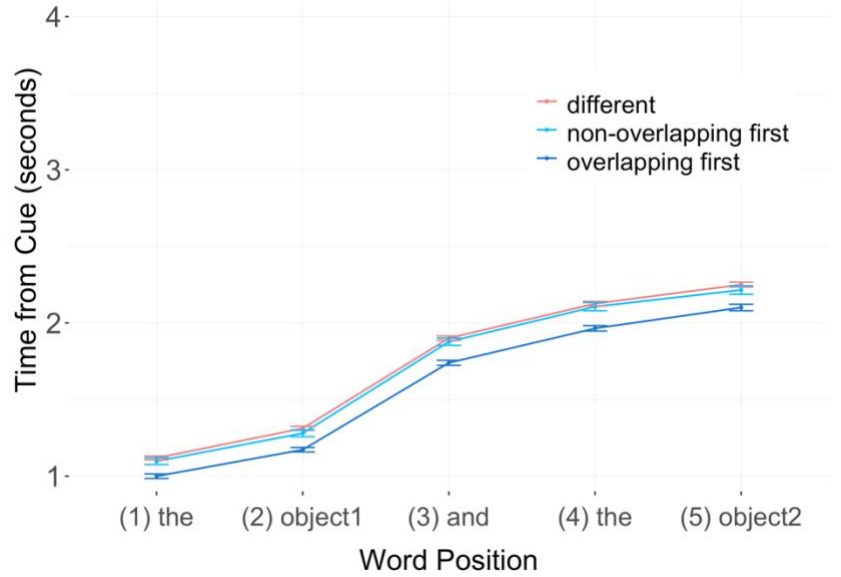


Fig 3. Initiation latencies in Exp 1. Data from the Overlap condition are divided into trials where participants placed the overlapping target first (dark blue line) or the non-overlapping target first (light blue line).

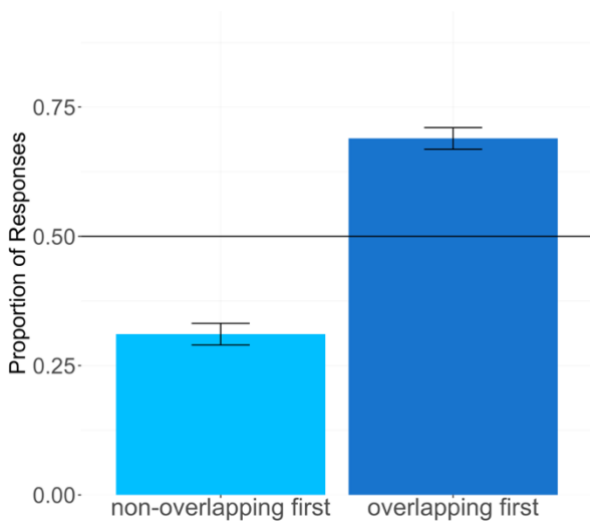


Fig. 4. Order choice in the Overlap condition in Exp 2.

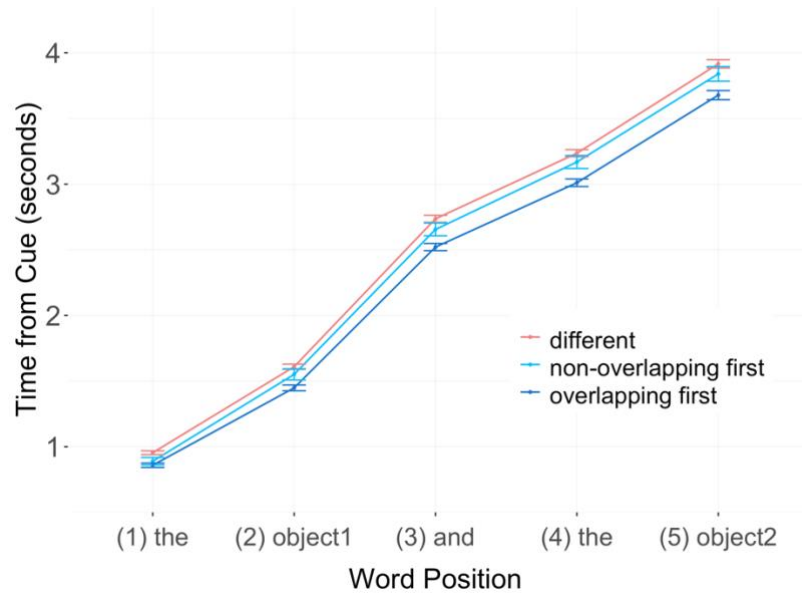


Fig 5. Initiation latencies in Exp 2.

Number attraction in pronoun production: evidence for antecedent feature retrieval

Cassidy Wyatt (UMD), Margaret Kandel (Harvard), Colin Phillips (UMD)

Pronoun production involves two processes: deciding to refer to a referent with a pronoun rather than a full NP and determining pronoun form. A speaker presumably decides to produce a pronoun after accessing the conceptual referent, and it is possible that this access provides the features required to determine pronoun form, e.g. by highlighting relevant features salient in the message or facilitating lemma activation [1]. However, agreement studies of pronoun number [2-4] and gender [5] show attraction from non-antecedent referents (1), suggesting that pronoun form is not derived directly from the message but rather through a feature retrieval process. Yet, these studies may bias speakers to such a process by requiring access of multiple referents to determine the message of the sentence to utter [5] or applying a preamble elicitation paradigm [2-5], which has been shown to influence anaphor planning [6]. Furthermore, the decision to refer to a referent with a pronoun is removed in these studies, as participants are explicitly instructed to begin or complete sentences with a pronoun or to produce sentence tags, in which a pronoun is required by the nature of the construction. In 3 scene-description experiments, we find reliable pronoun number attraction effects, in some instances leading to apparent Principle B violations [7], showing that pronoun encoding involves retrieval referencing items active in the linguistic representation, even when the relevant features could be accessed directly from the message. Timing data shows that this process occurs even in trials where errors are avoided.

Experiments: Participants were introduced to 3 types of alien and the action mimming: when an alien mims another, the other alien's antenna lights up (Fig 1). Participants viewed scenes of aliens mimming and described who mimmed whom, disambiguating the action by referencing the other aliens on the screen. We manipulated the number of aliens in the scenes so that the NPs in the responses either matched or mismatched in number (Table 1). In Exp 1, participants described scenes using either an object or reflexive pronoun (e.g. "The bluey above the greeny mimmed it/itself"); we report the object pronoun trial results here. Exp 2 elicited only the object pronoun trials from Exp 1. Exp 3 elicited sentences in the form "The bluey mimmed the greeny above it". In all experiments, speakers were significantly more likely to produce pronoun number errors in the mismatch conditions (Fig 2). The effect size was similar in Exp 1 and 2. In Exp 3 (where the effect was larger), speakers were more likely to pause before pronoun articulation in the mismatch conditions in error-free sentences, paralleling timing effects observed for verb number attraction with intervening attractors [e.g. 6, 8-10].

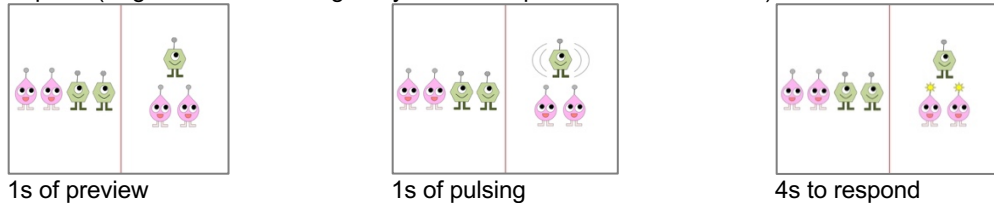
Discussion: The presence of attraction in our study suggests that pronoun form is determined through an agreement process referencing the features of the linguistic antecedent. We show that the effect occurs in a setting similar to natural speech when speakers make a choice about how to refer to the referent. We observed interference effects in timing even when no error was made, suggesting that this retrieval process is not limited to cases when agreement goes awry. In situations of intra-sentential pronominalization, decisions about pronoun use may depend on other items in the sentence (rather than the conceptual referent) because speakers must attend to these items to abide by constraints on anaphora use and NP repetition. We explain our results using a retrieval model of attraction [e.g. 11-13] within the context of a pronoun selection model in which an *in focus* feature of the conceptual referent cues the speaker to produce a pronoun instead of the full NP [e.g. 1]. The antecedents in our experiments had unambiguous number, so the observed effects cannot be attributable to a faulty or ambiguous number evaluation, as proposed by representational models of attraction [e.g. 4, 14-16]. We propose that after accessing the conceptual referent and noting its *in focus* feature, cueing need for a pronoun, the speaker uses a retrieval process to look for a corresponding in focus antecedent. In our sentences, there may be two linguistic representations in focus (salient in the discourse and active in working memory): the antecedent plus an NP lure (in Exp1-2, the sentence subject, recently activated for verb agreement; in Exp 3, the NP individuated by the PP modifier containing the pronoun). The presence of two in focus items may lead the agreement process to pick the number feature of the incorrect NP for agreement, resulting in a form error.

- (1) Agreement attraction occurs when nearby material interferes with normal agreement processes. This effect is typically studied within the context of subject-verb agreement.

Agreement type	Example attraction error
Verb number	*The key to the cabinets <i>are</i> on the table [17]
Pronoun number (reflexive)	*The actor in the soap operas watched <i>themselves</i> [2]
Pronoun number (tag)	*The actor in the soap operas rehearsed, didn't <i>they</i> ? [2]
Pronoun gender	Kijk, daar ligt een aardappelc bij een badpak _N . # <i>Het</i> _N is gaar. [5] (Look, there is a potato _[common gender] next to a backpack _[neuter gender] . It _[neuter] is cooked.)

Figure 1: Stills from experiment scenes

- a) Exp 1-2 (target sentence: “the greeny above the pinkies mimmed them”)



- b) Exp 3 (target sentence: “the bluey mimmed the greeny above it”)

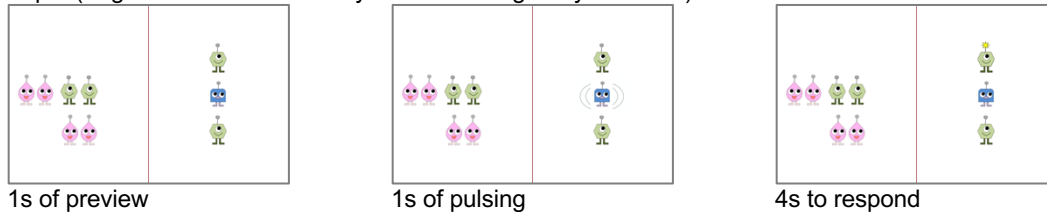
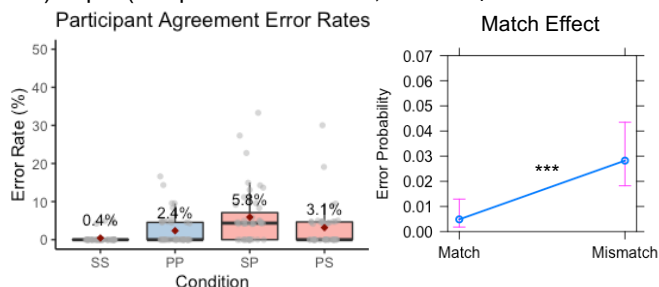


Table 1: Experiment conditions with example sentences

Condition	Sub-Condition	Exp 1-2 sentence	Exp 3 sentence
Match	SS	the pinky above the greeny mimmed it	the pinky mimmed the greeny above it
Match	PP	the pinkies above the greenies mimmed them	the pinkies mimmed the greenies above them
Mismatch	SP	the pinky above the greenies mimmed them	the pinky mimmed the greenies above it
Mismatch	PS	the pinkies above the greeny mimmed it	the pinkies mimmed the greeny above them

Figure 2: Participant error rates and match effect plots

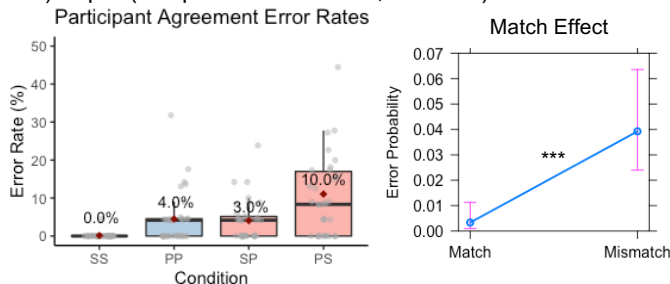
- a) Exp 1 (comparisons: SS – PS, PP – SP)



- b) Exp 2 (comparisons: SS – PS, PP – SP)



- c) Exp 3 (comparisons: SS – SP, PP – PS)



Example elicited errors

(intended antecedent underlined)

Exp 1-2: the pinky above the greenies mimmed it

Exp 3: the pinky mimmed the greenies above them

References: [1] Schmitt et al., 1999; [2] Bock et al., 1999; [3] Bock, Cutler, et al., 2004; [4] Bock, Eberhard, & Cutting, 2004; [5] Meyer & Bock, 1999; [6] Kandel et al., 2019; [7] Chomsky, 1981; [8] Staub, 2009; [9] Staub, 2010; [10] Veenstra et al., 2014; [11] Badecker & Kuminiak, 2007; [12] Wagers et al., 2009; [13] Dillon et al., 2013; [14] Bock & Eberhard, 1993; [15] Solomon & Pearlmutter, 2004; [16] Eberhard et al., 2005; [17] Bock & Miller, 1991

A computational model of reference production based on listener visual-search costs

Julian Jara-Ettinger (Yale University) and Paula Rubio-Fernandez (University of Oslo)

julian.jara-ettinger@yale.edu

A foundational assumption of human communication is that speakers should say as much as necessary, but no more [1,2]. The pressure to be efficient is typically formalized as an egocentric bias whereby speakers aim to minimize production costs. While intuitive, this view has failed to explain why people routinely produce redundant adjectives, particularly color, or why this phenomenon varies cross-linguistically. Here we propose an alternative view of referential efficiency, whereby speakers produce referential expressions designed to facilitate the listener's visual search for the referent. **We present a computational model of our account, the Incremental Collaborative Efficiency (ICE) model, which generates referential expressions by considering the listener's expected visual search in real time.** Under this formulation, cost estimation is not entirely *egocentric* (i.e. determined by speaker production costs), but is in fact partly *allocentric* (i.e. aimed to minimize listener costs) (see Model Equations on p.2). That means that amongst two equally informative descriptions (e.g., 'The red cup' vs 'The plastic cup'), the more efficient one would lead to faster identification of the referent. To achieve this, we implemented a model that simulates how a listener would search for an object in real-time as they process words incrementally, relying on the assumption that people can detect color from the periphery, but they must fixate on an object to evaluate its material or kind. A number of psycholinguistic studies support the view that over-specification aims to facilitate the listener's visual search for the referent [3-14], but no work to date has formalized the computations and cognitive capacities that might underlie an allocentric metric of efficient communication [cf. 15,16].

Here we (1) validate the principles behind our model empirically, and (2) test our model's predictions in a quantitative manner against published reference production data, and (3) in a novel acceptability task designed to test our model in a rigorous way. We began by confirming in an eye-tracking task that color is more visually salient than material, and that speakers prefer color-modified descriptions of the same visual targets over material descriptions. Crucially, we observed a strong, negative correlation between the mean description rating and the mean RT for each color and material description ($r = -.88$ (CI95%: $-.93 - -.80$)), confirming that speakers preferred those descriptions that led listeners to faster target identification (see Fig.1).

To evaluate the ICE model's capacity to explain reference production, we tested whether it could reproduce known qualitative patterns of over-specification: (i) speakers are more likely to over-specify color in denser visual displays [5,9]; (ii) this propensity, however, decreases as a function of the number of objects of the target's color [8,9,11]; and (iii) in identical visual displays, English speakers (prenominal modification) are more likely to use redundant color adjectives than Spanish speakers (postnominal modification) [8,9,12,13]. Fig. 2 shows the results of these analyses: like people, our model's preference for redundant color words (i) increases as a function of the number of objects in the scene, (ii) decreases with increasing monochromaticity, and (iii) is greater for prenominal adjectives than for postnominal adjectives. **Critically, our model predicts production patterns in a quantitative manner without having to fit the parameters to data.**

Finally, to evaluate our model in a more comprehensive way, we also designed a graded acceptability task in which we asked participants to rate how natural different color and material descriptions sounded, allowing us to evaluate our model not only based on its preferred expression, but also on the full distribution of expressions that it produces. Overall, our main (ICE) model showed a correlation of $r = .93$ (CI95%: $.91-.95$), while an alternative Brevity model that penalizes utterances based on utterance length (see Model Equations on p.2) showed a lower correlation of $r = .70$ (CI95%: $.63-.80$). Crucially, our ICE model showed a significantly higher correlation relative to the alternative model ($\Delta r = .22$; CI95%: $.14-.29$).

Supporting our theoretical account, these findings suggest that reference production is best understood as driven by a cooperative goal to help the listener identify the referent in the visual context, rather than by an egocentric bias to minimize utterance length.

Incremental Communicative Efficiency (ICE) Model

$$U(\text{expression}, \text{target}) = R(\text{target})pL(\text{target}|\text{expression}) - C(\text{time})$$

Brevity Model

$$U(\text{expression}, \text{target}) = R(\text{target})pL(\text{target}|\text{expression}) - C(\text{words})$$

Model Equations defining the utility of a referential expression to communicate a target. $pL(\text{target}|\text{expression})$ is the probability that the listener will correctly identify the target from the expression, and $R(\text{target})$ is the speaker's subjective reward for successfully conveying the target. Our model estimates the utility of

different expressions, and then assigns a probability to each expression by softmaxing this utility function.

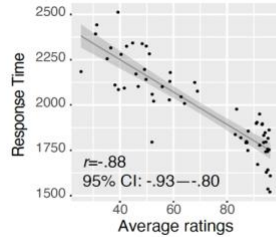


Fig.1 (above). Average description ratings against RT, showing that descriptions rated as more likely to be produced were the ones that led listeners to identify the referent faster.

Fig.2 (right). Model simulations and fits to documented effects on redundant color use.

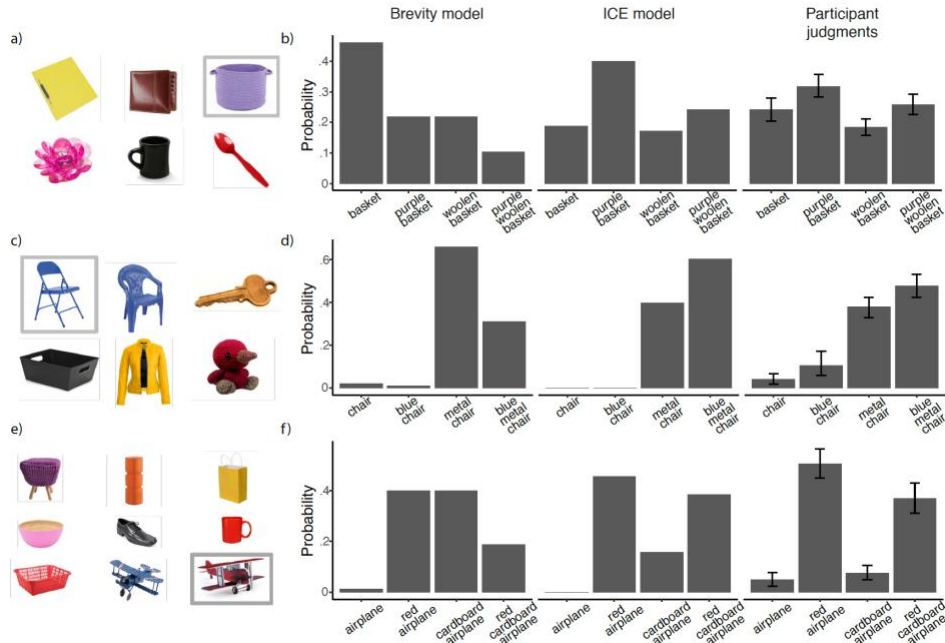
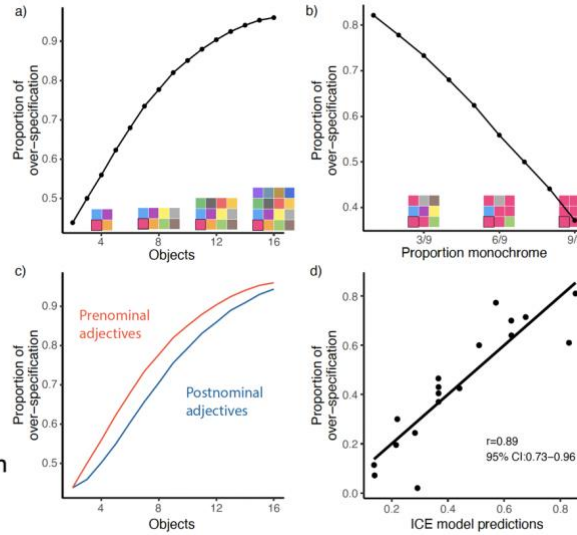


Fig.3: Sample trials from our main experiment along with model predictions and participant judgments.

References: [1] Zipf, 1949. *Human behavior and the Principle of Least Effort*. Addison-Wesley Press [2] Grice, 1975. Logic and conversation. In *Syntax and Semantics*. Academic Press [3] Sonnenschein & Whitehurst, 1982. *J Psycholing Res* [4] Mangold & Pobel, 1988. *J Lang Soc Psych* [5] Paraboni, Van Deemter & Masthoff, 2007. *Comp Lings* [6] Arts, Maes, Noordman & Jansen, 2011. *J Prags* [7] Paraboni & Van Deemter, 2014. *Lang, Cog & Neuro* [8] Rubio-Fernandez, 2016. *Front Psych* [9] Rubio-Fernandez, 2019. *Cog Sci* [10] Tourtouri, Delogu, Sikos & Crocker, 2019. *J Cult Cog Sci* [11] Long, Rohde & Rubio-Fernandez, 2020. *Sci Reps* [12] Rubio-Fernandez, Mollica & Jara-Ettinger, 2020. *JEP:G* [13] Wu & Gibson, 2020 (in press). *Cog Sci* [14] Rehrig, Cullimore, Henderson & F. Ferreira, 2020. *PsyArXiv* [15] van Gompel, van Deemter, Gatt, Snoeren & Krahmer, 2019. *Psych Rev* [16] Degen, Hawkins, Graf, Kreiss & Goodman, 2020. *Psych Rev*.

34th Annual CUNY Conference on Human Sentence Processing

Thursday March 4, 2021

Hour	Parallel Session	Time	Title	Authors
Hour 1	1	12:30	Attachment Preferences in Participle Constructions	Caroline Berg-Love and Masaya Yoshida
Hour 1	1	12:30	Revisiting attachment preferences in Spanish: is there a high attachment bias?	Noelia A. Stetie and Gabriela Mariel Zunino
Hour 1	1	12:30	Illusory NPI licensing and identification of universal quantifier restrictions in real time	Luis Hildebrandt and E. Matthew Husband
Hour 1	1	12:30	The missing VP effect in German: Effects of syntactic position and degree of embedding	Markus Bader
Hour 1	1	12:30	Mandarin argument structure processing: ERP reading data from reversible and irreversible NNV sentences with and without BA and BEI	Max Wolpert, Jiarui Ao, Hui Zhang, Shari Baum and Karsten Steinhauer
Hour 1	1	12:30	PPI Illusion Ignores Binding but is Facilitated by Reactivation	Wesley Orth and Masaya Yoshida
Hour 1	2	12:30	Second Language Processing of Information at the Syntax-Discourse Interface	Didem Kurt and Nazik Dinçtopal Deniz
Hour 1	2	12:30	English locative inversions are not special in terms of their discourse function	Giuseppe Ricciardi, Rachel Ryskin and Edward Gibson
Hour 1	2	12:30	#fitspo: Cognitive Implications of Interacting with "Fitspiration" Content on Social Media	Jordan Zimmerman, Angelica De Rezende, Anna Wright, Kaitlin Lord and Sarah Brown-Schmidt
Hour 1	2	12:30	The Role of Sensory Experience and Communication Modality in the Neural Mechanisms Supporting Social Communicative Processes: A fNIRS Hyperscanning Study	Lauren Berger, Clifton Langdon, Xian Zhang, Joy Hirsch and Ilaria Berteletti
Hour 1	2	12:30	Decomposing the focus effect: Evidence from reading	Morwenna Hoeks, Maziar Toosarvandani and Amanda Rysling
Hour 1	2	12:30	Syntactic focus activates mentioned and unmentioned alternatives in Samoan	Sasha Calhoun, Mengzhu Yan, Honiara Salanoa, Fualuga Taupi and Emma Kruse Va'Ai
Hour 1	3	12:30	Do islands affect only filler-gap dependencies? Evidence from Spanish	Alejandro Rodriguez and Grant Goodall
Hour 1	3	12:30	Acceptability of extraction out of adjuncts depends on discourse factors	Anne Abeillé, Barbara Hemforth, Elodie Winckel and Edward Gibson
Hour 1	3	12:30	The structural source of English Subject Islands	David Potter and Katy Carlson
Hour 1	3	12:30	Semantic interference in dependency formation: NP types in cleft sentences	Myung Hye Yoo and Rebecca Tollan

Hour	Parallel Session	Time	Title	Authors
Hour 1	3	12:30	Backgroundedness measures predict island status of non-finite adjuncts in English	Savithry Namboodiripad, Felicia Bisnath, Alex Kramer, Noah Luntzlara and Adele Goldberg
Hour 1	3	12:30	Oscillatory dynamics of complex dependency processing reveal unique roles for attention and working memory mechanisms	Shannon McKnight, Don Bell-Souder, Phillip Gilley, Akira Miyake and Albert Kim
Hour 1	4	12:30	It takes two the tango: Predictability and detectability affect processing of phrase structure errors	Anthony Yacovone, Paulina Piwowarczyk and Jesse Snedeker
Hour 1	4	12:30	Comparison of Structural and Neural Language Models as Surprisal Estimators	Byung-Doh Oh, Christian Clark and William Schuler
Hour 1	4	12:30	Lexical and partial prediction in a Brazilian Portuguese eye-tracking corpus	João Vieira, Sidney Leal, Erica dos Santos Rodrigues, Sandra Maria Aluísio, Denis Drieghe and Elisangela Nogueira Teixeira
Hour 1	4	12:30	Do children predict grammatical gender of nouns?	Katja Haeuser, Yoana Vergilova and Jutta Kray
Hour 1	4	12:30	Both semantic and form representations are pre-activated during sentence comprehension: Evidence from EEG Representational Similarity Analysis	Lin Wang, Trevor Brothers, Cheng Feng, Sophie Greene, Ole Jensen and Gina Kuperberg
Hour 1	4	12:30	Contributions of Propositional Content and Syntactic Categories in Sentence Processing	Byung-Doh Oh and William Schuler
Hour 1	5	12:30	German pronoun use follows Bayesian principles	Clare Patterson, Petra B. Schumacher, Bruno Nicenboim, Johannes Hagen and Andrew Kehler
Hour 1	5	12:30	"Good-enough" production: accessibility influences choice of taxonomic level	Crystal Lee, Casey Lew-Williams and Adele Goldberg
Hour 1	5	12:30	Choosing a Referring Expression: Intrasentential Ambiguity Avoidance in Romanian	Rodica Ivan, Brian Dillon and Kyle Johnson
Hour 1	5	12:30	Invisible, unmentioned entities affect referential forms	Si On Yoon, Breanna Pratley and Daphna Heller
Hour 1	5	12:30	Implicit Causality Can Affect Pronoun Use in Fragment Completion Tasks	Yining Ye, Kathryn C. Weatherford and Jennifer E. Arnold
Hour 1	5	12:30	Irregular and regular verbs elicit identical morphological decomposition ERPs	Arild Hestvik, Valerie Shafer and Richard Schwartz
Hour 1	6	12:30	Clefting and prosody affect pronoun processing in dialogue contexts	Abigail Toth, Liam Blything, Juhani Järvisikivi and Anja Arnhold
Hour 1	6	12:30	Comprehension meets production: null/overt subject pronouns in Italian and Spanish	Carla Contemori and Elisa Di Domenico
Hour 1	6	12:30	Cross-linguistic patterns in person systems reflect efficient coding	Mora Maldonado, Noga Zaslavsky and Jennifer Culbertson
Hour 1	6	12:30	Prosody modulates subjecthood and linear order effects in German pronoun resolution	Regina Hert, Anja Arnhold and Juhani Järvisikivi

Hour	Parallel Session	Time	Title	Authors
Hour 1	6	12:30	Adaptation to discourse patterns depends on the relative frequency of competing structures	Valerie Langlois and Jennifer Arnold
Hour 1	6	12:30	Are both syntactically and semantically-based pronoun dependencies stored in memory?	Jennifer E. Arnold, Avery Wall and Taylor Steele
Hour 2	7	13:30	Temporary ambiguity and memory for the context of spoken language use	Kaitlin Lord and Sarah Brown-Schmidt
Hour 2	7	13:30	Investigating suppletion with novel adjectives	Lyn Tieu and Nichola Shelton
Hour 2	7	13:30	A Dynamic Tree-Based Item Response Model for Visual World Eye-tracking Data	Sarah Brown-Schmidt, Matthew Naveiras, Sun-Joo Cho and Paul De Boeck
Hour 2	7	13:30	Processing referring expressions: Accessibility is not predictability	Wei jie Xu and Ming Xiang
Hour 2	7	13:30	Good-enough for all intensive purposes: Eggcorns and noisy channel processing	Gwendolyn Rehrig and Fernanda Ferreira
Hour 2	8	13:30	Systematicity in gesture production, perception may support sign language emergence	Chuck Bradley
Hour 2	8	13:30	The time course of sentence planning and production in two Australian free word order languages	Gabriela Garrido Rodriguez, Sasha Wilmoth, Rachel Nordlinger and Evan Kidd
Hour 2	8	13:30	Underlying clausal structure modulates lexical interference: Evidence from raising and control	Jeremy Doiron and Shota Momma
Hour 2	8	13:30	Transitioning to online language production: a direct comparison of in-lab and web-based experiments	Margaret Kandel, Cassidy Wyatt and Colin Phillips
Hour 2	8	13:30	Attribute Salience and Adjective Order Preferences	Monica Do
Hour 2	8	13:30	Flexibility in language production: insights from completion of fragmentary inputs	Peng Qian and Roger Levy
Hour 2	9	13:30	Lexical activation dynamics and interference in sentence processing: the effect of time	Carolyn Baker and Tracy Love
Hour 2	9	13:30	Accessibility-Based Constraints on Morphosyntax in Corpora of 54 Languages	Kyle Mahowald, Isabel Papadimitriou, Dan Jurafsky and Richard Futrell
Hour 2	9	13:30	Syntax guides sentence planning: Evidence from multiple dependency constructions	Shota Momma and Masaya Yoshida
Hour 2	9	13:30	Evidence for Early Application of Binding Theory and Late Intrusion Effects	Arild Hestvik and Myung Hye Yoo
Hour 2	9	13:30	Predicting binding domains: Evidence from fronted auxiliaries and wh-predicates	Keir Moulton, Cassandra Chapman and Nayoun Kim
Hour 2	9	13:30	Classifier as a cue for structure building in head-final relative clause	Zirui Huang and Matthew Husband

Hour	Parallel Session	Time	Title	Authors
Hour 2	10	13:30	ERPs reveal how semantic and syntactic processing unfolds across parafoveal and foveal vision in sentence comprehension	Chuchu Li, Katherine Midgley and Phillip Holcomb
Hour 2	10	13:30	A noisy channel model of N400 and P600 effects in sentence processing	Jiaxuan Li and Allyson Ettinger
Hour 2	10	13:30	The benefits and costs of language prediction: Evidence from ERPs	Jiaxuan Li, Jinghua Ou and Ming Xiang
Hour 2	10	13:30	Dissociating Effects of Predictability, Preview and Visual Contrast on Eye Movements and ERPs	Jon Burnsky, Franziska Kretschmar, Erika Mayer, Lisa Sanders and Adrian Staub
Hour 2	10	13:30	Modeling influences of coercion on N400 amplitudes as change in a probabilistic representation of meaning	Milena Rabovsky
Hour 2	10	13:30	Neural correlates of expectation violations and discourse updating: The case of Bulgarian object agreement	Paul Compensis and Petra B. Schumacher
Hour 2	11	13:30	Feature Reactivation in Minimalist Parsing	Aniello De Santo
Hour 2	11	13:30	Can English Idioms Undergo the Dative Alternation? A Priming Investigation	Breanna Pratley and Philip Monahan
Hour 2	11	13:30	The effect of representational complexity on working memory processes	Chi Dat Lam and Ming Xiang
Hour 2	11	13:30	Null nouns can trigger intervention in Spanish relative clauses' comprehension	Marisol Murujosa, Carolina Gattei, Diego Shalom and Yamila Sevilla
Hour 2	12	13:30	Prosodic Phrasing in English and the Processing of Agreement Attraction	Adam Royer
Hour 2	12	13:30	Prosody and eye movements on attachment in Brazilian Portuguese	Aline Fonseca, Andressa Christine da Silva and Marcus Maia
Hour 2	12	13:30	Two-dimensional parsing, the iambictrochaic law, and the typology of rhythm	Michael Wagner, Alvaro Iturralde Zurita and Sijia Zhang
Hour 2	12	13:30	Case interference and phrase length effects in processing Turkish center-embeddings	Özge Bakay and Nazik Dinçtopal Deniz
Hour 2	12	13:30	Prosody drives eye movements from early on in semantic comprehension	Petra Augurzky, Ruth Kessler and Claudia Friedrich
Hour 2	13	13:30	Preferences for shorter dependencies in miniature language learning are modulated by the statistics of learners' L1	Masha Fedzechkina, Charles Torres and Yiyun Zhao
Hour 2	13	13:30	Children's acquisition of new/given markers in English, Hindi, Mandinka and Spanish	Vishakha Shukla, Madeleine Long, Vrinda Bhatia and Paula Rubio-Fernandez
Hour 2	13	13:30	The role of language context in the acquisition of novel words	Anna Alberski and Kathryn Schuler
Hour 2	13	13:30	Effects of lifetime and fact knowledge in language comprehension	Daniela Palleschi and Pia Knoeferle

Hour	Parallel Session	Time	Title	Authors
Hour 2	13	13:30	Effect of referent lifetime in the processing of verbal morphology: a self-paced reading study	Daniela Palleschi, Camilo Rodriguez Ronderos and Pia Knoeferle
Hour 2	13	13:30	A protracted developmental trajectory for English-learning children's detection of consonant mispronunciations in newly learned words	Carolyn Quam and Daniel Swingley

Attachment Preferences in Participle Constructions

Caroline Berg-Love and Masaya Yoshida (Northwestern University)

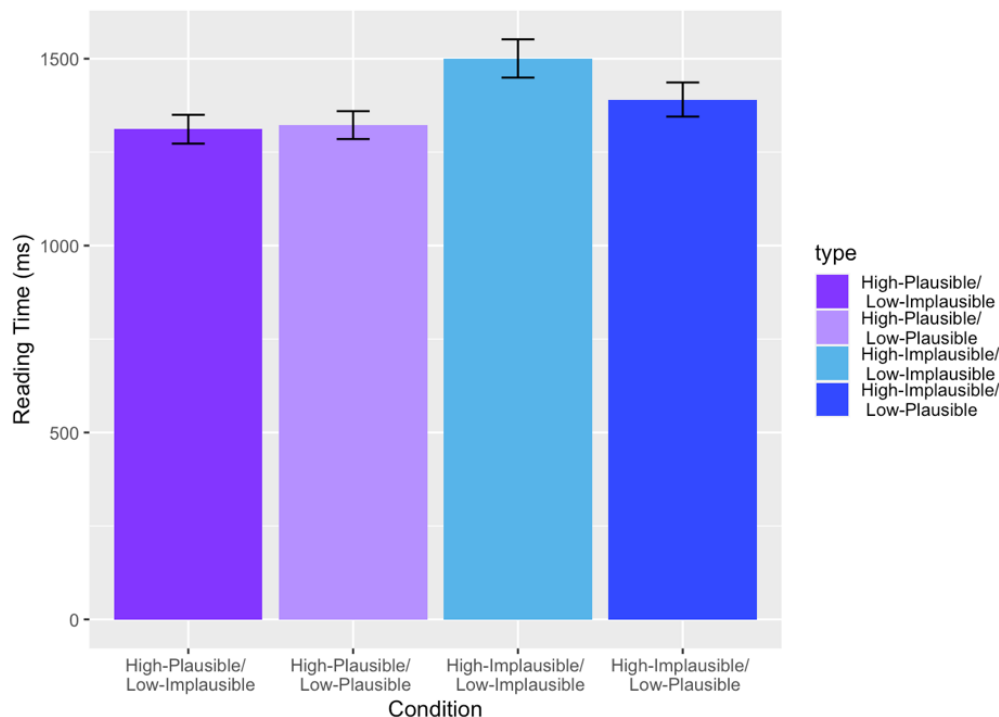
Introduction: Online ambiguity resolution processes are subject to various constraints. In English, studies have shown widespread biases for local/low attachment in sentences with attachment ambiguities [1-4], but a bias for high attachment is observed in some configurations [5]. Specifically, it has been suggested that high attachment is preferred when the low attachment structure is more complex [5]. This study investigates processing sentences with global ambiguity of the bare present participle clauses (PPCs) such as (1). The PPC, *wearing a hat*, can either attach low to the NP, *the girl*, or high to the VP, *met*. An A-Maze incremental reading experiment shows a bias for high attachment in online sentence processing. We argue that high attachment is preferred because it yields a simpler structure in the case of PPCs.

Processing PPCs: In an offline experiment, [6] suggested that the general preference for low attachment in English is also present in PPCs. This low attachment bias has not previously been tested in online processing. Considering structural complexity as a factor in attachment preferences, high attachment of PPC could be preferred in online processing. The structure of PPCs differs with high versus low attachment, as shown roughly in (2a) and (2b). (2a), high attachment, has an adjunct control structure, while (2b), low attachment, has a subject-gapped relative clause structure [7]. As in (2b), the NP-modifier structure involves movement of a silent relative pronoun and omission of the *be-verb*, making it more complex than the VP-modifier structure. If the parser obeys only the local attachment bias, there should be a bias for low attachment; but if a simpler structure is preferred, the parser would prefer high attachment.

Experiment: A Maze incremental reading experiment [2,8], where sentences are presented one word at a time and participants choose between two words as to which continues the sentence, was conducted with native English speakers (24 items: n=40). The semantic plausibility between the attachment site (*Attachment Site*: High vs. Low) and the PPC (*PPC*: Plausible vs. Implausible) were manipulated as independent factors in a 2x2 factorial design, in the following four conditions: High-Plausible/Low-Implausible (3a), High-Plausible/Low-Plausible (3b), High-Implausible/Low-Implausible (3c), and High-Implausible/Low-Plausible (3d). Thus, for example, in (3a), if the PPC is attached high, it yields the semantically plausible interpretation of *the coach holding the glove*, but if attached low, it has an implausible reading of *the padlock holding the glove*. This plausibility manipulation allows us to test the attachment preferences in PPCs: if a PPC attachment has an implausible interpretation, it should lead to reading time slowdown [9]. If the low attachment is preferred, when the parser reaches the embedded verb region, *holding* in (3), where the implausibility is recognized, the reading time should be slower in Low-Implausible conditions than Low-Plausible conditions. If, however, the high attachment is preferred, *holding* should be read significantly slower in High-Implausible conditions than in High-Plausible conditions. A linear-mixed effect model of log reading time revealed a significant main effect of *Attachment Site*, with low attachment conditions read significantly slower than high attachment ($\beta=.09$, $SE=.03$, $t=3.05$, $p<.01$) at the embedded verb region. Subset analysis revealed that the embedded verb in the High-Implausible/Low-Plausible condition is read significantly slower than the High-Plausible/Low-Implausible condition ($\beta=.05$, $SE=.03$, $t=2.96$, $p<.05$). The embedded verb in High-Implausible/Low-Implausible conditions was also read significantly slower than in High-Plausible/Low-Plausible conditions ($\beta=.12$, $SE=0.04$, $t=2.93$, $p<.01$).

Conclusions: The results of this experiment suggest that high attachment of PPC is preferred over low attachment. This supports that structural complexity influences ambiguity resolution and that high attachment is preferred when it yields a simpler structure. Additionally, the slower reading of High-Implausible/Low-Implausible than High-Plausible/Low-Plausible supports previous studies suggesting the ambiguity advantage [10, 11].

- (1) The boy met the girl [_{PPC} wearing a hat].
- (2) a. The boy [_{VP} met the girl [_{CP} [_{IP} PRO wearing a hat]]].
- (2)b. The boy met [_{NP} the girl [_{CP} Op [_{IP} t ~~was~~ wearing a hat]]]
- (3)a. High-Plausible/Low-Implausible
The coach locked the padlock holding a glove meanwhile the game went poorly.
- (3)b. High-Plausible/Low-Plausible
The coach locked the vehicle holding a glove meanwhile the game went poorly.
- (3)c. High-Implausible/Low-Implausible
The keys locked the padlock holding a glove meanwhile the game went poorly.
- (3)d. High-Implausible/Low-Plausible
The keys locked the vehicle holding a glove meanwhile the game went poorly.



References: [1] Traxler, Pickering, & Clifton. (1998). "Adjunct attachment is not a form of lexical ambiguity resolution." [2] Witzel, N., Witzel, J., Forster, K. (2012) "Comparisons of online reading paradigms: eye tracking, moving-window, and maze." [3] Pickering, M. & Traxler, M. (1998). "Plausibility and recovery from garden paths: an eye-tracking study." [4] Phillips & Gibson. (1997). "On the strength of the local attachment preference." [5] Ferreira, F. & Clifton, C. (1986). "The independence of syntactic processing." [6] Kang, S. & Speer, S. (2004). "Prosodic disambiguation of participle constructions in English." [7] Williams, E. (1992). "Adjunct control." [8] Boyce, V., Futrell, R., Levy, R. (2019). "Made made easy: better and easier measurement of incremental processing difficulty." [9] Pickering, M. & Traxler, M. (1998). "Plausibility and recovery from garden paths: an eye-tracking study." [10] Sloggett, Van Handel, Sasaki, Duff, Rich, Orth, Anand, & Rysling. (2020). "'Ambiguous' isn't 'underspecified': evidence from the maze task." [11] van Gompel, Pickering, & Traxler. (2001). "Reanalysis in sentence processing: evidence against current constraint-based and two-stage models."

Revisiting attachment preferences in Spanish: is there a high attachment bias?

Noelia A. Stetie & Gabriela M. Zunino (CONICET || University of Buenos Aires)

nstetie@conicet.gov.ar

Psycholinguistic studies carried out in the last decades have found that attachment preferences present crosslinguistic variation and that Spanish speakers usually prefer high attachment (Carreiras *et al.* 1993, Dussias 2001). Many theories, such as Construal (Frazier & Clifton 1996), Recency and predicate proximity (Gibson *et al.* 1996) and Implicit prosody (Fodor 2002), have been proposed to account for the crosslinguistic variation on relative clause attachment. However, another set of theories suggests that crosslinguistic differences in biases may be reduced to individual differences (Swets *et al.* 2007, Wells *et al.* 2009) or to a syntactic difference, namely the availability of pseudo relatives (Grillo *et al.* 2014).

We are conducting a series of experiments to deepen the study of the psycholinguistic processes carried out during sentence parsing in Spanish, specifically in the Rioplatense variety. Here we present the results of our first study, which will serve as a baseline for the following experiments. We conducted a reading task with comprehension questions. We presented ambiguous sentences with relative clauses in two positions (see Sample stimuli): object (ORC) and subject (SRC). After each sentence, participants had to answer a multiple-choice interpretation question, to verify attachment preferences. The items were presented in 3 counterbalanced lists: 18 items and 27 fillers each. The task was programmed and performed in IBEX and 147 people were tested (103 women, age: $M=34.41$, $SD=13.85$).

Regarding response types and attachment preferences, we found a bias towards the second noun-phrase (NP2) for both ORC and SRC (Figure 1). For the SRC, the preference towards low attachment is clear: 75% (low) vs 25% (high). For ORC, the preference is at the level of chance: 57% (low) and 42% (high). We used Generalized Mixed Effects Models for the analysis and find a statistically significant difference on the response types regarding the position of the RC ($\beta_0=0.4642$, $z=2.024$, $p=0.0429$; $\beta_{1_SRC}=1.1038$, $z=3.909$, $p=9.28e-05$).

When analyzing the response times (Figure 2), we found that participants took longer to attach to the first noun phrase (NP1) ($M=4494$, $SD=4035$) than to the second ($M=3767$, $SD=3601$) for both ORC and SRC. We used Linear Mixed Effects Models for the analysis and found an effect of attachment preference (high vs low): $\beta_0=4234.2$, $t=21.748$, $p<2e-16$; $\beta_{1_NP2}=-329.9$, $t=-2.149$, $p=0.0318$. We also found an interaction between the position of the RC: for the ORC there was no statistically significant difference between attaching to the NP1 or to the NP2 ($p=0.9799$). However, this difference was significant for the SRC ($p=0.0004$). Also participants took shorter times to attach the NP1 to an ORC ($M=4205$; $SD=3378$) than to a SRC ($M=4997$; $SD=4942$), the difference was statistically significant ($p=0.0218$).

Firstly, we found a preference for low attachment when the relative clause is in the subject position, as reported in previous studies (Hemforth *et al.* 2015). Secondly, these results show no offline preference for high attachment in Spanish, as suggested by some recent studies (Alonso-Pascua 2020, Hemforth *et al.* 2015). Moreover, we found longer response times for high attachment, which could indicate that, when it occurs, it's an offline and interpretative preference. The analysis of responses to determine attachment bias points out a statistically significant difference between SRC and ORC, but in both cases the bias is towards low attachment, although for the ORC the attachment preferences seem to be at the level of chance. One possible explanation lies on the syntactic characteristics of the stimuli: it could be the case that some sentences allow a pseudo relative (PR) interpretation, which, according to the Pseudo-Relative First Hypothesis (Grillo *et al.* 2014), will be preferred over a genuine RC, forcing thus a high attachment. This hypothesis was not considered in the confection of the stimuli, however, a posterior analysis of the results shows that 6 sentences allow a PR lecture and for 4 of them showed a strong bias towards high attachment. Taken together, these results would suggest that there is no clear Spanish attachment bias, however, further experiments should be done to test de Pseudo-Relative First Hypothesis and the processing of genuine RC in Spanish interpretation.

Sample stimuli

1. ORC:

El joven empujó al sobrino (NP1) de la maestra (NP2) que viajaba en el barco.
The young man pushed the nephew (NP1) of the teacher (NP2) who was traveling on the boat.
¿Quién viajaba en el barco? a. el marinero; b. el sobrino; c. la maestra; d. la lingüista
Who was traveling on the boat? a. the sailor; b. the nephew; c. the teacher; d. the linguist

2. SRC:

El asistente (NP1) del ministro (NP2) que hablaba tres idiomas tuvo un romance prohibido.
The assistant (NP1) of the minister (NP2) who spoke three languages had a forbidden romance.
¿Quién hablaba tres idiomas? a. el asistente; b. el ministro; c. el intérprete; d. el físico
Who spoke three languages? a. the assistant; b. the minister; c. the deputy; d. the physicist

Figures

Figure 1: Attachment preferences by RC position

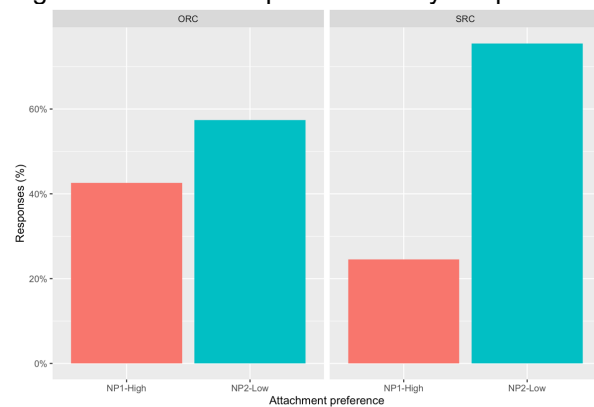
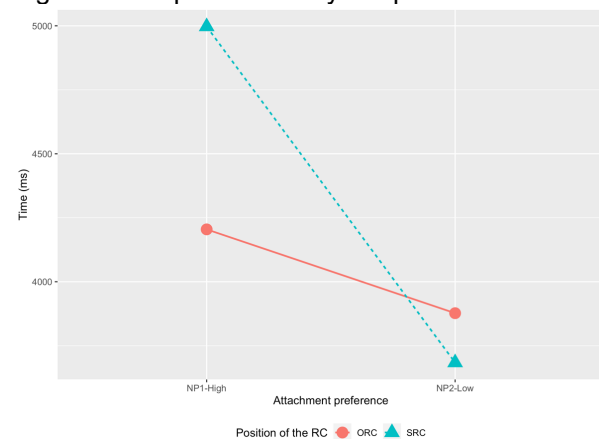


Figure 2: Response time by RC position



References

- Alonso-Pascua, B. (2020). New evidence on the Pseudorelative-First Hypothesis: Spanish attachment preferences revisited. *Topics in Linguistics*, 21(1), 15-44.
- Carreiras, M., & Clifton Jr, C. (1993). Relative clause interpretation preferences in Spanish and English. *Language and speech*, 36(4), 353-372.
- Dussias, P. E. (2001). Sentence parsing in fluent Spanish-English bilinguals. *One mind, two languages: Bilingual language processing*, 159, 176.
- Fodor, J. D. (2002). Prosodic Disambiguation in Silent Reading. *NELS*, 32, 113-32.
- Frazier, L. & Clifton, Jr., C. (1996). *Construal*. Cambridge, MA: MIT Press.
- Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, 59(1), 23-59.
- Grillo, N., & Costa, J. (2014). A novel argument for the universality of parsing principles. *Cognition*, 133(1), 156-187.
- Hemforth, B., Fernandez, S., Clifton Jr, C., Frazier, L., Konieczny, L., & Walter, M. (2015). Relative clause attachment in German, English, Spanish and French: Effects of position and length. *Lingua*, 166, 43-64.
- Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, 136(1), 64.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, 58(2), 250-271.

Identification of universal quantifier restriction and illusory NPI licensing

Luis A. Hildebrandt & E. Matthew Husband (University of Oxford)

Negative Polarity Items (NPIs) like *ever* must be licensed by downwards entailing operators (negation, *only*, etc) in structurally accessible configurations. Nevertheless, psycholinguistic research has found that the presence of potential licensors in structurally inaccessible locations can drive illusions of a licensed NPI (Parker & Phillips, 2016; Vasisht et al., 2008). There are currently two competing hypotheses for the source of these illusions. One hypothesis is that illusions are a result of incorrect retrieval of structurally inaccessible licensors due to noisy cue-based memory retrieval. A second hypothesis considers that these illusions may reflect the application of semantic/pragmatic processes (Xiang, Dillon, & Phillips, 2009; Xiang, Grove, Giannakidou, 2013).

Universal quantifiers like *every* offer an interesting but unexplored testbed for both of these hypotheses. *Every* can license NPIs within its restrictor clause (1a), which is a downward entailing environment (Ladusaw, 1980), but not within its scope which is not downward entailing (1b). NPI licensing with a universal quantifier requires the parser to identify the extent of the restrictor and determine the structural position of the NPI, a process that, due to the delicacy of real-time NPI licensing, may be prone to errors. We investigated whether illusory licensing of NPIs occurs in the scope of a universal quantifier.

Predictions. We predicted that **(P1)** if illusory NPI licensing is driven by faulty memory retrieval exclusively, these illusions should persist independently of manipulations to the restrictor clause. However, **(P2)** if these illusions are the result of difficulty in identifying the boundaries of the quantifier's restrictor, we predicted that the addition of modifiers to the quantified subject would allow the parser to identify the extent of the restriction clause by providing a suitable contrast set before parsing the NPI, thus reducing the illusory effect.

Prior: Speeded judgments. In prior research (Hildebrandt & Husband, 2019), four speeded acceptability judgments (summarized in Table 1) found A) illusory licensing of *ever* outside the restriction of *every* (2,3) that was B) not found with the existential quantifier *some* (2,4), suggesting that illusions are specific to universal quantifiers, not quantifiers in general. This illusory licensing effect was diminished when either C) a pre-nominal modifier (2,3,5a) or D) a post-nominal modifier (2,3,5b) was introduced into the quantifier's restrictor. These results are consistent with **(P2)**. Adding a modifier aided identification of the quantifier's restrictor, allowing the parser to more easily reject the unlicensed NPI, thus reducing the illusory effect.

Current: Self-paced reading. To observe the online effect of illusory licensing, we conducted two self-paced reading studies using the items from speeded judgements. **Study 1** (N=72, Item=40) compared the sentences in (2,3,4) [4 conditions]. Reading times for the Definite ($t=2.394$, $p=.017$) and Existential ($t=2.126$, $p=.034$) condition were significantly slower than Negation on the first Spill-over word. The Universal was not ($t=0.464$, $p=.642$), a result consistent with the illusory licensing effect found in speeded judgments (A, B).

Study 2 (N=72, Item=50) compared the sentences in (2,3,5) [5 conditions]. Reading times for the Definite ($t=2.436$, $p<.01$) were significantly slower than Negation on the first Spill-over word. The Universal and Universal+Pre-/Post-modification conditions were not (Uni: $t=0.927$, $p=.0354$; Uni+Pre: $t=1.111$, $p=.267$; Uni+Post: $t=0.181$, $p=.856$). Illusory licensing persisted with both modification conditions, a distinct effect from speeded judgments (C, D).

Conclusions. While speeded judgement results suggest that modification aids identification the universal quantifier's restriction **(P2)**, self-paced reading times continued to show illusory licensing effects even in the presence of modifiers. This suggests that the parser requires time online to identify a quantifier's restriction and close it off to further processing. This slow identification process can snag stray NPIs, leading them to appear to be licensed temporarily online. Further research is planned to investigate the fine-grained timing of this temporary illusory licensing effect.

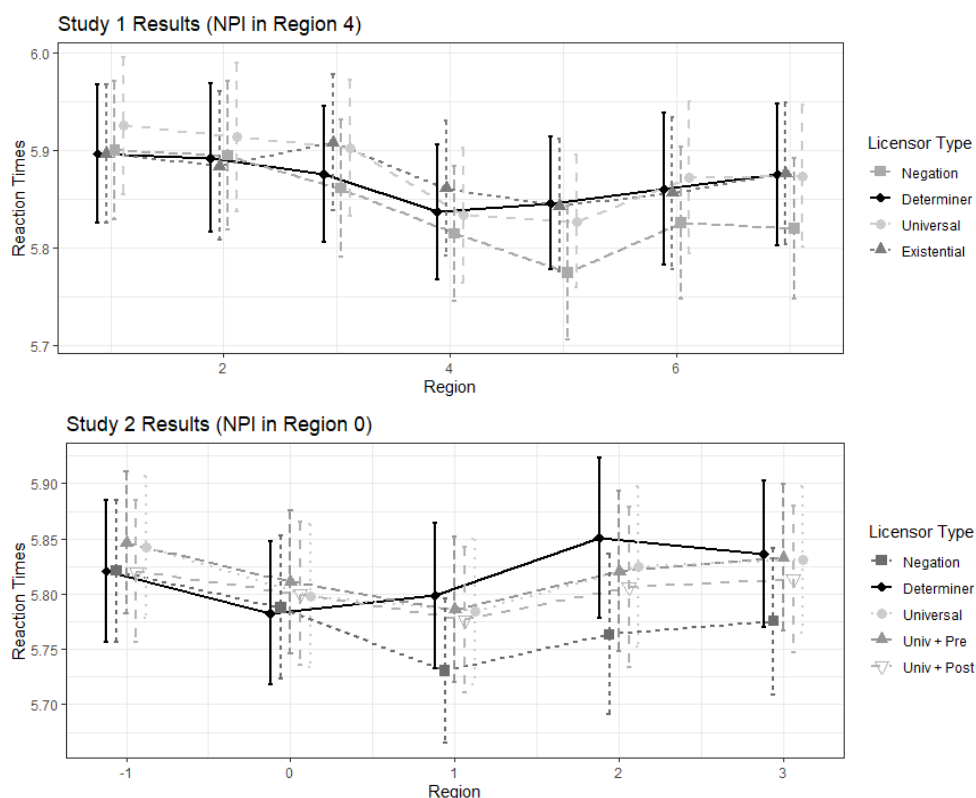
- (1) a. Every [_{RESTRICTOR} student [who has ever come to class]] [_{SCOPE} has received a good mark].
 b. Every [_{RESTRICTOR} student [who has come to class]] [_{SCOPE} has *ever received a good mark].

Example Stimuli

- (2) **No/The** journalist has ever been recognized for his online contributions. (Neg / Def)
 (3) **Every** journalist has ever been recognized for his online contributions. (Universal)
 (4) **Some** journalist has ever been recognized for his online contributions. (Existential)
 (5) a. **Every newspaper** journalist has ever been recognized for his online contributions. (Uni+pre-mod)
 b. **Every** journalist who was published on the website has ever been recognized for his online contributions. (Uni+post-mod)

Table 1: Summary of speeded judgement study results (significant effects in bold)

		<i>NPI_Q – NPI_The</i>	<i>z</i>	<i>p</i>
Study A	Every	7.66%	2.229	.026
Study B	Some	-0.65%	-0.161	.872
Study C	Every + pre-mod	4.92%	1.231	.218
Study D	Every + post-mod	5.94%	1.394	.163



Selected References. Hildebrandt, L., & Husband, E. M. (2019). Quantifiers, Restrictors, and Illusory NPI Licensing. Poster given at the 32nd CUNY Conference on Human Sentence Processing. Boulder, CO. Ladusaw, W. A. (1979). *Negative polarity items as inherent scope relations*. Ph.D. Dissertation, University of Texas at Austin. Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, 157:321-339. Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing Polarity: how the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32, 685-712. Xiang, L., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, 108, 40-55. Xiang, M., Grove, J., & Giannakidou, A. (2013). Dependency dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Frontiers in Psychology*.

The missing VP effect in German: Effects of syntactic position and degree of embedding

Markus Bader (Goethe University Frankfurt)

The missing VP effect is a syntactic illusion that has played a major role in recent discussions of sentence memory and processing complexity (see Futrell et al., 2020 for a recent summary). In contrast to other languages, a missing VP effect has been found for German in some experiments but not in others. This may be due to different syntactic positions in which relative clauses (RCs) exhibiting multiple center-embedding appeared in different experiments, but also to differences concerning experimental procedures and the particular sentence material. In order to better understand the missing VP effect in German, I ran three experiments using the same procedure and similar materials across experiments. All experiments were distributed as paper-and-pencil questionnaires and required participants – students with German as native language – to rate sentences on a scale from 1 (totally unacceptable) to 7 (totally acceptable).

The first experiment tested whether sentences with an incomplete RC are rated as more acceptable when the RC is in center-embedded position than when it is not. To this end, Experiment 1 compared sentences with a complex RC adjacent to its head noun (embedded) to corresponding sentences with the complex RC in extraposed position (see (1) and (2); only adjacent RCs are shown). The complex RC was either complete or missing the VP of the outer RC. In Experiment 1a (47 participants), the modified NP was in sentence-initial position, in Experiment 1b (47 participants), it was in sentence-internal position. The results are shown in Figure 1. When the head NP occurred sentence-initially, complete sentences were rated much better than sentences with a missing VP. When the head NP occurred sentence-internally, in contrast, complete RCs were judged slightly better when extraposed than when center-embedded. Incomplete RCs, in contrast, received very low ratings when extraposed, but were judged as acceptable as complete RCs when center-embedded. Experiment 1 thus confirms former findings that the occurrence of a missing VP effect in German depends on the position of the NP hosting the complex RC: A missing VP effect is observed when the NP and its RC occur sentence internally (Häussler and Bader, 2015) but not when they occur sentence initially (Vasishth et al., 2010).

Experiment 2 (24 participants) compared sentences containing double-center embedding (as in Experiment 1) to sentences with triple center-embedding, with the complex RC always modifying a sentence-initial NP (see (2) and (3)). The results for Experiment 2 are shown in Figure 2. The condition with double center-embedding replicates the finding from Experiment 1. Triple center-embedding, in contrast, showed no longer the pattern formerly found for RCs in sentence-initial position, but the pattern for RCs in sentence-internal position: The acceptability for complete sentences did not differ from the acceptability of incomplete sentences. Thus, under particularly high processing load, a missing VP effect is observed even for RCs modifying sentence-initial NPs.

Experiment 3 (40 participants) compared triple center-embedding in sentence-initial position (as in Experiment 2; see (3)) to triple center-embedding in sentence-internal position (combination of (1) and (3)). The results for Experiment 3, shown in Figure 3, reveal similar ratings for sentences with sentence-initial and sentence-internal RCs: Sentences with a missing VP received acceptability ratings that were slightly, although not significantly, below those for complete sentences. Thus, when processing load is sufficiently high, the position of the complex RC does no longer matter.

The above experiments show that in German a missing VP effect is observed across syntactic contexts, with the exception of doubly center embedded RCs modifying a sentence initial NP, for which no missing VP effect was found. The computational theory of Futrell et al. (2020) accounts for the complexity effect found for sentence-initial RCs, but fails to account for the missing VP effect found for double center-embedded RCs in sentence-internal position. The interference theory of Häussler and Bader (2015) is only informally stated so that no firm conclusions are possible. I will discuss how the two theories can be joined in order to account for the full range of data.

Sample sentences from Experiments 1-3; the bold-faced verbs were missing in the condition “Missing VP”.

- (1) *Sentence-internal complex RC, double-embedding (Experiment 1)*
 Ich glaube, dass der Musiker, den der Dirigent, der das Konzert mit vielen berühmten Solisten planen soll, **unterstützt hat**, interviewt wurde.
 ‘I believe that the musician who the conductor who has to plan the concert with many soloists supported was interviewed.’
- (2) *Sentence-initial complex RC, double-embedding (Experiment 1 and 2)*
 Der Musiker, den der Dirigent, der das Konzert mit vielen berühmten Solisten planen soll, **unterstützt hat**, wurde interviewt.
 ‘The musician who the conductor who has to plan the concert with many soloists supported was interviewed.’
- (3) *Sentence-initial complex RC, triple-embedding (Experiment 2 and 3)*
 Der Musiker, den der Dirigent, der das Konzert, das auf nächstes Jahr verschoben wurde, planen soll, **unterstützt hat**, wurde interviewt.
 ‘The musician who the conductor who has to plan the concert that had to be moved to the upcoming year supported was interviewed.’



Fig. 1: Acceptability ratings in Experiments 1a (sentence-initial RC) and 1b (sentence-internal RC).

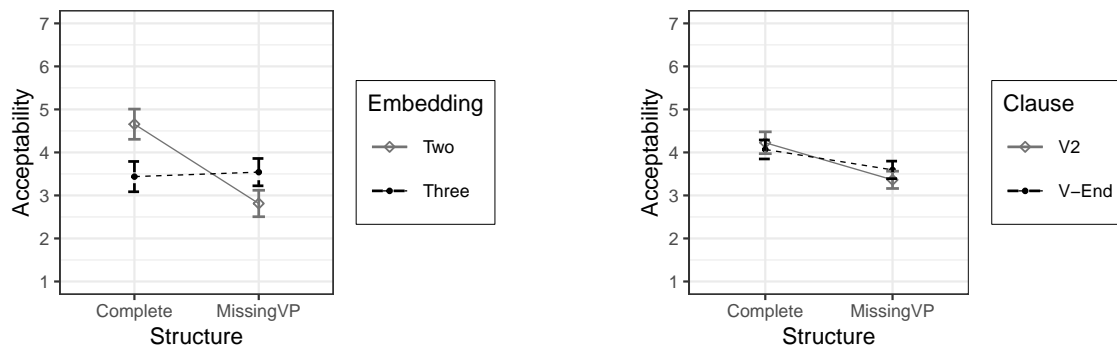


Figure 2: Acceptability ratings in Experiment 2.

Figure 3: Acceptability ratings in Experiment 3.

References

- Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3):e12814.
- Häussler, J. and Bader, M. (2015). An interference account of the missing-VP effect. *Frontiers in Psychology*, 6:1–16.
- Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Mandarin argument structure processing: ERP reading data from reversible and irreversible NNV sentences with and without BA and BEI

Wolpert M^{1,2}, Ao J³, Zhang H⁴, Baum S^{2,5}, Steinhauer K^{2,5}

¹McGill University, Integrated Program in Neuroscience; ²Centre for Research on Brain, Language, and Music; ³McGill University, Faculty of Science; ⁴Nanjing Normal University, School of Foreign Languages and Cultures; ⁵McGill University, School of Communication Sciences and Disorders

Introduction. Despite having no inflection for case or agreement, Mandarin Chinese has flexible word order. This presents a challenge for sentence processing models to explain how Mandarin speakers manage conflicting cues to assign agent and patient status. Behavioral data have shown word order plays a role in the absence of competing cues, but is overridden by animacy¹. Electrophysiological (EEG) experiments have reported both N400² and semantic P600³ effects for Mandarin semantic reversals. For noun-noun-verb (NNV) sentences specifically, experiments suggest that the preferred word order is object-subject-verb with an inanimate object and an animate subject and that animacy and word order interact in a complex way⁴; NNV sentences with no animacy contrast may be uninterpretable⁵. Questions still remain about 1) the relative strength of each cue and their interactions, especially in the case of conflicting cues, and 2) whether semantic reversals elicit N400 or P600 effects.

Methods. We recorded Mandarin monolinguals' (n = 30) EEG while they read transitive NNV sentences word-by-word (SOA 750 ms) and judged which noun was the agent. To create cue competition, we manipulated four cues: **1) Agent Animacy** and **2) Reversibility** (irreversible sentences had a single plausible agent with opposite animacy status of the patient; reversible sentences had two equally plausible agents with shared animacy status (see [Table 1](#) for examples)); **3) Word Order** (each noun could appear in 1st or 2nd position); and **4) Structure** (presence/absence of coverbs BA and BEI, which assign explicit agent status to the preceding (BA) or following (BEI) noun phrase, as shown in [Table 2](#)).

Results. For the behavioral data, logistic mixed effects models analyzing the proportion of first noun agent choice showed interactions for each cue manipulation (shown in [Tables 1 and 2](#)). Unlike in earlier Competition Model experiments¹, Word Order did not affect argument assignment for reversible sentences. When present, coverbs BA/BEI were the strongest cues, but slightly less so if resulting in implausible readings (*servant BEI mirror polished*). Word Order interacted with Reversibility, so plausible agents were more likely to be chosen in irreversible sentences. Inanimate agents were overall selected slightly less often than animate agents.

For the EEG data, we used linear mixed effects models to analyze ERP amplitudes in time windows at multiple sentence positions. Within the first noun time window, ERPs were not influenced by animacy. In the BA/BEI time window, the BA character elicited a smaller P200 than BEI or a noun, and the noun elicited a larger N400 than both BA and BEI ([Figure 1A](#)). On the verb, we found a significant, biphasic N400/P600 effect for BA semantic reversals ([Figure 1B](#)) and a significant frontal P600-like positivity for BEI semantic reversals ([Figure 1C](#)).

Conclusion. We adapted behavioral methods used in Competition Model studies to an ERP paradigm evaluating argument structure processing in Mandarin NNV sentences. Our behavioral results confirm a preference for animate over inanimate agents⁴ and that NNV sentences without contrasting animacy are ambiguous⁵. In line with predictions from the extended Argument Dependency Model², we found an N400 effect for BA semantic reversals, which could mean that the 750 ms SOA was sufficient for participants to predict the verb. In line with Chow & Phillips³, both BA and BEI reversals elicited a P600-like positivity. The difference between BA and BEI reversals indicates that each coverb impacts argument assignment differently; BA may confirm an agent-first default processing strategy⁶, while BEI requires reanalysis before processing the verb, which may contribute to the greater P200 amplitude. These results highlight the importance of crosslinguistic comparison of sentence processing.

Table 1. Effect of Reversibility, Agent Animacy, and Word Order collapsed across Structure. Dashed line shows chance level noun selection. Error bars show standard deviation.

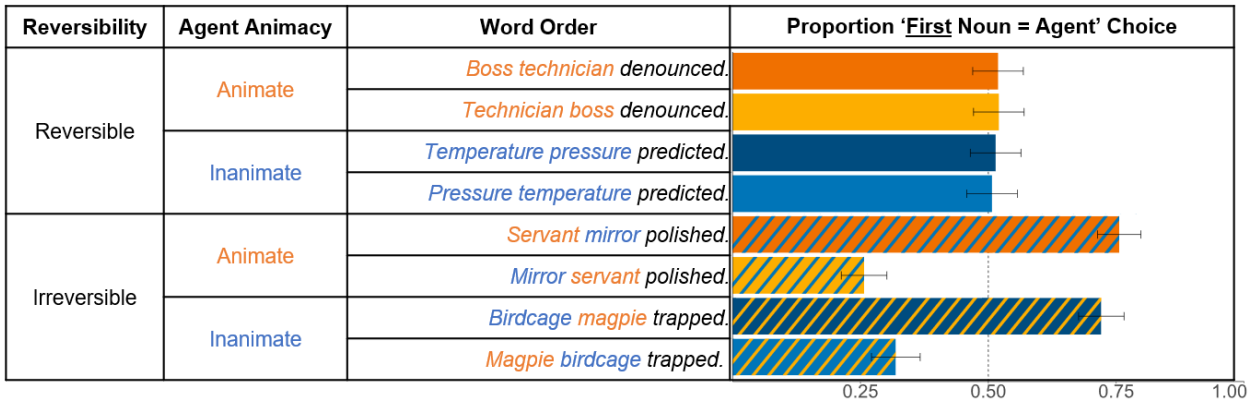


Table 2. Effect of Structure collapsed across other variables. Dashed line shows chance level noun selection. Error bars show standard deviation. LE is aspect marker in Mandarin.

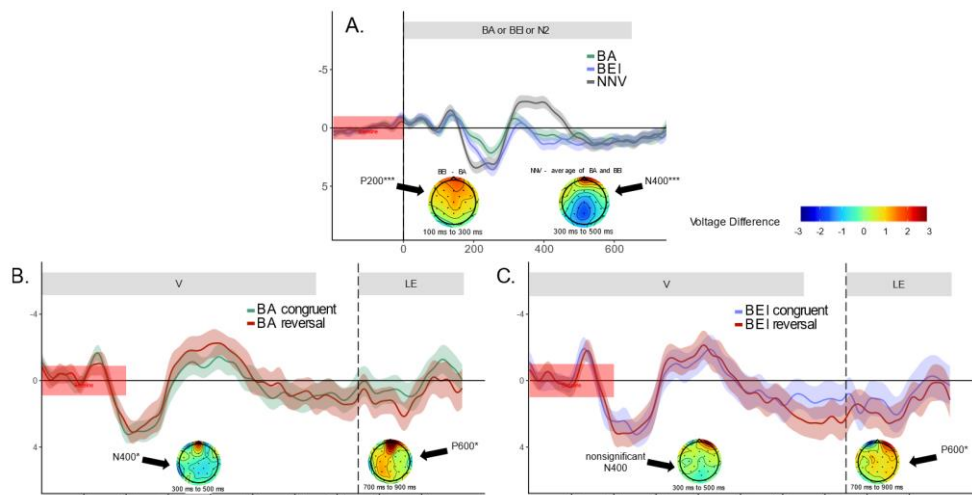
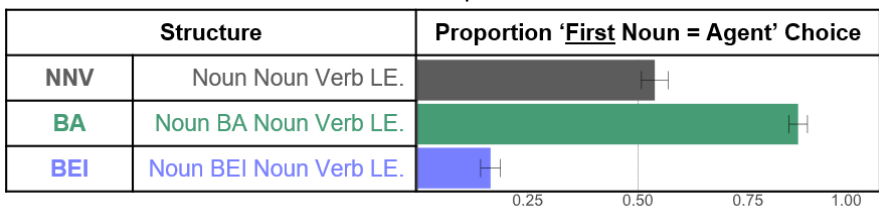


Figure 1. ERPs and voltage maps of select contrasts at Pz. Shading shows 95% confidence interval. **A.** Three Structure conditions at second word position (BA, BEI, or 2nd noun). **B.** Semantic reversal effect for BA irreversible sentences at verb onset, collapsed across animacy. **C.** Semantic reversal effect for BEI irreversible sentences at verb onset, collapsed across animacy. Verb-locked ERPs (B, C) were analyzed with pre- and post-onset baselines, and we determined that a post-onset baseline minimized spillover effects from preceding words. The post-onset baseline is shown here.

References: 1. Li, P et al. Cues as Functional Constraints on Sentence Processing in Chinese. in Language Processing in Chinese (eds Chen, HC & Tzeng, O) 207–234 (North-Holland, 1992). 2. Bornkessel-Schlesewsky, I et al. Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. Br. Lang. 117, 133–152 (2011). 3. Chow, WY & Phillips, C. No semantic illusions in the “Semantic P600” phenomenon: ERP evidence from Mandarin Chinese. Br. Res. 1506, 76–93 (2013). 4. Wang, L et al. The Role of Animacy in Online Argument Interpretation in Mandarin Chinese. in Case, Word Order and Prominence 40, 91–119 (2012). 5. Yu, S & Tamaoka, K. Age-related differences in the acceptability of non-canonical word orders in Mandarin Chinese. Ling. Sin. 4, (2018). 6. Wang, L et al. Exploring the nature of the ‘subject’-preference: Evidence from the online comprehension of simple sentences in Mandarin Chinese. Lang. Cogn. Pr. 24, 1180–1226 (2009).

PPI Illusion Ignores Binding but is Facilitated by Reactivation
Wesley Orth & Masaya Yoshida – Northwestern University

Introduction: NPIs are lexical items (e.g. “ever”, “any”) which are grammatically licensed by a negative element in a structural relation, c-command [1], as seen in the contrast between (1) and (2). There exists an illusion of grammaticality for NPI, such that the relative acceptability of sentences like (3), where a negative element (**no**) does *not* c-command the NPI (**ever**), is higher than the ungrammatical counterpart (2) containing no negative element [2-3]. Positive Polarity Items (PPI) are lexical items (e.g. “still”, “somewhat”) which are ungrammatical in environments that can host NPI as shown in (1) and (2) [4-6]. PPI are subject to an illusion of ungrammaticality in the environment where NPI are subject to the illusion of grammaticality [7].

- (1) **No** hunter who the fisherman believed to be trustworthy will **ever/still*** shoot a bear.
- (2) The hunter who the fisherman believed to be trustworthy will **ever*/still** shoot a bear.
- (3) The hunter who **no** fisherman believed to be trustworthy will **ever*/still?** shoot a bear.

PPI illusions are observed at the polarity item and are limited to negative elements which are also quantified expressions (e.g. “**no**”, “**not a single**”) [7-8]. In this series of studies, we aim to investigate if these illusions are sensitive to prior binding relationships involving the quantifier. We performed two experiments with a third follow-up experiment to be completed.

Experiment 1: To provide a negative quantified element that can generate illusions and bind a pronoun within the relative clause, we conducted a speeded acceptability judgment study with 71 participants comparing “**none of the NP**” and “**no NP**.” Participants viewed potential illusion sentences with these elements in the relative clause and baseline grammatical and ungrammatical controls following Orth Sloggett and Yoshida 2020. As shown in Table 1 and Figure 1, effects were found for grammaticality ($\beta = 0.462$, $t = 4.36$) and negative element presence ($\beta = 0.244$, $t = 2.73$). However, “**none of the NP**” and “**no NP**” were not statistically different, suggesting both phrases produce an illusion compared to the ungrammatical baseline.

Experiment 2: Having established the illusion generating ability of “**none of the NP**”, we conducted a maze task experiment with 39 participants to examine the role of established binding relationships in the PPI illusion [9-10]. The experiment employed a 2x2 gender mismatch paradigm, varying in negativity of the quantifier and pronoun gender as in (4).

- (4) The carpenter who $\begin{Bmatrix} \text{none} \\ \text{one} \end{Bmatrix}$ of the salesmen said believed $\begin{Bmatrix} \text{him} \\ \text{her} \end{Bmatrix}$ [about the tool] will **still**...

Log reaction time from the critical region “**still**” was analyzed using a deviation coded mixed effects model. With fixed terms for quantifier negativity and gender match and random intercepts for items and participants, we find that there is an interaction between the negativity of quantifier and the gender of the pronoun ($\beta = 0.122$, $t = 2.05$), such that a reading time penalty was observed when the quantifier was negative, and the pronoun matched the gender of the relative clause subject. Reading times of at the critical region visualized in Figure 2 and full model output is available in Table 2. The parser appears to be experiencing the PPI illusion of ungrammaticality, but only when the quantifier is binding the pronoun. Within a theory where the illusion is caused by the parser raising the relative clause quantifier to test for possible scope relations [8], this result suggests the parser performs raising recklessly without privileging existing binding relationships. One remaining question is why no illusion appears to occur when the quantifier is negative but does not bind the relative clause pronoun. This could be due to the distance between the negative quantifier and the polarity item, which has previously been shown to modulate the appearance of the NPI illusion [11]. If binding results in the reactivation of the quantifier, this could help preserve the negative quantifier in memory, allowing for the illusion to occur over greater distances than it otherwise would be able to.

Experiment 3: This follow-up experiment will test the role of distance utilizing items like (4), but with the manipulation being the gender of the pronoun and the presence of the prepositional phrase [about the tool]. In all sentences the relative clause quantifier will be negative, allowing us to observe if the binding relationship opens the possibility of a long-range PPI illusion.

[1] Ladusaw, W. A. (1980), [2] Drenhaus, H., Saddy, D., & Frisch, S. (2005), [3] Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008), [4] Homer, V. (2012), [5] Giannakidou, A. (2011), [6] Szabolcsi, A. (2004), [7] Orth, W., Yoshida, M., & Sloggett, S. (2020A), [8] Orth, W., Yoshida, M., Sloggett, S. (2020B), [9] Boyce, V., Futrell, R., & Levy, R. P. (2020), [10] Forster, K. I., Guerrera, C., and Elliot, L. (2009), [11] Parker, D., & Phillips, C. (2016)

Table 1: Fixed effects from logistic mixed effects regression. Helmert coded contrasts included for grammaticality (-1, 1/3, 1/3,1/3), negative element presence (0, 1/2, 1/2, -1), and illusion “no” vs illusion “none” (0, -1, 1, 0). Maximally convergent random slopes were also included.

Term	Estimate	Std.Error	Z value
(Intercept)	0.45261	0.14750	3.069
Grammatical	0.46233	0.10610	4.357
Negative Element Present	0.24457	0.08974	2.725
IllusionAvsB	0.04581	0.06847	0.669

Table 2: Fixed effects from linear mixed effects regression of log reading time at the critical region. Conditions were deviation coded with negativity conditions coded (one 1/2, **none** -1/2) and gender match coded (Match 1/2, Mismatch -1/2). Random intercepts for participant and item were also included.

Term	Estimate	Std.Error	T value
(Intercept)	-0.004	0.032	-0.151
Negativity	0.004	0.026	0.136
Gender Match	0.033	0.026	1.245
Negativity: Gender Match	0.122	0.060	2.045

Figure 1: **Proportion Acceptable NPI None/No Illusion**

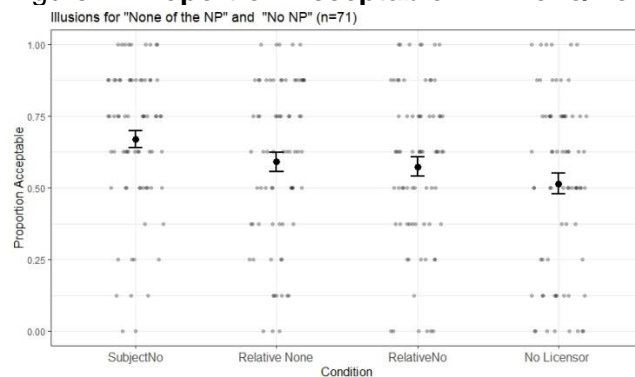
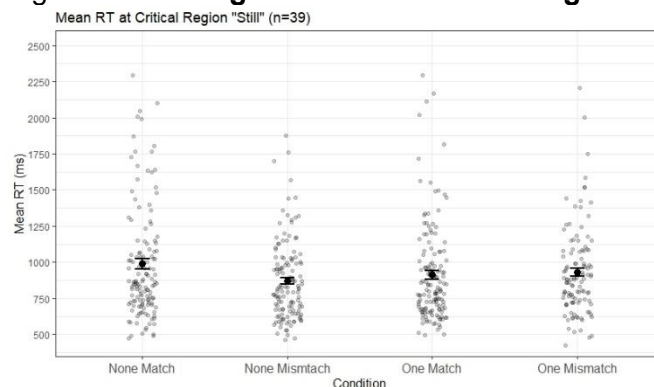


Figure 2: **Reading Time at the Critical Region: PPI “Still”**



Second Language Processing of Information at the Syntax-Discourse Interface

Didem Kurt & Nazik Dinçtopal Deniz (Boğaziçi University)

Background: This study examines how Turkish speakers of English process focus in English. Processing focus requires integrating syntactic and discourse-related information which second language (L2) learners may fail in due to insufficient resources for processing information at the interface of a sub-module (e.g., syntax) and an external domain (e.g., discourse) (Interface Hypothesis (IH), [1]). Broad focus, associated with sentential stress, [2] is placed mostly on the immediately preverbal constituent in Turkish [3] but on the rightmost constituent in English [4]. Turkish commonly marks narrow focus by moving a constituent to the immediately preverbal position [3]. Focus marking via word-order in English is observed in a very limited context, e.g., ditransitives where the rightmost constituent presents new information [5]. The experimental sentences in the present study include ditransitives in double-object or dative alternation forms. To distinguish narrow and broad focus, changes to the syntactic structure (i.e., word-order) need to be associated with discourse information (i.e., given-before-new principle). L2 speakers' insensitivity to word-order changes to mark narrow focus would indicate failure in using information at the syntax-discourse interface supporting the IH. If L2 speakers mark focus on the subject, the only preverbal constituent in English, that would indicate transferring an L1 constraint to the L2. An eye-tracking experiment and a sentence completion task were conducted with L1 speakers of English ($N = 8, 21$, respectively) and advanced Turkish speakers of English ($N = 47$). (L1 eye-tracking data collection paused due to the Covid-19 pandemic.) **Materials:** The experimental sentences included ditransitive verbs and replacive phrases as in (1). The focus structure of the main clause (broad/narrow) and congruency of the replacive phrase (congruent/incongruent) with the focused (i.e., rightmost) constituent were manipulated. Focus was manipulated via word-order: canonical as in (1a-c) where the rightmost/indirect object, *to the director*, would have broad focus, or non-canonical as in (1d-f) where the rightmost/direct object, *the flowers*, would have narrow focus. Congruency with the focused constituent was manipulated via the replacive phrase, *not*____, which contrasted either with the rightmost object as in (1a,d), and was congruent, or with another constituent, (the (in)direct object as in (1b,e) or the subject as in (1c,f)), and was incongruent. **Procedure:** In the eye-tracking experiment, the participants read the sentences and answered a comprehension question. The phrase following *not* was blank in the sentence completion task. The participants chose an option that could complete it with a phrase contrasting the (in)direct object or subject. **Results:** The eye-tracking data (see Table 1) were analyzed through mixed effects linear/logistic regression models. Processing narrow/broad focus was examined at the rightmost object (region 4) and its spill-over region (region 5). Sensitivity to the rightmost position as default focus position was examined at the replacive phrase (region 6) and its spill-over region (region 7). The L1 results were not reliable. The L2 data did not show any difference between broad and narrow focus conditions in any measure (t 's < 1.34). The analyses at the replacive phrase showed slow-downs for incongruency: contrast with the non-focused object for reading duration (RRD) and total duration (TD), t 's ≥ 2.05 , with the subject for RRD $t = 2.71$, $p < .05$. Both groups of participants preferred to complete the sentences with a phrase that would contrast with the rightmost constituent ($M_{L1} = 63\%$, $M_{L2} = 54\%$; see Table 2). **Discussion:** L2 speakers' sentence completions and sensitivity to the (in)congruity in the eye-tracking experiment show that they have acquired the syntactic information that focus is placed on the rightmost object in English. Their insensitivity to changes to word-order to mark broad/narrow focus indicates failure to use information at the syntax-discourse interface. This is in line with the IH [1]. But it may also be because the participants might have had insufficient input for scrambling as a strategy to mark focus in English. Or, they may, in general, over-rely on "good enough" processing [6] and fail to distinguish broad and narrow focus as the latter would require deeper processing [7].

References: [1] Sorace, A. (2011). *Ling. Appr. to Biling.*, 1, 1-33. [2] Kahnemuyipour, A. (2009). *The syntax of sentential stress*. [3] Göksel & Özsoy (2003). *Lingua*, 113, 1143-1167. [4] Carlson et al. (2009). *Quart. J. of Expt. Psych.*, 62, 114–139. [5] Brown et al. (2012). *JML*, 66, 194–209. [6] Ferreira et al. (2002). *Cur. Dir. in Psych. Sci.*, 11, 11–15. [7] Lowder & Gordon (2015). *Psych. Bullet. & Rev.*, 22(6), 1733–1738.

(1) Experimental Sentences: Conditions distinguish for broad (B) vs. narrow (N) focus and Congruent (C) vs. Incongruent (InC) replacive phrase with the direct object (DO), indirect object (IO) or the subject (SU). Regions are shown via “/” and subscripted numbers.

a.B-N/C-IO: The presenter/₁ gave/₂ the flowers/₃ to the director/₄ yesterday/₅, not to the actress/₆.

b.B-N/InC-DO: The presenter/₁ gave/₂ the flowers/₃ to the director/₄ yesterday/₅, not the prizes/₆.

c.B-N/InC-SU: The presenter/₁ gave/₂ the flowers/₃ to the director/₄ yesterday/₅, not the organizer/₆.

d.N-N/C-DO: The presenter/₁ gave/₂ the director/₃ the flowers/₄ yesterday/₅, not the prizes/₆.

e.N-N/InC-IO: The presenter/₁ gave/₂ the director/₃ the flowers/₄ yesterday/₅, not to the actress/₆.

f.N-N/InC-SU: The presenter/₁ gave/₂ the director/₃ the flowers/₄ yesterday/₅, not the organizer/₆.

*All conditions (1a-f) were followed by a content-neutral phrase (e.g., “It was/₇the procedure/₈).

Table 1. Mean values (and standard errors in parentheses) for first fixation duration (FFD), gaze duration (GD), regression path duration (RPD), re-reading duration (RRD), total duration (TD) (in milliseconds) and probability of regression out (PRO) for regions 5 and 7. Congruent conditions are in bold face. B-N focus is marked in green, N-N focus is marked in blue.

		FFD	GD	RPD	RRD	TD	PRO
Region 5	B-N/C-IO	242 (7.5)	286 (12.1)	392 (23.4)	150 (17.4)	419 (22.05)	.19 (0.03)
	B-N/InC-DO	244 (8.07)	296 (13.5)	374 (24.2)	144 (19.4)	417 (22.7)	.12 (0.02)
	B-N/InC-SU	240 (9.02)	284 (13.2)	352 (25.8)	182 (21.9)	449 (25.6)	.09 (0.02)
	N-N/C-DO	236 (8.29)	291 (12.5)	434 (29.7)	191 (24.1)	440 (24.8)	.23 (0.03)
	N-N/InC-IO	249 (9.96)	278 (13.4)	379 (24.0)	176 (21.9)	447 (24.4)	.20 (0.03)
	N-N/InC-SU	238 (7.36)	277 (11.5)	409 (26.5)	163 (17.9)	437 (20.6)	.25 (0.03)
		FFD	GD	RPD	RRD	TD	PRO
Region 7	B-N/C-IO	239 (7.31)	308 (13.7)	377 (29.0)	116 (22.4)	317 (20.7)	.11 (0.03)
	B-N/InC-DO	243 (7.15)	297 (13.6)	404 (34.2)	168 (24.7)	362 (22.3)	.08 (0.02)
	B-N/InC-SU	248 (8.37)	308 (14.4)	416 (40.3)	149 (26.2)	341 (24.2)	.08 (0.03)
	N-N/C-DO	250 (7.82)	311 (14.9)	383 (31.6)	95(20.4)	328 (23.8)	.09 (0.03)
	N-N/InC-IO	245 (8.02)	298 (11.9)	379 (34.5)	180 (27.3)	332 (23.0)	.06 (0.02)
	N-N/InC-SU	249 (7.98)	318 (15.7)	455 (35.4)	155 (26.2)	367 (26.7)	.18 (0.03)

Table 2. Percent sentence completion preferences.

Conditions		L1	L2
Canonical Order (Broad Focus)	C-IO	32.7	26.6
	InC-DO	12.6	15.8
	InC-SU	4.5	7.5
Non-Canonical Order (Narrow Focus)	C-DO	30.3	27.39
	InC-IO	14.6	14.47
	InC-SU	4.9	8.1

English locative inversions are not special in terms of their discourse function

Giuseppe Ricciardi (Harvard University), Rachel Ryskin (UC Merced), & Edward Gibson (MIT)

It is widely assumed that one of the factors underlying word order variation in (at least the Subject-Verb) languages is the tendency to place information assumed to be already known from the previous discourse before information assumed to be new (Halliday, 1967; Chafe 1976; Gundel, 1988; Prince, 1992). Experimental evidence for the existence of an ‘old-before-new principle’ has been extensively offered in the case of the English dative alternation which allows speakers to choose the relative order of the two post-verbal arguments (Arnold et al. 2000; Frazier, 2004; Brown et al. 2012). Here, we focus on a different case of word order alternation, i.e. the case of the English ‘locative inversion’ -- e.g. [_{PP} *Behind the box*] *lay* [_{NP} *a knife*] -- where the relative order of the locative prepositional phrase and the subject noun phrase is “inverted” with respect to the more canonical word order -- e.g. [_{NP} *A knife*] *lay* [_{PP} *behind the box*]. Birner & Ward (Birner, 1996; Birner & Ward, 1998; Ward & Birner, 2004, 2019; B&W) propose that the ‘old-before-new principle’ plays a different role across the two word order options: for the canonical NP-V-PP the ‘old-before-new principle’ represents just a tendency which doesn’t affect the felicity of the sentence, but for the non-canonical inverted PP-V-NP the same principle imposes a stricter requirement on the felicity of the sentence, i.e. the PP must not represent information that is less familiar in the discourse than that represented by the NP. In this work, we aim to test this hypothesis in a sentence acceptability rating study. **Methodology (see 1):** Participants read two sentences and were asked to rate the second sentence within the context of the first; we adopted a 2X2X2 design by manipulating the second (critical) sentence with respect to its word order (NP-V-PP vs PP-V-NP) and the information status of the preverbal constituent (new vs old) and of the postverbal constituent (new vs old). “Old” constituents were explicitly mentioned in the context sentence and preceded by the definite article “the”; “New” constituents were not previously mentioned and were preceded by the indefinite article “a.”. **Predictions:** If prior findings generalize to locatives, then sentences with “old” first constituents will be rated as more acceptable (e.g., Arnold, 2000) than sentences with “new” first constituents. We take B&W’s account as predicting a three-way interaction among word order, discourse status of the first constituent, and discourse status of the second constituent such that the rating of sentences with ‘new’ first constituents (but not the ‘old’ first constituents) will be lower when the second constituent is ‘old’ compared to when it is ‘new’ but only for the PPvNP word order. **Results:** Ratings from 2 experiments (E1: N=51, see Fig. 1; E2: N=57, see Fig. 2) were analyzed with mixed-effects linear models with three effects-coded fixed effects and their interactions. We found that NPvPP sentences were rated as more acceptable than PPvNP (E1: $b=0.70$, $p<.001$; E2: $b=0.59$, $p<.001$) and sentences with “old” first were rated as more acceptable than those with “new” first (E1: $b=-0.19$, $p<.01$; E2: $b=-0.32$, $p<.001$). Crucially, we did not find any significant three-way interaction - in either experiment - among word order, discourse status of the first constituent and of the second constituent (E1: $b=0.31$, $t=1.09$, $p=0.28$; E2: $b=-0.15$, $t=-0.77$, $p=0.44$). **Conclusion** These findings extend the ‘old-before-new’ principle to inverted locatives and fail to support B&W’s hypothesis that discourse constraints play a different role across the English canonical word order NP-V-PP and the inverted PP-V-NP. Though caution is warranted when interpreting null effects, these results suggest that English speakers may prefer that the first constituent of a sentence represent discourse ‘old’ information no matter the specific word order of the sentence.

(1) Sample test item

PPvNP-old-old

Paragraph(E1)/Context(E2): The police officer entered the room and saw a hunting **weapon**, a broken chair, a **box**, and a scary painting.

Behind the box lay the weapon.

PPvNP-old-new

P/C: The police officer entered the room and saw a half-empty whiskey bottle, a broken chair, a **box**, and a scary painting.

Behind the box lay a weapon.

PPvNP-new-old

P/C: The police officer entered the room and saw a hunting **weapon**, a broken chair, an open cupboard, and a scary painting.

Behind a box lay the weapon.

PPvNP-new-new

P/C: The police officer entered the room and saw a half-empty whiskey bottle, a broken chair, an open cupboard, and a scary painting.

Behind a box lay a weapon.

NPvPP-old-old

P/C: The police officer entered the room and saw a hunting **weapon**, a broken chair, a **box**, and a scary painting.

The weapon lay behind the box.

NPvPP-old-new

P/C: The police officer entered the room and saw a hunting **weapon**, a broken chair, an open cupboard, and a scary painting.

The weapon lay behind a box.

NPvPP-new-old

P/C: The police officer entered the room and saw a half-empty whiskey bottle, a broken chair, a **box**, and a scary painting.

A weapon lay behind the box.

NPvPP-new-new

P/C: The police officer entered the room and saw a half-empty whiskey bottle, a broken chair, an open cupboard, and a scary painting.

A weapon lay behind a box.

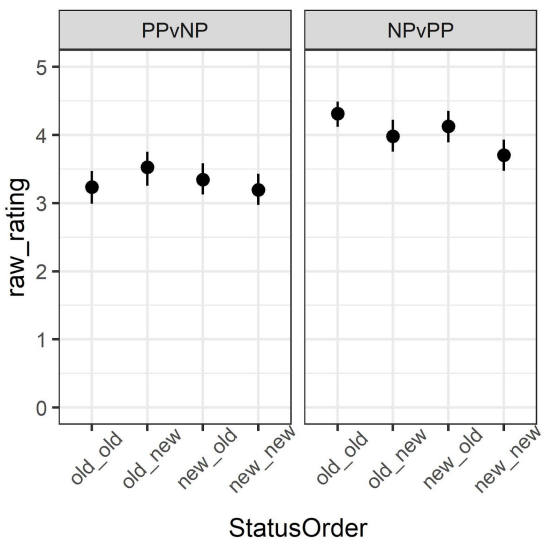


Fig. 1 Mean ratings for E1 by discourse status order condition. Error bars reflect bootstrapped 95% CI. Prompt for participants: "Rate how natural the bolded sentence is within the paragraph"

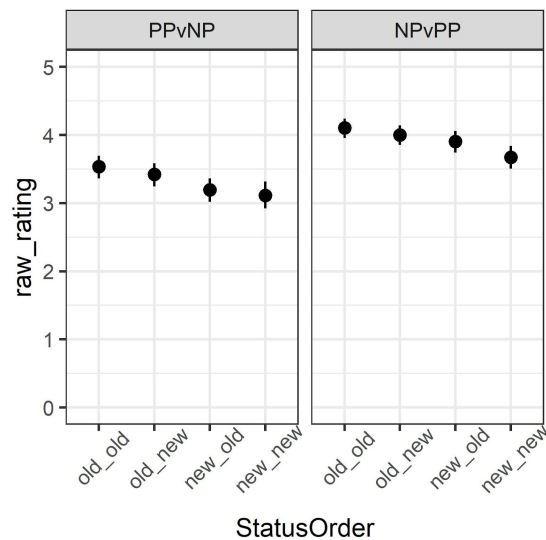


Fig. 2 Mean ratings for E2 by discourse status order condition. Error bars reflect bootstrapped 95% CI. Prompt for participants: "Rate how natural the bolded sentence is within the context"

#fitspo: Cognitive Implications of Interacting with “Fitspiration” Content on Social Media

Jordan Zimmerman,^{1,2} Angelica De Rezende³, Anna Wright¹, Kaitlin Lord¹, Sarah Brown-Schmidt¹

(1- Vanderbilt University, 2- Massachusetts General Hospital, 3- Florida International University)

Social media is a routine part of every-day life for hundreds of millions of people worldwide. Here we examine how communicating on social media shapes enduring memories for that experience¹. Describing an object boosts memory for that object and other related objects (e.g. a striped shirt when describing “*dotted shirt*”)². Two studies explore whether commenting also boosts memory for related content in the same context. Understanding the cognitive processes involved is important given the ubiquity of social media, and popularity of potentially problematic content such as imagery intended to invoke dieting and fitness inspiration, widely known as “fitspiration.” Prior findings indicate that exposure to #fitspo is associated with unhealthy behaviors³⁻⁴, and motivates exploratory analyses relating memory for food and fitness social media and individual differences in eating behaviors and self-image.

In **E1** (N=210) participants (Ps) were recruited online through Qualtrics Panels. Materials were real Instagram posts featuring “healthy” food, and men and women engaging in fitness activities. Posts featuring dogs, cats, nature were used as control. In total, Ps viewed 270 Instagram posts (60 food, 60 fitness, 150 control), in 30 3x3 arrays, similar to Instagram’s “explore” feature. For each of 30 arrays, Ps were asked to comment on a target image enclosed in a green box, as they would on their own feed. After a brief distractor task, Ps completed a recognition memory test for the critical stimuli (commented-upon and non-commented food and fitness images). They viewed 240 food and fitness images (half old and seen in the exposure phase, half new and closely matched to old items; counterbalanced across lists), and responding old/new. Lastly, Ps completed the Eating Disorder Examination Questionnaire (EDE-Q) to assess eating disorder symptomology. **Results:** Comments were coded for # of words (e.g. “*So yummy!*”, “*She’s really flexible*”). Memory data were analyzed with logistic mixed-effects models. Ps were more likely to correctly recognize images that they commented on vs. ones viewed in the same array ($b=2.05$, $p<.0001$), and correct recognition of targets (but not non-targets) increased with comment length ($b=.103$, $p<.01$). Non-targets from the same category as target (i.e., non-target fitness images when commenting on a different fitness image) were better recognized compared to control images ($b=.14$, $p<.01$). Exploratory analyses of individual differences revealed the memory boost for target images was *negatively* related to EDE global scores ($r = -0.71$) such that individuals with more symptomology had less boost.

E2 (N=300, MTurk) was a replication of E1 with a 2AFC paradigm. **Results:** As in E1, commented-upon images were better recognized than those viewed in the context ($b=1.88$, $p<.0001$), and correct recognition of targets (but not non-targets) increased with comment length ($b=.11$, $p<.0001$). Unlike E1 accuracy for non-targets from the same category as the target was not higher than unrelated control images ($b=.075$, $p=.07$). Similar to E1, the memory boost for target over context images was *negatively* associated with EDE scores ($r = -.27$).

Conclusion: We find that when browsing real Instagram images in arrays similar to the explore feature, that the act of commenting on a post boosts memory for that post, and the longer the comment, the larger the boost. Evidence for a commenting-related boost to related imagery in context (e.g. memory for fitness images in the context when commenting on a different fitness image) was equivocal, suggesting that the effect, if real, may be quite small. We speculate that unlike task-based conversation², social media arrays may demand less consideration of context when generating descriptions. Exploratory analyses of individual differences revealed the higher the severity of reported disordered eating behaviors, the *less* of a boost Ps experienced in memory for target over control images, possibly because persons with higher scores distributed attention more equally to the target and context. This leads to the intriguing possibility that individual differences in what is self-relevant in context may modulate attentional distribution in communication -- observed here in the context of social media.

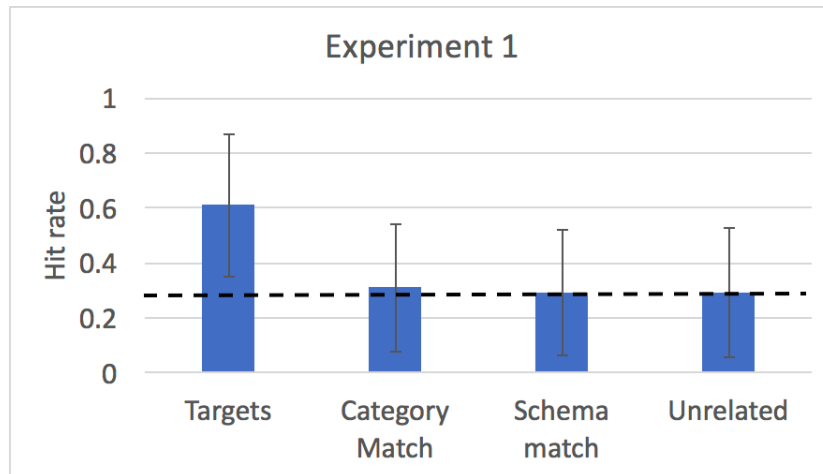


Figure 1. Experiment 1 hit rate on memory test by image type and target type; the false alarm rate (28%) is indicated by the dotted line. Error bars indicate by-participant standard deviation.

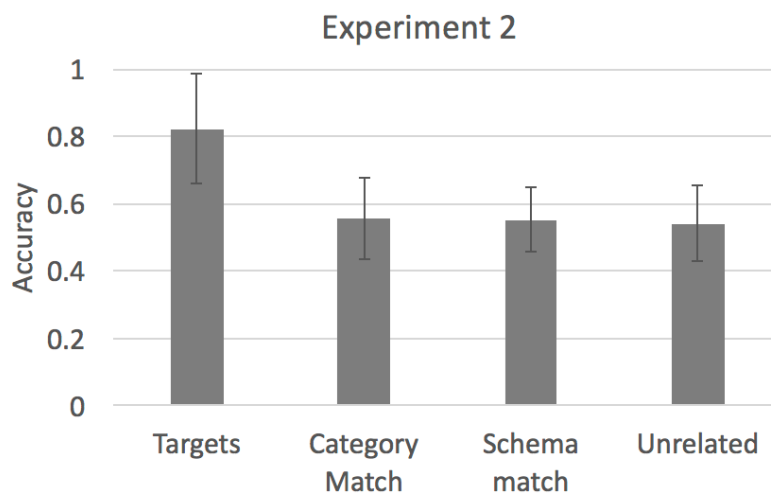


Figure 2. Experiment 2: Accuracy by image type. Error bars indicate by-participant standard deviation.

References

1. Zimmerman, J. & Brown-Schmidt, S. (2020). #foodie: Implications of interacting with social media for memory. *Cognitive Research: Principles and Implications*, 5:16. <https://doi.org/10.1186/s41235-020-00216-7>
2. Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, 154, 102-117.
3. Griffiths, S., & Stefanovski, A. (2019). Thinspiration and fitspiration in everyday life: An experience sampling study. *Body image*, 30, 135-144.
4. Holland, G., & Tiggemann, M. (2017). "Strong beats skinny every time": Disordered eating and compulsive exercise in women who post fitspiration on Instagram. *International Journal of Eating Disorders*, 50(1), 76-79.

The Role of Sensory Experience and Communication in the Neural Mechanisms Supporting Social Communicative Processes: An fNIRS Hyperscanning Study.

Introduction.

The neural networks for language processing have classically been considered as arising from modality-specific processes, until studies investigating signed languages demonstrated that these areas were responsible for carrying out linguistic functions regardless of language modality (i.e., functional specificity) (Nishimura, et al., 1999; Petitto et al., 2000; Cardin et al., 2016). In this study, we aim to broaden our understanding of the functional specificity of the high-level language processing neural networks by investigating the much less studied tactile language. We specifically test the extent of functional specificity of the left lateralized language network covering the inferior frontal gyrus (IFG) and superior temporal gyrus (STG) areas. If the language network is function specific rather than modality linked, we hypothesize that tactile signed languages perceived without sight or sound recruit the same canonical language regions as spoken and visual-based signed languages both for production, that is the IFG, and perception, that is the STG.

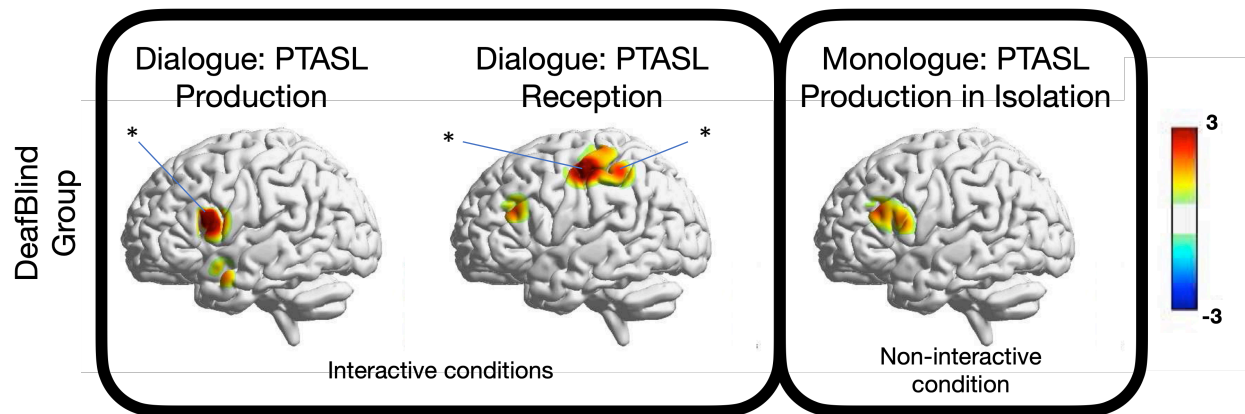
Method. We recruited 8 DeafBlind adults who use a tactile language: ProTactile ASL (PTASL). Neural activations were recorded with an fNIRS continuous-wave dual-brain imaging paradigm (i.e., hyperscanning), with time-locked recordings over the left hemispheres. This method has successfully been used for studying language processes, is portable and provides a greater ecological testing environment. Participants were divided in dyads and undertook a dialogue (production and perception) and a monologue (production only) task. For both tasks, participants were first given an object to explore before either naming and describing the object to themselves or before naming and describing the object to their partner. In the dialogue task, partners took turns in their roles: while one dyad member described an object, the other dyad member perceived the description.

Results. All contrasts were run against the baseline as statistical power was too low to compare two active tasks. The monologue task (production) showed activation in the left IFG (Figure 1). For the dialogue task, language production showed similar activation in the left IFG which is typically recruited in other language modalities. Language perception in the dialogue task showed activations in the somatosensory areas rather than the canonical left STG area.

Discussion.

Language production results support our hypothesis, and are consistent with research in visual languages, that PTASL shows functional specificity for high-level language processes and recruits the canonical left IFG. Language perception in the dialogue task shows novel and surprising results: perceiving language occurred in the somatosensory areas with no significant activations in Wernicke's area. This may be explained by several factors: First, language perception in PTASL requires a unique cooperative process where the listener's top-down anticipatory processing of linguistic utterances informs co-articulation of linguistic content. Interestingly, these somatosensory activations were found only in the listener role in the dialogue condition, and not the talker role, strengthening the idea that the somatosensory areas play a specific role related to tactile language perception. Second, the absence of activity in Wernicke's area (i.e., IFG) could be explained by low statistical power of our sample. However, activations for both production conditions (monologue and dialogue) resulted significant suggesting that we should have been able to detect activation in other language areas. Importantly, these results might be specific to our group of participants as none were native, from birth, PTASL users but varied in the time of acquisition of PTASL. It is possible that these activations resemble more those of second language learners who might show neural adaptation occurring later in life. Together, these findings are indicative of a functional specificity in the language *production* network, with novel adaptability of the human brain in order to perform the same language *perception* functions irrespective of modality differences.

Figure 1: Results for the contrasts against baseline showing IFG activation for the language production conditions and novel somatosensory recruitment for language perception. Due to the small power of our groups, maps are presented with a threshold of $p < .01$ to visualize the context in which peak activations occurred. Peaks at $p < 0.05$ are represented with *.



References:

- Cardin, V., Orfanidou, E., Kästner, L., Rönneberg, J., Woll, B., Capek, C. M., & Rudner, M. (2016). Monitoring Different Phonological Parameters of Sign Language Engages the Same Cortical Language Network but Distinctive Perceptual Ones. *Journal of Cognitive Neuroscience*, 28(1), 20–40. http://doi.org/10.1162/jocn_a_00872
- Hirsch, J., Noah, J. A., Zhang, X., Dravida, S., & Ono, Y. (2018). A cross-brain neural mechanism for human-to-human verbal communication. *Social Cognitive and Affective Neuroscience*, 13(9), 907–920. <http://doi.org/10.1093/scan/nsy070>
- Hirsch, J., Zhang, X., Noah, J. A., & Ono, Y. (2017). Frontal temporal and parietal systems synchronize within and across brains during live eye-to-eye contact. *NeuroImage*, 157, 314–330. <http://doi.org/10.1016/j.neuroimage.2017.06.018>
- Nishimura, H., Hashikawa, K., Doi, K., Iwaki, T., Watanabe, Y., Kusuoka, H., et al. (1999). Sign language “heard” in the auditory cortex. *Nature*, 397(6715), 116–116. <http://doi.org/10.1038/16376>
- Petitto, L. A., Zatorre, R. J., Gauna, K., Nikelski, E. J., Dostie, D., & Evans, A. C. (2000). Speech-like cerebral activity in profoundly deaf people processing signed languages: Implications for the neural basis of human language. *Proceedings of the National Academy of Sciences*, 97(25), 13961–13966. <http://doi.org/10.1073/pnas.97.25.13961>

Decomposing the focus effect: Evidence from reading
Morwenna Hoeks, Maziar Toosarvandani & Amanda Rysling
University of California Santa Cruz

Investigations of linguistic focus in reading have found mixed results. Some report a decrease in reading times on focused material [9, 4], while others report an increase [3, 2, 8, 12]; see Table 1. We show that these inconsistencies are clarified by a notion of focus that is more informed by formal semantics. While previous work explained slowdowns on foci by appealing to their newness, foci need not be new [11, 1], as in (1), where *article* is repeated but also focused in (1b).

- (1) a. Did Sarah read an article about penguins, or a book?
b. Sarah only read an [ARTICLE]_F about penguins.

A focus particle like *only* in (1b) contributes to the meaning of its sentence by negating alternate versions of that sentence that differ solely in the focus, i.e., (1b) conveys that Sarah didn't read a book. Some theories take this further, analyzing every focus as negating *alternatives*, all those expressions that contrast with the focus [10]. No previous study on focus in reading explicitly manipulated alternatives as such, but Table 1 shows that only studies in which alternatives were mentioned in preceding contexts found speed-ups in reading times on foci. In addition, theories like [10]'s do not treat newness and focus as coextensive, and instead determine focus in question-answer pairs by whether a word alone completely answers a preceding question.

E1: Question-answer pairs manipulated whether a target word (*lawyer* in 2) was focused (\pm FOC) or newly mentioned (\pm NEW). In +FOC conditions, the target was in focus, because it was a complete answer to the preceding question, while in -FOC conditions, the target was not; in (2), the reader can also verify the presence/absence of accent on target *lawyer* in response to different questions. In -NEW conditions, the target was mentioned in the question, and in +NEW conditions, it was not.

- | <p>(2) Speaker A: "This company often makes bad decisions, but..."</p> <p>a. ...did they hire a lawyer last fall, or an accountant?"</p> <p>b. ...did they hire a lawyer last fall?"</p> <p>c. ...did they hire an accountant last fall?"</p> <p>d. ...what did they announce this time?"</p> <p>Speaker B: "I think they announced they hired {\emptyset only} a lawyer last fall, but I'm not sure."</p> | <table border="0"> <tr> <th style="text-align: left; padding-right: 10px;">E1</th> <th style="text-align: left;">E2</th> </tr> <tr> <td>-NEW +FOC</td> <td>-NEW +ALT</td> </tr> <tr> <td>-NEW -FOC</td> <td>-NEW -ALT</td> </tr> <tr> <td>+NEW +FOC</td> <td>+NEW +ALT</td> </tr> <tr> <td>+NEW -FOC</td> <td>+NEW -ALT</td> </tr> </table> | E1 | E2 | -NEW +FOC | -NEW +ALT | -NEW -FOC | -NEW -ALT | +NEW +FOC | +NEW +ALT | +NEW -FOC | +NEW -ALT |
|--|---|----|----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| E1 | E2 | | | | | | | | | | |
| -NEW +FOC | -NEW +ALT | | | | | | | | | | |
| -NEW -FOC | -NEW -ALT | | | | | | | | | | |
| +NEW +FOC | +NEW +ALT | | | | | | | | | | |
| +NEW -FOC | +NEW -ALT | | | | | | | | | | |

E2: Sentences in E2 were identical to E1, except the focus particle *only* was added to unambiguously focus the target in all conditions. Identical preceding questions manipulated whether the target was newly mentioned (\pm NEW) or a contextual alternative was present (\pm ALT). Questions in +ALT conditions mentioned an alternative (*accountant*); questions in -ALT conditions did not.

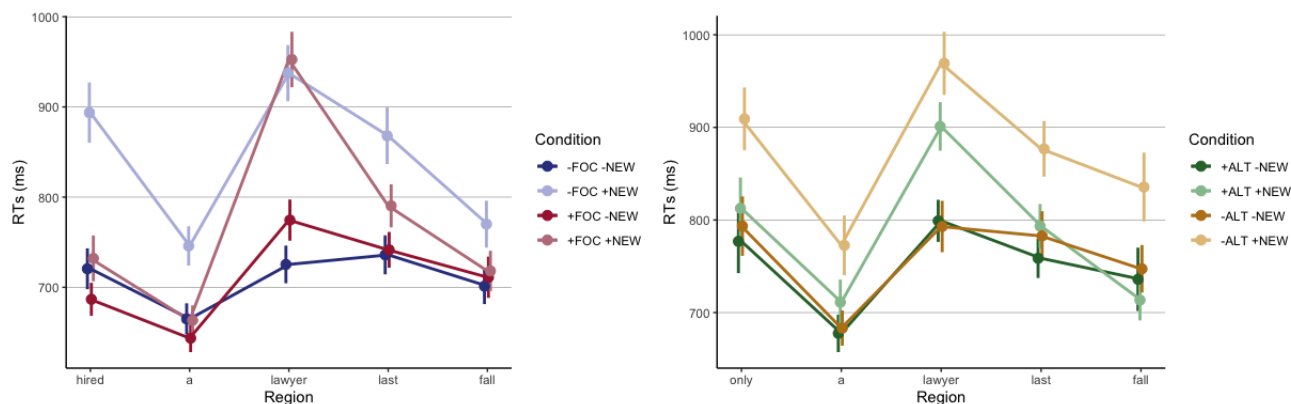
Method: For both E1/E2 ($n = 48$ each), 48 items like (2) were constructed, and target sentences were presented using the Maze task [5]. In this task, each word in the target sentence is shown alongside a foil and participants progress through the sentence by choosing correct continuations.

Results: Figs. 1 and 2 show RTs from E1 and E2. Mixed effects linear regressions with full random effects structure found a significant effect of \pm NEW on the target in both experiments, with longer RTs in +NEW conditions than -NEW conditions (E1: $t = 7.82$, E2: $t = 5.27$). In **E1** they revealed a significant main effect of \pm FOC, such that +FOC targets showed longer RTs than -FOC targets ($t = 3.23$), and in **E2**, a significant interaction between \pm NEW and the \pm ALT: RTs were longer in -ALT conditions than +ALT conditions only when the target was also +NEW ($t = 2.60$).

Conclusion: In line with [2, 8, 3, 12], we found an overall slowdown in RTs for foci compared to non-foci, suggesting a focus cost that does not reduce to newness. But this focus slowdown was modulated by context: RTs were longer on +NEW foci than on -NEW foci. When alternatives to foci were contextually mentioned, the slowdown on new foci was significantly reduced. This suggests that presenting information about alternatives aids reading of foci, thus providing converging evidence for the role of alternatives in focus processing [7, 6]. Controlling for newness versus focus and contextual mention of alternatives clarifies the earlier results summarized in Table 1: previous work only found a focus speed-up after contextual mention of alternatives with no newness contrast between foci and baselines, and only found a slowdown in the absence of alternatives.

		Inhibition		Facilitation		ALT	NEW	
		Early	Late	Early	Late		Focus	Baseline
Birch & Rayner (2010)		×	×	✓	✓	present	new	new
Morris & Folk (1998)		×	×	×	✓	present	new	new
Ward & Sturt (2007)		×	×	×	×	absent	new	new
Birch & Rayner (1997)	Exp 1	×	✓	×	×	absent	new	new
Lowder & Gordon (2015)		✓	✓	×	×	absent	new	new
Birch & Rayner (1997)	Exp2	✓	✓	×	×	absent	new	given
Benatar & Clifton (2014)	Exp 1 & 2	✓	✓	×	×	absent	new	given
Benatar & Clifton (2014)	Exp 3	✓	✓	×	×	absent	new	given
Sloggett et al. (2019)		✓	✓	×	×	absent	new	given

Table 1: Overview of context manipulations in previous work on focus in reading



References

- [1] D. Beaver, B. Z. Clark, E. Flemming, T. F. Jaeger, and M. Wolters. When semantics meets phonetics: Acoustical studies of second-occurrence focus. *Language*, 83:245–276, 2007.
- [2] A. Benatar and C. Clifton. Newness, givenness and discourse updating: Evidence from eye movements. *Journal of Memory and Language*, 71(1):1–16, 2014.
- [3] S. L. Birch and K. Rayner. Linguistic focus affects eye movements during reading. *Memory & Cognition*, 25(5):653–660, 1997.
- [4] S. L. Birch and K. Rayner. Effects of syntactic prominence on eye movements during reading. *Memory & cognition*, 38(6):740–752, 2010.
- [5] K. I. Forster, C. Guerrero, and L. Elliot. The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171, 2009.
- [6] N. Gotzner, I. Wartenburger, and K. Spalek. The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, 8(1):59–95, 2016.
- [7] E. Husband and F. Ferreira. The role of selection in generating focus alternatives. *Language, Cognition and Neuroscience*, 31(2):217–235, 2015.
- [8] M. W. Lowder and P. C. Gordon. Focus takes time: structural effects on reading. *Psychonomic bulletin & review*, 22(6):1733–1738, 2015.
- [9] R. K. Morris and J. R. Folk. Focus as a contextual priming mechanism in reading. *Memory & Cognition*, 26(6):1313–1322, 1998.
- [10] M. Rooth. *Association with Focus*. PhD thesis, University of Massachusetts, Amherst, 1985.
- [11] M. Rooth. On the interface properties for intonational focus. *Semantics and Linguistic Theory (SALT)*, 6:202–226, 1996.
- [12] S. Sloggett, A. Rysling, and A. Staub. Linguistic focus as predictive attention allocation. 2019.

Syntactic focus activates mentioned and unmentioned alternatives in Samoan

Sasha Calhoun¹, Mengzhu Yan², Honiara Salanoa³, Fualuga Taupi³ and Emma Kruse Va'ai^{1,3}
(¹Victoria University of Wellington, New Zealand; ²Huazhong University of Science and Technology, Wuhan, China; ³National University of Samoa, Samoa)

Key functions of focus-marking are to highlight focused words and contrastive alternatives to them. For example, “*The visitor ate the CAKE*” (caps mark accent) emphasizes the *cake* and implies alternatives the visitor didn’t eat, e.g. *sandwiches*, that are relevant to the interpretation of the sentence (e.g. Rooth 1992). Consistent with this, a growing body of psycholinguistic evidence shows referents and their alternatives are more strongly activated when they are focus-marked than when they are not, whether or not the alternatives have been explicitly mentioned (e.g. Braun & Tagliapietra 2010, Fraundorf et al. 2010, Gotzner et al. 2016, Yan & Calhoun 2019). However, this evidence draws from a small number of languages, mostly Germanic, which primarily use prosodic prominence to mark focus (although some studies have looked at combinations of prosodic and morphosyntactic marking in these languages). We present the results of a probe recognition experiment (Gotzner et al. 2016) looking at activation of contrastive alternatives in the Austronesian language Samoan, which primarily uses syntactic focus marking (Calhoun 2015).

56 native speakers in Samoa heard short stories (see Table 1), which were said with neutral prosody. The context introduced alternatives to the subject and object in the critical sentence (e.g. people and foods). Then a continuation sentence repeated alternatives from each set so the number of their mentions was balanced across the story. In the critical sentence, an alternative to the object (e.g. *le keke* ‘cake’) was either focused or not using the cleft-like ‘o-fronting construction, which we have previously shown to be the primary marker of focus in Samoan in production and perception experiments (Calhoun 2015, Calhoun et al. 2019). Participants then saw a probe which was one of the object word, a mentioned or unmentioned alternative, or an unrelated control, and had to respond as quickly as possible whether the probe was in the story. There were 40 critical items, plus fillers with different story structures and probes.

A linear mixed effects model was built with logged response time as the dependent variable for the 1,901 correct responses. The final model included fixed effects of probe type, $X^2(3) = 81.9$, $p < 0.0001$; focus condition, $X^2(1) = 0.01$, $p = 0.91$; their interaction, $X^2(3) = 8.25$, $p = 0.041$; and the position of the trial in the experiment, $X^2(1) = 9.63$, $p = 0.001$; as well as random intercepts for participants and the probe word. The *step* function in the *lmerTest* package (Kuznetsova et al. 2017) was used to remove non-significant effects, which were the participants’ gender, age and relative language dominance in Samoan versus English (almost all Samoan speakers are at least partly bilingual with English, Kruse Va’ai, 2011), as well as a random slope for the probe-focus interaction by participant. Figure 1 shows estimated RTs extracted from the model. Planned comparisons were carried out using *emmeans* (Lenth 2020) with the FDR correction.

The comparisons showed object probes were recognised faster than unrelated regardless of focus. However, for both mentioned and unmentioned probes, listeners were slower to correctly respond if the object was focused than not, compared to unrelated. Further, mentioned probes were faster than unmentioned only if the object was not focused. These findings show syntactic focus marking makes it harder to correctly distinguish mentioned and unmentioned alternatives; similar to what has previously been shown in Germanic (e.g. Gotzner et al. 2016). This is because focus-marking activates all plausible alternatives, including those not mentioned in the discourse.

This study contributes to psycholinguistic evidence that focus-marking is a common mechanism cross-linguistically to heighten activation of contrastive alternatives. This evidence has important implications for the mechanisms supporting resolution of implicature and referent tracking. Cross-linguistic differences lie in the types and relative importance of different focus markers in different languages. To our knowledge, this is the first time focus effects on activation of alternatives has been shown for a language that primarily uses syntactic focus marking and is one of a very small number of psycholinguistic studies involving Austronesian languages.

Table 1: Example story from the experiment

Context	
Sa fa'atau e le mālō ma lona to'alua meaa: 'o le pai, 'o le falaoa ma le keke.	
'The visitor and her husband bought some food: a pie, bread and cake.	
Continuation	
Sa u'u e le tamāloa le falaoa ma le pai.	
'The husband carried the bread and the pie.'	
Critical Sentence	
Focused	'O le keke sa 'ai e le mālō. 'It was the cake that the visitor ate.'
Unfocused	'O le mālō sa 'aia le keke. 'It was the visitor who ate the cake.'
Probe Words	
Object	le keke 'the cake'
Mentioned	le falaoa 'the bread'
Unmentioned	le ēsi 'the papaya'
Unrelated	le kolisi 'the college'

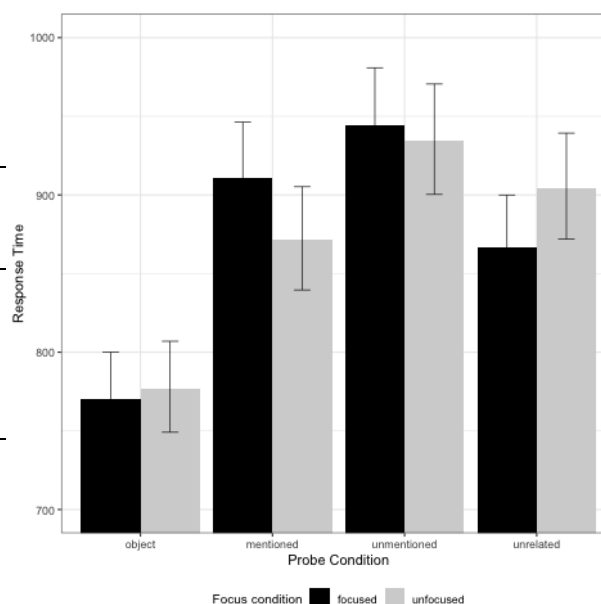


Figure 1: Back-transformed fitted RTs in milliseconds by Focus and Probe condition. Error bars show standard error of the means.

References

- Braun, B., & Tagliapietra, L. (2010). The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes*, 25(7–9), 1024–1043.
- Calhoun, S. (2015). The interaction of prosody and syntax in Samoan focus marking. *Lingua* 165: 205–229.
- Calhoun, S., Wollum, E., & Kruse Va'ai, E. (2019). Prosodic prominence and focus: Expectation affects interpretation in Samoan and English. *Language and Speech*. doi: 10.1177/0023830919890362
- Gotzner, N., Wartenburger, I., & Spalek, K. (2016). The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, 8(1), 59–95.
- Kruse Va'ai, E. (2011). *Producing the text of culture: The appropriation of English in contemporary Samoa*. Apia, Samoa: Govt of Samoa Printing Press.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26
- Lenth, R. (2020). emmeans: Estimated marginal means, aka least-squares means [Computer software manual]. Retrieved from <https://github.com/rvnlenth/emmeans>
- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics* 1(1): 75–116.
- Yan, M. & Calhoun, S. (2019). Priming effects of focus in Chinese. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2019.01985.

Do islands affect only filler-gap dependencies? Evidence from Spanish

Alejandro Rodríguez and Grant Goodall (UC San Diego)

There are many types of long-distance dependencies in natural language (e.g., anaphoric or cataphoric use of pronouns), but filler-gap dependencies have often been thought to be the only type that is sensitive to islands, perhaps because they are the only type that involves a gap (Ross 1967). If islands result from processing limitations, as sometimes thought, this would suggest that it is the processor's difficulty with gaps in particular that induces the island effect. It is thus noteworthy that some dependencies without gaps have sometimes been claimed to be sensitive to islands. Clitic Left-Dislocation (CLLD), as shown in (1), a kind of topicalization structure in Spanish and some other languages in which a clitic pronoun is used instead of the expected gap, is one prominent example of this.

- (1) **A la vecina**, la señora **la** invitó.
the neighbor the lady CL.3sf invited
'The neighbor, the lady invited her.'

CLLD has been claimed to be like *wh*-arguments with respect to island-sensitivity: very sensitive to strong islands (e.g., relative clause islands), but barely sensitive at all to weak islands (e.g., *wh*-islands) (Cinque 1990, López 2009). Formal acceptability experiments have nonetheless been able to detect an effect with *wh*-arguments and weak islands (Sprouse et al. 2016), so if CLLD is similar, we should be able to find an effect there as well, given an appropriately designed experiment. Here we do exactly that, comparing *wh*-dependencies and CLLD in Spanish with regard to two types of weak islands.

Experiment: Experimental items were prepared using a 2 x 3 x 2 design, crossing the factors DEPENDENCY TYPE (*wh*-dependency vs. CLLD), CLAUSE TYPE (non-island vs. *whether* island vs. *wh*-island), and DISTANCE (short vs. long). 48 lexicalized sets were distributed into 12 counterbalanced lists using a Latin square design (4 tokens per condition) plus 12 additional lists in reverse order. 50 native Spanish speakers, all living in their native country at the time of the experiment, rated the acceptability of experimental items plus 54 filler items using a 7-point scale. Sample stimuli are given in (2) and (3).

Results: We constructed a linear mixed-effects model (LME) using the "lmer" function in the "lme4" package in R. The results show a super-additive interaction between CLAUSE TYPE and DISTANCE (Sprouse et al. 2012) for both *whether* and *wh*-islands in *wh*-dependencies ($p < 0.01$ and $p < 0.001$ respectively), as shown in **Fig. 1**, but not in CLLD ($p = 0.449$ and $p = 0.859$ respectively), as in **Fig. 2**.

Discussion: As expected, *wh*-dependencies in Spanish exhibit clear effects with both *wh*- and *whether* islands. CLLD, on the other hand, shows no evidence of such sensitivity to islands. The standard view in the syntax literature has been that island effects arise with filler-gap dependencies and with a handful of similar structures. Here we have taken one of those other structures and shown that despite earlier claims, it does not display the sensitivity to weak islands that we would expect if it behaves similarly to *wh*-arguments. This is important, because the main distinguishing characteristic of CLLD is that there is no gap, so our results suggest that it is something about gaps (e.g. detecting them or integrating the filler into them) that induces island effects. Further work is necessary to know whether CLLD will be similarly insensitive to other types of islands, or whether other structures without gaps will be island-insensitive in the same way, but the results here lend credence to the idea that island effects arise if and only if there is a gap.

REFERENCES: Cinque 1990. Types of \bar{A} -dependencies. López 2009. A Derivational Syntax for Information Structure. Ross 1967. Constraints on variables in syntax. Sprouse et al. 2012. A test of the relation between working memory capacity and syntactic island effects. Sprouse et al. 2016. Experimental syntax and the variation of island effects in English and Italian.

SAMPLE STIMULI (ISLAND = *wh*-island or *whether* island)

(2) WH-DEPENDENCY

- a. ¿**Quién** __ cree [que la señora invitó a la vecina]?
 who __ think.3s [that the lady invited the neighbor]?
 'Who thinks that the lady invited the neighbor?' [NON-ISLAND | SHORT]
- b. ¿**A quién** crees [que la señora invitó __]?
 whom think.2s [that the lady invited __]
 'Who do you think the lady invited?' [NON-ISLAND | LONG]
- c. ¿**Quién** __ se pregunta [si la señora invitó a la vecina]?
 who REFL wonder.3s [whether the lady invited the neighbor]
 'Who wonders whether the lady invited the neighbor?' [ISLAND | SHORT]
- d. ¿**A quién** te preguntas [si la señora invitó __]?
 whom REFL wonder.2s [whether the lady invited __]
 'Who do you wonder whether the lady invited?' [ISLAND | LONG]

(3) CLLD-DEPENDENCY

- a. Creo [que la señora invitó a la vecina]
 think.1s [that the lady invited the neighbor]
 'I think that the lady invited the neighbor' [NON-ISLAND | SHORT]
- b. **A la vecina**, creo [que la señora la invitó]
 the neighbor, think.1s [that the lady CL.3sf invited]
 'The neighbor, I think that the lady invited her.' [NON-ISLAND | LONG]
- c. Me pregunto [si la señora invitó a la vecina]
 REFL wonder.1s [whether the lady invited the neighbor]
 'I wonder when the lady invited the neighbor.' [ISLAND | SHORT]
- d. **A la vecina**, me pregunto [si la señora la invitó]
 the neighbor, REFL wonder.1s [whether the lady CL.3sf invited]
 'The neighbor, I wonder whether the lady invited her.' [ISLAND | LONG]

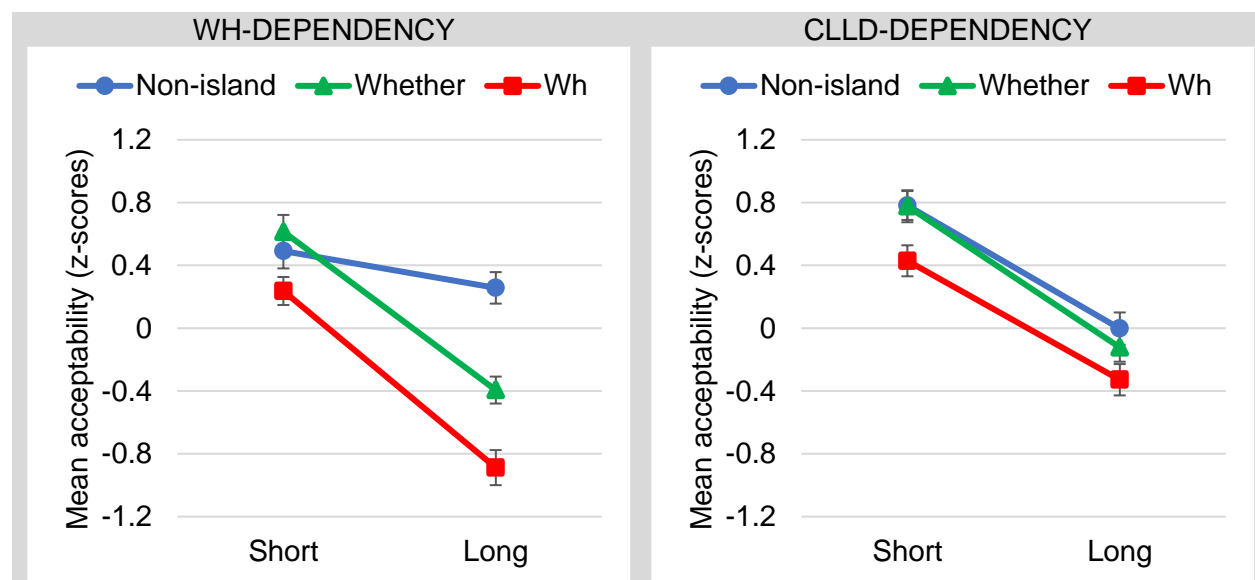


Fig. 1. Mean acceptability of experimental conditions in z-scores (Error bars show SE)

Fig 2. Mean acceptability of experimental conditions in z-scores (Error bars show SE)

Acceptability of extraction out of adjuncts depends on discourse factors

Edward Gibson (MIT), Barbara Hemforth (CNRS, U. Paris), Elodie Winckel (Erlangen U), Anne Abeillé (U. Paris) egibson@mit.edu, bhemforth@gmail.com, abeille@linguist.univ-paris-diderot.fr

Conditions on non-local dependencies usually referred to as “island constraints” have been at the center of much debate on the nature of language since Ross (1967). According to Boeckx (2012), “island effects are perhaps the most important empirical finding in modern theoretical linguistics.” For generative grammarians, these constraints are usually considered syntactic in nature (e.g., Huang, 1982, Chomsky, 1986) and should generalize across constructions (Schütze et al., 2015). Other linguists have argued that semantic and discourse factors play a role and that most examples reported to be ungrammatical are in fact pragmatically infelicitous, because they make salient elements that should belong to the background (Erteschik-Shir, 1973, 1981; Van Valin, 1986; Kuno, 1987; Takami, 1992; Goldberg, 2006, 2013). Abeillé et al. (2020) found a difference between *wh*-questions (1a) and relative clauses (1b) for PP extraction out of subjects in English and French, and propose that the differences in acceptability follow from differences in the discourse statuses of the two constructions: (a) in a *wh*-question, the extracted element is a focus, and interpreting it as the complement of the subject, which is a local topic, yields a clash of discourse status (and hence makes it less acceptable); (b) in a relative clause, the extracted element is not a focus (it corresponds to a local topic inside the relative clause), thus there is no discourse clash with the subject of the relative clause, and so it is more acceptable. They propose the **Focus-background conflict (FBC)** constraint: A focused element should not be part of a backgrounded constituent.

In this project, we seek to evaluate the FBC constraint on adjunct islands (Huang 1982, Stepanov 2007), across constructions. Most syntactic theories assume extraction out of an adjunct is worse than out of a complement, except for non-finite adjuncts denoting the same event as the main clause (2) (Truswell 2007, 2014). Some recent work has found differences across constructions that are as predicted by the FBC constraint, although not discussed in these terms (because the hypothesis did not yet exist): (a) in English, extractions from *if*-adjuncts are not islands in relative clauses, whereas they are in *wh*-questions (Sprouse et al., 2016); (b) in Norwegian, extractions from *if*-adjuncts are better in topicalizations in supportive context than in null-context *wh*-questions (Kush et al., 2017, 2019). These results are as predicted by the FBC constraint, but there were also confounds: e.g., in Sprouse et al., the RC involved an animate (*who*) while the *wh*-questions an inanimate (*what*).

We ran 3 acceptability experiments on English to directly test the FBC constraint, comparing *wh*-questions and relative clauses, for extraction out of a *that* complement clause and out of an adjunct *if* clause, with the same matrix predicates. According to syntactic theories of island effects, extraction out of *if*-clauses should be rated lower, and the same holds for frequency-based approaches since the verb+*that* frame was always more frequent than the verb+*if* frame (frequencies extracted from the COCA). According to the FBC constraint, on the other hand, a penalty is only expected with *wh*-questions, assuming that *if*-clauses are more backgrounded than *that*-clauses.

Experiment 1 was run on *wh*-questions, with \pm extraction and *that/if* clause, with the same predicates (5). We replicated the “island” effect from literature: extraction out of *if*-clause was rated lower than out of *that*-clause (Fig.2), and there was an interaction between extraction and clause type (lmer model on z-scores; $\beta = -.52$; $t = -4.12$; $p < .001$). (There was also a main effect of “*if*” clauses, which were rated better than “*that*” clauses, but this probably relates to the plausibility of the events described by the two kinds of clauses, which is orthogonal to our research question.)

Experiment 2 was run on relative clauses with the same design (6). Unlike E1, there was no island effect (interaction: $t = 0.65$; $p = .52$). Across experiments, there was a 3-way interaction, showing that the interaction in E1 was not present in E2 ($\beta = -0.38$; $t = -2.49$; $p = .013$). (Note that there was also a main effect of experiment, such that RC materials were rated lower, probably because of the extra clause in the RCs compared to the WHQs. This is again orthogonal to the effects of interest.) Following Abeillé et al. 2020, we propose that the differences in acceptability come from differences in the discourse status of the two constructions: *wh*-questions put the extracted element in focus position, which is incompatible with the FBC, but the relative clause does not change its discourse status, hence there is no adjunct penalty. This account predicts that an appropriate discourse context may ameliorate *wh*-questions. We ran E3 with the same *wh*-questions preceded by a supportive context, which made the questioned element less focal (7). Here, we did not find any adjunct penalty, and extraction out of the *if*-clause was rated as high as extraction out of the *that*-clause (Fig.3), resulting in no interaction between extraction and *if/that* in supportive contexts (interaction: $t = -1.54$; $p = .14$), and a 3-way interaction when compared with the null contexts ($\beta = -0.29$; $t = -1.92$; $p = .056$).

We conclude that extraction constraints are limited to focalizing constructions (wh-questions, topicalizations) and are due to the lack of an appropriate discourse context. Hence they pose no learning conundrum, contrary to the syntax-only hypothesis.

Abeillé, A., Winckel, E., Hemforth, B., Gibson, E. 2020. Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition* 204. Kush D.; Lohndal T.; Sprouse J., 2017. Investigating variation in island effects. *NLLT*, 1-37. Kush D.; Lohndal T.; Sprouse J., 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language* 95(3). Sprouse, J., Caponigro I., Greco C., Cecchetto C., 2016. Experimental syntax and the variation of island effects in English and Italian, *NLLT*, 34(1), 307-344 Truswell, R., 2007. Extraction from adjuncts and the structure of events. *Lingua* 117, 1355–1377.

- (1)a. Of which sportscar did [the color __] delight the baseball player? (Abeillé et al. 2020)
- b. The dealer sold a sportscar, of which [the color __] delighted the baseball player.
- (2) What did John come home [trying to understand __]? (Truswell 2007)
- (3)a. Wh-no-island: *What* do you think [that the lawyer forgot __ at the office] ? (Sprouse et al. 2016)
- b. Wh-island: *What* do you worry [if the lawyer forgets __ at the office]?
- (4)a. RC-no-island: I called the client *who* the secretary thought [that the lawyer insulted __].
- b. RC-island: I called the client *who* the secretary worries [if the lawyer insults __]. (Sprouse et al 2016)
- (5) Experiment 1. Wh-questions; N = 60; 16 items
 - a. +extract-that: Which concert would Paul worry [that I miss __]?
 - b. +extract-if: Which concert would Paul worry [if I miss __]?
 - c. -extract-that: Would Paul worry that I miss this concert?
 - d. -extract-if: Would he worry if I miss this concert?

- (6) Experiment 2. Relative clauses; N = 60; 16 items
 - a. +extract-that: Paul told me about a concert which he would worry [that I miss __].
 - b. +extract-if: Paul told me about a concert which he would worry [if I miss __].
 - c. -extract-that: Paul cares about my music training, and he would worry that I miss this concert.
 - d. -extract-if: Paul cares about my music training, and he would worry if I miss this concert.

- (7) Experiment 3. Wh-questions with a supportive context; N = 60; 16 items
 - a. +extract-that: Paul cares about my music training. Which concert would he worry [that I miss __]?
 - b. +extract-if: Paul cares about my music training. Which concert would he worry [if I miss __]?
 - c. -extract-that: Paul cares about my music training. Would he worry that I miss this concert?
 - d. -extract-if: Paul cares about my music training. Would he worry if I miss this concert?

Figure 1: E1 WHQs, null context; z-scores for (IF / That) x (WHQ, no-extract).
Reliable island effect (interaction t-value = -4.12), replicating the literature

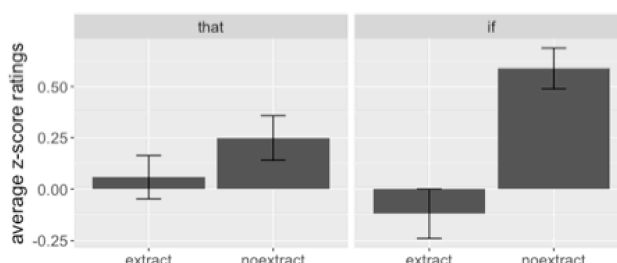


Figure 2: E2 RCs; z-scores for (IF / That) x (RC, no-extract).
No island effect (interaction t-value = -0.65)
3-way interaction with E1/E2: t = -2.49; p = .013

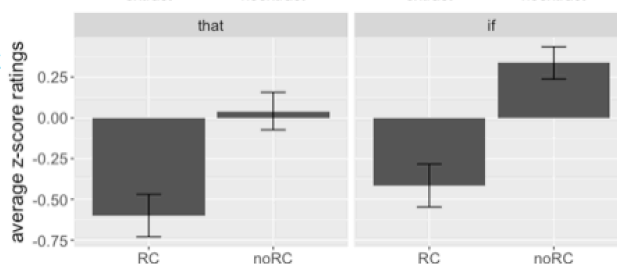
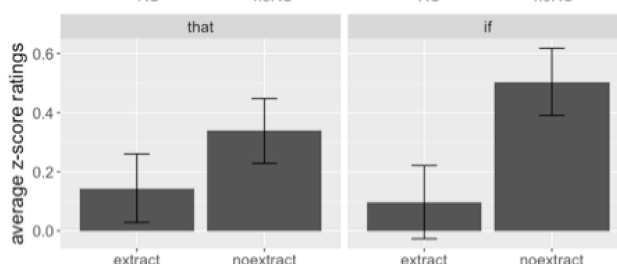


Figure 3: E3 WHQs, supportive context; z-scores for (IF / That) x (WHQ, no-extract).
No island effect (interaction t-value = -1.54)
3-way interaction with E1: t = -1.92; p = .056



The structural source of English Subject Islands

David Potter & Katy Carlson (Morehead State University; davidkpotter@gmail.com)

We argue that the islandhood of complex subjects, (1), arises from their syntactic properties^{1,2}, rather than their information structure or processing constraints^{3,4,5}. Key to this claim is the comparison between the behaviour of complex subjects in the context of it-clefts (1a) and the type of ellipsis known as *stripping* (1b). Previous research has argued that ellipsis is insensitive to island constraints that derive from structural sources while remaining sensitive to islands deriving from non-structural sources^{6,7}. The meaning and information structure of stripping fragments and it-clefts are very similar, and consequently if subject islands were ultimately the result of the backgrounded status of complex subjects, we would expect no difference in island sensitivity between these two constructions. On the other hand, if the islandhood of complex subjects were the result of their structural properties, we would expect stripping to be insensitive to island effects. We show, in two acceptability experiments, that stripping is indeed insensitive to subject islands, while it-clefts exhibit these island effects. We conclude that complex subjects are islands as a result of their syntactic properties.

- (1) [_{DP} Two paintings of Iggy Pop] were in the musician's office.
a. No, it was Sting who two paintings of _{DP} were in the musician's office.
b. No, Sting [_{DP} two paintings of _{DP}] were in the musician's office.

In experiment 1, participants (N=28) listened to short dialogues, in which an antecedent, containing a complex subject, e.g. (1), or the corresponding existential expletive, e.g. (2), was followed by a corrective it-cleft or a stripping fragment, e.g. (1a, 1b). Participants rated the acceptability of the continuation. Experimental paradigm and average ratings by condition are found in table 1. We found that stripping continuations were rated more acceptable than it-cleft continuations (β :-3.56+/-0.62; p < 0.001) and that complex subject conditions were rated worse than existential expletive conditions (β :1.91+/-0.28; p < 0.001). The interaction between these factors was also significant (β :4.32+/-0.73; p < 0.001): the complex subject it-clefts were rated worse than the existential expletive it-clefts (β :3.73+/-0.54; p < 0.001), while the ratings given to the complex subject stripping continuations were no different from those given to the existential expletive stripping continuations (p > 0.47).

- (2) There were [_{DP} two paintings of Iggy Pop] in the musician's office.
a. No, it was Sting who there were [_{DP} two paintings of _{DP}] in the musician's office.
b. No, Sting there were [_{DP} two paintings of _{DP}] in the musician's office.

Experiment 2 again had participants (N=62) listen to dialogues and to rate the naturalness of the continuation. See table 2 for full paradigm and average ratings by condition. In contrast to experiment 1, these are *sprout*-type continuations, in which the correlate to the cleft pivot and the stripping fragment, e.g. *Iggy Pop* above, is implicit rather than explicit⁸. Additionally, we manipulated whether the preposition was stranded or pied-piped. We found all main effects of continuation type, islandhood, and preposition position to be significant, as was the three way interaction (β :1.02+/-0.4; p = 0.01). In the stripping continuations, the p-stranding conditions were rated worse than the pied-piping conditions providing experimental support for the “no new words” constraint on sprout-type ellipsis⁹. The stripping conditions showed no island effect, however, with island conditions rated no worse than non-island conditions, corroborating non-experimental claims that sprout-type ellipsis is insensitive to subject islands¹⁰. The cleft conditions showed an interaction between islandhood and p-stranding; both p-stranding and pied-piping conditions showed an island effect, corroborating previous results¹¹, with the complex subject conditions rated worse than the existential expletive conditions, but with a larger island effect in the p-stranding conditions (β :1.89+/-0.22; p < 0.001) than in the pied-piping conditions (β :0.42+/-0.20; p = 0.038).

In summary: long distance dependencies, in the form of it-clefts, consistently show subject island effects, though the magnitude of this effect varies with the type of extracted element (PP vs. DP). Stripping, on the other hand, does not exhibit any subject island effects. This contrast is unexpected under processing and information structural accounts yet predicted by structural accounts of subject islands.

Table 1: Average Acceptability Ratings, Experiment 1

Experiment 1			Average Acceptability Rating
Complex Subject	Antecedent	Two paintings of Iggy Pop were in the musician's office.	
	It-Cleft	No, it was Sting who two paintings of were in the musician's office.	2.90
	Stripping	No, of Sting.	5.64
Existential Expletive	Antecedent	There were two paintings of Iggy Pop in the musician's office.	
	It-Cleft	No, it was Sting who there were two paintings of in the musician's office.	4.56
	Stripping	No, of Sting.	5.61

Table 2: Average Acceptability Ratings, Experiment 2

Experiment 2				Average Acceptability Rating
Complex Subject	Antecedent		Two paintings were in the musician's office.	
	It-Cleft	PS	Yeah, it was Sting who two paintings of were in the musician's office.	3.06
		PP	Yeah, it was Sting of whom two paintings were in the musician's office.	3.44
	Stripping	PS	Yeah, Sting.	3.91
		PP	Yeah, of Sting.	6.15
	Antecedent		There were two paintings in the musician's office.	
Existential Expletive	It-Cleft	PS	Yeah, it was Sting who there were two paintings of in the musician's office.	4.13
		PP	Yeah, it was Sting of whom there were two paintings of in the musician's office.	3.68
	Stripping	PS	Yeah, Sting.	3.68
		PP	Yeah, of Sting.	6.10
	Antecedent		There were two paintings in the musician's office.	

REFERENCES: [1] Chomsky, N. 1977, in *Formal syntax* (New York: Academic Press), 71–132. [2] Nunes et al. 2000, *Syntax*, 3, 20. [3] Goldberg, A. E. 2013, in *Experimental syntax and island effects* (Cambridge University Press), 221. [4] Chaves et al. 2019, *Journal of linguistics*, 55, 475. [5] Abeillé et al. 2020, *Cognition*, 204. [6] Merchant, J. 2001, *The syntax of silence: Sluicing, islands, and the theory of ellipsis* (Oxford University Press, USA). [7] Ross, J. R. 1969, in *Fifth regional meeting of the Chicago Linguistic Society*, Vol. 252286. [8] Chung et al. 1995, *Natural language semantics*, 3, 239. [9] Chung, S. 2006, in *Proceedings of the annual meeting of the Berkeley Linguistics Society*, Vol. 31, 73–91. [10] Lasnik et al. 2003, *Linguistic Inquiry*, 34, 649. [11] Abeillé et al. 2020, in *Poster presented at the 33rd Annual CUNY Conference on Human Sentence Processing*. UMass Amherst.

Semantic interference in dependency formation: NP types in cleft sentences

Myung Hye Yoo & Rebecca Tollan (University of Delaware)

[INTRODUCTION] We used similarity-based interference effects to test how NP types of an intervenor modulates the processing of filler-gap dependencies [1], under the cue-based retrieval mechanism [2]. This interference effect arises when a distractor that has partially or wholly matching features with a target noun phrase (NP) is retrieved in parallel to the filler, leading to processing overload. Warren & Gibson (2002, 2005)'s complexity rating study ([3,4]), meanwhile, observed that parsers were sensitive to the gradient status of a distractor in discourse, following the Givenness Hierarchy ([5,6]). For example, a distractor that is most central in the discourse (e.g. *pronouns*) caused the least processing cost, followed by less central NPs on the hierarchy (e.g. *definites*) [pronouns > first names > full names > definites > indefinites]. On this view, this paper explores whether the interference effect of a distractor is truly a similarity effect or is in fact a more fine-grained discourse-level of the semantic hierarchy, or both.

[EXPERIMENT] A self-paced moving window experiment had a 2 x 3 design (n=36), crossing two types of the filler in the clefted position (NP1) and three types of a distractor in the embedded NP position (NP2) as shown in (1): [definite descriptions, indefinite descriptions] x [pronouns, definite descriptions, indefinite descriptions]. Experimental materials consisted of 24 sets of 4 items in each 6 conditions, and each item was followed by a comprehension question.

(1) *It was {the actor/an actor} who {we/the director/a director} graciously thanked before the show.* The reading time on the critical verb (e.g. *thanked*) did not reveal a main effect of NP1 ($t=-0.62$, $p=.54$) but showed a reliable effect of NP2 type. The pronoun NP2 condition was read significantly faster than definite and indefinite conditions ($t=-3.60$, $p < .001$). Surprisingly, the reading time of the definite NP2 was slower than the indefinite NP2, which conflicts with the prediction of the givenness hierarchy. The statistical analysis showed a marginal effect of definiteness between the definite and indefinite conditions ($t=1.78$, $p=.07$). In addition, the definite-definite description took the slowest reading time ([Fig 1]). The response times to comprehension questions showed a similar pattern with the reading time on the verb in that (i) there was no main effect of NP1 ($t=0.34$, $p > .05$) and (ii) sentences involving a pronoun in NP2 were responded significantly faster than those with a definite or indefinite NP2 ($t=-5.34$, $p < .001$). The comparison between definite and indefinite conditions, however, revealed no significant difference ($t=-0.15$, $p > .05$) [Fig 2].

[DISCUSSION] The result showed that NP types of the filler (definite vs. indefinite) did not modulate the processing of the filler-gap dependencies in clefts sentences, unlike previous findings of the effect of semantic and syntactic status of the fillers in the processing of other filler-gap dependencies such as islands and *wh*-questions [7,8]. In terms of the NP types of the distractor, the givenness hierarchy predicted a faster reading time of the pronoun than definite and indefinite conditions. However, the slower reading time of a definite than an indefinite was not predicted by the givenness hierarchy. The similarity-based interference effect was also observed only in the definite-definite condition, but not in the indefinite-indefinite condition.

These overall patterns of reading times suggest that the definite NP type of distractors appears to be sensitive to the similarity-based interference effect, in addition to a definiteness effect. This observation could be attributed to the absence of contexts. A sentence without contexts may give rise to the processing load of definites, but not indefinites: definites tend to refer to old or established referents in the discourse [9]. Thus, parsers are likely to automatically look for a referent when they encounter a definite. Since no contexts were given, they would fail to find the referent, and this can be the source of increased processing difficulty. Indefinites, on the other hand, introduce a new referent and thus do not trigger a search for the referent. Parsers do not have to trace back and no additional processing load is required for indefinite. In terms of the response times to the comprehension questions, the similarity-based interference and definiteness effects of definite distractors disappeared. It suggests that working memory load due to these effects no longer affect post-sentence level processing. The extra memory load of

definiteness implies that the process of the accommodation of the definite attractors seem to arise during on-line building of sentence representations, but not post-sentence level processing.

Figure 1. Mean reading times in the critical verb (ms)

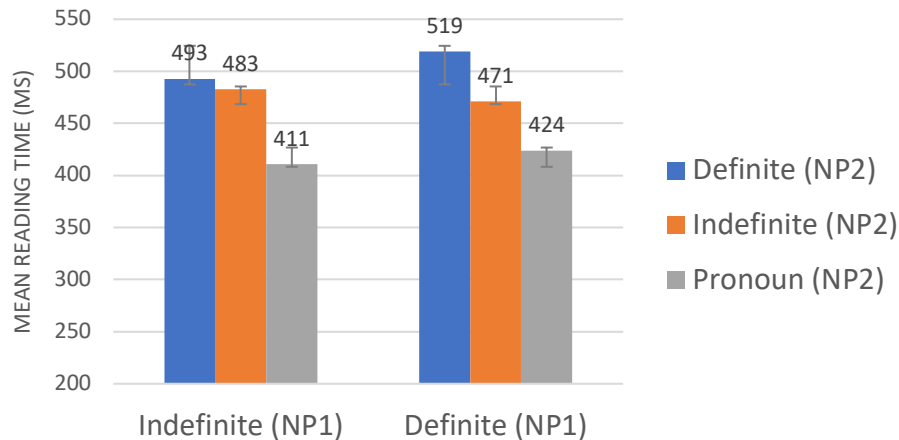
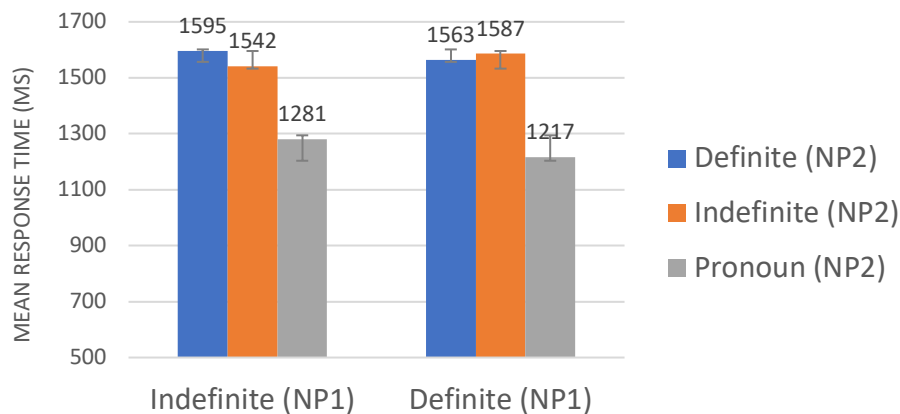


Figure 2. Mean response times to comprehension questions (ms)



References

- [1] Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory Interference During Language Processing Memory Interference During Language Processing. *Journal of Experimental Psychology*, 27(6), 1411–1423.
- [2] Lewis, R. L., & Vasissth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 375–419.
- [3] Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85, 79–112.
- [4] Warren, T., & Gibson, E. (2005). Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes*, 20(6), 751–767.
- [5] Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.
- [6] Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2), 274–307.
- [7] Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*, 26(3), 109–123.
- [8] Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, 86, 366–415.
- [9] Heim, I. R. (1982). *The semantics of definite and indefinite NPs*. University of Massachusetts.

Backgroundedness measures predict island status of non-finite adjuncts in English

Savithry Namboodiripad¹, Felicia Bisnath¹, Alex Kramer¹, Noah Luntzlara¹ and Adele Goldberg²

¹University of Michigan, Ann Arbor, ²Princeton University

Previous work has argued that the extent to which a construction is “backgrounded” in discourse predicts the extent to which it is an island for long-distance extraction (Erteschik-Shir 1979; Goldberg 2006). While the claim was supported by a study of verb complement clauses (Ambridge & Goldberg 2008), the interpretation has been challenged due to a lack of super-additive effects, indicating that verb complement clauses may not be islands after all (Liu et al. 2019). The current study investigates the case of non-finite *adjunct islands* and asks whether the degree to which they are backgrounded predicts their status as islands to wh-questions. **Backgroundedness measures:** We operationalized backgroundedness in two ways. (1) *Negation test*: the extent to which an adjunct interpretation is unaffected by main clause negation predicts the adjuncts degree of backgroundedness (Erteschik-Shir 1979; Goldberg 2013; negated adjunct interpretation = less backgrounded/more acceptable). In a preregistered norming study, 96 participants rated the extent to which main clause negation implied that the adjunct clause was negated. (2) *Temporal overlap*: 80 participants rated how likely the events in the main and adjunct clauses were to occur at the same time (one event = less backgrounded; cf. Truswell 2007). We used two types of non-finite adjunct clauses (Michel & Goodall 2013), *to* clauses and *ing* clauses; we expected the differences in event structure across items would ensure variation in both measures.

Acceptability study: Our preregistered experiment employed a 2x1 design, crossing SENTENCE TYPE (declarative vs. adjunct-extracted) with DEGREE OF BACKGROUNDEDNESS as described above. 32 declarative and 32 adjunct-extracted sentences were recorded and distributed across 4 lists pseudorandomly using a Latin Square design. 128 English-speaking participants were recruited via Prolific.co and asked to rate acceptability on a 1-7 Likert scale. Participants heard 16 items from each sentence type (no more than one type for any item), and 48 fillers which varied in acceptability. **Results and discussion:** As predicted, both backgroundedness measures predicted the acceptability of adjunct-extracted sentences more than they did declarative sentences. Specifically, linear mixed effects models were fit for each backgroundedness measure (fixed effects = z-scored rating, SENTENCE TYPE, & BACKGROUNDEDNESS MEASURE; random effects = PARTICIPANT, ITEM). Model comparison via ANOVA confirmed a significant interaction between judgments on the negation task and SENTENCE TYPE as compared to an additive model ($\chi^2 = 20.5$ $df = 1$ $p < 0.001$; Fig 1). Similarly, model comparison via ANOVA confirmed a significant interaction between temporal overlap ratings and SENTENCE TYPE compared to an additive model ($\chi^2 = 6.4848$ $df = 1$ $p < 0.011$; Fig 2). That is, the extent to which an adjunct was presupposed (not negated) was inversely correlated with independent judgments on the corresponding wh-question (adjunct extraction); the extent to which an adjunct was interpreted as a distinct event also inversely correlated with judgments on extractions. Since adjunct type varied categorically (Table 1), we tested whether the continuous backgroundedness measures predicted ratings above and beyond adjunct type, by including adjunct type as well as backgroundedness and sentence type as fixed effects; results showed the negation test predicted acceptability above and beyond adjunct type, but the temporal overlap measure did not. Variation across *to* adjuncts is driving this effect (Fig 3). This work supports the claim that non-finite adjunct clauses are islands for wh-questions to the extent they are backgrounded in discourse, and we show the first experimental evidence for systematic differences between *to*-infinitival and gerundive adjunct clauses. Additionally, the construction- and measure-specific variation seen here opens the door to ask how processing-relevant factors such as frequency (e.g. Chavez & Dery 2018; Liu et al. 2019; Dąbrowska 2013), type of extraction (Abeillé et al. 2019; Sag 2010), or working memory (Deane 1991; Hofmeister & Sag 2012) might contribute to the within- and across-language variation.

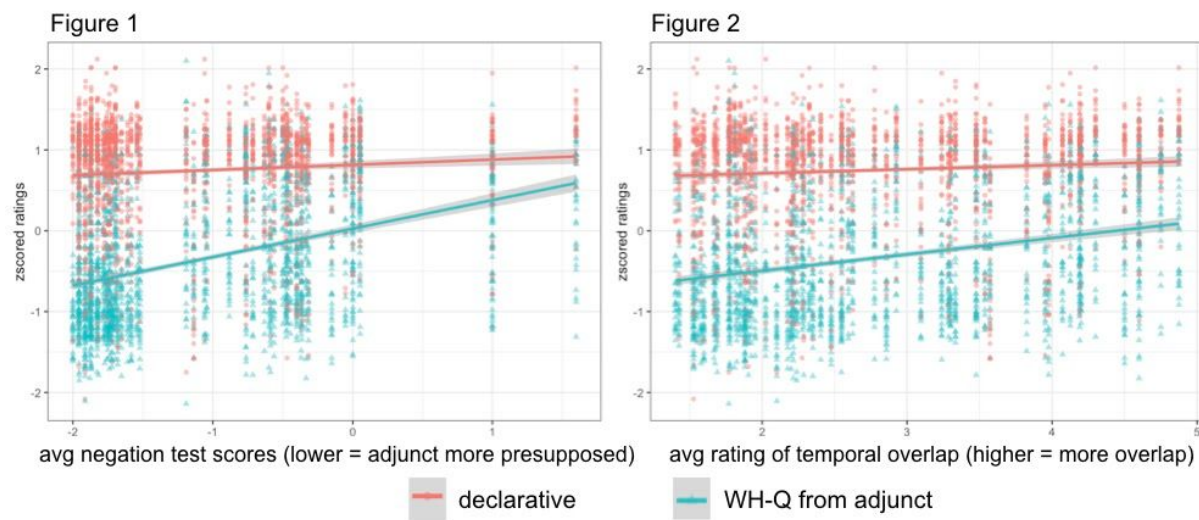


Fig 1: x-axis: the degree to which main clause negation was interpreted as negating the adjunct clause (higher = more negated/less backgrounded); **y-axis:** z-scores of acceptability ratings. **Fig 2: x-axis:** the degree to which the main clause and adjunct clause were interpreted as occurring at the same time (higher = more overlap/less backgrounded); **y-axis:** z-scores of acceptability ratings. **Green:** WH-Q extractions from adjuncts; **Red:** Declarative sentences. Lines represent smoothed linear model fits.

Table 1. Sample items; Sentence type (declarative vs. wh-question) and adjunct clause type. Backgroundedness measures were based on declarative sentences.

SENTENCE TYPE	Adjunct (<i>to</i>)	Adjunct (<i>ing</i>)
Declarative	The mechanic changed classes to meet the engineer.	The mechanic changed classes after the engineer.
Wh-Q from adjunct	Who did the mechanic change classes to meet?	Who did the mechanic change classes after meeting?

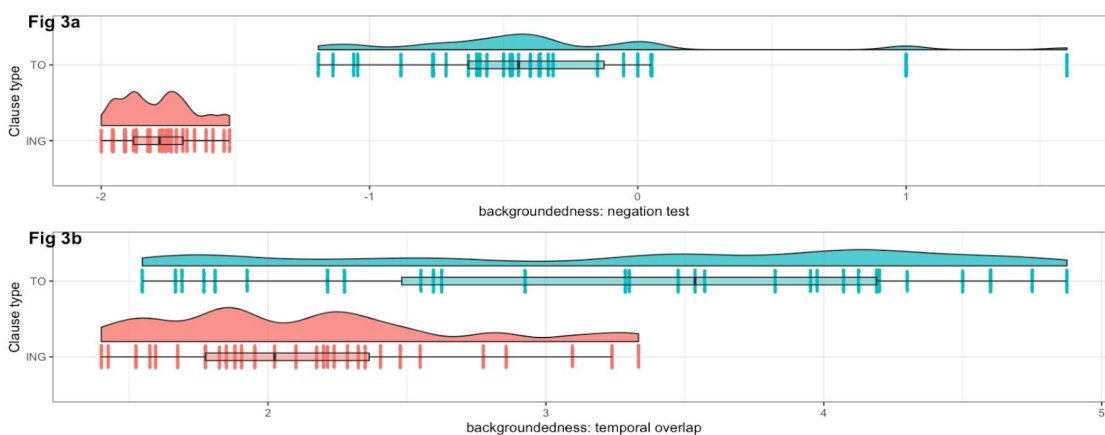


Fig 3: backgroundedness measures by clause type (blue/top = *to*, red/bottom = *ing*); *to* shows more by-item variation, and the negation test (3a) and temporal overlap (3b) differentially correspond to clause type

Oscillatory dynamics of complex dependency processing reveal unique roles for attention and working memory mechanisms

Shannon McKnight, Don Bell-Souder, Phillip Gilley, Akira Miyake, Albert Kim (CU Boulder)

Domain general cognitive processes, such as working memory (WM), play critical roles in sentence comprehension, but much uncertainty remains over the precise nature of these roles. Here we examined the neural oscillatory dynamics of a well-known processing asymmetry between object-extracted (ORC) and subject-extracted (SRC) relative clauses to provide new insight into the engagement of working memory and attention during complex dependency processing. Previous research evaluating these sentence types have mainly focused on minimally dimensional measures of processing difficulty, such as reading time and ERP. Neural oscillatory dynamics, on the other hand, provide a higher dimensional space to evaluate processing differences between these conditions which might be lost when averaging over multiple frequencies, locations, and time points such as in ERP analyses. Furthermore, neural oscillations have previously been uniquely linked to specific domain general cognitive processes, such as working memory recall and attention focus. Thus, we attempt to use these measures to evaluate two major theorized sources of relative clause processing discrepancies, working memory and frequency/expectation based accounts.

1. The lawyer that the_C judge_A disliked was_B fired for corruption. **ORC, more demanding**
2. The lawyer that disliked_C the judge_A was_B fired for corruption. **SRC, less demanding**

Sentence processing theories that ascribe WM demands to the processing discrepancy between relative clause types have hypothesized that: **A**) increased interference demands from unresolved dependencies^[1] due to the shared thematic role of agent between the main noun phrase and the embedded noun phrase, and **B**) increased integration costs at the main verb^[2]. An increase in WM interference would be most apparent when comparing activity at the embedded noun-phrases_A across sentences, the point at which the maintenance of two separate agents begins, and which would only occur for ORC sentences. Integration cost demands, on the other hand, would be most evident at the onset of the main verb phrase, as at this point agent/patient nouns need to be recalled from WM. Theories that propose frequency-based expectation violations underlie the processing asymmetry hypothesize that, since SRCs are more common than ORCs in English, ORCs will be more surprising, and differences in processing difficulties should be apparent at the word immediately following the complementizer *that* (**C**). It is unclear what underlying cognitive cost expectation violations would incur – so we evaluated an index of attention as well as WM.

We evaluated mid-frontal theta (4-7Hz) power and occipital alpha (8-12Hz) power in scalp recorded EEG collected while 205 participants read relative clause containing sentences within the context of a large-scale study designed to measure multiple facets of the neural dynamics of sentence processing. **We assumed increases in mid-frontal theta power would reflect increases in working memory recall^[3,4] and event-related de-synchronization of occipital alpha power would reflect increases in sustained attention^[4,5].**

We used the Morelet wavelet transform^[6] to create time-frequency representations of and determined individual theta and alpha power by identifying peak power spectra from a set of non-complex sentences presented during the experiment (see figure captions for additional description). Mid-frontal theta power increased for ORC sentences, mainly at the onset of the main verb phrase, and to a lesser extent within the embedded noun phrase (Fig1A). Additionally, alpha power decreased throughout the relative clause, resolving at the main verb phrase (Fig1B). Notably, this decrease reflects a lack of alpha synchrony during ORCs. **Overall, the lack of alpha synchronization we find in ORCs is consistent with ORCs requiring greater attentional demand. Furthermore, the results in the theta band provide support for a theory of increased integration (working memory recall demands) on the verb (B).**

REF: [1] Van Dyke & Lewis (2003) *JML*. [2] Gibson (2000) *Image, Language Brain* [3] Hsieh & Ragnanath (2015) *Neuroimage* [4] Klimesch (1999) *Brain Research Reviews* [5] Clayton, Yeung, Kadosh (2015) *TICS* [6] Torrence & Compo (1998) *Bulletin of the American Meteorological Society*

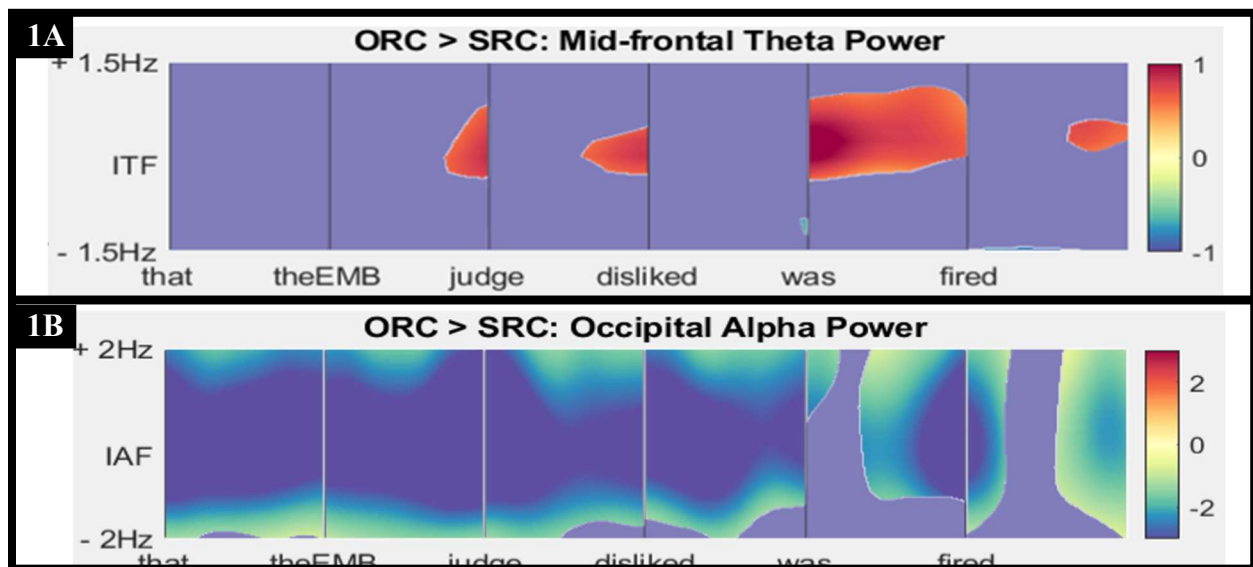
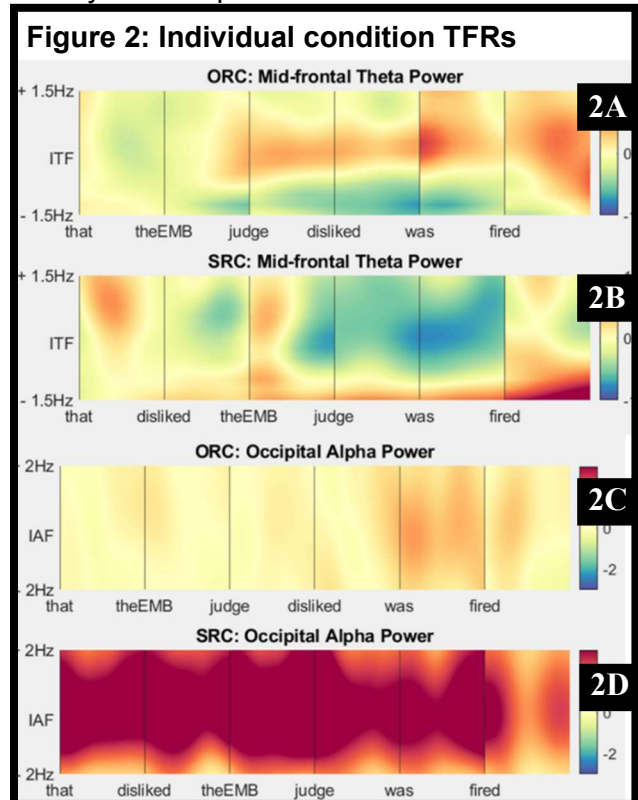


Figure 1: Power Contrast maps (ORC > SRC) Mid-frontal Theta (A) and Occipital Alpha (B)
 Both figures above represent power contrasts (more red, higher synchronization; more blue, higher desynchronization) between object-relative containing (ORC) and subject-relative containing (SRC) sentences, word by word. Word category was controlled for with the resulting contrast maps presented in ORC word ordering. Purple coloring represents time x frequency combinations did not have a significantly non-zero ORC > SRC contrast (FDR corrected). Individual theta frequency (ITF, 1A) was determined by a local spectra maximum between 4-7Hz with a ± 1.5 Hz bandwidth. Individual alpha frequency (IAF, 1B) determined by a local spectra maximum between 8-12Hz with a ± 2 Hz bandwidth. Mid-frontal theta synchronization was greater for ORC containing sentences (A), largest at the main verb and present to a lesser extent on the offset of the embedded noun phrase. Occipital alpha desynchronization was greater for ORC containing sentences (B), throughout the course of a relative clause. **Individual condition plots (Figure 2)** show that the mid-frontal theta synchronization effect (1A) was due in part to an increase in mid-frontal theta desynchronization for the main-verb during SRC containing sentences. Furthermore, the occipital alpha desynchronization displayed in the contrast TFR (1B) is due to a lack of alpha synchrony for ORC containing sentences. Taken together, these results suggest that working memory recall is taxed during the integration of information at the main verb of a sentence (theta synchronization), while attention is more focused during ORC processing overall.



It takes two **the** tango: Predictability and detectability affect processing of phrase structure errors
Anthony Yacovone, Paulina Piwowarczyk, & Jesse Snedeker (Harvard University)
anthony_yacovone@g.harvard.edu

Introduction. If you hear a sentence like “It takes two **the** tango,” how do you recognize that the speaker misspoke? Prior work with ERPs has found greater neural responses to phrase structure (PS) errors when they occur in sentences that create strong expectations about upcoming words (or syntactic categories).¹ PS errors elicit an early left anterior negativity (ELAN) and a P600—ERP components related to error detection.² The ELAN-P600 response, however, is only reported in studies where participants monitor sentences for errors and/or judge their correctness.^{2,3} In contrast, studies that de-emphasize errors (e.g. with a visual distractor) often find a sustained negativity.³ Crucially, many of these studies use stimuli that do not resemble natural speech; thus, their real-world implications are unclear. The present study explores how PS error detection occurs in naturalistic contexts when people listen to a larger discourse without error monitoring. Is predictability still a factor? Which ERP is elicited and when? We answer these questions below.

Method. We used a novel EEG paradigm called the *Storytime task*. In this task, we recorded 30 participants’ EEG responses while they listened to a 30-minute story with PS errors. Errors were created by swapping determiners for prepositions (and vice versa) in both predictable and unpredictable contexts (see **Figure 1**). This procedure ensured that each lexical item appeared in both grammatical and ungrammatical conditions. Both errors and controls were spliced into the story. We had 120 target sentences (60 predictable, 60 unpredictable) and participants heard 60 errors in total. Predictability was determined with a cloze probability task (predictable: 81-100%; unpredictable: 0-6.25%). We pre-registered a set of linear mixed models designed to detect an ELAN-P600 response: two models for mean amplitudes from 0-200 and 200-400ms at left anterior electrodes (for ELANs) and another from 500-800ms at Cz and Pz (for P600s).

Results and Conclusions. There were two notable findings: First, ERP waveforms revealed a weak sustained negativity (not an ELAN-P600) for the PS errors overall. However, this effect was driven by the predictable contexts (see **Figure 2**). These findings remain tentative, as our pre-registered models did not find any significant effects. Second, during the debriefing, participants reported hearing only a handful of the 60 errors in our story. Prior work has found that the salience or detectability of the errors influences comprehenders’ sensitivity to them.^{2,4} Thus, we quantified how detectable each error was by asking a new set of participants on MTurk (N=40) to listen to the story and push a button upon hearing an error. Results confirmed that some errors were very detectable, while others were not (range: 0-86%, median: 33%). The detectability of errors was not strongly confounded with the predictability of the context ($r_s(118) = -.17, p = .06$). Given these findings, we returned to our original analyses to see if detectability moderated the effects of predictability. We performed another set of linear mixed models—but this time, we looked at all electrode sites and tested for a 3-way interaction between Error, Predictability, and Detectability (using a median split categorization). This interaction was significant between 200-400ms and 500-800ms. Pairwise comparisons revealed main effects of PS errors only when the context was predictable, *and* the error was detectable (see **Figure 3** for waveforms and model results). These data are consistent with prior work showing that predictability influences comprehenders’ sensitivity to PS errors.¹ We also report a novel finding: PS errors in rich discourse contexts do not elicit ELAN-P600 responses but rather sustained negativities akin to the pattern in studies that de-emphasize syntactic errors.³ The implication of this finding is that listeners may be adopting different processing strategies depending on the task: When simply listening to a story, participants may prioritize understanding the discourse and down-weight speech errors (or at least prolong their resolution). Whereas, in artificial tasks, participants may prioritize resolving errors at the cost of their understanding. These strategies may be able to explain the different ERP effects across studies. However, future work will require replicating and then extending these findings. In particular, we will investigate whether detectability is a bottom-up feature of the stimulus (e.g. the salience of the acoustic difference) or the result of more shallow syntactic processing (e.g. emphasizing content words over function words) in discourse contexts.

	Original sentences	Violations created by swapping words
Predictable	I walked all along THE row of cages.	I walked all along OF row of cages.
	We've been missing a critical piece OF evidence.	We've been missing a critical piece THE evidence.
Unpredictable	She snapped THE book shut and held it behind her back.	She snapped FOR book shut and held it behind her back.
	I am leaving after lunch FOR a meeting.	I am leaving after lunch THE a meeting.

Figure 1: Examples of stimuli conditions. Items were paired based on predictability and then the target determiner and the target preposition were swapped to create the PS error conditions.

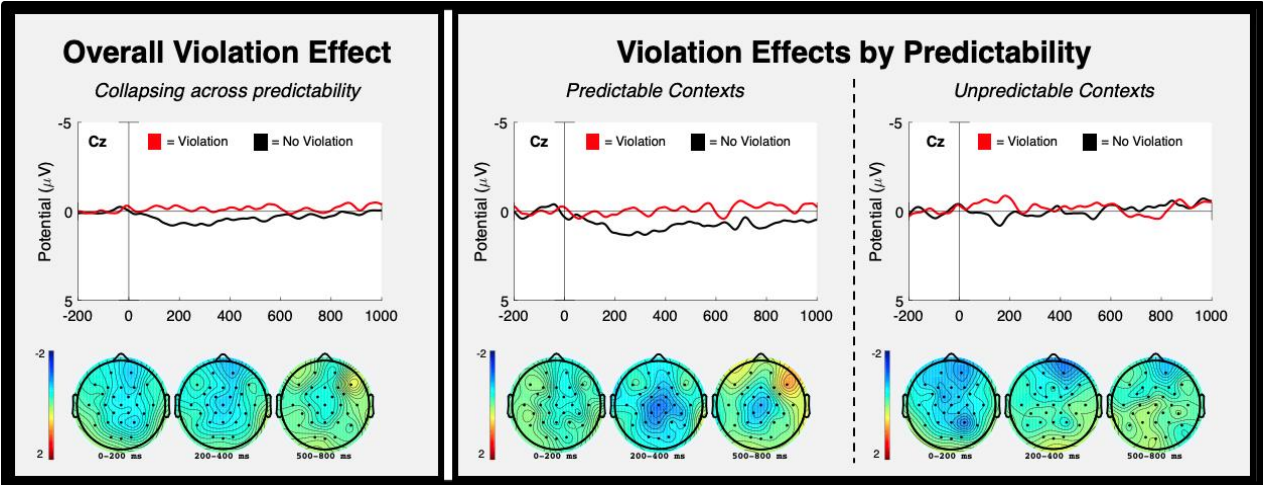


Figure 2: Results from pre-registered analyses. ERP waveforms revealed a weak sustained negativity for PS errors. The effect is primarily found in the predictable contexts. Scalp maps reveal that the effect is centrally located on the scalp in predictable contexts, and critically the negativity is long-lasting, and it is not left-lateralized (no ELAN).

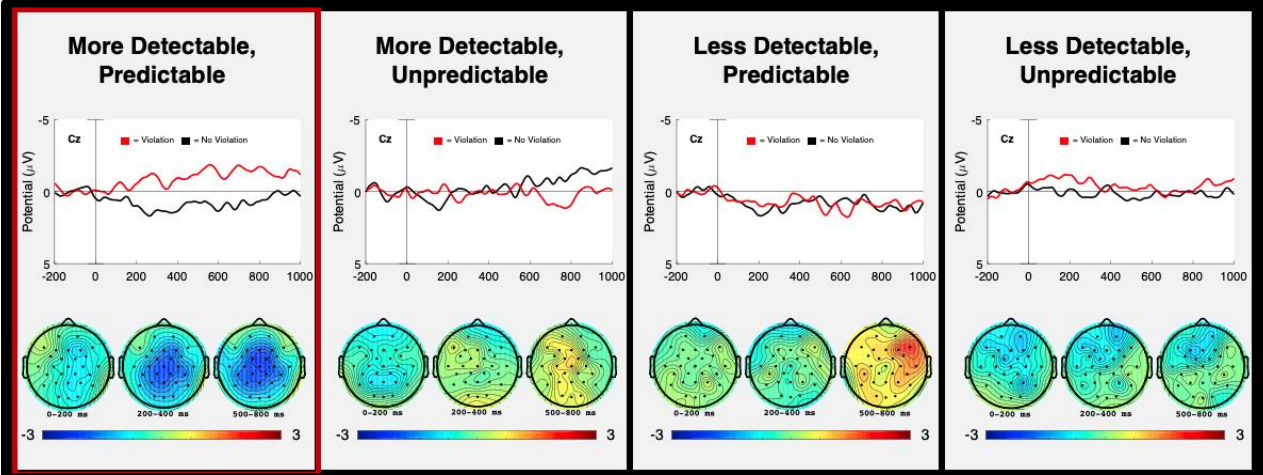


Figure 3: Results from exploratory analyses. When categorizing the effects by detectability, there were significant 3-way interactions between 200-400ms ($\beta = 1.87$, SE = 0.86, $t = 2.19$, $p < 0.05$) and 500-800ms ($\beta = 3.22$, SE = 1.05, $t = 3.06$, $p < 0.01$). Pairwise comparisons revealed main effects of the PS errors only when the context was predictable and the error was detectable ($\beta = 1.35$, SE = 0.51, $t = 2.65$, $p < 0.01$).

References: ¹Lau, Stroud, Plesch, & Phillips, 2006; ²Steinhauer & Drury, 2012; ³Hasting & Kotz, 2008; ⁴Gunter, Friederici, & Hahne, 1999.

Comparison of Structural and Neural Language Models as Surprisal Estimators

Byung-Doh Oh (oh.531@osu.edu), Christian Clark, and William Schuler (The Ohio State University)

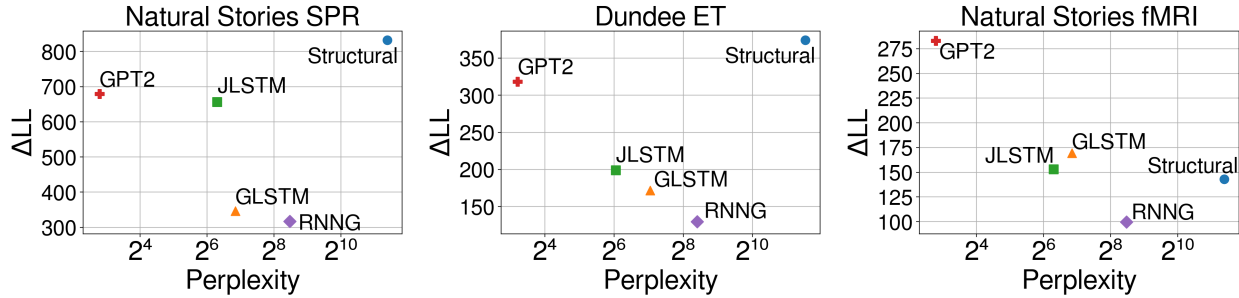
A recent trend in computational psycholinguistics has been to use large pretrained neural language models (NLMs) to generate surprisal estimates [2, 5, 8]. Although there is some evidence that NLM surprisal is predictive of human behavioral responses [2, 5], there has been very little work (see [4]) comparing its predictive power to that of surprisal from structural parser-based processing models. In this study, we conduct regression analyses on three different datasets and demonstrate that surprisal estimates from a sentence processing model informed by syntactic and morphological structure contribute to substantially better fits than those from widely-used pretrained NLMs [3, 6, 4, 9] on self-paced reading and eye-tracking data, but not on fMRI data.

To this end, we use a non-recurrent neural extension of a left-corner parser [12] that has a character-based model for estimating word generation probabilities at preterminal nodes. The proposed model defines a process of generating words w_t from underlying lemmas x_t and morphological rules r_t , which allows the processing model to capture the predictability of given word forms in a fine-grained manner.

In order to evaluate the quality of surprisal estimates from the sentence processing model informed by syntactic and morphological structure (*Structural Model*) as well as those from widely used pretrained NLMs (*GLSTM* [3], *JLSTM* [6], *RNNG* [4], *GPT2* [9]), linear mixed-effects regression analyses were conducted to evaluate model fit in terms of log-likelihood improvement on top of a baseline regression model. To this end, surprisal predictors for the Natural Stories self-paced reading corpus [1] and the Dundee eye-tracking corpus [7] were calculated from the structural model and the pretrained NLMs. The baseline predictors included were word length, word position, and unigram surprisal for Natural Stories, and word length, word position, and saccade length for Dundee. All predictors were z-transformed prior to fitting, and all surprisal predictors were spilled over by one position. All regression models included by-subject random slopes for all fixed effects. The results show that on both corpora, surprisal from the structural model made the biggest contribution to model fit in comparison to surprisal from the pretrained NLMs (Figures 1a and 1b, difference between structural model and other models significant with $p < 0.001$ by a permutation test). This finding, despite the fact that the pretrained NLMs were trained on much larger datasets (Table 1) and also show lower perplexities on test data, suggests that the structural model may provide a more human-like account of processing difficulty and may suggest a larger role of morphology, phonotactics, and orthographic complexity than was previously thought.

Additionally, to examine whether a similar tendency is observed in brain responses, we analyzed the time series of blood oxygenation level-dependent (BOLD) signals identified using functional magnetic resonance imaging (fMRI) with continuous-time deconvolutional regression (CDR; [11]). For this experiment, we used the fMRI data of the language network used in [10], which were collected from 78 subjects that listened to a recorded version of the Natural Stories Corpus. Similarly, a baseline CDR model and a series of CDR models that include each surprisal estimate were fitted to BOLD measures. The baseline predictors included were the index of current fMRI sample, unigram surprisal, and the deconvolutional intercept. Subsequently, the contribution of each surprisal estimate was examined by calculating the improvement in regression log-likelihood. The results show that in contrast to self-paced reading and eye-tracking data, surprisal from *GPT2* made the biggest contribution to regression model fit (Figure 1c, difference between *GPT2* and other models significant with $p < 0.001$ by a permutation test, other comparisons not significant).

Taken together, these results suggest that sentence processing is not purely driven by accurate next-word prediction that large NLMs are capable of. In addition, the differential contribution of surprisal from the structural model suggests that latency-based measures and blood oxygenation levels may capture different aspects of processing difficulty.



(a) Baseline log-likelihood: -17485.2 (b) Baseline log-likelihood: -60807.5 (c) Baseline log-likelihood: -269825.1

Figure 1: Perplexity measures from each model, and improvements in regression model log-likelihood from including surprisal estimates from each model. The perplexity of the structural model and the RNNG model is higher partly because they are optimized to predict a joint distribution over words and parse trees.

Model	Training corpus
<i>GPT2</i> [9]	>1B tokens
<i>JLSTM</i> [6]	~800M tokens
<i>GLSTM</i> [3]	~80M tokens
<i>RNNG</i> [4]	~950k tokens
<i>Structural Model</i>	~950k tokens

Table 1: The training corpus size for each model.

References

- [1] R. Futrell, E. Gibson, H. J. Tily, I. Blank, A. Vishnevetsky, S. Piantadosi, and E. Fedorenko. The Natural Stories Corpus. In *LREC*, pages 76–82, 2018.
- [2] A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *CMCL*, pages 10–18, 2018.
- [3] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. Colorless green recurrent networks dream hierarchically. In *NAACL-HLT*, pages 1195–1205, 2018.
- [4] J. Hale, C. Dyer, A. Kuncoro, and J. Brennan. Finding syntax in human encephalography with beam search. In *ACL*, pages 2727–2736, 2018.
- [5] Y. Hao, S. Mendelsohn, R. Sterneck, R. Martinez, and R. Frank. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *CMCL*, pages 75–86, 2020.
- [6] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. Exploring the limits of language modeling. *CoRR*, 2016.
- [7] A. Kennedy, R. Hill, and J. Pynte. The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.
- [8] G. Prasad, M. van Schijndel, and T. Linzen. Using priming to uncover the organization of syntactic representations in neural language models. In *CoNLL*, pages 66–76, 2019.
- [9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *ArXiv*, 2019.
- [10] C. Shain, I. A. Blank, M. van Schijndel, W. Schuler, and E. Fedorenko. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138, 2019.
- [11] C. Shain and W. Schuler. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *PsyArXiv*, 2019.
- [12] M. van Schijndel, A. Exley, and W. Schuler. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540, 2013.

LEXICAL AND PARTIAL PREDICTION IN A BRAZILIAN PORTUGUESE EYE-TRACKING CORPUS

João Vieira¹, Sidney Leal², Érica Rodrigues³, Sandra Aluísio², Denis Drieghe⁴, Elisângela Teixeira¹.

¹Universidade Federal do Ceará, ²Universidade de São Paulo, ³PUC-Rio de Janeiro, ⁴University of Southampton.

Introduction: Besides predicting exact upcoming words (lexical prediction) [1], readers also predict semantic and morphosyntactic information (partial prediction) [2]. Effects of high levels of predictability are regularly reported in the literature, but could be due to text manipulation, common practice in linguistic experimentation, such as in the use of sentences or contexts that trigger strong anticipation of specific words [3]. It has been argued that, in daily life, linguistic comprehension typically does not mirror such high levels of predictability. Corpora of verbal language usually differ from experimentally constructed materials in the sense that they are not built with manipulated stimuli, but with natural passages taken from books, magazines etc. In one such corpus [4], the authors used predictability norms from a Cloze Task for every word in 55 paragraphs in English and analyzed eye movements of participants who read the same paragraphs. The authors found that predictability was influential on language processing even when it was only partially correct, such as when a grammatical category was predictable, but the exact word was not. The authors also found that function words are generally more predictable than content words.

Materials and Methods: To further investigate the influence of predictability in languages in which nominal and verbal inflections differ from English, we built the first corpus of written language processing in Brazilian Portuguese using the eye movement methodology. We focused the analysis on function and content words, while examining both lexical (exact word prediction) and partial prediction. The corpus consists of predictability norms and reading measures of 50 short paragraphs from three different genres: News, Pop-Science and Literary. To calculate predictability norms, 286 participants answered an on-line word-by-word Cloze Task. Each participant answered five paragraphs, except the first word in each paragraph. Eye movements of different 37 participants were recorded using an EyeLink 1000 Hz while they read all paragraphs in a 19" monitor. Paragraphs were authentic and self-contained in meaning. In total, paragraphs had 2494 words (49 on average), out of which 1237 were unique. Target words (original words) and words answered in the Cloze Task were tagged for part of speech and divided into eight grammatical categories (nouns, adjectives, verbs, adverbs, determiners, prepositions, conjunctions and pronouns), and two classes (content and function).

Results and Discussion: Lexical predictability was measured by comparing the orthography of target and answered words (OrthographicMatch), and for partial predictability, the part of speech tag was compared (POSMATCH). Here, we report two eye movement measures closely related to early processing (Gaze duration and skip rates), expected to be sensitive to predictability effects. Lexical prediction was rare, but higher for function words (0.24) than for content words (0.13). Partial prediction was more common and higher for content words (0.44) than for function words (0.38) (Fig. 1). Lexical prediction was higher on News (0.17) and Pop-Science (0.15) texts than on Literary (0.09) texts. We ran linear mixed model analysis on Gaze Duration and logit linear mixed effects on skip rates (Tables 1 and 2). Predictability was facilitative in general, but lexical prediction was more influential than partial prediction. In Fig. 2, we see how Gaze duration dropped as both lexical and partial prediction increased, while Skip Rates increased as lexical prediction increased. Partial prediction did not influence Skip Rates. Lexical prediction had stronger effects when compared to partial prediction in general. Comparing these findings with previous research in English [4], lexical prediction is lower in BP, inviting further investigation. The Cloze Task results also indicate that predictability is involved in everyday language processing, not only when the context is highly restrictive.

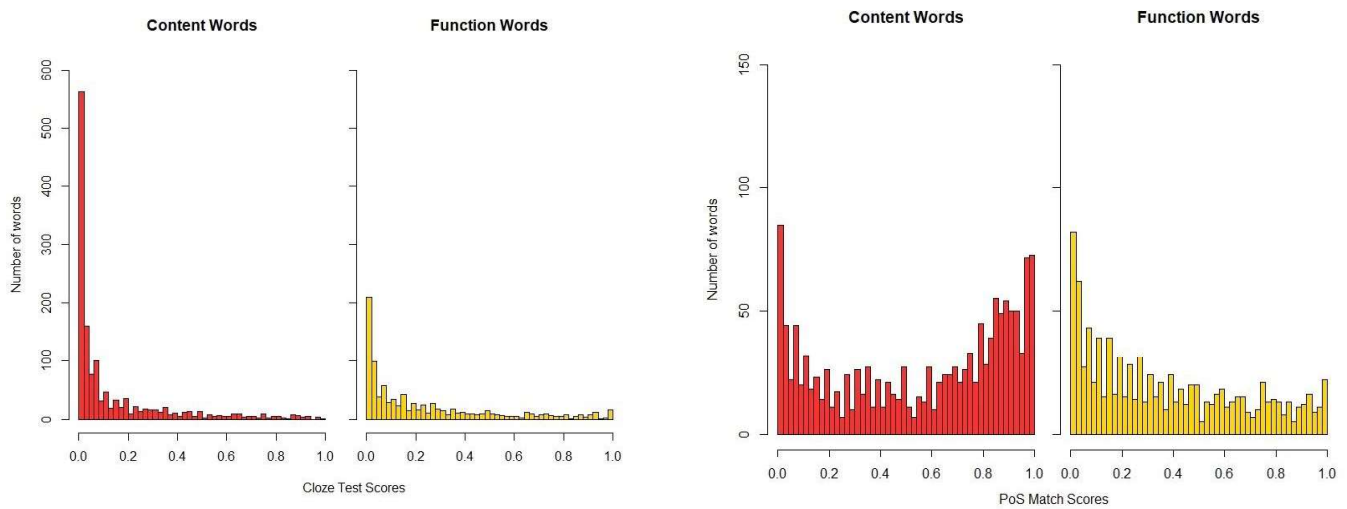


Figure 1. Histogram of lexical (left) and partial (right) predictability of content and function words.

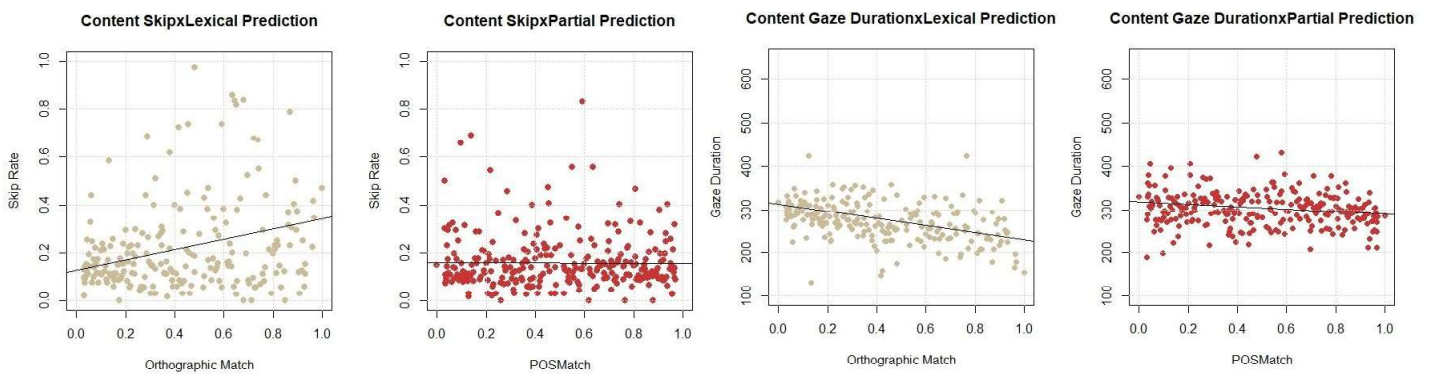


Figure 2. Influence of lexical (beige) and partial (red) predictability on Skip Rates and Gaze Duration.

[1] Calvo, M., & Mesenguer, E. (2002). Eye Movements and processing stages in reading: Relative contribution of Visual, lexical and contextual factors. *The Spanish Journal of Psychology*, 5, 66-77.

[2] Paczynski, M., & Kuperberg, G. R. (2011). Electrophysiological Evidence for Use of the Animacy Hierarchy, but not Thematic Role Assignment, During Verb Argument Processing. *Language and cognitive processes*, 26(9), 1402–1456.

[3] Huettig, F., & Mani, N. (2015) Is prediction necessary to understand language? Probably not. *Language, Cognition And Neuroscience*, 31, n. 1, p.19-31.

[4] Luke, S. G., & Christianson, K. (2016) Limits on lexical prediction during reading. *Cognitive Psychology*, 88, p.22-60.

		b	SE	df	t value	p value
Content words	(Intercept)	5,62	0,02	36,11	264,96	< 0.001
	Lexical Pred.	-0,28	0,01	41450,00	-24,14	< 0.001
	Partial Pred.	-0,04	0,01	41450,00	-5,14	< 0.001
Function words	(Intercept)	5,36	0,02	39,44	281,07	< 0.001
	Lexical Pred.	-0,12	0,02	13660,00	-5,76	< 0.001
	Partial Pred.	-0,04	0,02	13660,00	-2,09	0,03

Table 1 - Linear Mixed Model for Gaze Duration (First Run Dwell Time) on content and function words.

		b	SE	z value	p value
Content words	(Intercept)	-1,68	0,08	-22,07	< 0.001
	Lexical Pred.	1,91	0,1	19,34	< 0.001
	Partial Pred.	-0,67	0,07	-8,84	< 0.001
Function words	(Intercept)	0,11	0,06	1,81	0,07
	Lexical Pred.	1,23	0,07	17,766	< 0.001
	Partial Pred.	-0,05	0,06	-0,77	0,44

Table 2 - Logit Mixed Model for Skip Rates on content and function words.

Do children predict grammatical gender of nouns?

Katja Haeuser, Yoana Vergilova & Jutta Kray (Saarland University)
khaeuser@coli.uni-saarland.de

During language comprehension, readers anticipate upcoming words, including morpho-syntactic features such as grammatical gender [1, 2]. When reading gender-marked pre-nominal articles or adjectives that are not consistent with their prediction, young-adult readers normally incur a processing cost [2, 3], generally thought to indicate switching or updating costs. Crucially, gender prediction has been demonstrated for children too [4-7], often with the implication that predictive processing is more likely to emerge in children with above-average performance in tests of receptive and productive vocabulary [8, 9]. However, many developmental studies up to date cannot dissociate whether individual differences in predictive processing emerge as a consequence of facilitation for predictable target words or slowed reading for unpredictable target words.

We present data from an online at-home self-paced reading experiment investigating whether children aged 8-12 years ($n=36$) incur a processing cost when reading prediction-inconsistent gender-marked articles and adjectives. Stimuli were German sentences such as (translated), "When Paul finally got his driver's license, he was constantly driving around with the (German "dem") neuter/dative. old but reliable car / the (German "der") feminine/dative old but reliable group of friends", where the gender marking of the definite article and the spill-over region ("old but reliable") foreshadowed whether the most predictable noun would come up or not. Offline cloze probability ratings from 55 young- and old-adult speakers of German (ratings on children are being collected) showed high and low cloze probabilities for predictable and unpredictable gender-marked nouns and articles (> 0.8 vs < 0.01 , respectively). Sentences were presented word-by-word; participants controlled their own pace during reading. After the self-paced reading task, all children completed a standardized measure of receptive vocabulary, the German version of the Peabody Picture Vocabulary Test (PPVT) [10].

We found no evidence for disconfirmed predictions at prenominal targets when examining the full sample of 36 children (see Table 1). However, when vocabulary skill was entered into the models as an interaction variable, there were significant interactions between predictability and the scaled continuous variable of the PPVT score at the level of the second and third spill-over word after the article ($b = 237.6$, $SE = 104.7$, $t = 2.27$, $p = .03$; $b = 237.2$, $SE = 103.1$, $t = 2.30$, $p = .03$). Children with higher vocabulary skills showed effects of disconfirmed predictions prenominally, whereas children with lower vocabulary skills scores did not (Figure 1).

In order to examine whether these effects were driven by facilitation or slowing for predictable and unpredictable targets, respectively, we ran follow-up models that estimated the contribution of vocabulary score separately for reading times of predictable and unpredictable items. According to these models, children with high PPVT scores showed slowed reading times for unpredictable items (both at the second and third spill-over word: $b = 250.9$, $SE = 118.2$, $t = 2.12$, $p = .04$; $b = 244.9$, $SE = 116.6$, $t = 2.10$, $p = .04$), but not for predictable items ($b = 143.1$, $SE = 92.1$, $t = 1.3$, $p = .1$; $b = 157.8$, $SE = 89.9$, $t = 1.6$, $p = .09$).

Our data suggest that German-speaking primary and middle-schoolers, especially those with high vocabulary skills, actively anticipate predictable continuations based on preceding gender-marked definite articles and adjectives (in line with adult reader findings [1-3]). These effects appear to reflect the cost of disconfirmed predictions for unpredictable target words, as opposed to a facilitation for highly predictable targets. Our outstanding goals are to substantiate previous findings arguing that especially receptive (but not productive) vocabulary skills determine prediction costs in children.

	<i>the</i>	<i>old</i>	<i>but</i>	<i>reliable</i>	<i>car</i>
<i>b</i>	8	9	52	57	114
<i>SE</i>	10	14	37	36	22
<i>p</i>	.8	.5	.2	.1	< .001

Table 1. Parameter estimates (*b*'s), standard errors (SE) and *p*-values from models estimating the RT difference (in ms) between unpredictable and predictable target words in the full sample of 36 children.

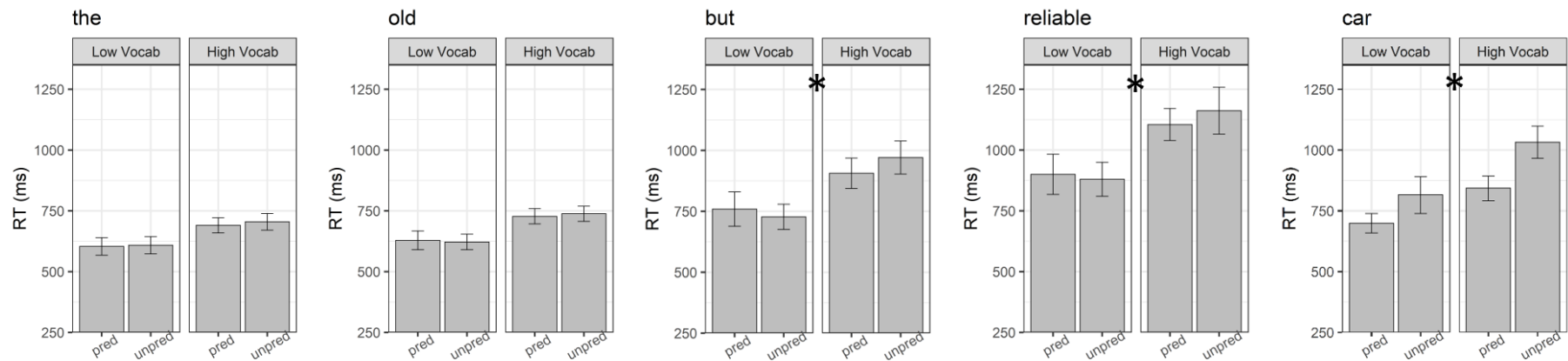


Figure 1. Average RTs (\pm SE) on target words in predictable and unpredictable items in children with low and high vocabulary scores, based on a median split of their PPVT scores (all statistical models were run with the scaled continuous variable). Asterisks indicate statistical significance at $p < .05$.

References

- [1] Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of cognitive neuroscience*, 16(7), 1272-1288.
- [2] Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 443-467.
- [3] Haeuser, K. I., Kray, J., & Borovsky, A. (2020). Great expectations: Evidence for graded prediction of grammatical gender. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 1157-1163). Cognitive Science Society.
- [4] Cholewa, J., Neitzel, I., Bürsgens, A., & Günther, T. (2019). Online-processing of grammatical gender in noun-phrase decoding: An eye-tracking study with monolingual German 3rd and 4th graders. *Frontiers in Psychology*, 10, 2586.
- [5] Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3), 193-198.
- [6] Brouwer, S., Sprenger, S., & Unsworth, S. (2017). Processing grammatical gender in Dutch: Evidence from eye movements. *Journal of Experimental Child Psychology*, 159, 50-65.
- [7] Van Heugten, M., & Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Developmental Science*, 12(3), 419-425.
- [8] Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, 38(4), 843-847.
- [9] Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, 112(4), 417-436.
- [10] Lenhard, A., Lenhard, W., Segerer, R., and Suggate, S. (2015). *Peabody Picture Vocabulary Test – Revision IV (Deutsche Adaption)*. Pearson Assessment.

Both semantic and form representations are pre-activated during sentence comprehension: Evidence from EEG Representational Similarity Analysis

Lin Wang (Tufts University, Harvard Medical School), Trevor Brothers (Tufts University, Harvard Medical School), Cheng Feng (Tufts University), Sophie Greene (Tufts University), Ole Jensen (University of Birmingham), Gina Kuperberg (Tufts University, Harvard Medical School)

It is well established that incoming words are facilitated in proportion to their predictability during language comprehension^[1]. However, it remains unclear whether upcoming linguistic information is *pre-activated* before new bottom-up input becomes available, and if so, whether such pre-activation occurs at semantic and/or form levels of representation^[2]. In the present study, we used Representational Similarity Analysis (RSA) in combination with EEG to address this question. The basic assumption of RSA is that unique representations are encoded as distinct spatial patterns of neural activity, and so representationally similar items (e.g. the same words) produce neural patterns that are more similar to each other than representationally distinct items (e.g. different words)^[3]. By combining RSA with EEG, it is possible to determine *when* representationally specific information is encoded prior to the onset of incoming word^[4]. In order to dissociate the time-course of form-based and meaning-based pre-activation, we capitalized on the ambiguity of homonyms — words that have the same orthographic and phonological form but distinct meanings (e.g. *bank*). Participants read highly constraining sentences that were predictive of either: (1) a homonym's subordinate meaning (e.g. a river *bank*), (2) its dominant meaning (e.g. a financial *bank*), or (3) a word that was semantically related to the dominant meaning (e.g. *loan*). Spatial RSA was conducted on EEG data at each time point prior to word onset to determine whether/when readers pre-activated *semantic* or *word-form* representations.

Design: We developed 84 triplets of highly constraining sentences (Table 1) (cloze: mean \pm SD = 88% \pm 8%). Each triplet contained a form-related homonym pair (*bank-bank*), with one member constraining for the homonym's subordinate meaning and the other constraining for its dominant meaning. Each triplet also contained a semantically related pair, with one member constraining for the homonym's dominant meaning and the other constraining for a word that was semantically related to this dominant meaning (*bank-loan*). In the EEG experiment, sentences in each triplet were presented in pseudorandom order and separated by at least 30 sentences. Each sentence was presented word by word (300ms per word + 400ms ISI). Participants (N=33) answered True/False comprehension questions following 1/6th of the sentences.

RSA Analysis: At each time point from -700ms before until the onset of critical words, we correlated spatial patterns of EEG activity (across 64 channels) *within* form related homonym pairs (e.g. *bank-bank*) and *within* semantically related pairs (e.g. *bank-loan*), and subtracted these values from the correlations produced *between* unrelated pairs (e.g. *bank-foot*, *bank-toes*, *loan-toes*) (Fig. 1). This difference reflects the increase in neural similarity associated with items with overlapping vs. non-overlapping representations. We then conducted cluster-based permutation tests (10,000 permutations) across the full prediction time window (-700ms to 0ms relative to critical word onset) to identify significant differences in spatial similarity across conditions.

Results: The semantically related pairs showed greater similarity effects (within-pairs > between-pairs) between -391ms and -309ms ($p = .003$) prior to the critical word onset, while the form related homonym pairs showed greater similarity between -53ms and -8ms ($p = .025$) (Fig. 2).

Discussion: These findings provide clear neural evidence for semantic and form *pre-activation* during the incremental comprehension of predictable sentences. Moreover, the earlier pre-activation of semantic than form information is consistent with a hierarchical generative framework^[5], which posits that top-down pre-activation is propagated from higher to successively lower levels of the linguistic hierarchy over time.

Table 1. Examples of sentences

1a	The muddy sides of a river are called a <u>bank</u> .	Subordinate] Form related
1b	James went to deposit the check at his <u>bank</u> .	Dominant	
1c	To pay for college the student took out a <u>loan</u> .	Dominant-related	
2a	There are twelve inches in a <u>foot</u> .	Subordinate] Between-pairs
2b	He put a shoe on his left <u>foot</u> .	Dominant	
2c	He had healthy nails on all his fingers and <u>toes</u> .	Dominant-related] Semantically related

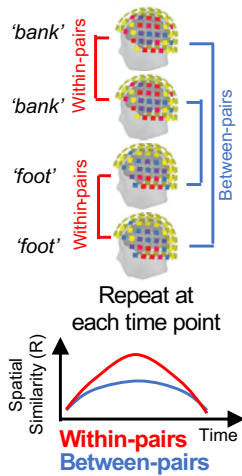
RSA methods

Fig. 1. A schematic illustration of the spatial RSA analysis stream. First, for each trial, and at each time point, we extracted a vector of EEG data that represented the spatial pattern of activity across all 64 EEG channels. Second, we quantified the degree of spatial similarity of EEG activity produced by pairs of trials by correlating their spatial vectors. Third, we averaged the spatial similarity R-values separately for sentence pairs that predicted words with overlapping representations (“within-pairs”) and for sentence pairs that predicted words without overlapping representations (“between-pairs”). Finally, we repeated this process at each time point, yielding time-series of R-values that reflected the degree of spatial similarity at each time sample between sentence pairs that predicted words with or without overlapping representations. The spatial similarity difference between the within-pairs and between-pairs reflected the increase in neural similarity associated with items with overlapping vs. non-overlapping representations.

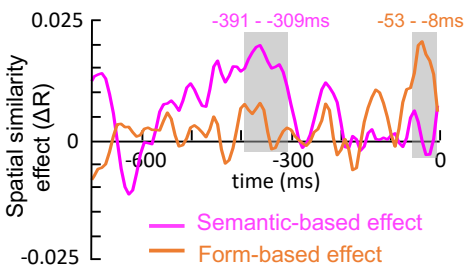
RSA results

Fig. 2. Time course of semantic-based and form-based spatial similarity effects. The semantic-based effect was obtained by subtracting the spatial similarity/correlations within the semantically related pairs from the between-pair correlations, and the form-based effect was obtained by subtracting the spatial similarity/correlations within the form related homonym pairs from the between-pair correlations. Relative to the between-pairs, the spatial similarity was greater when the predicted words were

semantically related ($p = .003$) between -391 and -309ms, and when the predicted words had the same word forms ($p = .025$) between -53ms and -8ms prior to the critical word onset.

References

- [1] DeLong, Urbach, & Kutas. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117-1121.
- [2] Nieuwland. (2019). Do ‘early’ brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews*, 96, 367-400.
- [3] Kriegeskorte, Mur, & Bandettini. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4.
- [4] Wang, Kuperberg, & Jensen. (2018). Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. *ELife*, 7, e39061.
- [5] Kuperberg, & Jaeger. (2016). What do we mean by prediction in language comprehension?. *Language, Cognition and Neuroscience*, 31(1), 32-59.

Contributions of Propositional Content and Syntactic Categories in Sentence Processing

Byung-Doh Oh (oh.531@osu.edu) and William Schuler (The Ohio State University)

Expectation-based theories of sentence processing posit that processing difficulty is determined by predictability in context [3, 6]. While predictability quantified via surprisal has gained empirical support, this representation-agnostic measure leaves open the question of how to best approximate the human comprehender’s latent probability model. One factor related to memory usage that has received less attention in psycholinguistic modeling is the influence of *propositional content*, or meaning that is being conveyed by the sentence. Early psycholinguistic experiments have demonstrated that the propositional content of utterances tends to be retained in memory, whereas the exact surface form and syntactic structure are forgotten [1, 4]. This suggests that memory costs related to incrementally constructing a representation of propositional content might manifest themselves in behavioral responses during online sentence processing.

This study uses a generative, incremental, and differentially content-sensitive processing model to estimate surprisal predictors that capture the influence of *propositional content* differentially with that of *syntactic categories*, which are devoid of propositional content. The processing model extends a left-corner parser [5, 9] to incorporate propositional content by augmenting each node in a parse tree to consist not only of a syntactic category label but also a *predicate context vector*, which consists of $\langle \text{predicate}, \text{role} \rangle$ pairs that specify the content constraints on a variable over discourse entities. These predicate context vectors are obtained by reannotating the training corpus using a generalized categorial grammar of English [8], which is sensitive to syntactic valence and non-local dependencies. The parser is implemented as a series of feedforward neural network submodels that make parsing decisions using predicate context vectors and syntactic category labels as features. An advantage of this formulation is that this processing model can be trained to make parsing decisions without conditioning on either predicate context vectors or syntactic categories, which allows a clean ablation of their contribution to the probability model.

In order to evaluate the contribution of propositional content and syntactic categories to predicting behavioral responses, surprisal predictors for the Natural Stories self-paced reading corpus [2] were calculated from the content-sensitive processing model and its two ablated versions, which were trained on sections 02 to 21 of the WSJ corpus [7] using three different random seeds. Subsequently, a series of ablative likelihood ratio tests with nested linear mixed-effects models were conducted to test whether surprisal estimates from the full processing model (*FullSurp*) improve regression model fit over those from a processing model that lacks propositional content information (*NoConSurp*) or syntactic category information (*NoCatSurp*). As there were three variants of each surprisal predictor, a total of nine (3×3) LRTs were performed for each ablated surprisal predictor. The regression models also included baseline predictors for word length, word position, and 5-gram surprisal. All predictors were z-transformed prior to fitting, and all surprisal predictors were spilled over by one position. All regression models included by-subject random slopes for all fixed effects and random intercepts for each word and subject-sentence interaction. The results in Table 1 show that *FullSurp* made a statistically significant contribution to model fit over *NoConSurp* in six out of nine LRTs, which is highly significant according to a binomial test ($p < 0.001$). The significant contribution of *FullSurp* over *NoCatSurp* was observed as well, with six out of nine LRTs indicating significantly improved model fit ($p < 0.001$).

To explore the extent to which integration costs associated with filler-gap constructions could be explained by the influence of propositional content, we replicate the same experiment on filler-gap verbs. The results in Table 2 show that *FullSurp* made a significant contribution to model fit over *NoConSurp* in three out of nine LRTs ($p = .008$). This indicates that the full processing model captures the influence of propositional content and syntactic categories differentially, both of which contribute to predicting self-paced reading times, suggesting their role in sentence processing.

NoConSurp	FullSurp			NoCatSurp	FullSurp		
	1	2	3		1	2	3
1	ConvFail	0.035*	0.018*	1	ConvFail	<0.001***	ConvFail
2	0.004**	ConvFail	0.047*	2	<0.001***	<0.001***	<0.001***
3	0.003**	0.058	0.036*	3	ConvFail	<0.001***	<0.001***

Table 1: p -values from LRTs testing the contribution of *FullSurp* over *NoConSurp* (left) and *NoCatSurp* (right) to regression models predicting self-paced reading times. Any LRT in which either the base or full regression model failed to converge (*ConvFail*) was considered as a null result.

* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

NoConSurp	FullSurp		
	1	2	3
1	0.095	0.046*	0.037*
2	0.119	0.058	0.049*
3	0.186	0.097	0.081

Table 2: p -values from LRTs testing the contribution of *FullSurp* over *NoConSurp* to regression models predicting self-paced reading times of filler-gap verbs. * : $p < 0.05$.

References

- [1] Bransford, J. D. and Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, 2:331–350.
- [2] Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., and Fedorenko, E. (2018). The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 76–82.
- [3] Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8.
- [4] Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10:409–416.
- [5] Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA.
- [6] Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- [7] Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- [8] Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2125–2140.
- [9] van Schijndel, M., Exley, A., and Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

German pronoun interpretation follows Bayesian principles

Clare Patterson (University of Cologne), Petra B. Schumacher (University of Cologne), Bruno Nicenboim (Tilburg University), Johannes Hagen (University of Cologne), Andrew Kehler (University of California, San Diego)

The Bayesian Model for pronouns (Kehler et al 2008 et seq.) predicts that pronoun production and comprehension are related by Bayes' rule: $P(\text{referent} | \text{pronoun}) \propto P(\text{pronoun} | \text{referent})P(\text{referent})$. $P(\text{referent} | \text{pronoun})$ represents the comprehension bias: the probability that a particular referent is being referred to by a pronoun. The likelihood term $P(\text{pronoun} | \text{referent})$ represents the production bias: the hearer's estimate of the probability that speaker will use a pronoun to refer to a particular referent. The prior term, $P(\text{referent})$, represents the next-mention bias: the probability that a particular referent will get mentioned next, regardless of the referring expression used. Values for the prior and likelihood terms are estimated from passage completion experiments with free prompt conditions, yielding a predicted comprehension bias that can be compared to the actual comprehension bias measured using pronoun-prompt conditions with the same contexts. The Bayesian Model has been quantitatively examined in English (Rohde & Kehler 2014) and Mandarin Chinese (Zhan et al 2020) by comparing its predictions against the predictions from two competing models: a Mirror Model, a normalized $P(\text{pronoun} | \text{referent})$, and the Expectancy Model, a normalized $P(\text{referent})$. In this study, we further the cross-linguistic support for the Bayesian Model by applying it to German personal and demonstrative pronouns, and provide novel quantitative support for the model by assessing model performance in a Bayesian statistical framework that allows implementation of a fully hierarchical structure, providing the most conservative estimates of uncertainty. Applying the Bayesian model to German provides new cross-linguistic evidence because both personal and demonstrative pronouns can refer to human entities. Additionally, the referential biases for the demonstrative *dieser* are not well understood, but demonstratives are thought to be more rigid in their interpretation than the personal pronoun (Kaiser 2011, inter alia), making them a good test for the Bayesian Model.

Two passage completion studies were conducted with items consisting of a context sentence followed by one of three prompt types: personal pronoun (*er*), demonstrative pronoun (*dieser*), and free prompt (a blank line). To explore the effects of syntactic and semantic context factors, Experiment 1 (N=48) compared contexts with active-accusative verbs (1) and dative-experiencer verbs (2) and Experiment 2 (N=40) compared contexts with experiencer-stimulus verbs (3) and stimulus-experiencer verbs (4). Each model (Expectancy, Mirror, and Bayes) was fit with Bernoulli likelihoods for the referent and categorical likelihoods for the expression type, with weakly regularizing priors. Observation-level predictions for each model were made based on the free-prompt data and fitted against the held out observations from the pronoun-prompt data. Model fit was evaluated graphically with holdout predictive check, and numerically using holdout validation (Vehtari & Ojanen 2012).

Overall, the Bayesian Model makes more accurate predictions than both the Expectancy and Mirror Models in both experiments (see table and figures, which compare the predictive accuracy of the models with respect to pronoun interpretation). Furthermore, the model accounts for the demonstrative pronoun *dieser* as well as the personal pronoun, despite its more rigid resolution preferences. We further confirmed that semantic factors (implemented as a verb-type contrast) affect the prior term $P(\text{referent})$ to a much greater extent than the likelihood term $P(\text{pronoun} | \text{referent})$, underlining the separation of pronoun-related biases from form-independent expectations about the upcoming referent (Kehler & Rohde 2013).

As an ensemble, the results for German pronouns strongly support the predictions of the Bayesian Model, according to which comprehenders reverse engineer the speaker's referential intentions using Bayesian principles.

- (1) Vorletzte Nacht hat der Hund den Papagei geärgert. Er/Dieser/____
The night before last the dog (nom.masc.) annoyed the parrot (acc.masc.). He/DEM/____
- (2) Gestern ist dem Feuerwehrmann der Polizist aufgefallen. Er/Dieser/____
Yesterday the firefighter (dat.masc.) noticed the police officer (nom.masc.). He/DEM/____
- (3) Der Dieb fürchtete den Polizisten. Er/Dieser/____
The thief (nom.masc.) feared the police officer (acc.masc.). He/DEM/____
- (4) Der Fußballer erstaunte den Manager. Er/Dieser/____
The footballer (nom.masc.) astonished the manager (acc.masc.). He/DEM/____

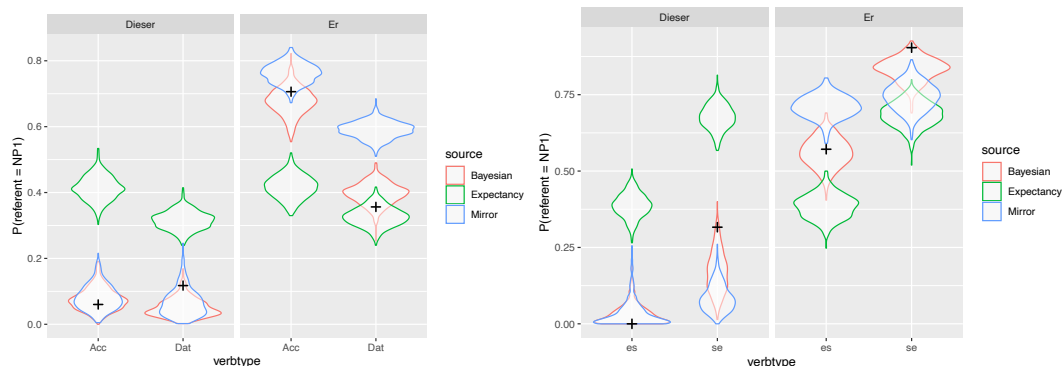


Figure 1. Crosses show observed proportion of NP1 interpretations for Experiment 1 (left plot) and Experiment 2 (right plot) (from held out data); violin plots depict distribution of simulated proportions based on model predictions.

Experiment 1						Experiment 2				
	elpd	SE elpd	elpd_diff	SE diff	weight	elpd	SE elpd	elpd_diff	SE diff	weight
B	-728	27	0	0	0.89	-368	19	0	0	0.9
M	-860	27	-132	14	0.00	-467	24	-98	13	0.0
E	-966	16	-238	24	0.11	-578	16	-209	23	0.1

Table 1. B = Bayesian Model, M = Mirror Model, E = Expectancy Model. A higher expected log-predictive density (elpd) indicates better predictive accuracy. The highest scoring model is the baseline for elpd difference (elpd_diff) and difference Standard Error (SE). Weight columns represent weights of the individual models that maximize the total elpd score of all the models.

- Kaiser E (2011). On the Relation between Coherence Relations and Anaphoric Demonstratives in German. In Reich I et al (Eds.), *Proceedings of Sinn & Bedeutung* 15 (pp. 337-351). Saarland University Press.
- Kehler A, Kertz L, Rohde H, & Elman J (2008). Coherence and Coreference Revisited, *Journal of Semantics* 25:1, 1-44.
- Kehler A & Rohde H (2013). A Probabilistic Reconciliation of Coherence-Driven and Centering-Driven Theories of Pronoun Interpretation, *Theoretical Linguistics* 39, 1-37.
- Rohde H & Kehler A (2014). Grammatical and Information-structural Influences on Pronoun Production. *Language, Cognition and Neuroscience* 29, 912-927.
- Vehtari A, & Ojanen J (2012). A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison. *Statistics Surveys* 6, 142-228.
- Zhan M, Levy R & Kehler A (2020). Pronoun Interpretation in Mandarin Chinese Follows Principles of Bayesian Inference. *PLoS ONE* 15:8, e0237012.

“Good-enough” production: accessibility influences choice of taxonomic level

Crystal Lee, Casey Lew-Williams, and Adele Goldberg (Princeton University)

Speakers often have a choice in how to label referents (e.g., *flower* vs. *rose*), and the most informative or ideal descriptions are not always used. For example, a sieve maybe called a *strainer*, or a caterpillar, a *bug*. We hypothesize that accessibility (ease of retrieval) predicts the use of such under-informative language. Specifically, we predict that there are conditions that restrict accessibility, under which speakers will produce less ideal but more accessible constructions that are “good enough” to convey the intended message [1]. Thus we suggest “good-enough production” exists in a way that parallels “good-enough comprehension” [2-4]. Critically, we predict that speakers produce descriptions that are “good enough” but not ideal, even when they have the requisite knowledge required to produce the ideal option.

In a preregistered study (<https://osf.io/r2t5y/>), we taught online participants ($n=100$) specific and general category names (e.g., *lantana* and *flower*) associated with images of 6 unfamiliar flowers and 6 unfamiliar weeds. Participants had to successfully produce at least 75% of the newly learned labels after a maximum of three learning cycles to continue to the main task, which required them to label images of the flowers or weeds they had just learned. Participants earned a small monetary reward for correctly using general labels (*weed* or *flower*), and earned twice the reward for correctly producing the newly learned specific labels. No reward was given for incorrect responses. Thus, specific labels were the ideal responses, and general labels were “good-enough.”

We manipulated the accessibility of labels in three ways. Half of participants were required to respond in under 3 seconds, which was intended to simulate naturalistic communicative demands; the other half had no time constraint (Speeded vs. Un-speeded conditions). Between the initial exposure and the main production task, all participants performed an intermediate filler task that required them to produce *flower* and *weed*, one three times as often as the other (Primed vs. Un-primed). Finally, half of participants learned visually unambiguous weeds and flowers, and half were tested on a subset of weeds that could be mistaken for flowers and *vice versa* (Interference vs. Non-interference). All items were normed separately.

We found a strong effect of time pressure on “good-enough” productions (Figure 1): participants produced significantly more category responses in the Speeded condition than the Un-speeded condition ($\beta = 1.07$, $z = 4.8$, $p < .001$). The priming manipulation yielded null results, likely because both category labels (*weed*, *flower*) were highly accessible, regardless of the priming manipulation. Few errors were produced (30 out of 1199 responses) and were almost entirely restricted to the subgroup who learned plants that were ambiguous between weeds and flowers (Interference: $\beta = 1.85$, $z = 3.3$, $p < .01$) and had to respond under time pressure ($\beta = 1.04$, $z = 3.0$, $p < .01$), with zero errors in the Non-interference, non-speeded subgroup.

After the main task, participants performed a two-alternative-forced-choice task on the specific labels they had been taught to ensure that they were familiar with the newly learned terms, even if they had produced good-enough (general) descriptions. For this, participants were given a specific label and two familiar images and were asked to identify the correct image. As intended, accuracy was very high ($M = 0.97$).

The current results suggest that speakers tend to produce a “good-enough” description when an ideal description is not sufficiently accessible at the moment of speaking. Good-enough production is particularly influenced by the time-pressure involved in natural, conversational dynamics, where the limited time between conversational turns creates a bottleneck on lexical retrieval. This work offers new insight into why it is so common for even fluent speakers to produce non-optimal words and sentences. Future work will test the same design with children, who are expected to rely more heavily on good-enough production, as they are likely to find it even more effortful to access ideal choices under naturalistic communicative demands.

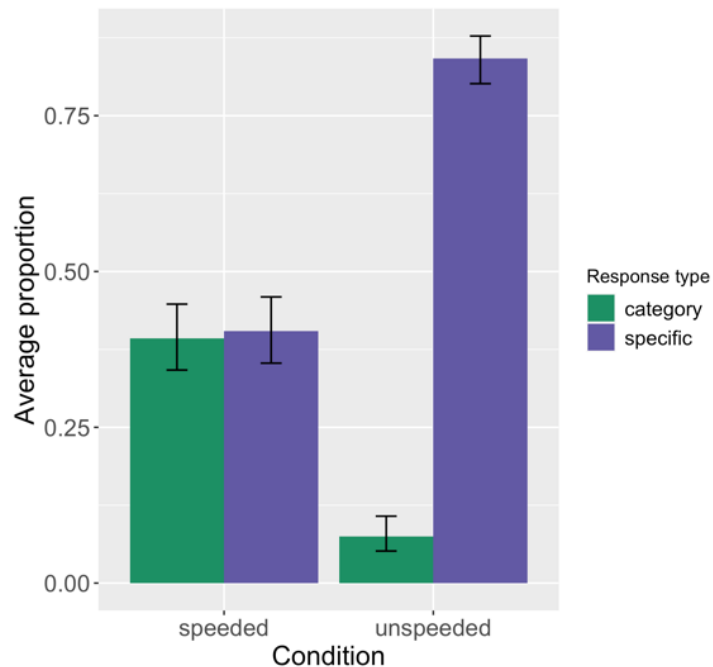


Figure 1. Average proportion of responses (category or specific) by condition (Speeded and Unspeeded).

- [1] Koranda, M., Zettersten, M., & McDonald, M. (2018). Word frequency can affect what you choose to say. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- [2] Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive psychology*, 42(4), 368-407.
- [3] Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1), 11-15.
- [4] Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1(1-2), 71-83.

Choosing a Referring Expression: Intra-sentential Ambiguity Avoidance in Romanian

Rodica Ivan, Brian Dillon, & Kyle Johnson (University of Massachusetts, Amherst)

Much work shows that ambiguity avoidance guides speakers' choice of referring expression when these forms refer to discourse antecedents introduced in previous clauses [1,2,3,4,5,8]. Here we investigate whether similar pressures apply to pronouns which have clausemate antecedents. We test this in four experiments in Romanian, a language which allows both reflexives (complex *el însuși* 'him himself', simplex *sine* 'self') and regular pronouns (*el/ea* 'him/her') to refer to syntactically local antecedents. We test whether the production and interpretation of these two forms is influenced by ambiguity avoidance both for *referential* (Exp 1 & 3) and *quantificational* (Exp 2 & 4) antecedents (see (1) and (2)). The semantic processes responsible for co-valuing a pronominal with a referential antecedent engage discourse information that is not exploited in co-valuing a pronominal with a quantificational antecedent. Some proposals reserve ambiguity avoidance effects to just those processes that involve discourse information [7, 10, 11]. Our findings do not support this view. We test (i) whether speakers produce pronouns *el/ea* 'him/her' less frequently in contexts in which they are ambiguous between a reflexive and a non-reflexive reading (Exp 1/2), and (ii) whether listeners interpret *el/ea* as non-reflexive more often when listening to speakers who regularly use reflexive pronouns (Exp 3/4). Our data support (i) and (ii).

Production (Exp 1, 2): We manipulated contextual ambiguity by providing contexts where all characters had matching or mismatching gender [4,5]. Participants continued a sentence fragment with a visually-provided context (Fig. 1). **Participants:** 68 native speakers of Romanian participated in each experiment. **Materials:** 16 items in 4 conditions: PICTURE TYPE (*Reflexive/Disjoint*) x AMBIGUITY (*Gender Match / Mismatch*) and 20 fillers. **Results:** The rate of production for all response types for each condition is given in Table 1 and Table 2. Regular pronouns *el/ea* were preferred in unambiguous *Gender Mismatch* scenarios for all reference relations. Logistic mixed-effects regression revealed a clear effect of AMBIGUITY (Exp 1 (*Referential DPs*): $z=5.13, p<0.001$, Exp 2 (*Quantified DPs*): $z=6.654, p<0.001$), and a main effect of PICTURE TYPE (Exp 1: $z=-2.68, p<0.01$, Exp 2: $z=-3.1, p<0.01$). **Speakers used unambiguous reflexives more often in ambiguous contexts.**

Comprehension: Exp 3, 4 test whether the interpretation of an ambiguous pronoun is sensitive to the availability of alternative referring expressions [1,2,3,4,5,6,8,9]. We gave participants a picture-matching task with the within-subjects factor of AMBIGUITY (*Ambiguous/Reflexive/Disjoint*). We manipulated the availability of unambiguous reflexive forms in the experiment in a between-subjects GROUP factor: the *Gender* group of subjects only heard sentences with regular pronouns *el/ea* (gender cues disambiguated), while the *Form* group heard sentences with unambiguous reflexives and demonstratives (referring expression form disambiguated). In both groups, the critical ambiguous stimuli were identical. **Participants:** 68 native speakers of Romanian per experiment. **Materials:** 15 items and 20 fillers per experiment. **Results:** The rate of choosing a reflexive interpretation, i.e. the dependent variable, is given by condition in Tables 3 and 4. Logistic mixed-effects regression revealed the rate of reflexive interpretation in the *Ambiguous* condition was significantly different from the rate of reflexive interpretation in the *Reflexive* (Exp 3 (*Referential DPs*): $z=5.98, p<0.001$, Exp 4 (*Quantified DPs*): $z=5.16, p<0.001$) and the *Disjoint* (Exp 3: $z=-8.18, p<0.001$, Exp 4: $z=-6.07, p<0.001$) conditions. Nested mixed-effects regression models revealed no significant effect of GROUP on the rate of reflexive interpretation in the *Ambiguous* condition in Exp 3 ($z=-1.72, p=0.08$), but a significant effect in Exp 4 ($z=-1.98, p<0.05$). **Listeners interpreted ambiguous pronouns as reflexive less often when speakers regularly used unambiguous reflexives.**

Discussion. Our results provide some evidence of ambiguity avoidance in production and comprehension for local coreference and bound variables alike. Broadly, our results support the hypothesis that ambiguity avoidance is a general (but not the only) constraint on reference. Contra [7, 10, 12], coreference and binding dependencies may be similarly affected by discourse context.

- (1) *Referential DP Subject (Experiments 1 & 3): 2 character context*
 Acasă la Mihai, Andreia vorbit despre el / el însuși / acesta
 home at Mihai, Andrei has talked about him / him himself / this one
 'At Mihai's house, Andrei talked about him(self) / himself / this one.'
- (2) *Quantified DP Subject (Experiments 2 & 4): 4 character context*
 Acasă la bunicul Radu, fiecare băiat a vorbit despre el / el însuși / acesta
 home at grandpa Radu, every boy has talked about him / him himself / this one
 'At grandpa Radu's house, every boy talked about him(self) / himself / this one.'

Figure 1. Sample Item by Condition in Production Experiment 1. (Exp. 2 has 4 characters)



Table 1. Exp. 1: Referential DPs (2 characters)

RESPONSE TYPE	PRON.	REFLEXIVE		OTHER	
	<i>el</i>	<i>el însuși</i>	<i>sine</i>	<i>acesta</i>	NAME
REFL. MISMATCH	54.5%	34%	5%	0%	4%
REFL. MATCH	38.8%	49.3 %	6.6%	0.5%	3.5%
DISJ. MISMATCH	50%	0%	0%	4.4%	45.5%
DISJ. MATCH	24%	0%	0%	3.7%	72.3%

Table 2. Exp. 2: Quantified DPs (4 characters)

RESPONSE TYPE	PRON.	REFLEXIVE		OTHER	
	<i>el</i>	<i>el însuși</i>	<i>sine</i>	<i>acesta</i>	NAME
REFL. MISMATCH	52.6%	35.6%	11%	0%	0%
REFL. MATCH	32.5%	48.7 %	16.5%	0%	0%
DISJ. MISMATCH	34.2%	0%	0%	11.8%	54%
DISJ. MATCH	16.6%	0%	0%	13.3%	70.1%

Figure 2. Sample Item by Condition in Comprehension Experiment 3. (Exp. 4 has 4 characters)

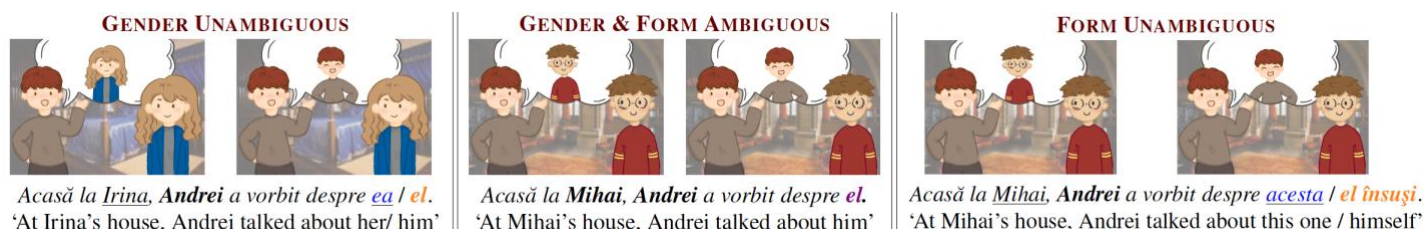


Table 3. Exp. 3: Referential DPs (2 characters)
 Rate of Reflexive Interpretation by Condition.

	FORM		GENDER	
	%Refl.	Pronoun	%Refl.	Pronoun
AMBIGUOUS	52.2%	<i>el / ea</i>	62.1%	<i>el / ea</i>
REFLEXIVE	95.7%	<i>el însuși / ea însăși</i>	96.5%	<i>el / ea</i>
DISJOINT	15.9%	<i>acesta / aceasta</i>	3.5%	<i>el / ea</i>

Table 4. Exp. 4: Quantified DPs (4 characters)
 Rate of Reflexive Interpretation by Condition.

	FORM		GENDER	
	%Refl.	Pronoun	%Refl.	Pronoun
AMBIGUOUS	42%	<i>el / ea</i>	59.5%	<i>el / ea</i>
REFLEXIVE	100%	<i>el însuși / ea însăși</i>	99.3%	<i>el / ea</i>
DISJOINT	14.6%	<i>acesta / aceasta</i>	0.6%	<i>el / ea</i>

[1] Ariel, M., 1990. Accessing NP antecedents [2] Ariel, M., 2001. Accessibility theory [3] Arnold, J.E., 1998. PhD Thesis. [4] Arnold, J.E., 2010. Ling. & Ling. Compass 4 [5] Arnold, J.E., Griffin, Z.M., 2007. JML 56 [6] Dowty, D., 1980. CLS [7] Grodzinsky, Y., Reinhart, T., 1993. LI 24 [8] Gundel, J.K., Hedberg, N., Zacahrski, R., 1993. Language 69. [9] Levinson, S.C., 1987. Journal of Linguistics 23 [10] Reinhart, T., 1983. Ling. & Phil 6 [11] Reinhart, T., 2006. Interface Strategies [12] Reuland, E., 2011. MIT Press.

Invisible, unmentioned entities affect referential forms

Si On Yoon (U. of Iowa) Breanna Pratley (U. of Toronto) and Daphna Heller (U. of Toronto)

Referential Expressions (REs) are subject to multiple influences. One such influence is discourse history, whereby speakers tend to reuse structures and concepts that were said earlier [e.g. 1,2], and even more so if the noun overlaps [3]. But speakers are also rational: they normally include *just enough* information to allow the addressee to pick out the intended referent [e.g., 1,4]. It is therefore surprising that speakers sometimes include information that distinguishes the intended referent from an entity that is no longer present: in contexts like Fig. (1a), speakers sometimes say “*the open umbrella*” to refer to a single umbrella after referring to a different umbrella on an earlier trial [e.g., 5]. This behavior is not rational because the umbrella from the earlier trial is no longer a potential referent. Here we demonstrate an even more surprising effect: REs are influenced by an entity that is not just no longer visible, but was not even described earlier.

General Method. Participants (n=24) viewed virtual grids of 15 “cards” each. On each trial, 4 of the 15 cards were “flipped” to show their images, and the participant described a target card for the experimenter to click. Participants completed 8 trials with each grid before moving to a new grid: 1 ENTRAINMENT, 1 TEST, and 6 interspersed fillers (trials order varied by grid, but the test trial always followed the entrainment trial).

Exp. 1. The test trial was constant, and always included one object (e.g., a striped open umbrella). The entrainment trial included (i) the same noun (e.g. umbrella) or a different noun (e.g. bottle), and (ii) one or two such objects. The same and different objects contrasted in the same property (e.g., open vs. closed) so as to elicit the same modifier. Indeed, speakers produced the modifiers at ceiling for two objects (same: 100%, diff: 97%), and much less for a single object (same: 33%; diff: 20%). Our main question is how referential forms at TEST are influenced by the ENTRAINMENT trials. We calculated how likely speakers were to say “open umbrella” after they said “*closed N*” in entrainment. Due to the difference in the production of modified expressions across conditions in the ENTRAINMENT trials (speakers had more of an opportunity to be primed by their own modified REs in the pair conditions), we examined this behavior relative to the “priming potential”. Thus, we asked how much of the priming potential was fulfilled, by examining the likelihood of priming out of those trials where priming was possible. We find, first, that more of the priming potential is fulfilled when the noun is repeated [cf. 3], but, strikingly, this measure reveals that priming was much less likely in Same-Pair (31%), where the primed form (“*open umbrella*”) could also describe an unmentioned entity from the entrainment trial, compared to Same-Single (64%), where such an object was not seen earlier. This effect cannot be explained by priming alone, and instead shows the need to represent the visual context, even after it is no longer visually available and the relevant memories possibly fade with time.

Exp. 2 was designed to further explore this effect while minimizing priming. We exploited the fact that the intermediate object in a set of three is called “*medium*” (pilot: 94%), but the same object would be called “*big(ger)*” when paired with just one object (pilot: 97%). Here (i) the TEST contained either a PAIR of objects or a SINGLE object – participants always described the object of intermediate size, and (ii) the ENTRAINMENT trial either completed the set of 3 (Critical), or had one less object (Baseline). The effect of the historical context was observed: the likelihood of comparatives (e.g., bigger) was higher (72%) when a third object of the same category was seen earlier than when it was not (59%). However, speakers rarely produced “*medium*” in the Critical conditions, revealing that the local physical context takes precedence over the historical context.

Conclusions. We observe a novel effect where an entity can influence the form of a referring expression, even though it is not a potential referent in the current context nor was it mentioned earlier in the discourse. This reveals that speakers do not just represent the language previously uttered, but also aspects of the non-linguistic context that has given rise to their utterance. However, speakers do exhibit rational behavior in that the past context has a weaker influence in shaping current referential forms.

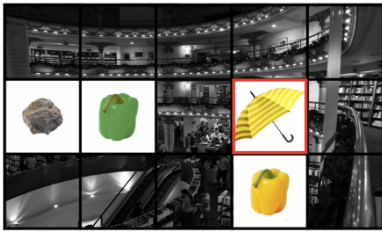
Exp. 1

(a)
Same-noun
Single-history



umbrella: 67%
closed umbrella: 33%

TEST TRIAL

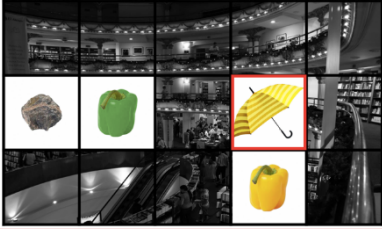


umbrella: 52%
PRIMED
open umbrella: 21%
PRIMING POTENTIAL
→ 21/33 = 63.6%

(b)
Same-noun
Pair-history

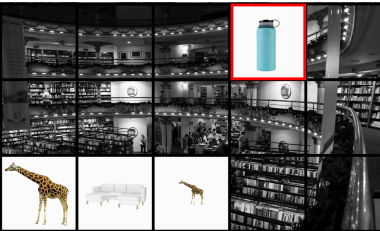


umbrella: 0%
closed umbrella: 100%

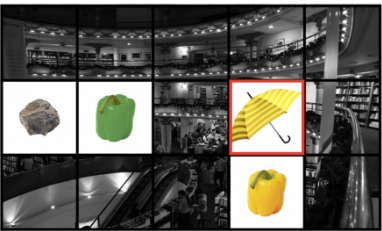


umbrella: 44%
PRIMED
open umbrella: 31%
PRIMING POTENTIAL
→ 31/100 = 31%

(c)
Diff-noun
Single-history

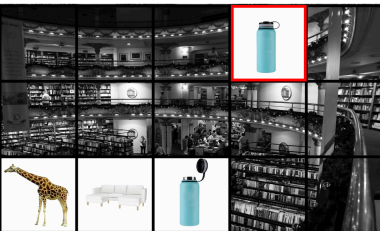


bottle: 80%
closed bottle: 20%

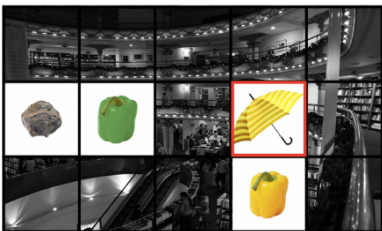


umbrella: 65%
PRIMED
open umbrella: 1%
PRIMING POTENTIAL
→ 1/20 = 5%

(d)
Diff-noun
Pair-history



bottle: 3%
closed bottle: 97%



umbrella: 55%
PRIMED
open umbrella: 14%
PRIMING POTENTIAL
→ 14/97 = 14.4%

Exp. 2

(a)
Pair
critical



flower: 76%
big flower: 6%
bigger flower: 0%
medium flower: 0%



flower: 0%
big flower: 24%
bigger flower: 72%
medium flower: 3%

(b)
Pair
baseline

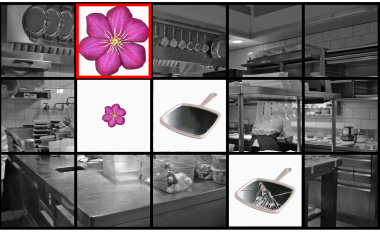


garlic: 77%
big garlic: 6%
bigger garlic: 1%
medium garlic: 0%



flower: 1%
big flower: 35%
bigger flower: 59%
medium flower: 4%

(c)
Single
critical



flower: 0%
big flower: 54%
bigger flower: 40%
medium flower: 0%

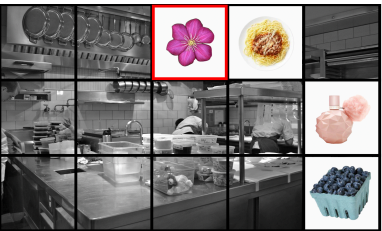


flower: 85%
small flower: 2%
smaller flower: 3%
medium flower: 6%

(d)
Single
baseline



Flower: 83%
big flower: 3%
bigger flower: 0%
medium flower: 0%



flower: 83%
small flower: 2%
smaller flower: 6%
medium flower: 4%

Implicit Causality Can Affect Pronoun Use in Fragment Completion Tasks

Yining Ye, Kathryn C. Weatherford & Jennifer E. Arnold (UNC-Chapel Hill)

An unresolved debate surrounds the question of whether speakers pay attention to predictability when choosing referential expressions. Every act of referring requires speakers to choose between explicit expressions (e.g., the *professor*) or attenuated ones (e.g., *she*). Several production theories suggest that less-explicit forms are used when a word conveys information that is already predictable from the context (e.g., Aylett & Turk, 2004; Tily & Piantadosi, 2009). Yet there is mixed evidence about whether this generalization applies to pronoun production.

Most of this work examines pronoun use in contexts where semantic constraints make one character more predictable – that is, more likely to be re-mentioned in the discourse. For example, in *Amanda amazed John because...*, Amanda is predictable because she is considered the likely cause of John's amazement, and "because" signals an upcoming explanation (e.g., Kehler et al., 2008). By contrast, for verbs like "admire" (e.g., *Amanda admired John because...*), the object (*John*) is the implicit cause. Numerous studies have examined these contexts with a fragment completion task. Results show that people tend to re-mention the implicit cause (i.e., it is predictable), but do not use pronouns more frequently to refer to the implicit cause (Fukumura & Van Gompel, 2010; Kaiser et al., 2011; Kehler et al., 2008; Rohde & Kehler, 2014). Rather, pronoun use is driven by a syntactic bias, where pronouns are used more for the subject than the object.

By contrast, implicit causality did affect pronoun use in a recent study using a different task, where participants memorized facts about different characters, and then filled in the more plausible fact to finish the sentence (Weatherford & Arnold, 2019). This study provided a richer context by introducing a set of characters appearing in all the stories, and with a context sentence for each story, e.g. *The maid and the cook put away the dishes on the top shelves. The cook appreciated the maid because {the maid/she} was tall.* In this task, people did use more pronouns for the implicit cause. This finding is consistent with evidence that semantic biases also guide pronoun use for a different verbtype (Arnold, 2001; Rosa & Arnold, 2017). The conflict between the above findings is critical to resolve, because it bears on a fundamental question about whether predictability affects referential form choices. This raises a question: for fragment-completion task, would adding a richer context be enough to observe an implicit causality effect on pronoun use?

We test this question using Weatherford & Arnold's stories, but in a fragment-completion task. Participants (24 for Exp. 1, 24 for Exp. 2) were introduced to the story setting and 6 characters (3 male, 3 female) with pictures. Then they read fragments (see Fig. 1) and provided a natural ending. Each story included a context sentence and a fragment with an implicit causality verb. For the 12 critical items, we manipulated verbs so that half the time the implicit cause was in subject position, and half in object position. As a control, the subject was first-mentioned in the context sentence half the time. Participants were instructed to begin their continuation with the character we underlined. The target was manipulated within each item, such that each of the two lists had 3 items in each of the four conditions resulting from the 2 (subject vs. non-subject) by 2 (implicit cause vs. non-cause) design. In Exp. 1 the critical stimuli had two same-gendered characters; in Exp. 2 the two characters had different gender. We examined pronoun use for the targets and expected more pronouns when the pronoun was unambiguous in Exp. 2. The critical question is whether implicit causality would increase pronoun use.

Results critically showed that subjects used more pronouns for the implicit cause, but only when the pronoun was ambiguous (Exp. 1) and not in Exp. 2. In both experiments people used pronouns more for the subject. This shows that the predictability effect of implicit causes can be observed in sentence-completion task. However, this effect is fragile. We speculate that our context-rich stimuli encouraged speakers to make inferences about referential predictability, supporting this effect. But even so, when gender made pronouns unambiguous, pronouns became more attractive, which wiped out the subtle effect of semantic bias. Furthermore, participants showed strong individual biases, raising concerns about how the fragment completion task relates to natural language performance.

Exp.1 (Gender-Ambiguous)	Exp.2 (Gender-Unambiguous)
Context: 1. Non-subject & cause continuation: <u>The duke</u> and the butler played pool. 2. Subject & Non-cause continuation: The duke and <u>the butler</u> played pool.	Context: 1. Non-subject & cause continuation: The maid and <u>the duke</u> played pool. 2. Subject & non-cause continuation: The maid and the duke played pool.
Prompt: The butler admired the duke because...	Prompt: The maid admired the duke because...
Sample Response: 1. He/the duke played well. 2. He/the butler could never beat the duke.	Sample Response: 1. He/the duke played well. 2. She/the maid was impressed by the duke.

Figure 1. Examples of experimental stimuli in Exp.1 (left) and Exp.2 (right).

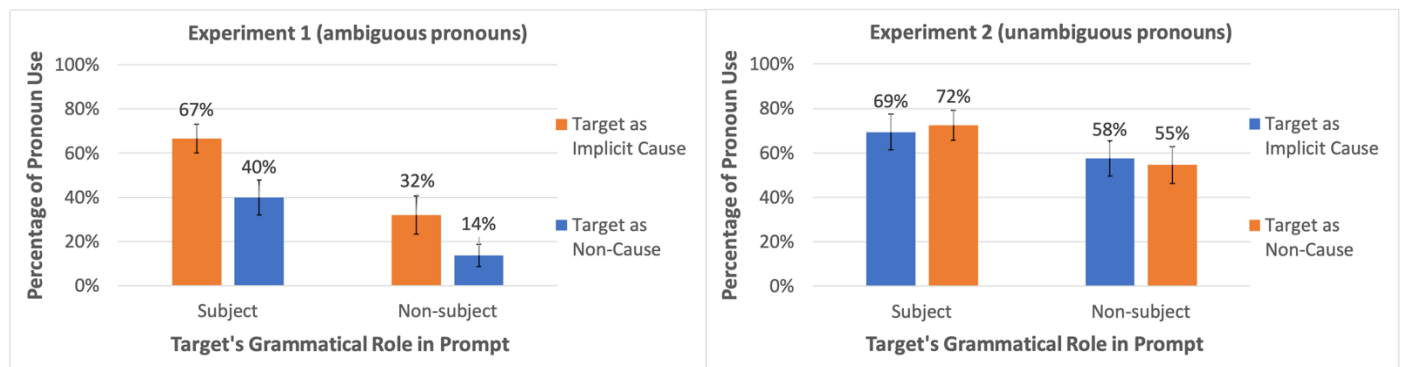


Figure 2. Percentage of pronoun use for the underlined character (target) in prompt as subject vs. non-subject and implicit cause vs. non-cause.

Experiment 1 Solutions for Fixed Effects					Experiment 2 Solutions for Fixed Effects				
Effect	Estimate	Standard Error	t Value	Pr > t	Effect	Estimate	Standard Error	t Value	Pr > t
Intercept	-0.4795	0.3088	-1.55	0.1385	Intercept	0.7664	0.3217	2.38	0.0258
Implicit Causality	0.9161	0.3300	2.78	0.0060	Implicit Causality	-0.2510	0.3327	-0.75	0.4513
Subjecthood	1.4149	0.3261	4.34	<.0001	Subjecthood	0.8493	0.3731	2.28	0.0314
Implicit Causality*Subjecthood	-0.08386	0.8479	-0.10	0.9232	Implicit Causality*Subjecthood	-0.1073	0.6594	-0.16	0.8709

Figure 3. Summary of effects of implicit causality and subjecthood on pronoun production.

References:

- Arnold, J. E. (2001). The effect of thematic roles ... *Discourse processes*, 31, 137-162.
- Aylett, M., & Turk, A. (2004). The smooth ... *Language and speech*, 47, 31-56.
- Fukumura, K., & Van Gompel, R. P. (2010). Choosing anaphoric expressions: ... *Journal of Memory and Language*, 62, 52-66.
- Kaiser, E., Li, D.CH., & Holsinger, E. (2011) Exploring ... In I. Hendrickx, D.S. Lalitha, A. Branco, & R. Mitkov, (Eds.) *Anaphora Processing and Applications*, 171-183.
- Kehler, A., et al. (2008). Coherence ... *Journal of semantics*, 25, 1-44.
- Rohde, H., & Kehler, A. (2014). Grammatical ... *Cognition and Neuroscience*, 29, 912-927.
- Rosa, E. C., & Arnold, J. E. (2017). Predictability ... *Journal of Memory and Language*, 94, 43-60.
- Stevenson, et al. (1994). Thematic roles ... *Language and Cognitive Processes*, 9, 519-548.
- Tily, H., & Piantadosi, S. (2009, July). Refer efficiently ... *Cogsci Proceedings*.
- Weatherford, K., & Arnold, J. E. (under review). Semantic predictability ... Ms., UNC.

Irregular and regular verbs elicit identical morphological decomposition ERPs

Arild Hestvik (University of Delaware), Valerie Shafer and Richard G. Schwartz (CUNY)

After almost two decades of studies examining the predictions of the Dual Route theory of verb inflection, the experimental record still contains contradictory conclusions. Newman et. al., (2007) presented written sentences, such as “Yesterday, I go to the store”. Regular verbs with unexpected tense (“Yesterday, I kick the ball”) elicited a left anterior negativity (LAN) event-related potential (ERP) (suggesting activation of rule computation) followed by P600 (perhaps indicative of repair processes). For irregulars, the authors observed *absence* of a LAN but presence of P600; they interpreted the absence of LAN as evidence that past tense irregulars are computed differently than regulars (look-up instead of rule). Note, however, that the Dual Route predicts that irregulars should generate an N400 effect, which was not observed. In contrast, Stockall & Marantz (2006) report an identical time course and priming pattern for regular and irregular verbs using magnetoencephalography; they argue that irregulars verb are fully decomposed into a root and an abstract tense suffix, in parallel to regular verbs because of the similar response pattern (see also Morris & Stockall (2012)). Here, we reassess the Newman et al. findings with two new experiments that extend the study design and complement subsequent literature. The Dual Route Model predicts that LAN will index morpho-syntactic rule violations and N400 will index lexical access violations, whereas Single Route predicts that regular and irregular violations will both be parsed as rule violations and elicit LAN responses.

ERP-methodology: ERPs were recorded time-locked to verb onsets and offsets using EGI systems and electrode nets while participants judged congruency. Data were re-referenced to the average; averaged ERPs were computed for incongruent and congruent tense for all relevant contrasts. Dimensionality reduction (from high-density scalp electrodes), component isolation and data-driven identification of brain responses were derived via temporo-spatial PCA/ICA on the subtraction (Incongruent-Congruent). Temporal-Spatial Factors were the dependent measures in ANOVAs.

Experiment 1: We replicated Newman et al., but used auditory stimuli, because written present tense form looks like a stem compared to the written past tense form, and thus, auditory stimuli minimize this factor; see Table 1 for the full design. Result: 25 participants’ data (out of 30 tested) showed a LAN response to both regular and irregular verbs, but no N400 was observed. Contra Newman et al., no P600 was observed, but note that neither theory makes specific predictions about P600 that can serve to differentiate between Dual vs. Single Route.

Experiment 2: Another criticism of Newman et al. was that they measured the brain response only to *present* tense verbs, which have no overt inflection signal. In Experiment 2 we controlled for this by replacing “yesterday” with “now”, making the overt inflected past tense form incongruent (see Table 2). Result: 31 (out of 33 tested) participants showed a LAN response to both regular and irregular verbs, similar to Experiment 1.

Conclusions: The results provide evidence that both irregular and regular verbs, when encountered with the “incorrect” tense, triggers LAN, which we interpret as reflecting morphosyntactic violation and re-computation. We also observed that the “LAN” was bilateral, and thus may be better termed “AN”. In addition, the results show that the direction of the tense predicted by the adverb did not matter: Whether present tense or past tense is unexpected, the same brain response for correctness computation is elicited. This provides new support for the basic methodology in Newman et al.’s study. When a listeners encounters a present tense verb when past is expected, this activates the computations required for the correct form, and therefore provide insight into whether irregulars are processed by lexical look-up or rule. The findings support the proposal that irregular verbs have compositional structure (Halle & Marantz, 1994), e.g. [went] is psychologically decomposed and represented as /go/ + [PAST].

Table 1: design of Experiment 1 and 2; adverb tense is between-subject variable.

ADVERB TENSE	VERB TENSE	VERB TYPE	congruency	example stimulus	# of trials
past (Exp 1)	past	irregular	congruent	I ate a sandwich	56
past (Exp 1)	past	irregular	congruent	Yesterday, I ate a sandwich	56
past (Exp 1)	past	regular	congruent	I walked to school	56
past (Exp 1)	past	regular	congruent	Yesterday, I walked to school	56
past (Exp 1)	present	irregular	congruent	I eat a sandwich	56
past (Exp 1)	present	irregular	INCONGRUENT	Yesterday, I eat a sandwich	56
past (Exp 1)	present	regular	congruent	I walk to school	56
past (Exp 1)	present	regular	INCONGRUENT	Yesterday, I walk to school	56
present (Exp 2)	past	irregular	congruent	I ate a sandwich	40
present (Exp 2)	past	irregular	INCONGRUENT	Now, I ate a sandwich	40
present (Exp 2)	past	regular	congruent	I walked to school	40
present (Exp 2)	past	regular	INCONGRUENT	Now, I walked to school	40
present (Exp 2)	present	irregular	congruent	I eat a sandwich	40
present (Exp 2)	present	irregular	congruent	Now, I eat a sandwich	40
present (Exp 2)	present	regular	congruent	I walk to school	40
present (Exp 2)	present	regular	congruent	Now, I walk to school	40

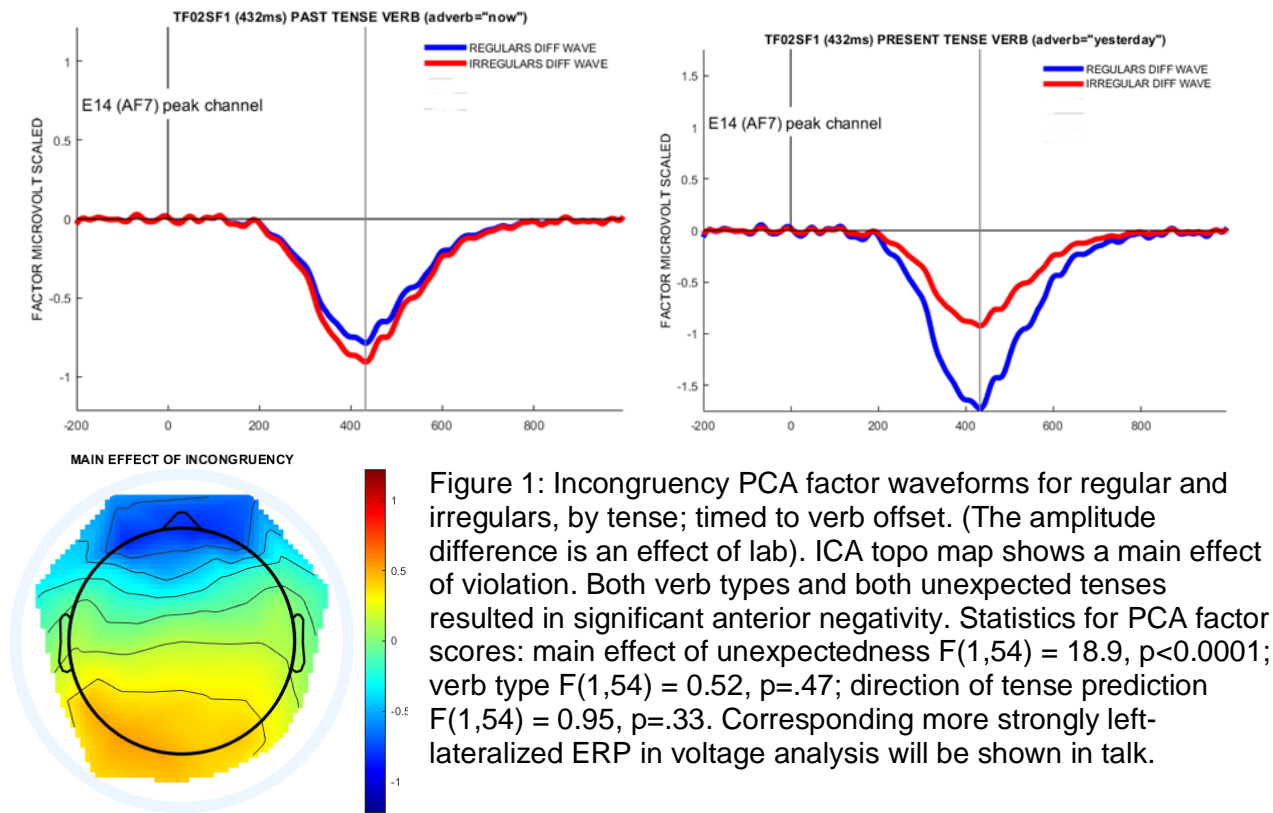


Figure 1: Incongruency PCA factor waveforms for regular and irregulars, by tense; timed to verb offset. (The amplitude difference is an effect of lab). ICA topo map shows a main effect of violation. Both verb types and both unexpected tenses resulted in significant anterior negativity. Statistics for PCA factor scores: main effect of unexpectedness $F(1,54) = 18.9, p < 0.0001$; verb type $F(1,54) = 0.52, p = .47$; direction of tense prediction $F(1,54) = 0.95, p = .33$. Corresponding more strongly left-lateralized ERP in voltage analysis will be shown in talk.

Halle, M., & Marantz, A. (1994). Some key features of Distributed Morphology. In A. Carnie & H. Harley (Eds.), *MIT Working Papers in Linguistics* 21.

Morris, J., & Stockall, L. (2012). Early, equivalent ERP masked priming effects for regular and irregular morphology. *Brain and Language*, 123(2), 81–93.

Newman, A. J., Ullman, M. T., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007). An ERP study of regular and irregular English past tense inflection. *NeuroImage*, 34(1), 435–445.

Stockall, L., & Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *Mental Lexicon*, 1(1), 85–123.

Clefting and prosody affect pronoun processing in dialogue contexts

Abigail Toth (University of Groningen) Liam Blything (University of Alberta), Juhani Järvikivi (University of Alberta), Anja Arnhold (University of Alberta)

Ambiguous personal pronouns in English are typically interpreted as co-referring with the subject and first mentioned referent; however, this interpretive preference is also guided by interactions with multiple discourse and pragmatic cues [1]. Although it is well established that linguistic focus marking can guide listeners' attention and memory for the focused part of the utterance [2], it is unclear whether this is used to help process ambiguous pronouns [3]. Using the visual world eye-tracking paradigm, we investigated the influence of linguistic focusing on both online and offline personal pronoun processing in English spoken dialogues. Linguistic focus was operationalized as prosodic marking additionally in the presence or absence of it-clefts. Crucially, this is the first study to do so whilst providing a felicitous discourse context that served to qualify the contrastive function of linguistic markers, namely to focus a referent relative to presupposed/established information. This reflects real-world use of linguistic focus.

Adults (N=58) listened to 20 spoken dialogues. In the experimental conditions, prosodic focus marking was either applied to the subject or object (8 and 8), with the focused character either being additionally it-clefted or not (Table 1). A fifth broad focus condition was included as a baseline. For all dialogues, Speaker A provided an introduction sentence (1) that named the subject, object, and two distractor characters (all depicted on the screen). Speaker B then asked a question that provided a felicitous context for each of the conditions; (2i) for the subject conditions, (2ii) for the object conditions and (2iii) for the broad focus condition. Speaker A's answer (3) provided the crucial focus sentence and was followed by the target pronoun *he*. With respect to the felicitous context, sentence (2i) for example, sets up a scenario where the new information in (3) is the subject, whereas for (2ii) the new information in (3) is the object.

- (1) *Last month at the meadow I saw a caterpillar, a bee, a spider, and a butterfly.*
- (2i) *Yeah, I heard someone tickled the caterpillar by the flower. Do you know who?*
- (2ii) *Yeah, I heard the caterpillar tickled someone by the flower. Do you know who?*
- (2iii) *Yeah I heard something happened. Do you know what?*
- (3) *The butterfly tickled the caterpillar by the flower. He wanted to lie down in the warm sunshine* (broad focus condition; see Table 1 for each condition).

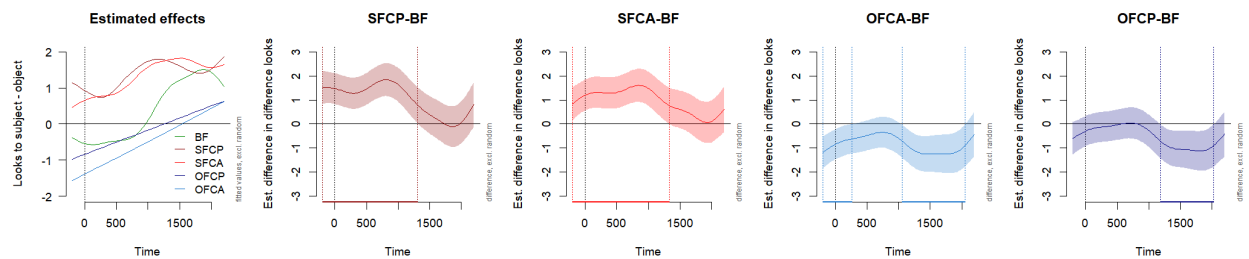
We conducted a GAMM analysis [4] fit to subject advantage looks (looks to subject minus looks to object). Our online data (see Figure 1) revealed that linguistic focusing via prosodic marking enhanced subject advantage in the case of subject focus, and overrode it in the case of object focus, regardless of clefting. As can be seen in the left panel of Figure 1, these focusing effects were present prior to the pronoun (-200 to 400ms). In the case of subject focusing, subject advantage looks further increased upon the processing of the pronoun (at least 400ms onward). In the case of object focusing, subject advantage looks linearly increased across the analysis time window. These findings are in line with previous work showing that focused parts of utterances are boosted in terms of attention and memory representation [2], and further show that rather than a mere additive continuation, these effects combine with constraints specific to the pronoun itself. It should also be noted that the inclusion of object clefts meant that the object was fronted, thereby disentangling subject and first mention preference in English: a subsidiary analysis with the response variable set to first mention advantage supports the presence of both subject and first mention cues, and that preferences are more robust when aligned.

Offline interpretations showed no effects of focus. There was a ceiling preference for the subject in all conditions apart from when the object was fronted by a cleft. This suggests that, while multiple cues are processed, adults may have developed such robust preferences for subjecthood and first mention that these cues dominate in cases of conflict.

Table 1. Test sentences for each condition; focused referents in bold print.

Focus Condition	Example: Speaker A answer (test sentence and pronoun)
Broad focus	The butterfly tickled the caterpillar by the flower. He wanted to lie down in the warm sunshine
Subject focus-cleft absent	The butterfly tickled the caterpillar by the flower. He wanted to lie down in the warm sunshine
Subject focus-cleft present	It was the butterfly that tickled the caterpillar by the flower. He wanted to lie down in the warm sunshine
Object focus-cleft absent	The butterfly tickled the caterpillar by the flower. He wanted to lie down in the warm sunshine
Object focus-cleft present	It was the caterpillar that the butterfly tickled by the flower. He wanted to lie down in the warm sunshine

Figure 1. Visualization of the summed effects derived from the GAMM of fixation patterns, with the random effects set to zero.



Notes. Left panel: Smooth terms for each time by condition term (0ms = pronoun, but effects due to pronoun constraints should be seen from at least 400ms onward). Other panels: Difference plots visualizing the difference between the broad focus condition with each other condition. A positive value indicates that the subject preference was greater relative to the broad focus condition.

References: [1] Arnold et al., (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, 76, B13-B26. [2] Foraker, S., & McElree, B. (2007). The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, 56, 357-383. [3] Cowles et al. (2007). Linguistic and cognitive prominence in anaphor resolution. *Topoi* 26, 3-18. [4]. Van Rij et al. (2015). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.2. <https://cran.r-project.org/package=itsadug>

Comprehension meets production: null/overt subject pronouns in Italian and Spanish

Carla Contemori (University of Texas at El Paso) & Elisa Di Domenico (Università per Stranieri di Perugia)

Although Italian and Spanish are two null-subject languages, they may present distinct discourse-pragmatic biases on the interpretation of anaphoric subject pronouns. The Position of Antecedent Hypothesis (PAH, Carminati, 2002) proposed that null pronouns are interpreted towards antecedents in a prominent syntactic position, while overt pronouns prefer antecedents in lower positions. In Spanish, it is not clear if the PAH can explain null and explicit pronoun interpretation preferences, as the existing evidence is mixed (e.g., Filiaci et al., 2014; Chamorro, 2018). For example, previous comparative research has shown differences in the interpretation of explicit pronouns in Italian and Spanish. However, it is unclear from the existing corpus studies whether differences in the interpretation of anaphora in the two languages may be linked to production patterns. The present study aims at contributing to fill this gap and tests the validity of the PAH in Italian and Spanish by comparing for the first time the two languages on comprehension (Experiment 1) and production (Experiment 2 and 3).

In Experiment 1, we compare the interpretation of overt and null pronouns in Italian and Mexican Spanish, by using an offline sentence comprehension task. We manipulate the type of pronouns (Null vs. Explicit) and the position of the pronoun (anaphoric vs. cataphoric). Thirty-three speakers of Italian and thirty-three speakers of Mexican Spanish interpreted sentences in which null and explicit pronouns are potentially ambiguous (Table 1). Participants answered a three-choice comprehension question, where the possible answers are the subject antecedent (George), the object antecedent (Lewis) and an external referent (someone else). A Logistic Mixed-effects Regression Modeling analysis revealed a clear division of labor between null and overt pronouns in both languages, as demonstrated by the Language Group*Type of Pronoun interactions that emerged in the null pronoun and explicit pronoun analyses (all $p < .0001$). This result suggests that the PAH can explain anaphora resolution biases both in Italian and the variety of Mexican Spanish tested here. In addition, the analysis revealed that: (i) Italian speakers chose the subject interpretation significantly more often for null pronouns than Spanish speakers ($p < .0004$), (ii) Italian speakers chose the object interpretations for explicit pronouns significantly more often than Spanish speakers ($p < .0001$).

With two production tasks, we measured referential choice in controlled discourse contexts, linking the production pattern to the differences observed in comprehension.

In Experiment 2, we adapted a picture-description task used by Arnold & Griffin (2007) to Spanish and Italian. We measured reference to a preceding subject referent when the number and gender of the referents in the pictures was manipulated (Table 2). The results indicated that Spanish speakers produced significantly fewer null subject pronouns and more overt pronouns than the Italian group to refer to subject antecedents, as indicated by main effects of Language (all $p < .01$; no Language*Condition interaction). In Experiment 3, we analyze production biases in Italian and Spanish further, including a comparison of references to subject/object antecedents in contexts of intra-sentential anaphora, using a sentence completion task with implicit causality verbs (Table 3). The results show that Italian-speaking participants produced more null pronouns than Spanish speakers, to refer to subject antecedents (as in Experiment 2) and object antecedents (main effects of Language, all $p < .01$). An object antecedent, thus, appears as a suitable antecedent for a null pronoun in Italian if it is the 'expected' antecedent (Calabrese, 1986) due to the verb implicit causality. Altogether, our study shows micro-variation in Italian and Spanish, with Spanish following the PAH but to a lesser degree than Italian. More specifically, in Spanish the weaker object bias for overt pronouns parallels with a higher use of overt pronouns (and with fewer null pronouns) in contexts of topic maintenance in production. The present study suggests that subtle differences in production patterns are in line with anaphora resolution patterns in comprehension in the two languages.

Table 1. Subject (he=George), object (he=Lewis) and external referent (he=someone else) interpretations in the four conditions of the comprehension study in Italian and Spanish.

	Italian (N=26)			Spanish (N=33)		
Intra-sentential anaphora and cataphora	Subject	Object	External	Subject	Object	External
Anaphora / Null pronoun (1) George saw Lewis when (he) was going to the coffee shop	0.73	0.19	0.05	0.62	0.36	0.015
Anaphora / Explicit Pronoun (2) George saw Lewis when he was going to the coffee shop	0.19	0.76	0.01	0.37	0.59	0.035
Cataphora / Null pronoun (3) When (he) was going to the coffee shop, George saw Lewis	0.86	0.06	0.07	0.64	0.06	0.28
Cataphora / Explicit Pronoun (4) When he was going to the coffee shop, George saw Lewis	0.39	0.38	0.19	0.46	0.11	0.41

Table 2. Proportion of null pronouns, explicit pronouns and full NPs (intra-sentential and inter-sentential) produced by Italian and Spanish speakers in the conditions with one or two referents, with similar or different gender.

	Italian (N=32)			Spanish (N=26)		
	Null Pronoun =(he) was tired	Explicit Pronoun =he was tired	NP=Mickey was tired	Null Pronoun =(he) was tired	Explicit Pronoun =he was tired	NP=Mickey was tired
Context: Mickey went for a walk (with Daisy/Donald) in the hills...						
1 Referent	0.87	0.03	0.10	0.66	0.11	0.24
2 Ref - different gender	0.41	0.06	0.52	0.20	0.10	0.70
2 Ref - gender ambiguous	0.27	0.00	0.73	0.15	0.06	0.79

Table 3. Proportion of (intra-sentential) null/explicit pronouns and NPs produced by Italian (N=24) and Spanish (N=24) speakers in reference to a preceding subject and object referent.

Subject-reference (Mary scared John because....)	Null Pronoun	Explicit Pronoun	NP
Italian	0.98	0.005	0.005
Spanish	0.93	0.06	0
Object-reference (Mary liked John because...)			
Italian	0.85	0.13	0.005
Spanish	0.74	0.25	0.003

Cross-linguistic patterns in Person systems reflect efficient coding

Mora Maldonado,*¹ Noga Zaslavsky,*² and Jennifer Culbertson¹

¹Centre for Language Evolution, University of Edinburgh; ²Department of Brain and Cognitive Sciences and Center for Brains Minds and Machines, MIT; * Equal contribution.

Person systems refer to individuals as a function of their conversational role: there is a speaker (e.g., ‘I’), an addressee (e.g., ‘you’), and others (e.g., ‘they’). Like other semantic domains, person systems exhibit constrained cross-language variation (Cysouw, 2003). For example, while many languages express the *you and us* inclusive meaning as a form of first person (1st-inclusive, e.g., ‘we’), Zwicky (1977) observed that no known language expresses that meaning as a form of second person (2nd-inclusive), which suggests an asymmetry in the representation of the speaker and addressee. Current linguistic theories account for this by positing strong grammatical constraints on possible systems (Harbour, 2016). However, a recent study (Maldonado and Culbertson, 2020) challenged this view by showing that the unattested 2nd-inclusive system is learnable in artificial settings, while the unattested 3rd-inclusive system is not. This finding is not explained by the aforementioned theories, leaving open the question of why these cross-linguistic patterns emerge.

Here, we address this open question by testing an alternative, information-theoretic hypothesis (Zaslavsky et al., 2018), which argues that languages efficiently encode meanings into words by optimizing the Information Bottleneck (IB: Tishby et al., 1999) tradeoff between the complexity and accuracy of the lexicon. This approach is grounded in Rate–Distortion theory (RDT: Shannon, 1948), and has gained empirical support in several semantic domains, e.g, color and containers. It is also closely related to other notions of efficiency (Kemp et al., 2018) that are not grounded in RDT but have been applied to domains such as kinship and indefinites (Kemp et al., 2018; Denic et al., 2020), which are qualitatively more similar to person. Therefore, the person domain poses an important test case for the applicability of RDT to the lexicon.

First, we show that the framework of Zaslavsky et al. (2018) allows us to formulate an ‘egocentric’ bias towards a distinct representation of the speaker, and test the proposal that Zwicky’s observation stems from this bias (Maldonado and Culbertson, 2020). Specifically, we derive two compression models: (i) an egocentric model that predicts that languages efficiently encode the domain in the presence of this bias; and (ii) an unbiased model that predicts that languages efficiently encode the domain given that all entities are equally salient. If an egocentric bias shapes person systems, in addition to pressure for efficiency, then the egocentric model should provide a better account of our data and distinguish between attested and unattested systems. For each model, we computed the IB theoretical limit of efficiency, defined by the set of optimal systems for different complexity–accuracy tradeoffs. We also evaluated the tradeoffs attained by ten commonly attested person systems, the two unattested systems mentioned above, and 1,500 hypothetical systems. The results show that the attested systems are near-optimally efficient according to the egocentric model, in contrast to most hypothetical systems. In addition, the egocentric model predicts a substantial efficiency gap between the attested and unattested systems, whereas the unbiased model predicts that the unattested systems are as efficient as attested systems. This suggests that Zwicky’s observation may be explained by functional pressure for efficient coding in the presence of an egocentric bias, and this explanation is consistent with the findings of Maldonado and Culbertson (2020). Finally, an initial analysis of a larger typological dataset (Cysouw, 2003) suggests that our result generalize well across languages.

This work shows that person systems across languages achieve near-optimal compression, providing converging evidence for the applicability of RDT to the lexicon. Furthermore, it suggests a principled way to study how cognitive biases may influence the lexicon, and may explain typological tendencies, such as Zwicky’s observation, which previous theories have struggled to explain.

References

- Cysouw, M. (2003). *The Paradigmatic Structure of Person Marking*. OUP Oxford, Oxford, UK.
- Denic, M., Steinert-Threlkeld, S., and Szymanik, J. (2020). Complexity/informativeness trade-off in the domain of indefinite pronouns. In *Proceedings of the 30th Semantics and Linguistic Theory Conference*.
- Harbour, D. (2016). *Impossible Persons*. Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.
- Kemp, C., Xu, Y., and Regier, T. (2018). Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, 4(1):109–128.
- Maldonado, M. and Culbertson, J. (2020). Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, pages 1–71.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27.
- Tishby, N., Pereira, F. C., and Bialek, W. (1999). The Information Bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.
- Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.
- Zwicky, A. M. (1977). Hierarchies of person. In *Proceedings from the Chicago Linguistic Society*, volume 13, pages 714–733, Chicago, MI. Chicago Linguistic Society.

Prosody modulates subjecthood and linear order effects in German pronoun resolution

Regina Hert, Anja Arnhold, Juhani Järvi­kivi (University of Alberta)

While it is often found that grammatical role is a strong predictor for referentially ambiguous pronouns - subject pronouns typically prefer the subject of the preceding sentence as their antecedent - it has long been debated whether this is best characterized as a subject preference (Frederiksen, 1981) or a first-mention effect (Gernsbacher et al., 1989). In English, both factors are typically conflated due to fixed word order, but languages with flexible word order allow disentangling syntactic structure and linear order (e.g., Järvi­kivi et al., 2005). However, changes in word order often signal variation in information structure, which has also been shown to affect pronoun resolution (e.g., Colonna et al. 2012). In this study, we crossed manipulations of prosodically-marked information structure and word order in the flexible word order language German to tease apart the three factors.

60 university students from the University of Konstanz and the University of Oldenburg took part in this visual world eye-tracking experiment. Participants listened to dialogues where the target pronoun was preceded by a critical sentence in SVO (1a) or OVS (1b) order, with prosody marking the sentence-initial constituent as either a focus or a given topic (and, conversely, assigning the other role to the sentence-final constituent), resulting in four conditions, all well-formed in German. In addition to prosody, dialogue context enforced information structure (full example dialogue in Table 1).

1. (a) SVO: Der Schauspieler (Given Topic/Focus) hat den Koch angerufen
'The actor (NOM) has called the cook (ACC)'
- (b) OVS: Den Koch (Given Topic/Focus) hat der Schauspieler angerufen
'The cook (ACC) has called the actor (NOM)'

After each dialogue, participants answered a question regarding to whom they thought the subject pronoun was referring. We analyzed these offline responses with generalized linear mixed-effects models. The best model showed that both word order and prosody, as well as the interaction between them, were significant. Overall there was a subject preference in all four conditions, which however decreased when the object was focused.

The eye gaze data for the segment with the critical manipulation (Fig. 1a) showed an increase in looks towards the focused referent for both subject and object referents. For the segment with the pronoun (Fig. 1b), there was an increase in looks towards the subject if the subject referent was focused in the preceding sentence (Given Topic + OVS and Focus + SVO). When the object referent was focused in the preceding segment, there were more looks towards the object during the initial part of the pronoun segment, but more looks towards the subject later. Statistical analyses using Generalized Additive Mixed Models confirmed that the differences in looks described here are significant.

These results show that prosody guides visual attention to the focused referent and that prosody and information structure can partially override the subject preference in the interpretation of pronouns when the referent in focus is the object. Nonetheless, the subject preference is stable across conditions. It has to be noted that the subject preference was also always a preference for the agent, since these were not disentangled in the current study. In conclusion, by clearly separating subject- and first-mention effects while controlling for information structure, the present study provides evidence that subjecthood / agentivity outweighs order of mention in German pronoun resolution.

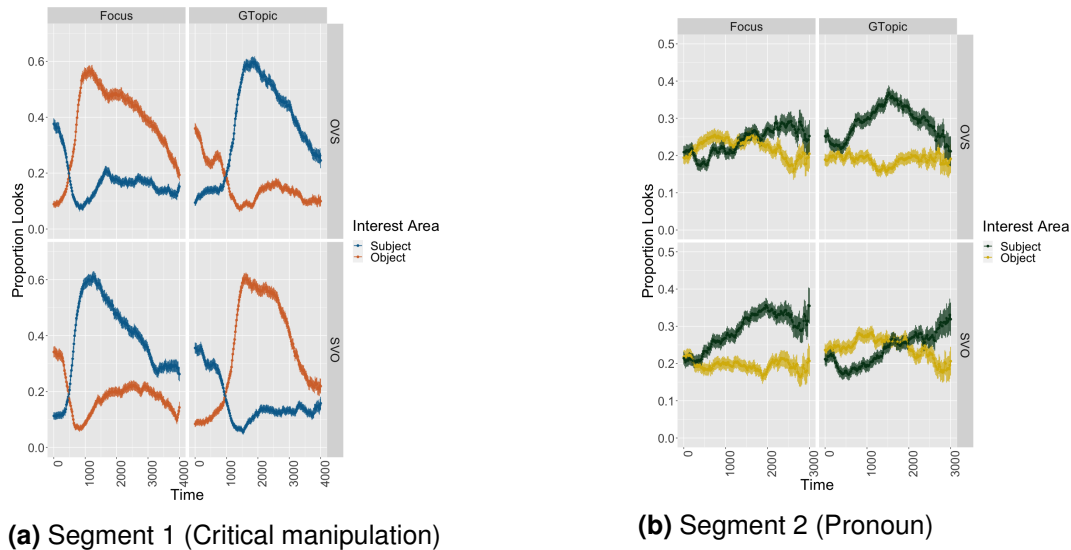


Fig. 1. Proportion of looks by condition for two time segments

Table 1. Example dialogue with critical sentence in SVO word order with either subject question (B1) and focus on the subject in critical sentence or object question (B2) and focus on the object in critical sentence. Critical sentence with manipulation of word order and information structure in italics, pronoun bold.

German	English translation
A: Ich habe gerade Ärger in meiner Strickgruppe, in der auch der Koch, der Schauspieler, der Maurer und der Detektiv sind. Wir haben einen Termin verschoben und ziemlich viel rumtelefoniert. Als letztes hat jemand den Koch angerufen.	A: I have some problems in my knitting group which also includes the cook, the actor, the bricklayer, and detective. We postponed an appointment and called back and forth. Lastly, someone called the cook.
B1: Und wer hat den Koch angerufen?	B1: And who called the cook?
B2: Und wen hat der Schauspieler angerufen?	B2: And who did the actor call?
A: <i>Der Schauspieler hat den Koch angerufen</i> , und zwar mit einem Handy. Er war zu diesem Zeitpunkt schon ziemlich müde.	A: <i>The actor called the cook</i> , namely with a mobile phone. He was already pretty tired at this point.
B: Das ist aber schade.	B: That is too bad.

Frederiksen, J. R. (1981). *Discourse Processes*, 4, 323–347.

Gernsbacher, M. A., Hargreaves, D. J., & Beeman, M. (1989). *Journal of Memory and Language*, 28, 735–755.

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). *Psychological Science*, 16(4), 260–264.

Colonna, S., Schimke, S., & Hemforth, B. (2012). *Linguistics*, 50, 991-1013.

Adaptation to discourse patterns depends on relative frequency of competing structures

Valerie J. Langlois & Jennifer E. Arnold (University of North Carolina – Chapel Hill)

valeriel@live.unc.edu

Comprehenders quickly interpret ambiguous third-person pronouns by following contextual constraints. In *Ana is cleaning up with Liz. She needs the broom*, there is a bias to assign the pronoun to the subject character *Ana* (e.g., Gernsbacher & Hargreaves, 1989; Järvikivi et al., 2005). There is evidence that this bias is modulated by experience, suggesting it may be learned from exposure to the more frequent patterns of pronoun reference. First, individuals with greater print exposure tend to follow the subject bias more consistently (Arnold et al., 2018). Second, exposure within a short (10-minute) experiment modulates interpretation biases. Williams & Arnold (CUNY 2019) exposed readers to stories with unambiguous pronouns that either always referred to the subject or always referred to the nonsubject, and people adapted to this pattern when interpreting ambiguous pronouns (for similar effects see Contemori, 2019; Kaiser, 2009). This demonstrates a causal link between exposure and pronoun comprehension. But Williams & Arnold used exposure sentences that all followed the same structure. Natural language is more variable. Can comprehenders adapt to partially predictive referential patterns? We test this by using Williams & Arnold's task, and manipulate the relative frequency of sentences where the pronoun refers to the non-subject or the subject character.

Our key test items probed interpretation preferences for ambiguous pronouns, e.g. *Liz planted flowers with Ana. She watered the seeds*. Participants answered two comprehension questions, one of which measured pronoun comprehension ("Did Ana water the seeds?" 2AFC: Yes, No). The question always asked about the non-subject character (here, Ana). Thus, responding "No" signals that participants assigned the pronoun to the subject character (here, Liz). We know that people have a "yes" bias with this task, which means that the question format works against the general bias for people to assign the pronoun to the subject character, Liz, and increases variability in responses. Our question was whether this bias would vary as a function of the filler stories, and whether the consistency of the fillers would matter. The fillers were disambiguated by gender and referred to either the subject (e.g. *Liz ate french fries with Matt. She spilled ketchup on the table*) or the non-subject referent (e.g. *Liz ate french fries with Matt. He ...*). To control for previous linguistic experience, we measured print exposure with the Author Recognition Task (Stanovich & West, 1989), where participants selected the authors they knew from a list of real and fake authors.

In each experiment, Mturk participants (100 for Exp. 1; 99 for Exp. 2) read 12 critical, ambiguous sentences and 40 filler sentences. We manipulated the frequency of the fillers referring to the subject and non-subject. Exp. 1 compared the 95-5 condition (95% subject fillers ($n=38$); 5% non-subject fillers ($n=2$)) with the 5-95 condition. Exp. 2 compared the 75-25 (75% subject fillers ($n=30$); 25% non-subject fillers ($n=10$)) and 25-75 conditions. Thus, the proportion of subject to non-subject fillers was more extreme in Exp. 1 than Exp. 2.

Results: In Exp. 1, participants were less likely to select the subject referent when 95% of the fillers had non-subject interpretations (Fig. 1a, $p = .025$). However, this was not the case for Exp. 2. Participants in both the 25-75 & 75-25 condition were equally as likely to interpret the pronoun as the subject referent, even though there was a numeric trend in the expected direction (Fig. 1b). There was an overall main effect of ART in both experiments (Exp. 1: $p < .01$; Exp. 2: $p < .01$), replicating previous findings where participants with higher print exposure were more likely to interpret the pronoun as the subject (see Fig. 2a&b).

Conclusions: Exp. 1 replicated Williams & Arnold (2019), demonstrating that even in a short experiment, people learn to follow the dominant pronoun interpretation pattern. While this adaptation is impressive, Exp. 2 shows that it disappears when the filler items have more than a couple items in the competing structure. This raises questions about how people learn about the frequency of discourse patterns, and whether longer exposures can counteract the kind of variability encountered in natural language.

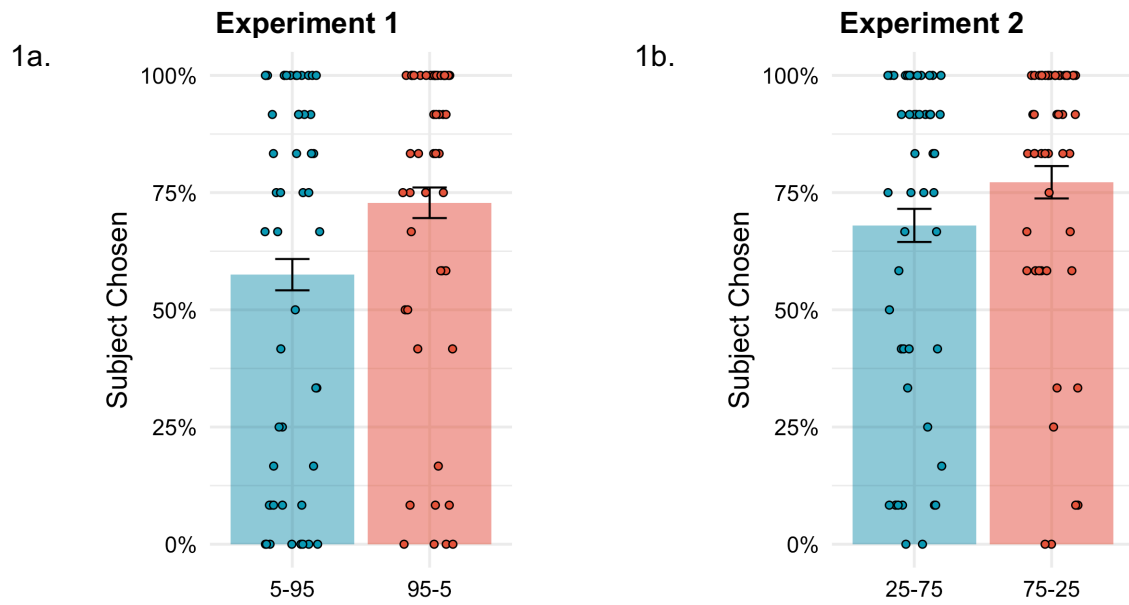


Fig 1a&b. Percentage of subject chosen for the different between-subject conditions (subject % to non-subject %). Each point represents the average subject chosen for a participant within the condition. Error bars represent 95% within-subject CIs.

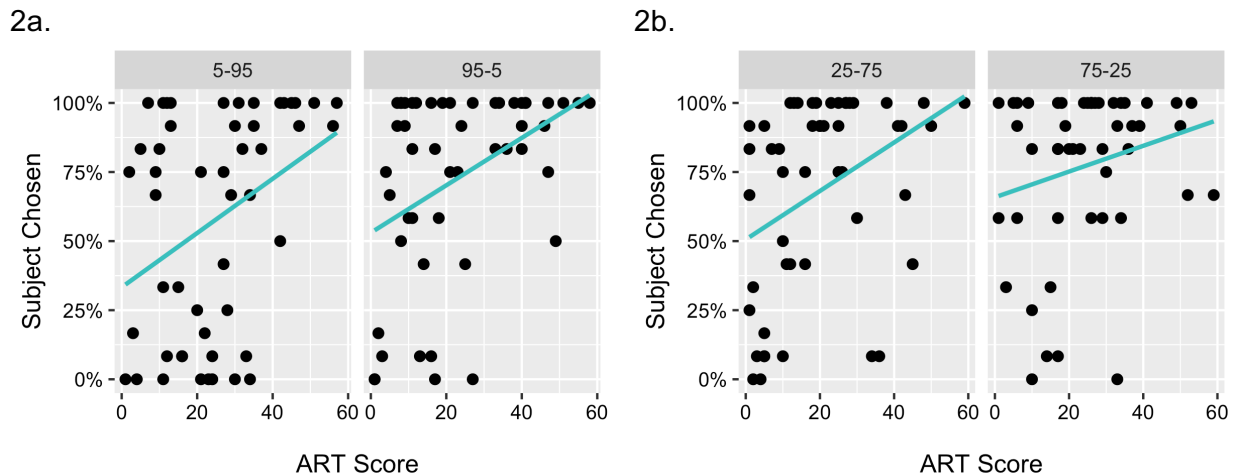


Fig. 2a&b. ART score predicts subject responses across condition and experiment

References:

- Arnold, J. E., Strangmann, I. M., Hwang, H., Zerkle, S., & Nappa, R. (2018). Linguistic experience affects pronoun interpretation. *Journal of Memory and Language*, 102, 41–54.
- Contemori, C. (2019). Changing comprehenders' pronoun interpretations: Immediate and cumulative priming at the discourse level in L2 and native speakers of English. *Second Language Research*.
- Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of memory and language*, 27(6), 699-717.
- Järvikivi, J., Van Gompel, R. P., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first-mention and subject-preference accounts. *Psychological science*, 16(4), 260-264.
- Kaiser, E. (2009). Effects of anaphoric dependencies and semantic representations on pronoun interpretation. In *Discourse Anaphora and Anaphor Resolution Colloquium* (pp. 121-129).
- Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly*, 24(4), 402.
- Williams & Arnold (2019). Priming discourse structure guides pronoun comprehension. CUNY 2019 Poster. University of Colorado Boulder

Are both syntactically and semantically-based pronoun dependencies stored in memory?

Jennifer E. Arnold, Avery Wall, & Taylor Steele (UNC Chapel Hill)

What representations are activated during language use? To answer this core question about the language system, one method is to test whether a structure can be primed. E.g., we know that both syntactic and semantic structures can be primed during language comprehension (e.g., Ziegler & Snedeker, 2018). Here we ask whether comprehension also stores long-distance dependencies, such as referential connections. E.g., in *Biden criticized Trump. He won the election*, do people store the connection between “he” and “Biden”, and use that structure to guide future pronoun processing? If yes, at what level of generalization is this link stored? Priming naturally involves some generalization, because it requires encoding a structure in such a way that it can apply to new instances. Perhaps people remember that a third-person pronoun was used to refer to the subject of the previous sentence (a syntactic generalization). Or perhaps they specifically represent a link between the pronoun and the agent of a judgment verb (a semantic generalization).

We test whether people store a representation of long-distance dependencies between a pronoun and its referent, and whether the type of referent is encoded at a syntactic level, a semantic level, or both. We examine pronoun interpretation in the context of transfer verbs, e.g. *Will took the popcorn from Matt and then he...* (Table 1). People tend to assign an ambiguous pronoun to the subject character (Will), following the well-known subject bias (e.g., Jarviki et al., 2005). But this bias is stronger with the verb “took” than “passed”, revealing a simultaneous bias toward the semantic role of “goal” (Langlois & Arnold, 2020).

We ask whether pronoun interpretation in these contexts is influenced by recent exposure to unambiguous pronouns, and if so, how. For example, in *Matt got the ketchup from Ana and then he...*, Matt is both the subject and the semantic goal of the transfer event. Do people remember this as a link between the pronoun and the prior subject (a syntactic generalization), or as a link between the pronoun and the prior goal (a semantic generalization)?

Methods. Both experiments tested pronoun interpretation in 12 critical stories about a transfer event with two same-gender characters, followed by an ambiguous pronoun (Table 1). Verb type was manipulated: 6 goal-source and 6 source-goal items. A question probed interpretation of the pronoun. As a control manipulation, the question either asked about the first or second character. In a heavy-handed priming manipulation, all 24 fillers had the same unambiguous pronoun structure, half in each verbytype. In Exp. 1 (118 participants), fillers used pronouns that were either subject-linked (Table 2 A&B) or non-subject-linked (Table 2 C&D'). In Exp. 2 (120 participants) filler pronouns were either Goal-linked (Table 2 A&D) or Source-linked (Table 2 B&C). Thus, both experiment used the same materials, but the fillers were re-combined to encourage either a syntactic (Exp. 1) or a semantic generalization (Exp. 2). We asked whether pronoun interpretation would follow the priming pattern of the filler sentences.

Results. Priming modulated results in both experiments (see Fig. 1). Exp. 1 categorized responses in terms of % selection of the subject character; subject selection was higher in the subject-prime than nonsubject-prime condition. Exp. 2 categorized responses in terms of % selection of the goal character; goal selection was higher in the goal-prime than source-prime condition. Verbytype effects revealed that for Exp. 1, there were more subject responses when the subject was the goal than when it was the source; for Exp. 2 there were more goal responses when the goal was the subject than when it was the nonsubject. An effect of question type showed a Yes bias (not pictured in Fig. 1). There were no interactions.

Conclusions. Results provide strong evidence that long-distance dependencies are activated and stored, and people tend to follow recently-encountered patterns when comprehending ambiguous pronouns (see also Author & Author, 2019). People can learn generalizations at both syntactic and semantic levels when recent input is strongly biased toward one level of generalization. Findings point to a role for the statistical frequency of structures at the discourse level in models of language comprehension.

Table 1. Example Ambiguous test item for both Exp. 1 and Exp. 2

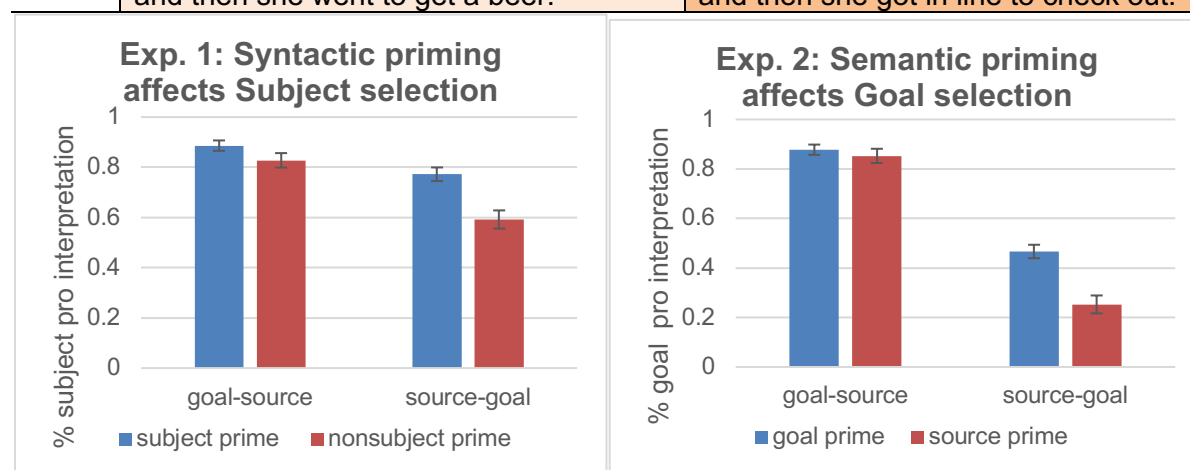
Goal-source verbs	Source-goal verbs
Will and Matt were watching a movie. Will took the popcorn from Matt and then he drank some soda.	Will and Matt were watching a movie. Will passed the popcorn to Matt and then he drank some soda.

Test questions:

- Subject question: Did Will drink some soda? (Yes / No) – Yes signals Subject interpretation
- Nonsubject Q: Did Matt drink some soda? (Yes / No) – No signals Subject interpretation

Table 2. Example priming stories (fillers with unambiguous pronouns)

A. Goal-source verbs Subject/Goal-linked pronoun fillers	B. Source-goal verbs Subject/Source-linked pronoun fillers
Will and Liz were watching TV. Will took the remote from Liz and then he changed the channel.	Will and Liz were grocery shopping. Will gave the credit card to Liz and then he browsed the magazines.
C. Goal-source verbs Nonsub./Source-linked pronoun fillers	D. Source-goal verbs Nonsub./Goal-linked pronoun fillers
Will and Liz were watching TV. Will took the remote from Liz and then she went to get a beer.	Will and Liz were grocery shopping. Will gave the credit card to Liz and then she got in line to check out.



Effect	Experiment 1 (Syntactic priming)			Experiment 2 (Semantic priming)		
	Est. (SE)	t	p	Est. (SE)	t	p
Priming	0.67 (0.23)	2.97	0.0036	0.74 (0.24)	3.12	0.0092
Verbtype	1.07 (0.19)	5.73	<.0001	3.11 (0.31)	10.19	<.0001
Question	1.26 (0.22)	5.66	<.0001	1.4 (0.29)	4.8	<.0001
Verbtype * Question	0.33 (0.34)	0.99	0.3437	0.37 (0.59)	0.63	0.5465
Priming * Question	-0.21 (0.37)	-0.58	0.5655	-0.44 (0.58)	-0.76	0.4592
Priming * Verbtype * Q	0.42 (0.87)	0.48	0.6398	0.4 (0.9)	0.44	0.668

Figure 1. Results from Exp. 1 and Exp 2. Exp. 1 uses Subject selection as the dependent measure; Exp. 2 uses Goal selection as the dependent measure.

References: Author & Author (2019). CUNY poster. ♦ Jarvikivi et al. (2005). Ambiguous pronoun ... *Psych. Science* 16, 260–4. ♦ Langlois & Arnold (2020). Print exposure explains ... *Cognition*, 197, 104155. ♦ Ziegler & Snedeker (2018). How broad... *Cognition* 179, 221-240.

Temporary ambiguity and memory for the context of spoken language use

Kaitlin M. Lord & Sarah Brown-Schmidt (Vanderbilt University)

Spoken language is interpreted incrementally, with listeners considering multiple potential referents as words unfold over time¹⁻². When interpreting an expression like *the yellow banana* in a scene with potential referents, upon hearing *the yellow*, listeners look at objects matching the initial words (yellow banana, yellow candy), and following *banana*, fixate objects matching subsequent words (brown banana), before identifying the referent³. The impact of incremental processing on enduring memory for linguistic experience, however, is poorly understood.

Measures of recognition memory following conversation reveal that speakers and listeners correctly recognize both past referents and contrasting items in the context (e.g. yellow & brown banana when referencing a yellow banana)⁴. Further, listeners form memorial representations of words that were predicted but not actually read⁵. The locus of the memorial boost for items that partially match the unfolding expression is unknown. Modifying words and phrases like *yellow* and *strawberry flavored* activate corresponding referential representations. Yet, the form of the referential phrase may circumscribe an initial set of candidate referents, ruling out items that only match subsequent words (e.g. chocolate flavored cake when hearing strawberry flavored cake). Two experiments test the hypothesis that it is the temporary activation of potential referents that modulates memory for the context of language use, with both early and late competitors encoded in memory better than items that never matched the unfolding phrase. We predict the longer the period of temporary activation, the more likely an item in the context will be remembered. Alternatively, if memory for items in the context is driven by temporary referential activation, items temporarily consistent with the initial part of a phrase will be better remembered than those that are ruled out by the initial words, and only match later words.

In **Exp1** (E1, N=147, mTurk), Ps viewed a series of 6-image grids and heard instructions to click on an image in the grid (**Fig1**). Referring expressions were pre-nominally or post-nominally modified (*Click on the strawberry cake* vs. *Click on the cake that's strawberry flavored*). Grids had a target, a competitor matching the *initial* part of the phrase (early-c), one matching the *latter* part (late-c), two images that did not match but matched one competitor (no-c), and two unrelated fillers. In the pre-nominal condition (*strawberry cake*) the early-c matched early (e.g. strawberry muffin); in the post-nominal condition (*cake that's strawberry flavored*), the early-c matched the noun (e.g. cake that's chocolate flavored), and vice-versa for the late-c. A 2AFC memory test followed: Ps saw an old image (seen in reference task), and a similar, new image, and were asked to click the old image. **Results:** Mixed-effects analysis of 2AFC data (**Fig2**) revealed recall was significantly higher for targets than non-targets ($z = -27.02$), for competitors (early-c & late-c) vs. non-c ($z = 9.70$); and early-c more than late-c ($z = 3.91$). These competitor effects (C vs no-c, and early-c vs. late-c) interacted with utterance form: both were larger with post-nominally modified phrases (z 's > -2.4). One explanation is that the period of temporary activation of competitors was longer for post- vs. pre-nominal modifiers (~1000 vs 800ms).

E2 (N=128, mTurk) added a speech rate manipulation. If the memory boost for competitors in E1 was due to the length of temporary activation, competitors should be better remembered in the slow vs. fast condition. **Results:** Memory (**Fig3**) for targets $>$ non-targets ($z = -20.01$), for competitors $>$ non-competitors ($z = 4.28$), and for early-c $>$ late-c ($z = 5.85$). Overall, memory was better for slow than fast speech ($z = 2.21$). Critically, speed interacted with the competition effect ($z = 3.17$), such that enhanced memory for early vs. late competitors was magnified when speech was slow (this effect was similar for pre/post mod).

Conclusion: Temporary activation of potential referents shapes memory for the context in which language is used. Items that temporarily matched the unfolding expression were better remembered than those that did not, indicating that temporary activation can support context memory. The longer the period of temporary activation, the stronger the boost, particularly for items that were temporary referential candidates. This indicates that both temporary activation, and temporary consideration as a referent improve memory for the context of language use.

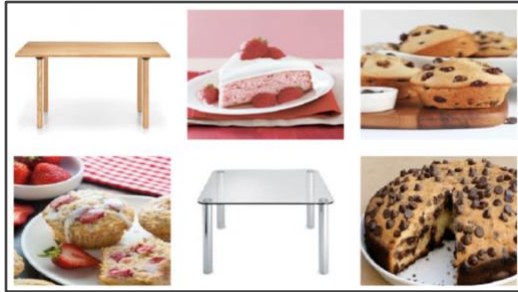
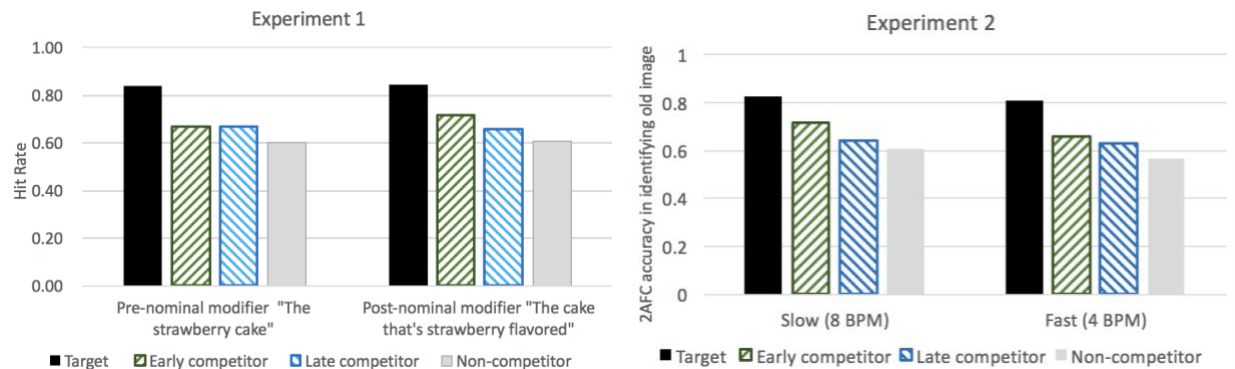


Figure 1: Example stimulus grid for test sentence “Click on the strawberry cake” (pre-nominal condition), and “Click on the cake that’s strawberry flavored” (post-nominal condition).



Figures 2-3: Hit rate (E1) and accuracy (E2) on memory test by condition. For E2, data in figure are collapsed across pre/post-nominal modification. Given the example in Figure 1, for pre-nominal modifiers (*Click on the strawberry cake*), the Target corresponds to memory for the strawberry cake, the Early competitor is the strawberry muffins, the Late competitor is the chocolate cake and the Non-competitor corresponds to the chocolate muffins. For post-nominal modifiers (*Click on the cake that's strawberry flavored*), the Target corresponds to memory for the strawberry cake, the Early competitor is the chocolate cake (i.e. cake that's chocolate), the Late competitor is the strawberry muffins (i.e. muffins that are strawberry flavored) and the Non-competitor corresponds to the chocolate muffins.

References:

- [1] Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C. et al. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *J Psycholinguist Res* 24, 409–436. <https://doi.org/10.1007/BF02143160>
- [2] Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, 38(4), 419-439.
- [3] Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, 32(1), 3-23.
- [4] Yoon, Benjamin, and Brown-Schmidt (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *International Journal of Cognitive Science*.
- [5] Hubbard, R. J., Rommers, J., Jacobs, C. L., & Federmeier, K. D. (2019). Downstream Behavioral and Electrophysiological Consequences of Word Prediction on Recognition Memory. *Frontiers in human neuroscience*, 13, 291. <https://doi.org/10.3389/fnhum.2019.00291>

Investigating suppletion with novel adjectives

Lyn Tieu (Western Sydney University), Nichola Shelton (University of Sydney)

English comparatives and superlatives are typically formed by adding *-er* and *-est* to adjectives, respectively (e.g., *tall-taller-tallest*). Yet there are exceptions involving suppletion (*good-better-best*). Surveying more than 300 languages, Bobaljik (2012) observes the ‘Comparative-Superlative Generalization’ (CSG): if the comparative degree is suppletive (*good-better*), the superlative is also suppletive (*best*), and if the superlative degree is suppletive, then so is the comparative; thus AAA and ABB are possible patterns, but *ABA and *AAB are not. According to Bobaljik, certain types of meaning, including the superlative, cannot be expressed monomorphemically. For this reason, the superlative structurally contains the comparative: [[[Adj]Comp]Sup]. Building on a poverty of the stimulus argument, Bobaljik proposes that the CSG is a linguistic universal. This leads us to predict that people may adhere to the CSG even for forms that they have not encountered before. Indeed, adults have been shown to follow the CSG when producing novel forms (Donegani 2016); but adults have learned suppletive patterns like *good-better-best*. We turn to children, who have considerably less experience with suppletion.


Exp.1 (Tested generalizations: AAA/ABB allowed, ABA disallowed): 48 adults and 21 children ($M=4;04$) were provided with an adjective (e.g., *tazzy*) describing a cartoon alien with a salient gradable property, and a comparative describing another alien with more of the same property (regular *tazzier* [AAX] or suppletive *wimmier* [ABX]); they then had to choose between two superlatives to describe a third alien (*the tazziest/wimmiest*) (Fig.1). Participants received 8 AAX targets and 8 suppletive ABX targets. Logistic regression models revealed **the comparative stem significantly predicted superlative stem choices** (adults: AAX: 99.7% ‘A’ choices, ABX: 93% ‘B’ choices; children: AAX: 68% ‘A’, ABX: 59% ‘B’).

Exp.2 (Comprehension of AAA/ABB): 48 adults and 22 children ($M=4;03$) saw an alien described with a novel adjective (e.g., *tazzy*); they were then presented with additional aliens that had more of the same property and had to choose the ones that matched the novel comparative and superlative (Fig.2). Participants received 8 comparative-first ‘AdjCompSup’ targets and 8 superlative-first ‘AdjSupComp’ targets; half were regular (*tazzy-tazzier-tazziest*) and half involved (potential) suppletion (*tazzy-wimmier-wimmiest*). For both groups, **the interpretation of the novel superlative matched the interpretation of the corresponding comparative, and vice versa**. Adults were at ceiling; logistic regression models on the children’s data revealed comparative choices significantly predicted superlative choices (AdjCompSup: $\chi^2(1)=5.7$, $p<.05$) and vice versa (AdjSupComp: $\chi^2(1)=7.3$, $p<.01$).


Exp.3 (Tested generalization: AAB disallowed): Exp.3 tested whether participants would allow a suppletive superlative following a non-suppletive comparative. The task and materials were the same as in Exp.2 except that participants were provided with adjective-comparative pairs and only had to choose the alien matching the superlative (or were given the adjective-superlative pairs and only had to choose the alien matching the comparative). On AAA and suppletive ABB controls, the 24 adults and 21 children ($M=4;08$) stuck with the original adjectival property; on AAB targets, they switched away from the original property to the second pictured property for the ‘B’ superlative, **reflecting the unavailability of a suppletive AAB pattern** (Condition was significant for both AdjCompSup ($\chi^2(1)=33$, $p<.001$) and AdjSupComp ($\chi^2(1)=41$, $p<.001$), though the difference was bigger for adults (significant interactions, $p<.01$)).

The experiments reveal that 4-year-olds, despite having less experience with suppletive forms than adults, are similarly sensitive to the CSG in their production and comprehension of novel comparatives and superlatives – providing additional support for a universal morphological constraint (Bobaljik 2012).


Example stimulus from Experiment 1 (forced choice task)



This alien has two blue hearts on its body! It is called a tazzy alien!



Non-suppletive target: This alien is even tazzier!
Suppletive target: This alien is even wimmier!



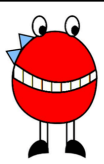
Of all three, this alien is the...

tazziest

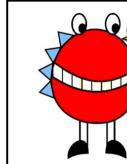
wimmiest

Example stimulus from Experiment 2 (picture selection task)


Comparative trial



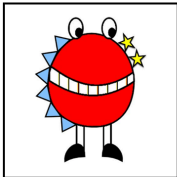
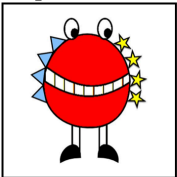
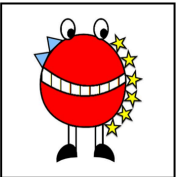
This alien has two blue triangles on its body! It is called a wezzy alien!



Non-suppletion target: Which of these two aliens is wezzier?
Suppletion target: Which of these two aliens is tebbier?

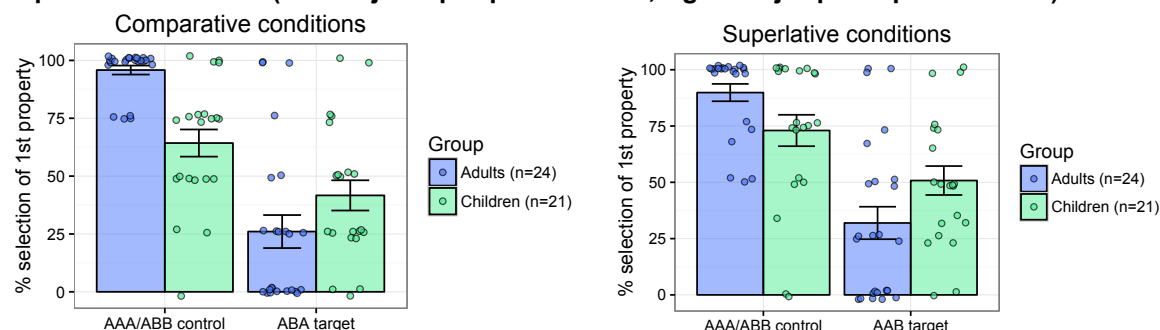


Superlative trial

Non-suppletion target: Which of these three aliens is the wezziest?
Suppletion target: Which of these three aliens is the tebbiest?

Experiment 3 results (left: AdjCompSup conditions; right: AdjSupComp conditions)



References

- Bobaljik, Jonathan David. (2012). *Universals in Comparative Morphology: Suppletion, Superlatives, and the Structure of Words*. Cambridge, MA: The MIT Press.
- Donegani, Josh. (2016). In support of the comparative-superlative generalisation: Experimental evidence from an artificial grammar experiment. Manuscript, University College London.

A Dynamic Tree-Based Item Response Model for Visual World Eye-tracking Data

Sarah Brown-Schmidt¹, Matthew Naveiras¹, Paul De Boeck², Sun-Joo Cho¹ (1-Vanderbilt University; 2-The Ohio State University; KU Leuven, Leuven, Belgium)

In complex scenes, eye gaze is probabilistically directed to different fixation locations, with the likelihood of a fixation to any particular location driven by several competing or complementary cognitive processes. In cases where gaze is in service of performing a task, one of the locations can be considered a task-relevant “target” location (e.g., an object that a person will select), a “competitor” may be similar to the target on some dimension, resulting in potential confusion, and other locations may be “unrelated” to the target and less likely to receive visual attention. We expect that multinomial processing will guide the likelihood of fixating different types of object categories, with one cognitive process increasing the likelihood of fixations to the target and competitor, and a separate process that selects the target and rules out the competitor.

Analysis of binary time-series data considers visual attention to a single interest area, whereas polytomous (e.g., target, competitor, other) time-series data considers visual attention given to several competing options that may be associated with different cognitive processes. The motivation for the present work is a research question for which multiple cognitive processes are assumed to differentially map onto one or more competing response options.

A dynamic generalized linear mixed effect model (GLMM) provides a flexible framework for modeling the heterogeneity and dependencies in observations and allowing the inclusion of trend and serial autocorrelations in intensive binary time series data. Here we present a dynamic tree-based item response (IRTree) model as a novel extension¹ of the dynamic GLMM². Unlike a dynamic GLMM, a dynamic IRTree model is capable of modeling differentiated processes indicated by intensive polytomous time series eye-tracking data. We illustrate a dynamic IRTree model using visual world eye-tracking data. A simulation study resulted in satisfactory parameter recovery and showed that the omission of trend and autocorrelation effects can result in biased estimates and standard errors of experimental condition effects.

We apply the dynamic IRTree model to an empirical dataset³. The motivating example concerns listeners’ interpretation of instructions, e.g., “Click on the small elephant” in scenes containing seven objects, including a small elephant (the Target, T), a small envelope (the Competitor, C), and five other Unrelated objects (Fig1-2). It is assumed that the likelihood of fixating the three object categories is guided by multinomial processing (Fig3): lexico-semantic processing narrows the set of candidate referents to T and C (e.g., small elephant and small envelope). Then, ambiguity resolution processes narrow down the search space, picking out the T (small elephant) over C (small envelope) in one of the experimental conditions. Lexico-semantic information concerns the meaning of words, and in this data set this information differentiates T&C vs. U. Ambiguity between T&C can be resolved using different sources of information, including the speaker’s perspective. To model these multinomial processes, we use a nested design with nested contrasts. The first node in the tree distinguishes objects that match the lexico-semantic information in the unfolding expression vs. those that do not (e.g., small elephant & small envelope vs. everything else). Among the items that match the unfolding expression, the second node in the tree distinguishes the target object from the competitor object (e.g., small elephant vs. small envelope). The dynamic IRTree approach allows us to disentangle complex relationships among different cognitive processes and different factors of interest. For example, it is possible that a given factor has an effect only on the first node of the tree (lexico-semantic processing), but not on the second node (ambiguity resolution), or vice versa. Separate consideration of the distinct cognitive processes involved is possible by a response tree approach, leading to new, more differentiated findings vs. other approaches.

This new method supports differentiation of hypothesized cognitive processes that guide eye-gaze, and testing of distinct predictions regarding the mechanisms driving each process.

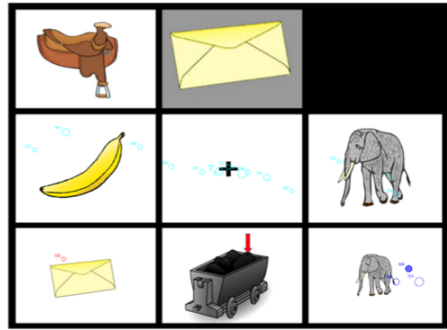


Figure 1. Example display from the empirical dataset³, featuring images of a saddle, envelopes, elephants, banana, and coal (indicated by red arrow). Display shown from the perspective of one participant (P); their partner viewed a similar scene. Images in white visible to both Ps; images in gray visible to only one P (the other P saw a black box in this spot). Ps received instructions about which images they could both see (shared), and which images only they could see (non-shared); this afforded the critical manipulation of visual perspective. Superimposed on the example display are circles corresponding to individual fixations on one trial (dark blue = target; red = competitor; light blue = unrelated objects).

light blue = unrelated objects).

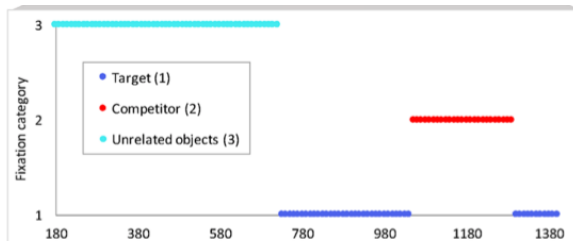


Figure 2. Example gaze data in the time region of interest (180ms after adjective onset in the small elephant) on one example trial, illustrating the polytomous nature of the data with the participant on this trial looking at an “other” unrelated object, then the target, the competitor and back to the target at the very end.

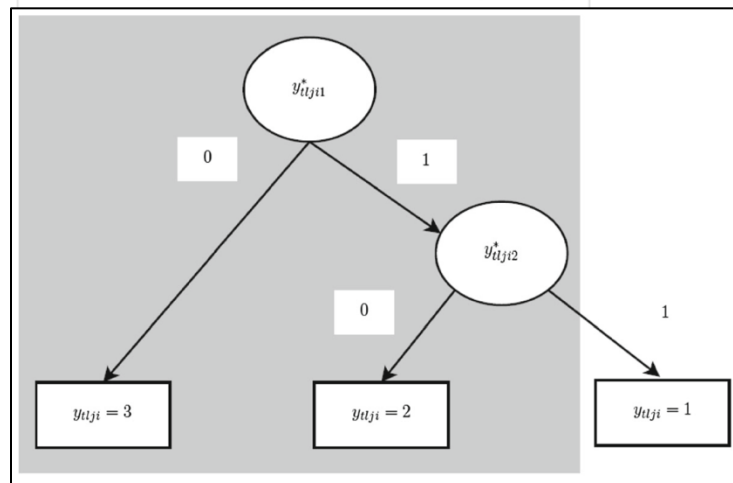


Figure 3. Tree diagram illustrating binary processes (two branches at each node in the tree) at each of two nodes within a three-category paradigm. In the empirical study, Node 1 captures lexico-semantic processing and Node 2 captures ambiguity resolution. Outcome 3 ($y_{t|j|1} = 3$) indicates a fixation to U for a particular timepoint (t), trial (l), participant (j), and item (i). Outcome 2 ($y_{t|j|1} = 2$) indicates a fixation to C at $t|j|i$. Outcome 3 indicates T fixation ($y_{t|j|1} = 1$) at $t|j|i$. At node 1 ($y_{t|j|1}^*$), fixation to U is coded 0, and fixation to either T or C coded 1. At node 2 ($y_{t|j|2}^*$), fixation to C coded 0, and fixation to T coded 1; at node 2,

fixations to U are considered missing at random (MAR⁴).

References

1. Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Shen, J. (2020). Modeling Intensive Polytomous Time Series Eye Tracking Data: A Dynamic Tree- Based Item Response Model. *Psychometrika*, 85, 154–184. [Supplemental Materials](#). [Tutorial](#).
2. Cho, S.-J., Brown-Schmidt, S., & Lee, W.-Y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: an application to intensive binary time-series eye tracking data. *Psychometrika*, 83, 751-771. [Raw data](#), [R code](#), [Model implementation details](#): <https://osf.io/fz9j6/>.
3. Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, 144(5), 898.
4. Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592. <https://doi.org/10.2307/2335739>.

Processing referring expressions: Accessibility is not predictability

Weijie Xu, Ming Xiang (The University of Chicago)

Introduction. The concept of *accessibility* is often assumed to be underlying factor in reference resolution. According to the Givenness Hierarchy (GH) Theory [1], a referent's accessibility in the mental state of a comprehender is encoded in the form of the reference (RF) as part of its lexical semantic representation. In the example in Table 1, therefore, pronouns encode the highest accessibility level, and definite descriptions the lowest. However, the current literature has not reached a consensus on what accessibility exactly means and how to best quantify it. The factors that modulate accessibility, however, show a great extent of overlap with another independently motivated concept of *predictability* [2-6], raising the possibility that the two could be unified. Unlike accessibility, there is a formalized metric of predictability: the likelihood that a given referent is to be mentioned next given the current discourse context. It is theoretically desirable if predictability could serve as the approximation of accessibility. In a self-paced reading study, the current study examines whether the two theoretical constructs are empirically equivalent.

Hypothesis. If accessibility in GH theory is exchangeable with predictability, each RF should encode a certain level of predictability, in the same order as the GH. For example, in Table 1, pronouns encode the highest predictability and definite descriptions the lowest. A plausibility-violation effect is therefore expected when the comprehender encounters a referent whose actual discourse predictability mismatches the predictability implied by the reference form.

Experiment. We evaluated whether each RF in Table 1 encodes a certain level of predictability in the same order as theorized by GH with a self-paced reading experiment. Given the hierarchy, from “the N” to the pronoun, the above mentioned violation effect should be gradually dampened for highly predictable referents and be enhanced for referents that are less predictable, resulting in an interaction effect between predictability and RF.

Method. Native English speakers recruited on Amazon MTurk ($n=112$) read a context passage and then self-paced read a one-sentence continuation, as in (1). We manipulated the form of the target referent in the continuation sentence (as shown in the curly bracket in (1)). Since the experimental materials were adapted from the corpus constructed by [7], the predictability of the target referents measured with a referent cloze game in the original study was available to us.

Results. LMEMs over log RTs were performed for the critical referent region and the spill-over region. The critical fixed effects predictors are the Reference Form (RF) and the Predictability of the referent. The regression model also controls for a number of other effects (see (2)). When comparing each RF in Table 1 with the previous RF on the GH, on neither the critical region nor the spill-over region, did we find step-by-step RF x Predictability interaction from the pronoun to “the N”, indicating that the RFs are not forming a hypothesized “Predictability Hierarchy”. However, in the spill-over region, there is a RF x Predictability interaction when comparing “the N” (Figure 1, Right) to the pronoun ($\beta = 0.188$, $p = 0.018$) and to “that N” ($\beta = 0.169$, $p = 0.034$). This provides some evidence that at least “the N” encodes a different degree of predictability of the referent, distinguishable from other reference forms

Conclusion. While there is no robust support to approximate the Givenness Hierarchy with a “Predictability Hierarchy”, there is some preliminary evidence for a partial correlation between the form of a referent and the predictability of a referent.

- (1) **Sample Experiment Stimuli** (only the continuation sentence was read in the SPR paradigm), critical region in the curly bracket.

Context Passage: Today, in Rich's Kitchen we'll learn about the fine attributes of baking a cake. Since I am not a phenomenal baker we will be assisted by the use of Little Debbie in using one of their fine cake mixes.

Continuation: In order to/ properly make/ {it/this cake/that cake/the cake}/ we/ will/ need/ some vegetable oil/ and/ a couple of eggs.

in focus	>	activated	>	familiar	>	uniquely identifiable
{it}		{this N}		{that N}		{the N}

Table 1: The GH investigated in the current study. The hierarchy is in descending order: the simplex pronoun encodes the highest accessibility level; the proximal “this N” encodes the second highest accessibility level, followed by the distal “that N” and the definite “the N”.

- (2) LMEMs over logRT with the maximal random effects that allow the model to converge.

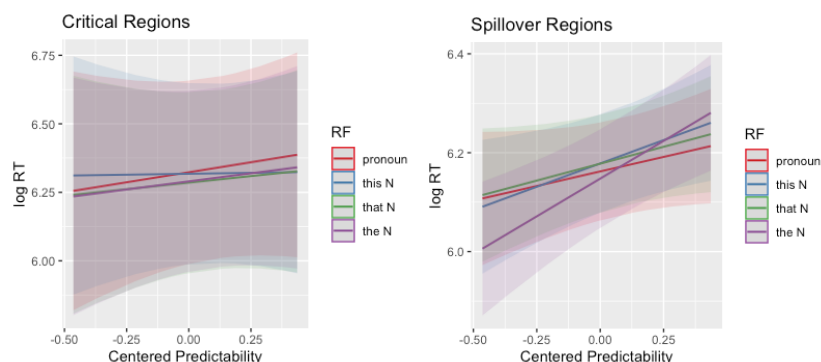
Fixed effects: Predictability * Reference.form + Word.length + Chunk.position + RT.previous + Phi.featured.ref + Recency + Frequency + Intervening.ref + Previous.ref + Gram.role + Previous.gram.role + If.in.SPR

Random effects: *Critical region:* (Predictability|participant) + (1|item)

Spill-over region: (1|participant) + (1|item)

Note: “RT.previous” is the logRT of the previous chunk; “Phi.featured.ref” is the number of referents with the same phi features as the target referent; “Recency” is the distance between the last antecedent and the target referent; “Frequency” is the number of mentions of the target referent in the discourse; “Intervening.ref” is the number of referents between the last antecedent and the target referent; “Previous.ref” is the number of referent appeared so far in the discourse; “Gram.role” is the grammatical role of the target referent; “Previous.gram.role” is the grammatical role of the most recent antecedent; “If.in.SPR” indicates whether the most recent antecedent is in the SPR sentence.

Figure 1. Model predicted interaction between Predictability and Reference Form



References. [1] Gundel et al. (1993) *Language* [2] Ariel (2001) *Text representation, linguistic and psycholinguistic aspect* [3] Kaiser & Trueswell (2008) *Language and Cognitive Processes* [4] Tily & Piantadosi (2009) *Proceedings of the workshop on the production of referring expressions* [5] Arnold (2001) *Discourse Processes* [6] Arnold & Zerkle (2019) *Language, Cognition and Neuroscience* [7] Modi et al. (2017) *Transactions of the Association for Computational Linguistics*

Good-enough for all intensive purposes: Eggcorns and noisy channel processing

Gwendolyn Rehrig (glrehrig@ucdavis.edu) and Fernanda Ferreira (UC Davis)

Linguistic communication occurs over a noisy channel that can distort the signal (Gibson et al., 2013; Levy, 2008); these distortions may be rational such that the distorted interpretation is more plausible to the receiver than the original message (Ferreira, 2003; Ferreira & Lowder, 2016). Psycholinguistics has largely overlooked a form of naturalistic data that may inform language processing models: eggcorns. Eggcorns are misperceptions of a source word or phrase (e.g., *up and coming* → *up incoming*; Liberman, 2003) that become codified in the lexicon—as evidenced by their repeated usage without self-correction—suggesting eggcorns do not register as errors to the speaker. Although eggcorns do not constitute sentences by themselves, they can be multiword sequences (18% of eggcorns in our dataset), and often occur in sentential contexts that help accommodate the mistaken forms. We posit that eggcorns may be ‘good-enough’ representations that approximate the source phrase signal well, can substitute for the source phrase in conversation (eggcorns are usually detected in written form), and may arise from rational language processing (Fig. 1). The current corpus study analyzes the characteristics of attested eggcorns in the context of noisy channel and good-enough language processing.

Method. We scraped 632 unique entries from The Eggcorn Database (Waigl, 2005). Syllables in each source and eggcorn were counted, and the difference in syllable counts was computed. Levenshtein distance between IPA transcriptions of the source phrase and resulting eggcorn approximated phonological similarity. Each pair was automatically transcribed to IPA using the ‘eng_to_ipa’ package in Python. Semantic relatedness and frequency were obtained from ConceptNet 5 (Speer et al., 2017) and COCA (Davies, 2008-), respectively. To assess whether frequent words form eggcorns, the difference in log frequency between the eggcorn and its source phrase was calculated. Pairs with a Levenshtein distance of 0 (misspellings; $N = 146$) and pairs for which the source and eggcorn were not both present in either ConceptNet ($N = 17$) or COCA ($N = 73$) were excluded. The remaining 396 pairs were analyzed.

Results. The number of syllables were equal in the majority of the pairs (93%; $N = 370$); few of the eggcorns either added (4%, $N = 17$) or deleted (2%, $N = 9$) a syllable. Levenshtein distance for most source-eggcorn pairs (58%, $N = 229$) was 1; an additional 31% ($N = 121$) had a Levenshtein distance of 2 (Fig. 2). Semantic relatedness between source and eggcorn was low ($M = 0.23$, $SD = 0.29$), though 8% were synonymous (relatedness = 1), and the difference in log frequency was negative on average ($M = -0.76$, $SD = 3.24$), indicating eggcorns were less frequent than their corresponding source. We conducted an ordered probit regression using Levenshtein distance as the dependent variable to characterize the relationship between phonological similarity, semantic relatedness, and the change in log frequency from source phrase to eggcorn. Larger Levenshtein distance was associated with greater relatedness ($\beta = 1.42$, $t = 3.22$, $p = .001$) and negative changes in frequency such that eggcorns were less frequent than sources ($\beta = -0.10$, $t = -2.52$, $p = 0.01$), and there was a marginal interaction between relatedness and frequency change ($\beta = 0.20$, $t = 1.95$, $p = 0.05$). The results suggest eggcorns tend to closely match the source phrase in sound, but may compromise sound similarity to better fit the context.

Eggcorns overwhelmingly matched the sound of the source signal at the expense of both frequency and semantic similarity. However, when phonological similarity was low, semantic relatedness was higher, suggesting a trade-off when the closest sound match does not fit the context well. We suggest that speech segmentation processes optimize first for similarity to the source signal and second for fit with the surrounding context. These processes operate in a good-enough fashion that is faithful to the input signal most of the time, but occasionally can deviate from the input in principled ways. We suggest that psycholinguists should take eggcorns seriously as naturalistic data points that can inform theories of language processing.

- Davies, M. (2008-). The Corpus of Contemporary American English (COCA): One billion words, 1990-2019. <https://www.english-corpora.org/coca/>
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203.
- Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 217–247). Academic Press.
- Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051–8056.
- Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243.
- Liberman, M. (2003). Egg corns: Folk etymology, malapropism, mondegreen, ??? <http://itre.cis.upenn.edu/~myl/language-log/archives/000018.html>
- Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of AAAI* 31, 4444–4451.
- Waigl, C. (2005). The eggcorn database. <https://eggcorns.lascribe.net/>

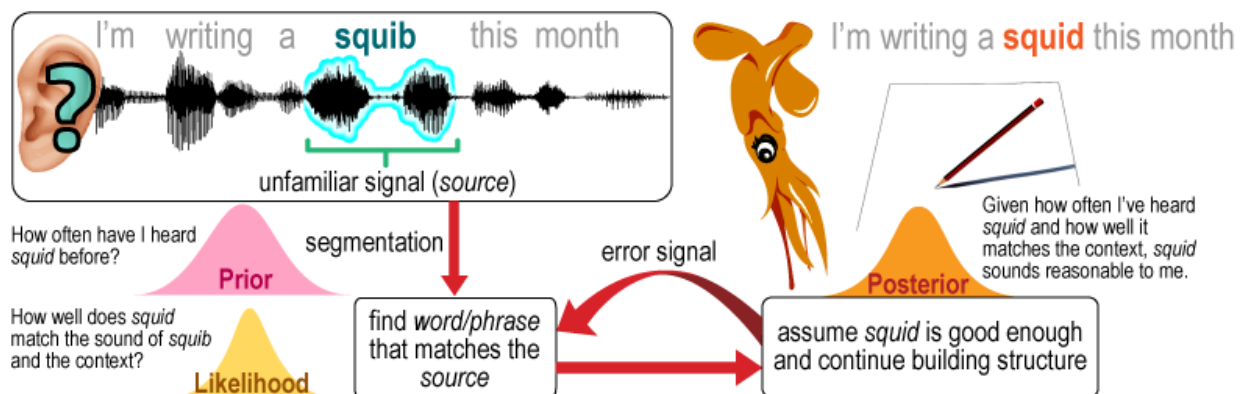


Figure 1. Schematic illustrating how unfamiliar linguistic signal (*squib*) could be misacquired as an eggcorn (*squid*).

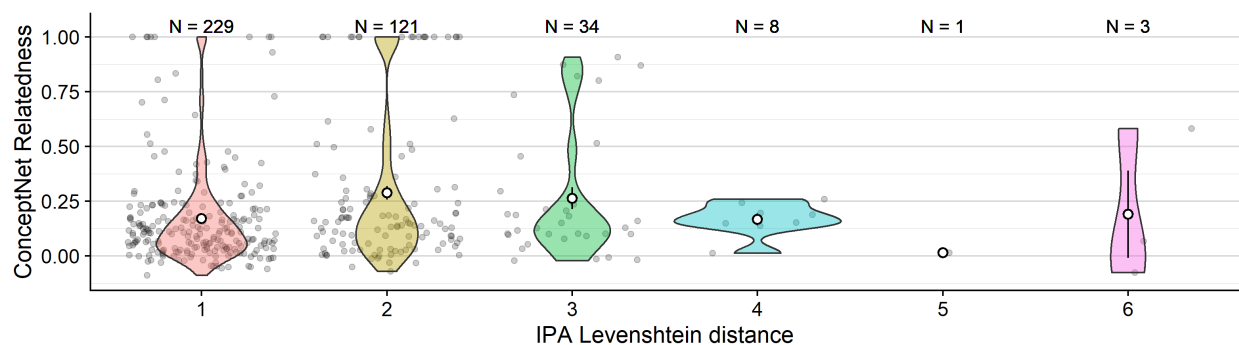


Figure 2. Scatter violin plots showing ConceptNet relatedness values (points, y-axis) plotted against the Levenshtein distance between source and eggcorn IPA transcriptions (violins, x-axis). White points superimposed over each violin plot indicate the mean and standard error.

Systematicity in gesture production, perception may support sign language emergence

Chuck Bradley (Independent scholar)

When communicating in a new medium, like silent gesture, people must adopt novel strategies to ensure successful communication. It has been argued that initial productions are inconsistent and unstructured, with systematicity emerging through interaction and transmission (Motamedi et al., 2019). In support, studies on sign language emergence have shown that homesigners, and signers of young and established sign languages systematically vary handshape to code transitivity in production, but gesturers do not (Brentari et al., 2017). However, perception studies show that non-signers can resolve abstract syntactic-semantic information, like distributivity, telicity, and phi-features (Marshall & Morgan, 2015; Strickland et al., 2015; Schlenker & Chemla, 2018) from gesture and sign on first exposure, suggesting that some aspects of the visual signal are immediately analyzable. Further, the recurrent emergence of handshape as a transitivity marker across unrelated sign languages suggests that this strategy is systematic. To reconcile these disparate findings, we conducted silent gesture production and perception experiments. We modeled handshape to uncover specific visual aspects of the signal that may undergird transitivity categorization.

Methods: We elicited silent gestures from 6 non-signing participants who portrayed 46 unique events involving the manipulation (transitive) or movement (intransitive) of a variety of objects ($6 \times 46 = 276$ gestures). Gestures representing transitive events were considered transitive, otherwise intransitive (*inherent transitivity*). Next, we collected 20 descriptions of the meanings of these gestures from 95 non-signers on Amazon Mechanical Turk (Turkers; $276 \times 20 = 5,520$ sentences; Fig. 1a). Gestures were annotated for 6 handshape features, each linked to transitivity marking in sign languages (Fig. 1b). We then labeled the sentences for transitivity (1='transitive'). A gesture was considered transitive if its proportion of transitive responses was greater than the median proportion of all transitive responses, otherwise intransitive (*perceived transitivity*). We performed two analyses: We trained linear support vector classifiers to predict (1) whether a given gesture is inherently in/transitive and (2) whether it is perceived in/transitive. In each analysis, we used a 6-fold leave-one-out paradigm: The data were randomly split into 6 partitions, trained on 5 of the partitions and tested on the 6th, producing an accuracy score. This was repeated 6 times, such that each partition was the test set once. We computed mean accuracy and compared it against chance using the probability mass function of the binomial distribution. Finally, we averaged the weights for each predictor across all 6 folds in each analysis to assess handshape parameter importance.

Results: Turkers were 91.3% accurate at guessing the transitivity of the gestures (chance=50%, $p < 0.001$). Likewise, classifiers trained on production and perception data were equally good at predicting the inherent and perceived transitivity of the silent gestures: 71.38% and 73.91% accurate, respectively ($p < 0.001$; Figs. 2a,2b). Three handshape features characterized both the production and perception of transitivity distinctions. Further, these features had the same relative weighting: *One- or two-handed* > *Flexion* > Finger Complexity (Fig. 2c).

Interpretation: Non-signers produce transitivity cues that are perceived accurately by other non-signers. This suggests that transitivity contrasts are more systematic than previously assumed, even in absence of a communicative history (interaction, transmission). Specifically, handshape features predict a significant amount of both the production and perception of transitivity distinctions across a diversity of events, indicating handshape variation as a general strategy for transitivity marking in gesture. Further, the same handshape features are informative in both production and perception, with the same relative weighting, consistent with the high interpretation accuracy observed. We suggest that transitivity contrasts in gesture involve the recruitment of stored representations subserving manual action production and perception (Rumiati et al., 2010), and that these representations may then be repurposed to mark transitivity contrasts in emerging sign languages.

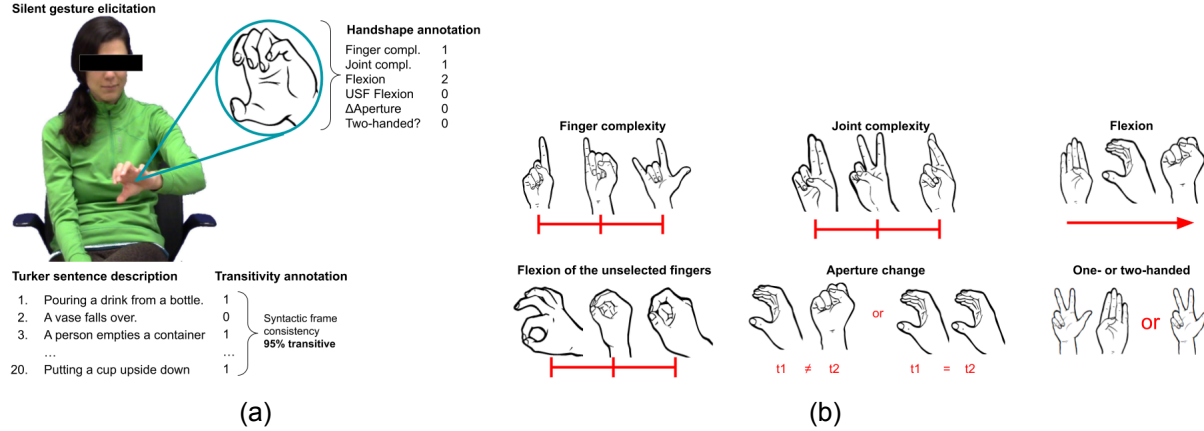


Figure 1: (a) **Experimental design:** An inherently transitive gesture, depicting *Someone put a book on its side*, with Turker response sentences annotated for transitivity. Handshape was annotated for features in (b); (b) **Handshape features:** 'Finger complexity' & 'Joint complexity' = measures of ease of articulation w.r.t. fingers and joints (each scored 1 to 3); 'Flexion' = degree of curvature of the profiled fingers (1 to 6); 'Flexion of unselected fingers (USF flexion)' = degree of curvature of the backgrounded fingers (-1 to 1); 'Aperture change' = whether the hand opens/closes (categorical); 'One- or two-handed' = whether the production involved one or two hands (categorical).

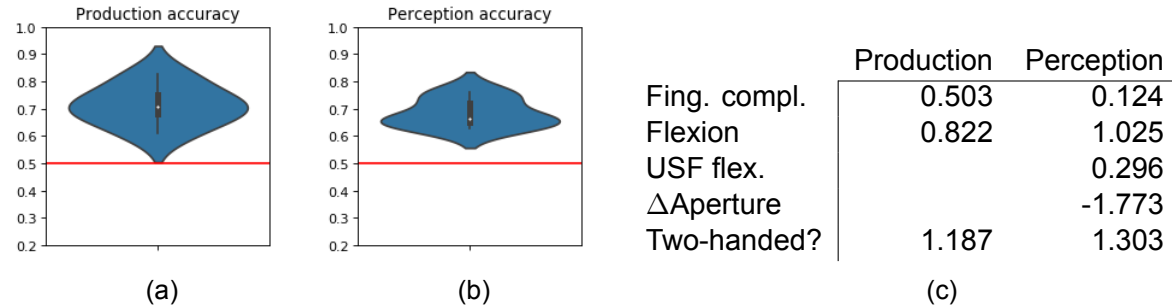


Figure 2: Violin plots showing distribution of classifier accuracies for the production (a) and perception (b) analyses. Red line indicates chance in both. (c) Average model coefficients for the best predictors. Three were most informative for the production analysis, five for the perception analysis. Positive values correspond with 'transitive' items. Some features, like 'Joint complexity' had near-0 weights (uninformative) and were omitted.

References

- Brentari, D., Coppola, M., Cho, P. W., & Senghas, A. (2017). Handshape complexity as a precursor to phonology: variation, emergence, and acquisition. *Language Acquisition*, 24(4), 283–306.
- Marshall, C. R., & Morgan, G. (2015). From gesture to sign language: Conventionalization of classifier constructions by adult hearing learners of BSL. *Top. Cogn. Sci.*, 7(1), 61–80.
- Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, 192, 103964.
- Rumiati, R. I., Papeo, L., & Corradi-Dell'Acqua, C. (2010). Higher-level motor processes. *Annals of the New York Academy of Sciences*, 1191(1), 219–241.
- Schlenker, P., & Chemla, E. (2018). Gestural agreement. *NLLT*, 1–39.
- Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *PNAS*, 112(19), 5968–5973.

The time course of sentence planning and production in two Australian free word order languages

Gabriela Garrido Rodriguez (ARC CoEDL, The University of Melbourne, Australia), Sasha Wilmoth (ARC CoEDL, The University of Melbourne, Australia), Rachel Nordlinger (ARC CoEDL, The University of Melbourne, Australia), & Evan Kidd (ARC CoEDL, The Australian National University, Australia; Max Planck Institute for Psycholinguistics, The Netherlands)

Australian Indigenous languages are well-known for having highly flexible word order, and while this and other properties have been central to debates in linguistics [1,2], there is virtually no psycholinguistic data from these languages. In this paper we present the results from two eye-tracking studies that investigated sentence planning and production in Murrinhpatha (non-Pama-Nyungan, Southern Daly) and Pitjantjatjara (Pama-Nyungan, Western Desert language), two unrelated free word order languages. We ask: (i) what influences the production of different word orders, and (ii) how does speaking a free word order language influence sentence planning?

While both Murrinhpatha and Pitjantjatjara have been described as having flexible word order, they differ significantly on several relevant typological dimensions. Notably, Murrinhpatha is polysynthetic and head-marking, containing only vestigial dependent marking via the optional use of ergative marking in some contexts [3-5]. In contrast, Pitjantjatjara is ergative and dependent-marking, with no verbal agreement morphology [6].

Native speakers of both languages (Murrinhpatha, N=43; Pitjantjatjara, N=49) completed a picture description task while their eye-movements were recorded. Our method closely followed Norcliffe et al. (2015). There were 48 target pictures that depicted two-participant events (e.g., a crocodile biting a man), which were manipulated for agent and patient humanness (+/- human). The target pictures were interspersed amongst 93 filler pictures, which mostly depicted intransitive events. The resulting picture descriptions were transcribed and coded for word order, and participants' eye movements were analyzed using multilevel logistic regression [8-10].

The results show that, consistent with the suggestion that the languages are free word order, participants from both languages produced all possible orderings of S, O and V in the experimental corpus and no word order occurred more than 50% of the time. As in past studies [7, 11-13], differences in word order were sensitive to the different configurations of Agent and Patient humanness. Specifically, the humanness of patients plays an important role in A-initial sentences. In contrast, human agents were more likely to condition P-initial and V-initial sentences, but in interaction with P humanness. Our analyses of the eye-movement data suggest that sentence planning in these languages is best described as a weakly hierarchical process [14, 15], with no evidence to suggest that bottom-up perceptual cues drive word order selection¹. Notably, the results suggest that speaking a free word order language results in a rather different pattern of sentence formulation than in languages with fixed word orders: speakers' gaze was more evenly distributed across the two characters, providing evidence of very early relational encoding during event apprehension that differed across A-initial and P-initial word orders. This result suggests that Murrinhpatha and Pitjantjatjara speakers lay down a very early conceptual representation of the event, which guides their subsequent linguistic encoding and production (see Figure 1). This pattern of early relational encoding is consistent across the two languages, despite their typological differences, although some differences emerged during linguistic encoding which may be attributable to differences between the languages.

Our results suggest that sentence planning is significantly affected by typological properties such as free word order and support the growing body of research revealing significant cross-linguistic differences in sentence production that are linked to grammatical properties of languages.

¹ cf. Gleitman et al., 2007 [16]

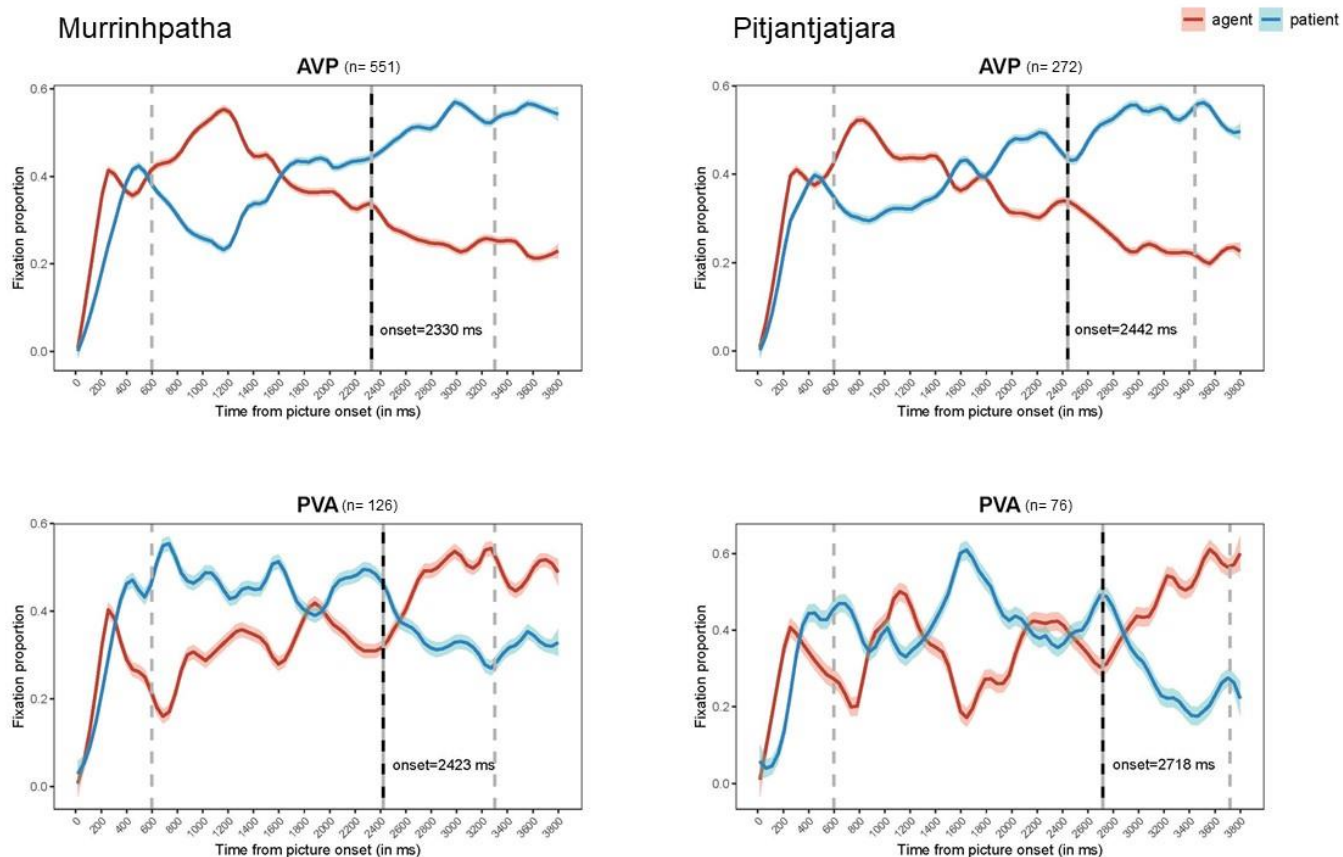


Figure 1. Proportion of agent- and patient- directed fixations in AVP and PVA sentences in Murrinhpatha and Pitjantjatjara after smoothing with LOESS method (span at 0.01). Ribbons indicate standard errors; dashed lines indicate analysis time windows.

References

- [1] Austin, P et al. *Nat Lang Linguistic Th.* 1996, 14, 215-268; [2] Hale, K. *Nat Lang Linguistic Th.* 1983, 1, 5-47; [3] Walsh, M. *AIAS.* 1976). [4] Nordlinger, R. *Morphology*, 2010, 20, 321-341; [5] Mansfield, J., *de Gruyter Mouton*, 2019; [6] Bowe, H.J., *Routledge.* 1990; [7] Norcliffe et al. *Lang Cogn Neurosci*, 2015, 30, 1187-1208; [8] Baayen et al. *J Mem Lang*, 2008, 59, 390-412; [9] Barr, D.J. *J Mem Lang*, 2008, 59, 457-474; [10] Jaeger, T. *J Mem Lang*, 2008, 59, 434-446; [11] Sauppe et al. In *CogSci*, 2013; [12] Ferreira et al. *J Psycholinguist Res*, 2003, 32, 669-692; [13] Christianson et al. *Cognition*, 2005, 98, 105-135; [14] Griffin et al. *Psychol Sci*, 2000, 11, 274-279; [15] Konopka et al. *Cognitive Psychol*, 2014, 73, 1-40; [16] Gleitman et al. *J Mem Lang*, 2007, 57, 544-569.

Underlying clausal structure modulates lexical interference: Evidence from raising and control

Jeremy Doiron and Shota Momma (University of Massachusetts Amherst)

Speaking requires effectively managing the interference between words in an utterance (e.g., Dell et al. 2008). In previous studies, it has been suggested that words within the same clause interfere with each other more than the words across two different clauses because words in the same clause are more likely to be planned simultaneously (Garrett, 1975). Thus, clausal boundaries may limit interference between words. Here we examine how clausal structures that are not necessarily transparent on the surface modulate interference between words in a single utterance to better understand how structural and lexical processes interact in speaking.

We investigated the production of sentences involving Raising-To-Object (RtO) and Object Control (OC) (see Table 1). In RtO, *the donkey* starts off in the underlying subject of the verb *follow*, then raises to the object position of the matrix verb *want*. In comparison, in OC, *the donkey* is not the subject of the verb *follow*, because the subject of the verb *follow* is a null pronoun coreferential to *the donkey* (Postal 1974 a.o; see Polinsky 2013). Therefore, in the underlying structural representations, *the donkey* and *the horse* belong to the same clause in RtO but not in OC. Given the previous finding that nouns in the same clause interfere with each other (Smith & Wheeldon, 2004) more than nouns in different clauses (Garrett, 1975), it is predicted that speakers show more interference between *donkey* and *horse* in RtO sentences than in OC sentences. If this prediction is borne out, it would suggest that sentence planning involves fine-grained structure building processes that distinguish between RtO and OC, and that the non-surface structures of sentences affect the time-course of lexical planning in production, which in turn affect how much words interfere with each other.

We used a sentence-recall task, where speakers ($n = 69$) memorized a sentence presented in RSVP fashion, read aloud 2-4 random verbs, and recalled the sentence upon seeing another (random) verb that was presented in red font (Fig. 1). The random verbs served to inhibit rote memorization thereby encouraging conceptual encoding. Speakers recalled 24 sentences like in Table 1. Our working assumption is that sentence recall involves the regeneration of sentences from conceptual memory (Potter & Lombardi, 1990), and thus it involves the usual processes of grammatical encoding. We measured the duration of the matrix verb of their utterance (e.g., *wanted/taught*), which we predicted to reflect the difficulty of selecting the upcoming noun (e.g., *donkey*). This choice of dependent measure was pre-registered. Because *donkey* and *horse* are semantically related and should interfere with each other (Levelt, 1999), to the extent *donkey* and *horse/man* are planned simultaneously, speakers should show longer duration in the matrix verb production in the related conditions (where the object of the embedded verb is *the horse*) than in the unrelated conditions (where the object of the embedded verb is *the man*). If clauses constitute planning domains such that elements in different clauses are not planned concurrently, then we should observe the interference effect only in the RtO condition and not the OC condition. We fit linear models, with SentenceType and Relatedness as fixed effects and maximal random effect structures that allowed model convergence, and with the number of syllables in the matrix verb as a covariate. The result shows that speakers indeed showed longer matrix verb duration in the related compared to unrelated conditions, but only in the RtO sentences (Fig. 2, interaction $p < .05$; pairwise comparison in the raising condition: $p = .01$), confirming the pre-registered prediction.

The current study suggests that the underlying clausal structures of sentences modulate how much words interfere with each other in a sentence. This in turn suggests that speakers construct syntactic structures detailed enough to distinguish between RtO and OC during planning and use these fine-grained structural representations to control the time-course of lexical access. We thus argue that the temporal dynamics of lexical planning are modulated by underlying syntactic structures even when these structures are not apparent on the surface.

Table 1. Example sentences used in the experiment in each condition. The underlined words are either similar (in the related condition) or dissimilar (in the unrelated condition).

MatrixVerb	Relatedness	Sentence
Raising to Object	Related	The rancher wanted <u>the donkey</u> to follow <u>the horse</u> .
Raising to Object	Unrelated	The rancher wanted <u>the donkey</u> to follow <u>the man</u> .
Object Control	Related	The rancher taught <u>the donkey</u> to follow <u>the horse</u> .
Object Control	Unrelated	The rancher taught <u>the donkey</u> to follow <u>the man</u> .

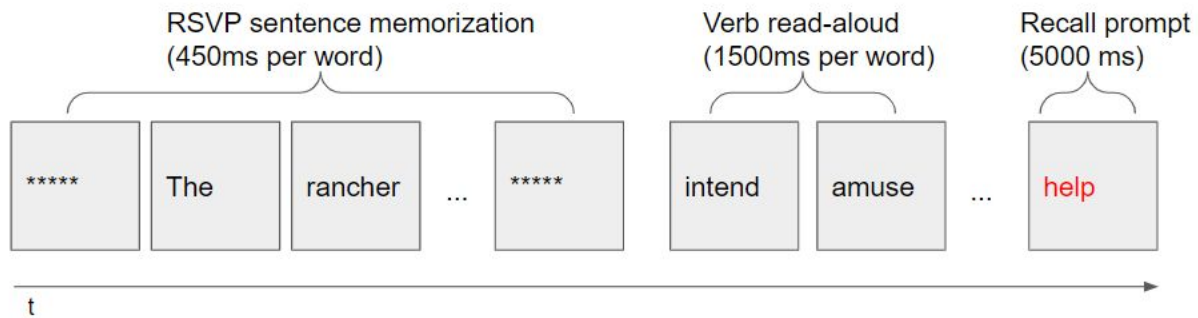


Fig 1. A schematic illustration of the sentence recall task.

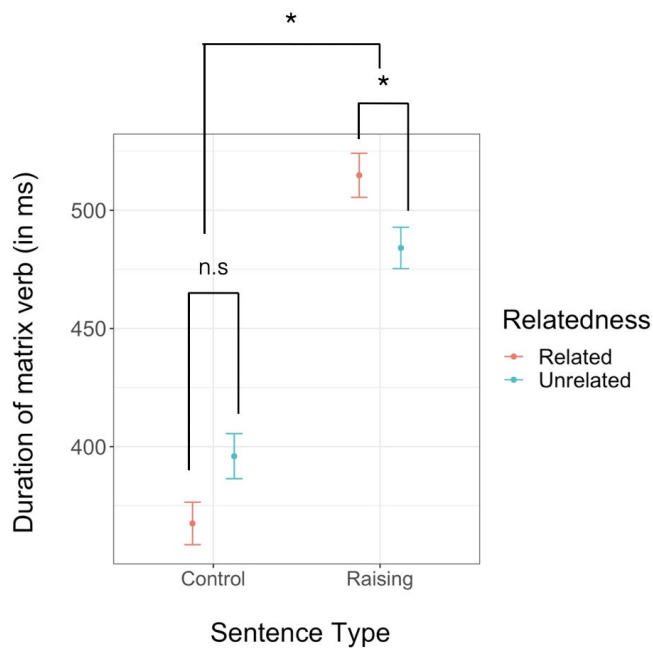


Fig 2. The mean duration of the matrix verb by condition.

Transitioning to online language production: a direct comparison of in-lab and web-based experiments

Margaret Kandel (Harvard), Cassidy Wyatt (UMD), Colin Phillips (UMD)

At a time when much in-person human subjects experimentation has been halted, the ability to collect data from web-based sources is increasingly valuable to language scientists. While some language tasks are already frequently executed online (e.g. self-paced reading, surveys, typed sentence completion; [e.g. 1-3]), there have been fewer web-based studies eliciting recorded speech. The collection and quality of production data may be more susceptible to limitations of online research [cf. 3] than other linguistic data. Variations in internet connections, software, and hardware may make it difficult to collect consistent data or obtain representative participant samples, and recorded speech may be more variable or noisier when elicited and recorded outside of a controlled lab environment. To assess the quality of web-collected production data and how well it can detect phenomena and measure variables of interest to production research, we performed a direct comparison of in-lab and web-based experiments analyzing speech errors and the production time-course of responses. The experiments investigated a robust language production phenomenon: verb agreement attraction (1) [e.g. 4-6].

Method: We used a speeded scene-description task to elicit responses. This task elicits speech through a process that more closely resembles natural production than the traditional preamble paradigm [e.g. 4-6]. Participants were introduced to three aliens (blueys, greenies, pinkies) and described scenes of these aliens *mimming* (lighting their antennae) (Fig 1). Each scene contained two groups of aliens to encourage participants to disambiguate the subject using spatial prepositions (e.g. “the pinky above the greenies”). We manipulated the number of aliens in the scenes so that the NPs in the target SubjPs either matched or mismatched in number (Table 1). 1s was added to the response window of the web experiment to accommodate the online setting and more diverse subject pool. We looked for evidence of attraction in both the distribution of errors and the time-course of error-free sentences (using a forced-aligner; [7]).

Exp 1: The in-lab experiment had 45 participants (34F; $M_{age} = 21$, $SD = 4.5$). We found standard agreement attraction effects, reflected in higher error rates (Fig 2a) and greater probability of producing errors in the mismatch conditions ($p < 0.0001$). Sentences with no errors displayed slowdowns prior to verb articulation in these same environments (Table 2): participants were more likely to pause before the verb ($p < 0.0001$), and these pauses tended to be longer ($p = 0.058$). We saw a plural markedness effect [e.g. 4] on error likelihood ($p < 0.0001$). Singular attraction errors (PS condition) were more common than typically observed in preamble studies [cf. 8], though elevated PS error rates have been seen in other elicitation paradigms [e.g. 9, 10].

Exp 2: The online experiment had 37 participants (26F; $M_{age} = 41$, $SD = 9.97$) recruited from Amazon Mechanical Turk. The experiment was conducted on PCibex Farm [11] and was unsupervised. The audio quality of the responses was sufficient to identify agreement errors and to forced-align. We again found evidence of agreement attraction in error rates (Fig 2b) and probabilities ($p < 0.0001$) in addition to corresponding slowdowns in production time-course (p 's < 0.0001) (Table 2). The distributions of errors and pre-VP delays were comparable to Exp 1, though with fewer errors and more pauses, suggesting a tradeoff between errors and delays in articulation, perhaps due the longer response window. We again observed high PS error rates.

Discussion: The similarities in the results of our experiments indicate that web-based experimentation is a viable and attractive avenue for language production research. Data collection for the in-lab experiment took 3 months to complete, whereas the online experiment took only 9 days of data collection. Using a web-based platform allowed us to recruit a more geographically and age diverse subject sample. We employed several successful measures to minimize drop-out and trial loss and to reduce effects of equipment variation. Nevertheless, there were some differences in our online experiment, with slightly higher participant omission rates and effects of context variability on the forced-aligner's ability to detect utterance onset. We believe that web-based experimentation will allow production research to proceed more flexibly and efficiently and provide easier access to the global population than ever before.

(1) Verb agreement attraction errors occur when nearby material interferes with normal agreement processes, as in the sentence **The key to the cabinets are on the table* [4]

Figure 1: Example scene with target sentence “the pinky above the greenies is mimming”

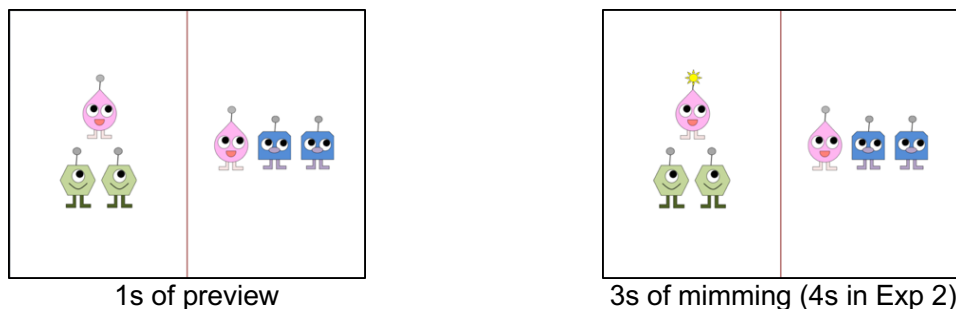


Table 1: Experiment conditions

Condition	Sub-Condition	Sample Sentence
Match	SS	the pinky above the greeny is mimming
Match	PP	the pinkies above the greenies are mimming
Mismatch	SP	the pinky above the greenies is mimming
Mismatch	PS	the pinkies above the greeny are mimming

Figure 2: Participant agreement error rates by sub-condition

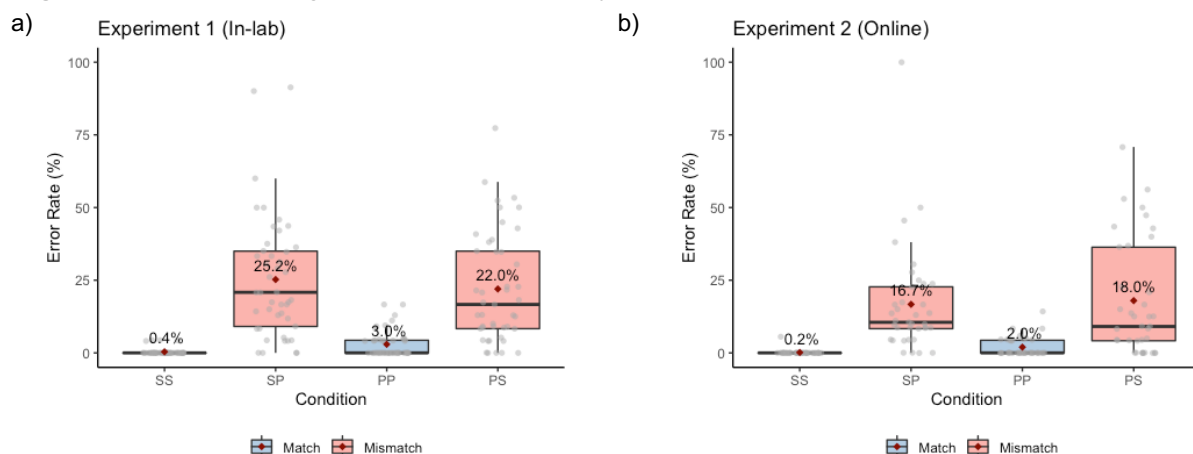


Table 2: Proportion of responses with pre-VP pauses & estimated pause durations

Exp	Condition	Proportion	Duration	Sub-Condition	Proportion
1	Match	0.05	73ms	SS	0.05
				PP	0.06
	Mismatch	0.15	98ms	SP	0.15
				PS	0.18
2	Match	0.12	57ms	SS	0.11
				PP	0.13
	Mismatch	0.31	85ms	SP	0.29
				PS	0.32

References: [1] Corley & Scheepers, 2002; [2] Enochson & Culbertson, 2015, [3] Sitka & Sargis, 2005, [4] Bock & Miller, 1991; [5] Bock & Cutting, 1992; [6] Bock & Eberhard, 1993; [7] McAuliffe et al., 2017; [8] Eberhardt et al., 2005; [9] Staub, 2009; [10] Veenstra et al., 2014; [11] Zehr & Schwartz, 2018

Attribute Saliency and Adjective Order Preferences

Monica L. Do (University of Chicago)

In pre-nominal languages like English, Hungarian, or Dutch, where adjectives linearly precede the noun, “*small blue box*” is vastly preferred over “*blue small box*”. In post-nominal languages like Spanish, Vietnamese, or Hebrew, this same preference emerges, albeit in mirrored form: “*box blue big*” is preferred over “*box big blue*”. These Adjective Ordering Preferences (AOP), while not without exception, are well-attested for a range of adjective classes cross-linguistically. [1] Nevertheless, AOPs continue to pose problems for formal linguistic theories and theories of language processing because – despite numerous accounts [2] – the roots of this apparent universal remain unclear.

We provide evidence from two experiments that AOPs may be rooted in speakers’ conceptual representation of to-be-described objects in non-linguistic cognition. **Exp 1** tested whether the AOP patterns in language would also surface in a memory task. Critically, if AOPs in language and the representation of objects and their attributes in non-linguistic cognition are homologous, then we should find corresponding evidence of AOPs in a fully non-linguistic task. We used a change detection paradigm and manipulated the size, color, shape, and material of novel objects (Fig.1). Participants ($n=134$) examined objects one-by-one, saw a second object, and decided whether that second object was exactly the same as the first. In between the first and second objects, participants performed math problems to block verbal encoding. [3] **Results** (Fig. 2) show a step-wise reduction in saliency that *closely matches the ordering of adjectives observed cross-linguistically*: Participants were statistically worst at detecting changes to size ($\beta=-1.89$, $SE=.28$, $|z|=6.66$), followed by color and shape, though these two did not differ statistically ($\beta=-.26$, $SE=.22$, $|z|=1.19$). Accuracy was highest for material changes ($\beta=-.57$, $SE=.25$, $|z|=2.23$).

In **Exp 2**, we see how well findings from our memory task predict AOPs among native English-speakers ($n=54$). Participants indicated their preference for pairs of Adj-Adj-Noun phrases using a sliding scale (Fig.3). Adjectives (e.g., size, color, shape, material) for the first member of each pair appeared in the order predicted by Exp 1’s memory task (Memory Predicted Order); adjectives for the second member of each pair were inverted (Memory Inverted Order). To minimize typicality and/or frequency of co-occurrence effects, the referents of each string were plausible, but not necessarily prototypical exemplars of the noun entity. Whenever possible, adjectives within each phrase had the same number of syllables. **Results** (Fig.4) showed a main effect of Order Type reflecting a significant preference for Memory Predicted Orders ($\beta=57.91$, $SE=4.56$, $|t|=12.66$). Also, preferences for the Memory Predicted Order were weaker in the Color Shape NP condition than in other conditions ($\beta=-23.42$, $SE=6.68$, $|t|=3.51$); this is in line with the non-significant differences between color and shape conditions found in Exp 1.

In conclusion, we provide initial evidence for an **Attribute Saliency Account** of Adjective Order in multi-adjective strings: Attributes that tend to be more conceptually privileged in speakers’ non-linguistic representations of an entity correspond to adjectives which tend to appear closer to nouns cross-linguistically. This account captured not only the relative order of adjectives, but also which deviations from AOPs would be more permissible than others. These findings also have implications for the distribution of pre- versus post-nominal adjective orders cross-linguistically. Like work from the domain of events showing that entities which are more conceptually salient are privileged syntactically (e.g., Agents tend to be syntactic subjects), we conclude that speakers’ conceptual representations can have direct effects on word order.

References: [1] Dixon, 1982; Hetzron, 1978; 1991; Sproat & Shih, 1991 [2] Sweet, 1898; Wharf, 1945; Sproat & Shih, 1991; Cinque, 1994; Truswell, 1999; Svenonius, 2008; Scontras et al., 2017, 2018 [3] Lakusta & Landau, 2012; Papafragou, 2010

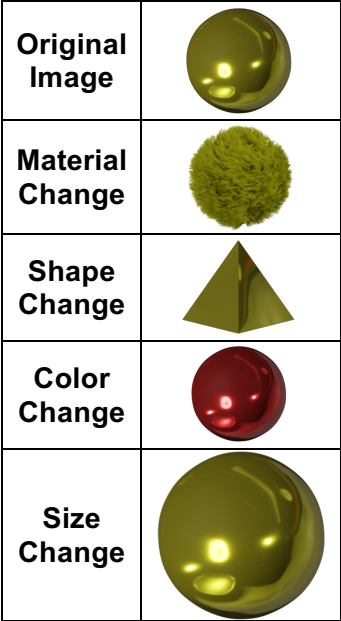


Figure 1 Experiment 1 sample (rescaled) items in each condition of the memory task.

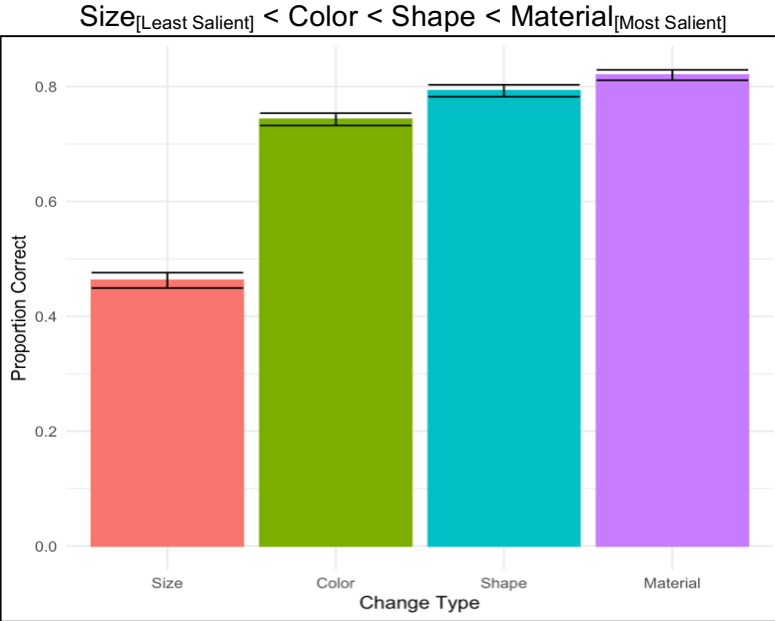


Figure 2 Mean accuracy rates for each change type condition ascending order of accuracy. Error bars indicate +/- 1 standard error.

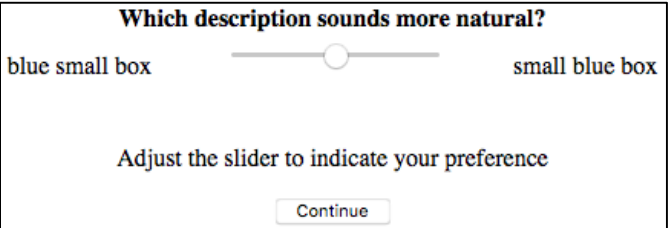


Figure 3 Sample sliding scale task in the Size-Color NP condition of Experiment 2.

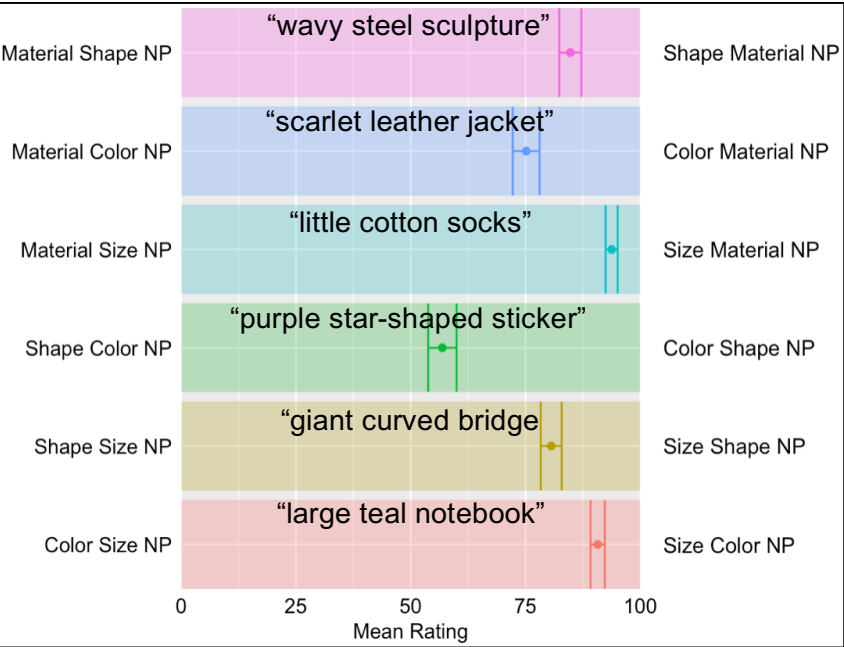


Figure 4 Mean preference scores from Exp 2's sliding scale task. Memory-Predicted Orders are given on right side of scale and Memory-Inverted Orders on the left. (Orders were left-right counterbalanced in experiments.) On this scale, 0 = complete preference for Memory Inverted Orders; 100 = complete preference for Memory Predicted Orders; 50 = no preference. Sample phrases from Memory Predicted Orders are given in middle. Error bars indicate +/- 1 standard error.

Flexibility in Language Production: Insights from Completion of Fragmentary Inputs

Peng Qian, Roger Levy (MIT)

Naturalistic human language production is predominantly left-to-right, but we can generate and predict language in much more flexible ways. Literate speakers use this ability regularly in text editing, which often involves changing part of a sentence while respecting the constraints imposed by the parts left unchanged. In utterance planning, speakers may commit to use a particular word or phrase, forcing them to navigate from a sentence’s beginning to arrive at it successfully. Although this flexible language generation ability is intuitively likely to be closely related to the mechanisms of language comprehension and production studied in psycholinguistics, it has thus far received comparatively little attention in sentence processing research. Here we report initial steps in advancing our understanding of this capability, under the hypothesis that these mechanisms of constrained linguistic generation are scaled-up versions of the same simple computational “motifs” that allow robust processing for degraded inputs (Samuel 1981; Dilley & Pitt, 2010), and follow principles of noisy-channel probabilistic inference (Levy, 2008; Gibson et al., 2013; Keshev et al., 2020).

We focus on the empirical problem of completing fragmentary linguistic input: for example, given an incomplete sentence such as “____ *easy* ____ *problem* ____”, native speakers can quickly come up with reasonable completions for the missing pieces, and can even handle more challenging inputs like “*Vineyards were found scattered throughout* ____ *visited grew any grapes*”. To gain insight into the mechanisms underlying these abilities, we use a reverse-engineering approach, evaluating the quality of a theory by its qualitative and quantitative fit to human behavioral data. We formalize the task of generating completions B from fragments C as Bayesian computation of the posterior $P(B|C)$, assuming a generative model over the space of all possible linguistic utterances. As a concrete instantiation of the “motif hypothesis”, we built a neurally-guided sampling-based inference algorithm, GibbsComplete, consisting of a masked language modelling motif (BERT; Devlin et al., 2018) as the proposal distribution $P(B_i|B_{-i}, C)$ and a next-word language modelling motif (GPT-2; Radford et al., 2019) as the scoring function $\phi(B, C)$, inspired by Wang & Cho (2019). Neither of the computational motifs is optimized for solving the exact target sentence completion task, in contrast to an alternative “fine-tuning hypothesis” of specialized mechanisms for fragmentary input completion, which we implement by tuning pretrained language models (ILM, Donuhue et al., 2020; BART, Lewis et al., 2019; T5, Raffel et al., 2019) to directly predict the completions B conditioned on a neural encoding of the input fragments C .

Our Study 1 evaluates models’ abilities to follow global syntactic context subject to the grammatical constraints, using 26 sets of targeted tests featuring structural reasoning. Here, GibbsComplete’s performance is comparable to fine-tuned models despite no specific training for the task (Figs 1, 2). Study 2 quantitatively compares models’ match to item-level patterns of fragment completion, using 120 stimuli of the form “____ w_1 ____ w_2 ____.” where w_1 and w_2 are single words (40 each Noun–Noun, Adj–Adj, Adj–Noun). We use the syntactic category of the least common ancestor of w_1 and w_2 in parsed completions as a statistic for human–model comparison. Here, GibbsComplete outperforms all the fine-tuned models (Figures 3, 4). These results provide initial support for our “motif” hypothesis, and open the door to new future investigations of how linguistic knowledge can be flexibly deployed by the human mind.

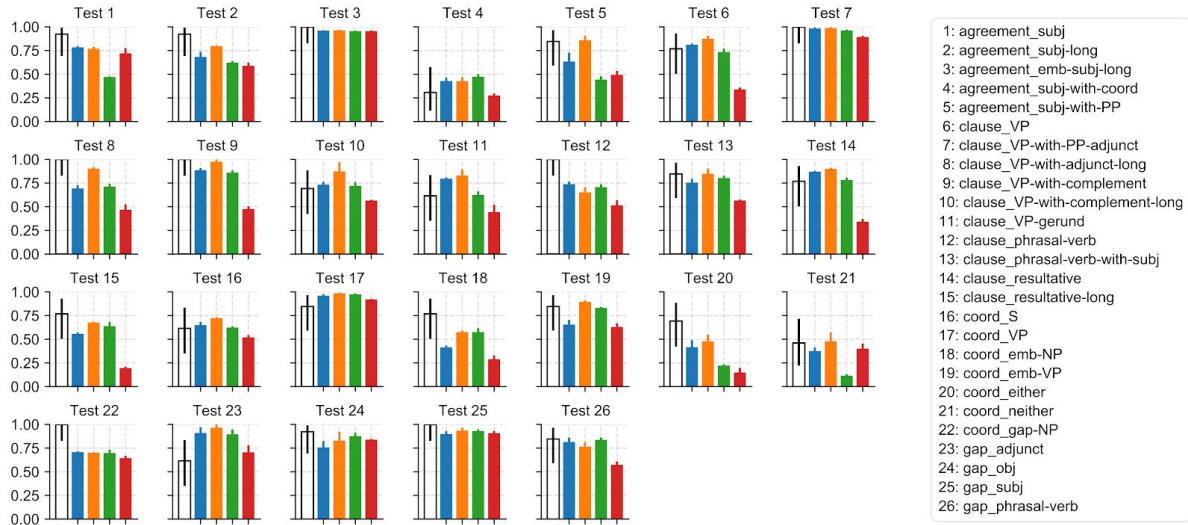


Figure 1: Models' performance in respecting grammatical constraints from fragments (Study 1)

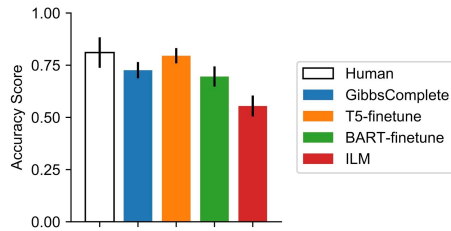


Figure 2: Aggregate performance (Study 1)

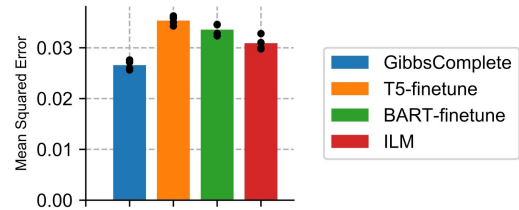


Figure 3: MSE to human completions (Study 2)

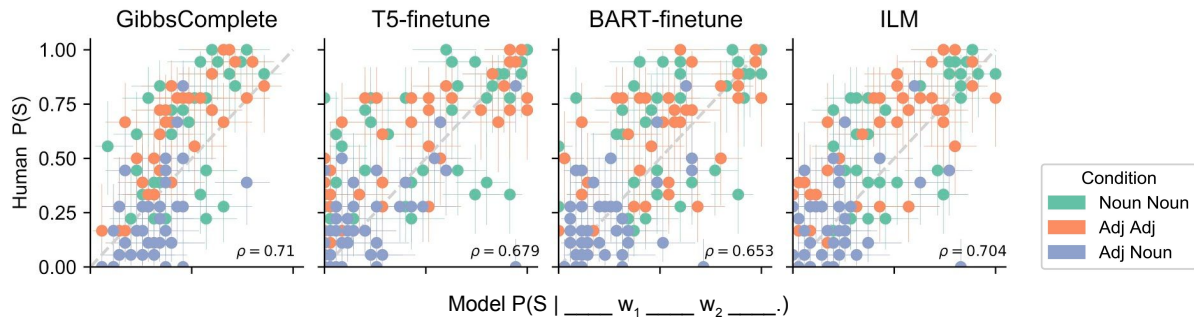


Figure 4: Comparing model output to human completions on the statistic of S as lowest common ancestor

References: Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. • Dilley & Pitt (2010). Altering context speech rate can cause words to appear or disappear. Psychological Science. • Donahue, C., Lee, M., & Liang, P. (2020). Enabling Language Models to Fill in the Blanks. ACL. • Gibson et al. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. PNAS. • Keshev & Meltzer-Asscher. (2020). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. Cognitive Psychology. • Levy (2008). A noisy-channel model of human sentence comprehension under uncertain input. EMNLP. • Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. • Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog. • Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683. • Samuel (1981). Phonemic restoration: insights from a new methodology. JEP: General. • Wang, A., & Cho, K. (2019). BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation.

Lexical activation dynamics and interference in sentence processing: the effect of time

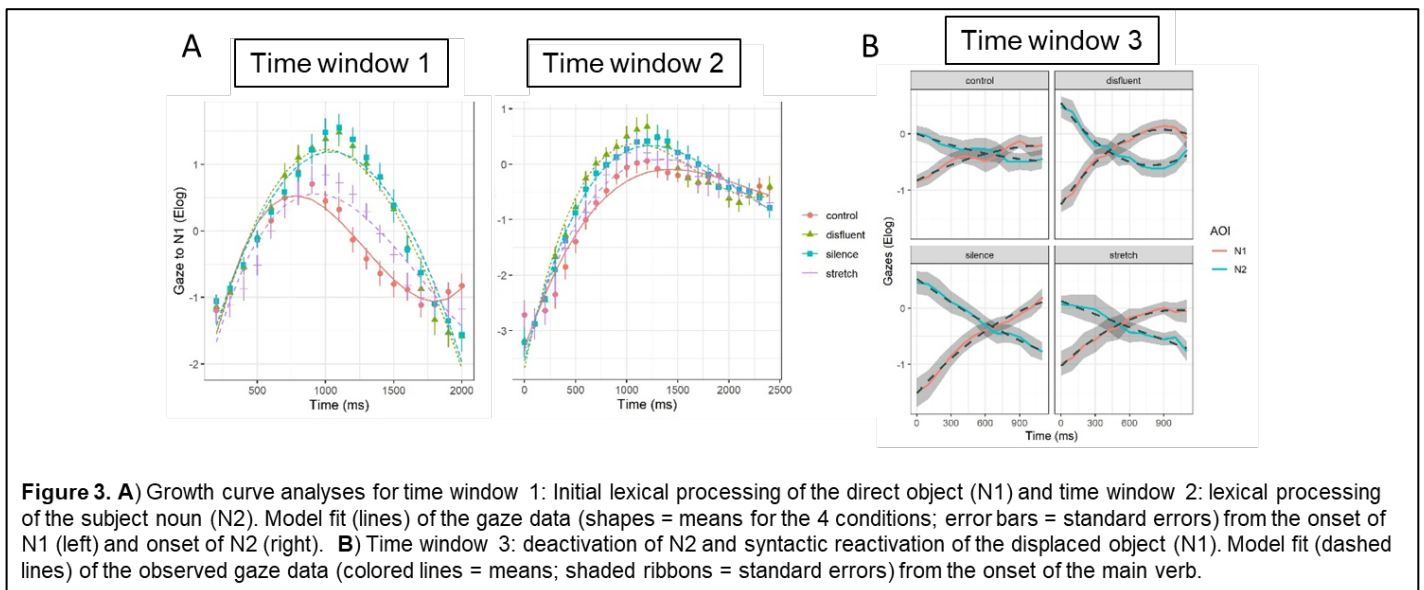
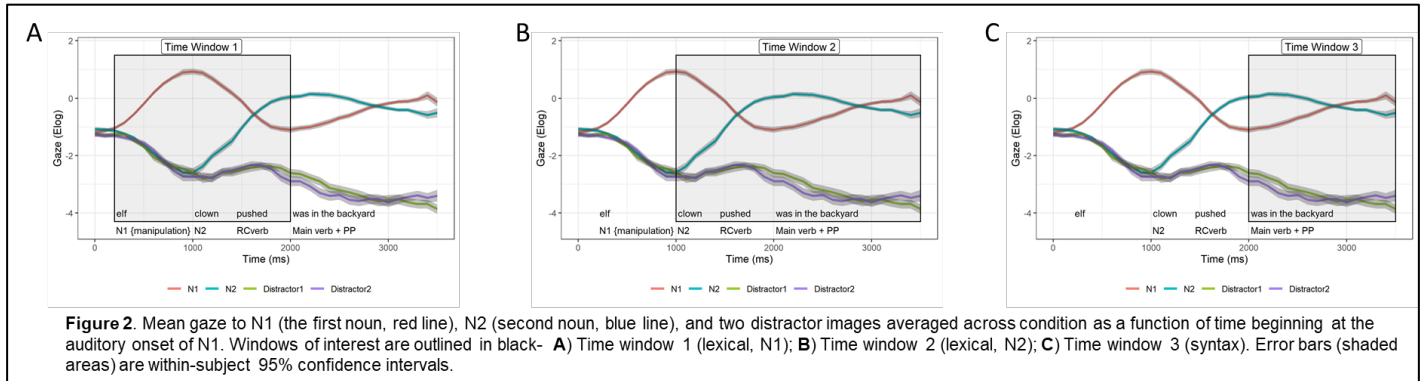
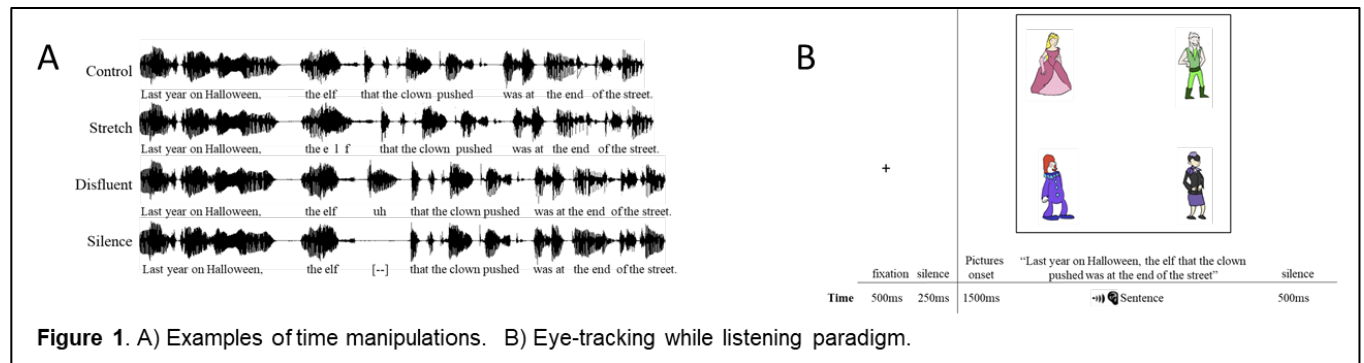
Carolyn Baker and Tracy Love (SDSU/UCSD Joint PhD Program in Language and Communicative Disorders)

Speech is characterized as a transient acoustic signal rapidly unfolding to convey a message. During auditory sentence processing, the listener must analyze, segment, and process the input to build a syntactic structure and arrive at the correct meaning. As sentences become more complex, more demands are placed on the system. Processing object-relative sentence constructions, for example, requires the listener to link non-adjacent but linguistically dependent information within the temporal constraints of the auditory input. This process may be further complicated by interference arising from similarity between competing constituents which has been shown to lead to longer and more costly processing^{1,2}. In this study we ask if manipulating a temporal aspect of speech input affects lexical and syntactic processing (including interference resolution). Here we show that the addition of time via focal rate manipulation (~460ms) after the direct object noun (N1) changes its lexical activation dynamics with a downstream effect on the subsequent subject noun (N2), dependency linking, and interference resolution.

METHODS: Design. We use eye-tracking-while-listening to examine the time course of lexical level processing (activation and deactivation) and dependency linking (reactivation) during the processing of object-relative sentence constructions (see control example in Figure 1A). We explore how manipulating temporal aspects of the direct object noun affects these processes using three manipulations (see Figure 1A). A natural recording served as the control condition with an average rate of speech (4.94 syllables/ second). These sentences served as the base to which the time manipulations were made. The stretch condition was created by increasing the duration of the direct object noun (+260ms). In the disfluent condition, the disfluency *uh* was inserted after the noun (+460ms), and in the silent condition the disfluency was replaced with a silent pause (+460ms). **Procedure.** During the experiment, the participants ($n=24$; $M_{age}=21$, $SD=3.3$) listened to sentences while presented with an array of four pictures on a computer screen (two depict referents in the sentence and two are distractors, Figure 1B). It is hypothesized that increased looks towards the picture of the referent recently processed indicates lexical activation, looks away from a referent indicate deactivation and looks back to the displaced noun after processing the verb indicate reactivation (i.e., syntactic dependency linking)^{3,4,5}. Interference can occur as a result of competition between the subject noun (N2 *clown*) and reactivation of the direct object noun (N1 *elf*) at verb offset. We link this to eye-tracking data as overlapping activation of both nouns during the post-verb portion of the sentence⁴. To ensure attention to each sentence, participants were instructed to respond to a yes/no comprehension question (e.g., “Did the clown push someone?”) at the end of each trial. **Data analysis.** To explore the time-course of lexical activation and the effects of the temporal manipulations, we employed growth curve analyses^{6,7}. Separate analyses were conducted on the three time windows [TW] of interest to explore aspects of sentence processing at hypothesized points: TW1 [lexical] encompassed the full time course of processing N1, TW2 [lexical] captured processing of the N2, and TW3 [syntax] captured reactivation of N1 and resolution of interference post-verb (see Figure 2).

RESULTS: In TW1, the temporal manipulations of disfluencies and silent pauses increased the overall magnitude of activation of N1 and increased the rate of activation and deactivation (see Figure 3A). Similar effects of these manipulations were found on the subsequent noun (N2) in TW2. In TW3, disfluent and silence conditions also resulted in increased rates of reactivation of the direct object (see Figure 3B). Interestingly, all three manipulations enhanced the deactivation of the competing N2 when compared to the control condition and resulted in a more rapid resolution of interference.

CONCLUSION: Additional time modulated the activation dynamics of lexical items and syntactic reactivation, possibly through enhanced lexical focus/attention. We argue that deactivation may play an important, beneficial role by mitigating interference during dependency linking.



References

1. Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3), 285–316.
2. Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2), 157–166. <https://doi.org/10.1016/j.jml.2006.03.007>
3. Akhavan, N., Blumenfeld, H. K., & Love, T. (2020). Auditory Sentence Processing in Bilinguals: The Role of Cognitive Control. *Frontiers in Psychology*, 11.
4. Sekerina, I. A., Campanelli, L., & Van Dyke, J. A. (2016). Using the visual world paradigm to study retrieval interference in spoken language comprehension. *Frontiers in Psychology*, 7(JUN), 1–15.
5. Thompson, C. K., & Choy, J. J. (2009). Pronominal resolution and gap filling in agrammatic aphasia: Evidence from eye movements. *Journal of Psycholinguistic Research*, 38(3), 255–283.
6. Mirman, D. (2017). Growth Curve Analysis and Visualization Using R. In *Statistical Methods in Medical Research* (Vol. 26, Issue 3). Chapman and Hall/CRC.
7. Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4), 475–494.

Accessibility-Based Constraints on Morphosyntax in Corpora of 54 Languages

Kyle Mahowald (UC Santa Barbara), Isabel Papadimitriou (Stanford University),
Dan Jurafsky (Stanford University), Richard Futrell (UC Irvine)

Introducing new information into a clause is cognitively costly and is often restricted to specific linguistic environments, a factor in many sentence processing models (Givón, 2001; Arnold, et al. 2003; MacDonald, 2013). Here we investigate a particular aspect of these costs: the difference between transitive and intransitive subjects. The theory of Preferred Argument Structure (PAS) predicts that new lexical content is less likely to appear as transitive subjects (A, in morphosyntactic notation) than intransitive subjects (S) and transitive objects (O) (Du Bois, 1987; Du Bois et al., 2003). Specifically, the claim is that the transitive subject (A) is a dispreferred location for new information since the object also often introduces new information, and it is cognitively costly to introduce new information in two core argument slots at the same time. Here, we operationalize these constraints in terms of referential form, which has been argued to correlate with accessibility in production (Ariel, 2001). The hypothesis is that more accessible nominals (null arguments, pronouns, proper nouns) are more likely to occur in A argument positions, whereas less accessible nominals (nominals with determiners, modified nominals) are more likely to occur in S and O positions. We run a reproducible, large-scale, cross-linguistic analysis, to evaluate the extent to which these claims about subjecthood and accessibility constitute a universal feature of language.

Our main experimental contribution consists of using the Universal Dependencies corpus of 54 languages from 11 families to extract and correlate two pieces of information about core verb arguments: (a) the accessibility of the argument, and (b) whether it is A, S or O. We investigate accessibility rather than the new/given information distinction because there are very few corpora across languages annotated specifically for information structure. We use UD annotations to classify core verb arguments into five classes of decreasing accessibility: empty subjects, pronouns, proper nouns, lexical items (with no modification other than a determiner), and modified lexical items. For (b), we use the Universal Dependencies parses to determine whether each argument is an A, S, or O.

We found that accessibility asymmetries between A, S, and O broadly hold across languages. Transitive subjects (A) are the least likely to be lexical, followed by intransitive subjects (S), and transitive objects (O). For 93% of languages in our sample, O was more likely to contain a lexical argument than S, and S was more likely to contain a lexical argument than A.

We show a fine-grained breakdown of our results for A and S in Figure 1, comparing how likely it is for arguments in different accessibility classes to appear as A rather than S. Each point in the graph indicates a different language. The downward trend from left to right for all languages shows that more accessible items (empty or pronouns) are more likely to be A, while less accessible items (eg. modified lexical items) are more likely to be S. Assessing significance by fitting a logistic maximal mixed effect model (predicting whether the argument is lexical as a function of argument role) with a random effect for language, we found a significant difference ($\beta = .59, p < .0001$) between A and S in probability of containing a lexical argument. O was even more likely to consist of a lexical argument, and more likely to have that argument modified.

Overall, we show that, cross-linguistically, less accessible (or newer) information is more likely to appear as S than A, and most likely to appear as O. Moreover, our experimental method is easily reproducible and generalizable to more languages. While previous support for theories such as Preferred Argument Structure relied on small, spoken corpora of a handful of languages, we hope that our analysis can lay the groundwork for supporting the empirical cross-lingual universality of such claims about information processing.

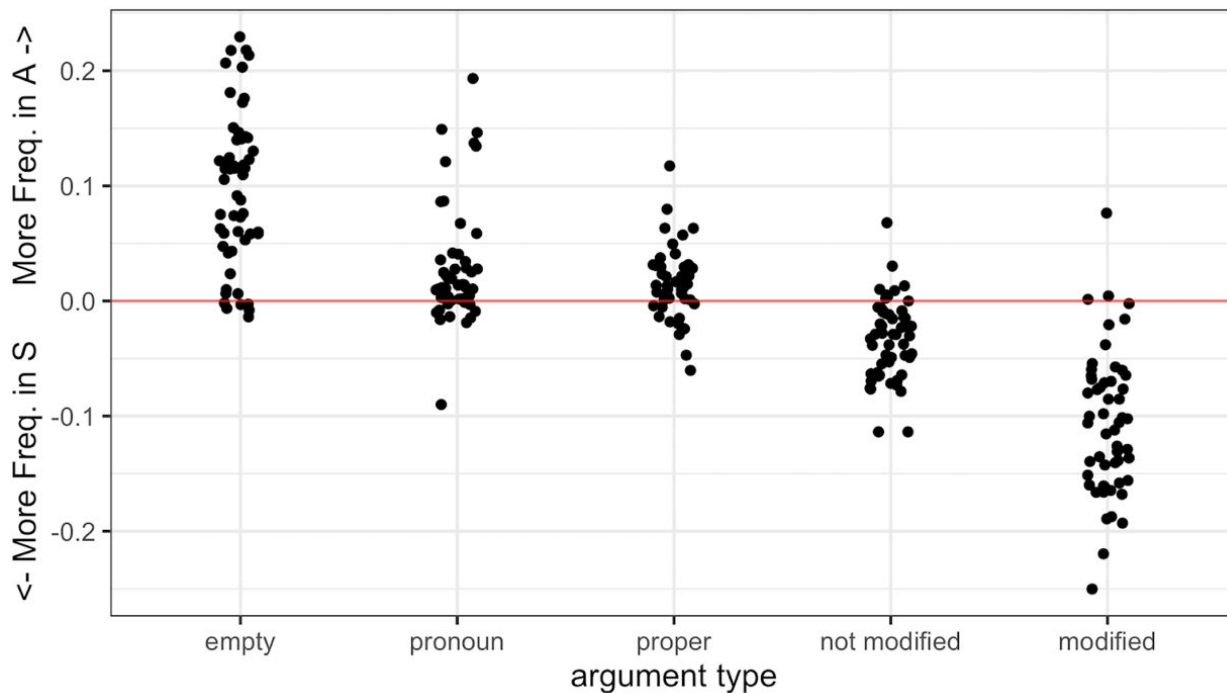


Figure 1 For each argument type, its relative frequency in A relative to S. Each dot is a language. Across languages, empty subjects are more common as transitive subjects, whereas modified subjects are more common in intransitive subjects.

References

- Ariel, M. (2001). Accessibility theory: An overview. Text representation: Linguistic and psycholinguistic aspects, 8, 29-87.
- Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research*, 32(1), 25-36.
- Du Bois, J., Kumpf, L., & Ashby, W. (2003). *Preferred argument structure: Grammar as architecture for function*. (Vol. 14) John Benjamins Publishing.
- Du Bois, J. (1987). The discourse basis of ergativity. *Language*, 805–855.
- Givón, T. (2001). *Syntax: an introduction*. (Vol. 1) John Benjamins Publishing.
- MacDonald, M. (2013). How language production shapes language form and comprehension *Frontiers in psychology*, 4, 226.

Syntax guides sentence planning: evidence from multiple dependency constructions

Shota Momma (UMass, Amherst) & Masaya Yoshida (Northwestern U.)

In production, it has been suggested that speakers plan verbs before starting to speak their patient (or theme) arguments but not their agent arguments (Momma & Ferreira, 2019 a.o.). However, it is unclear *why* speakers plan verbs selectively before the production of patient arguments. One possibility is that speakers plan verbs before their patient arguments because the (conceptual correlate of) patient role but not the agent role critically depends on verb meaning (Kratzer, 2002) (*the conceptual account*). An alternative possibility is that speakers plan verbs before their patient arguments because they are directly selected by verbs as their complement (*the syntactic account*). To evaluate these two accounts, here we examine the timing of verb planning in the production of sentences involving *Across-The-Board* (ATB) and *Parasitic Gap* (PG) constructions (Table 1). ATB and PG are very similar. Most relevantly, the conceptual dependency between fillers and verbs are identical in the example sentences; in both ATB and PG, the initial filler (*which article*) is the theme/patient of the event denoted by the second verb (*criticize*). In contrast, under some theories (Chomsky, 1982 a.o.), the filler is directly selected by the second verb only in ATB, because the syntactic object of the second verb in PG is a null pronoun (or operator) coreferential with the filler. That is, the filler is directly selected by the second verb in ATB (so the second gap is obligatory) but not in PG (so the second gap can be replaced with an overt pronoun). Therefore, the conceptual account predicts that the second verb is planned before the filler in both ATB and PG. Meanwhile, because the filler is directly selected by the verbs only in ATB, the syntactic account predicts that it is planned before the filler in ATB but not in PG. We tested these predictions in two experiments.

In both experiments, we used a new variant of sentence recall task, where participants read a sentence in the RSVP fashion, read aloud 2-4 random verbs, and recalled a sentence as soon as they saw a distractor verb in red font (Fig. 1). Our working assumption is that sentence recall involves the *regeneration* of sentences from conceptual memory (Potter & Lombardi, 1990), and thus it involves the usual processes of grammatical encoding. The distractor words, which also served as a recall prompt (indicated by the font color), were sometimes semantically related to the second verb of the target sentences (e.g., *recommend* for the target *criticize*). Related distractors were used as unrelated distractors in other trials, so the set of related and unrelated distractors were identical. By examining where speakers slow down in their utterances due to the interference from related distractors, we can make an inference about when speakers plan verbs. **Exp. 1** ($n = 47$) ensured that this new task works and that the distractors we chose specifically interfere with the second verb in ATB and PG. Speakers recalled 64 sentences like in Table 2, where the critical verb is the verb that will be used as either the first or the second verb in ATB and PG in Exp. 2. Only correctly recalled sentences were analyzed. Speakers were slower to start utterances given the related distractor, but only in sentences with verbs that were used as the second verb in Exp. 2. This result establishes that speakers indeed plan verbs before the filler in this particular task context and that distractors are effective at eliciting the semantic interference effect specifically on the verb that will be used as the second verb in the ATB and PG sentences. **Exp. 2** ($n = 155$) tested the main predictions. Example target sentences, created by reusing filler NPs and verb-distractor pairs as in Exp. 1, are shown in Table 1. Given related distractors, speakers were slower to start speaking the filler in ATB but not PG (interaction $p = .01$), suggesting that speakers plan the second verb before sentence onset in ATB but not in PG. In comparison, speakers were slower to say the pre-second verb word in PG but not in ATB given related distractor (interaction $p = .03$), suggesting that speakers plan the second verb right before they say it in PG, but before utterance onset in ATB.

These results suggest that speakers plan verbs before the filler in ATB but just-in-time in PG (note that this does not suggest that all words between the filler and the gap are planned in ATB, because planning is likely not sequential. Momma et al. 2019). This timing contrast supports the syntactic account. More broadly, the results suggest that sentence planning is guided by the syntactic dependency (between verbs and their object) that is not reducible to a conceptual dependency (between verbs and their patients).

Table 1: Example ATB and PG sentences (these are also stimuli used in Experiment 2)

Sentence type	Target sentence
ATB	Which article did you read and criticize?
PG	Which article did you read before criticizing?

Table 2: Example stimuli used in Experiment 1. *First* and *Second* refer to the relative position of the verbs in the ATB and PG sentences used in Experiment 2.

Verb position	Target sentence
First verb	Which article did you read?
Second verb	Which article did you criticize?

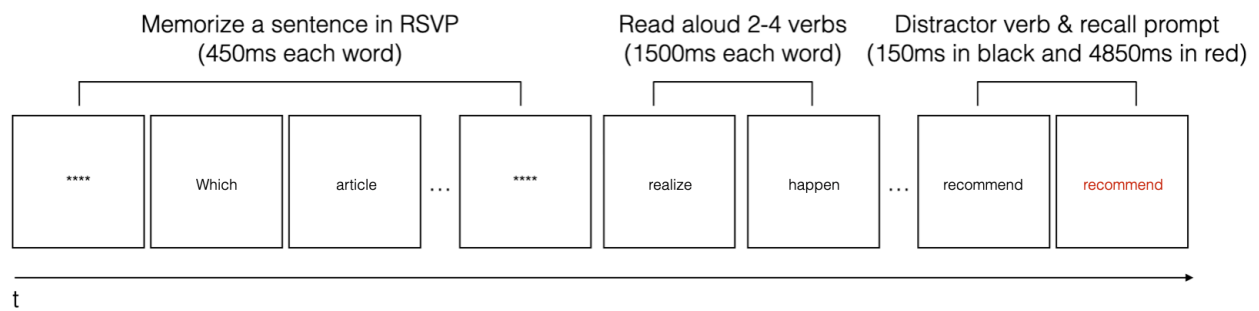


Fig. 1: A schematic illustration of the experimental task. Note that the number of random verbs to be read aloud between the memorization and the recall was variable across trials. To prevent speakers from predicting when they were to recall sentences. Note also that the distractor verb was initially in black; it turned red 150ms after the onset of the presentation. This is to increase the chance that speakers register the distractors.

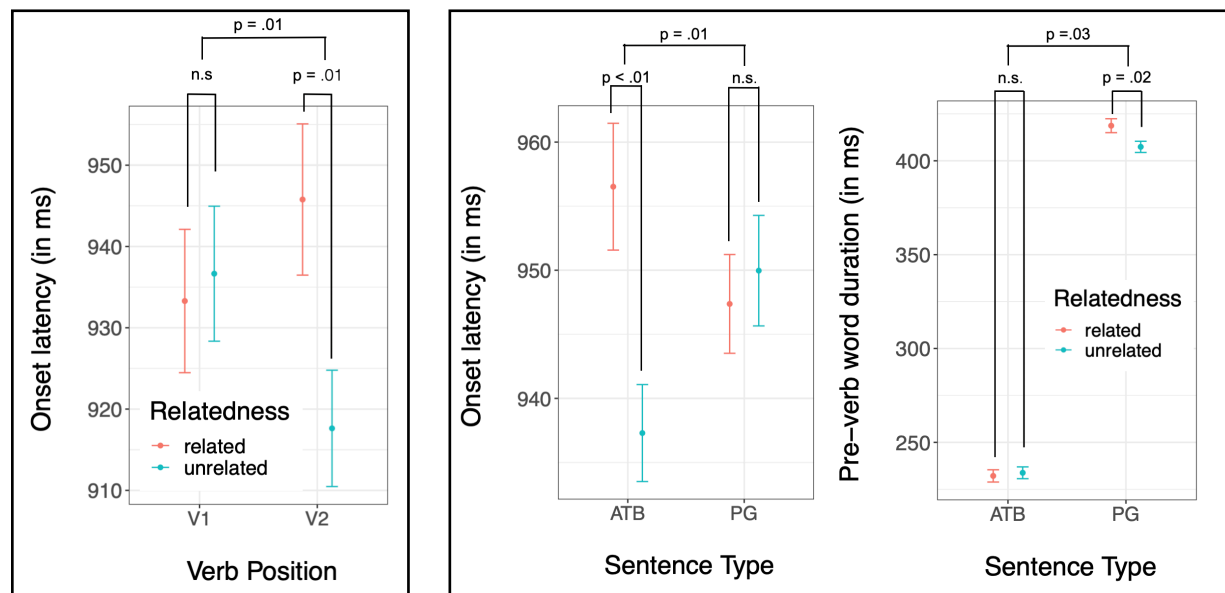


Fig. 2 (left): Results of Experiment 1. Fig. 3 (right): Results of Experiment 2 (onset latency and pre-second verb word duration).

EVIDENCE FOR EARLY APPLICATION OF BINDING THEORY AND LATE INTRUSION EFFECTS

Arild Hestvik & Myung Hye Yoo (University of Delaware)

INTRODUCTION: Authors have previously argued that the parser “knows and obeys” binding theory during antecedent selection for reflexives (Nicol & Swinney, 1989). Sturt’s “Defeasible binding” theory predicts that BT-grammatical antecedents are established during a “first-pass,” and that intrusive antecedents can “take over” as illicit antecedent only in a temporally subsequent stage (Sturt, 2003). However, recent cue-based theories of memory retrieval predict that structurally illicit but feature matching (“intrusive”) antecedents should immediately interfere with antecedent assignment (Jäger et al., 2017; Lewis & Vasishth, 2005; Parker et al., 2015, 2016; Parker & Phillips, 2017), and that intrusive antecedents are facilitated when the BT-grammatical antecedent fails to agree in some feature (Patil et al., 2016). The extant literature has relied on reading time and eye-tracking studies, but ERPs are well suited for measuring time course of processing and should converge with eye-tracking data. To date, only one ERPs study has measured intrusion effects in reflexive binding (Xiang et al., 2009), but it employed an unbalanced design and was primarily about negative polarity licensing. We turned Xiang et al into a 2x2 design, as in Patil et. al., (2016), to obtain an ERP time course test of BT-grammatical vs. intrusive binding.

METHODS: Two ERP experiments were conducted (see table 1 and 2). Sentences were presented word-by-word centered on the screen (300ms duration+200ms ISI). Each sentence was followed by a comprehension question. Experiment 1 (N=24) sought to establish a baseline measure of the time course of agreement violation detection between BT-grammatical antecedent and reflexive, in the absence of intrusive antecedents (e.g. “The male soldier that the team treated in the military hospital introduced himself/herself to all the nurses”). In Experiment 2 (N=23) we introduced an intrusive antecedent (Table 1). We again measured (i) whether we observed the same BT-grammatical ERPs as in Exp 1, and also (ii) whether feature mismatches between the intruder and the reflexive modulated the same ERP as in Exp 1 or showed up in a separate (later) ERP, and (iii) whether there was an interaction such that the intrusive effect was facilitated by failure of BT-grammatical binding.

RESULTS: After artifact correction, main effects were constructed as difference waves (matching minus mismatching antecedent). The temporal and spatial dynamics of the brain response to agreement violations was factored with a temporo-spatial sequential PCA/ICA analysis (Dien, 2010, 2012). The “factor ERP” scores were used as dependent measures, but also used to constrain selection of time windows and electrode regions in the undecomposed voltage data, which was also analyzed as dependent measures, for convergence. In Experiment 1, in addition to a LAN factor (500ms), the BT-grammatical agreement violation was reflected in a modulation of the N170 visual cortex response (Vogel & Luck, 2000), analogous to Dikker et. al., (2009), see Fig 1. This effect was exactly replicated for BT-grammatical violations in Experiment 2 (Fig 2). However, there was no signal in the N170 component of intrusive antecedent agreement violation. Rather, this condition elicited two later ERP components (388ms and 496ms), none of which reached statistical significance (mirroring the results in (Xiang et al., 2009)). Analysis of the cue-based theory’s predicted interaction between failed BT-grammatical binding and intrusive binding only revealed a main effect of BT-grammatical antecedents (Fig 3).

CONCLUSION: BT-grammatical binding is visible as early as 170ms. This is a new finding and shows that BT-grammatical binding is established much earlier than previously reported (Osterhout & Mobley, 1995). It is interpretable as grammatical predictions driving top-down sensory expectations about visual word forms (Dikker et al., 2009). Intrusive binding elicited later ERP effects (~500ms), but was variable across individuals (and therefore statistically weaker). There was no facilitation of intrusive binding when the BT-grammatical antecedent failed to agree, contra the predictions of cue-based memory retrieval models. The results support Sturt’s 2-stage defeasible binding process theory.

		BT-GRAMMATICAL ANTECEDENTS	
		A. <i>Incongruent grammatical</i>	B. <i>Congruent grammatical</i>
INTRUSIVE ANTECEDENTS	C. <i>Incongruent intrusive</i>	The <u>male soldier</u> that Fred treated in the military hospital introduced <u>herself</u> to all the nurses.	The <u>male soldier</u> that Katie treated in the military hospital introduced <u>himself</u> to all the nurses.
	D. <i>Congruent intrusive</i>	The <u>male soldier</u> that Katie treated in the military hospital introduced <u>herself</u> to all the nurses.	The <u>male soldier</u> that Fred treated in the military hospital introduced <u>himself</u> to all the nurses.

Table 1:
Design of
Exp 2. Each
cell had 30
trials.

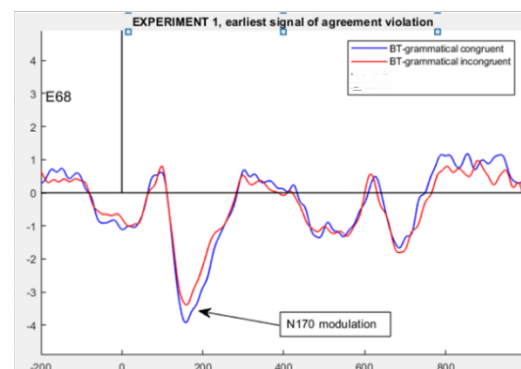
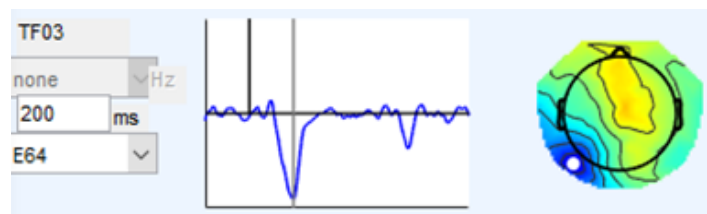


Figure 1: Left panel: Temporal PCA factor for Exp 1 BT-grammatical agreement violation difference wave (two spatial subfactors both sign. by t-test against 0). Right panel: corresponding undecomposed grand average absolute voltage waveforms; the difference incongruent-congruent, mean voltage 160-224ms peak channel E68 (defined by PCA/ICA) was statistically significant with t-test against zero ($t(23)=-5.69$, $p<0.00001$).

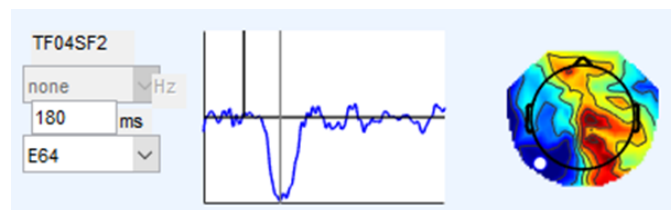


Figure 2, left: N170 effect in Exp2, BT-grammatical antecedents ($t(19)=4.79$, $p<0.001$). Corresponding voltage effect: $t(19)=-4.22$, $p<0.001$, t-tests against 0.

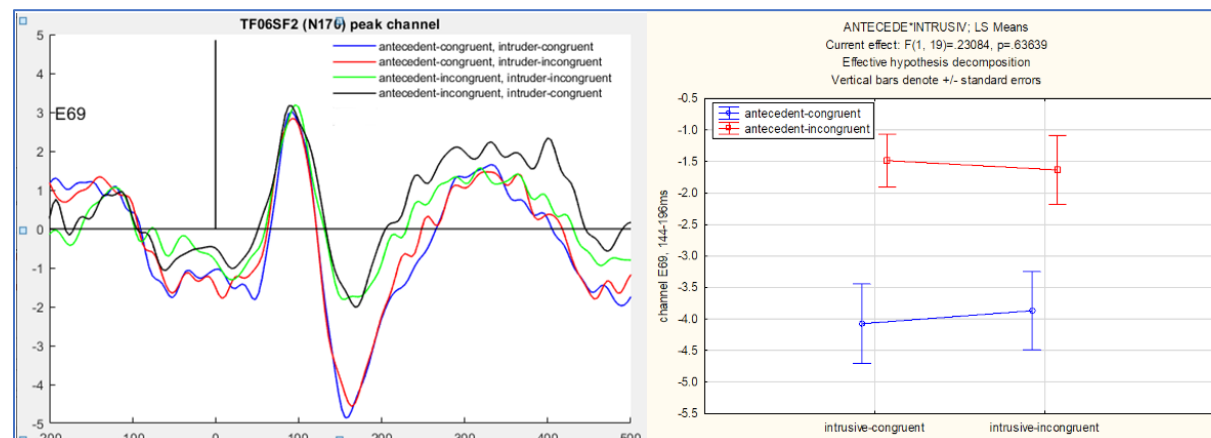


Figure 3: Analysis of absolute waves, full 2x2 ANOVA. Only a main effect of BT-grammatical antecedents was observed in the N170 component, no main effect of intruder or interaction between the two factors were observed. Intruder agreement violations had a later effect (~500ms), not shown here.

Predicting binding domains: Evidence from fronted auxiliaries and wh-predicates
Keir Moulton (U. of Toronto), Cassandra Chapman (U. of Toronto), and Nayoun Kim (Sungkyunkwan U.)

Online anaphoric dependency resolution has been argued to be immediately guided by structural constraints such as the Binding Theory (BT, [1,2,3,4]). In a sentence completion study and a self-paced reading (SPR) experiment, we investigate the role of predicted structures in antecedent retrieval [5]. The results suggest that structural expectations arising from fronted auxiliaries influence the retrieval of antecedents for pronouns in fronted wh-predicates.

Key Manipulation: was vs. did Pronouns in predicate wh-phrases (*how proud of him*) are subject to BT constraints at the gap site [6]: a matrix clause gap (1a/2a, Table 1) puts the matrix subject and pronoun in the same binding domain and co-reference is precluded by Principle B [1]. When the gap is in a different binding domain (1b/2b), co-reference is possible. When presented with auxiliary *was* (1), we expect readers to pursue the simpler (1a) over (1b), eliminating *the boy* as an antecedent. Auxiliary *did* (2) does not allow a matrix gap (**How proud did John*), so we predict that a continuation introducing a new binding domain is more likely than with *was*. Consequently, the matrix subject is more likely to be retrieved as an antecedent.

Sentence Completion Study 60 participants completed sentence fragments like (1/2) ending at *the*. Of 274 grammatical continuations (of 300) provided in the *was* condition, no completions (0%) involved a new binding domain (like 1b). Of 252 grammatical continuations provided in the *did* condition, participants provided 66 completions with a new binding domain (like 1b) (26%).

SPR experiment: Using a Gender Mismatch Effect paradigm (GMME) [7], we tested whether the different expectations triggered by *was* vs. *did* have any impact on online antecedent retrieval. A SPR experiment (n=127) tested items shown in Table 2, crossing Gender (whether the pronoun in the wh-predicate Matches or Mismatches the matrix subject) and Auxiliary (*was* vs. *did*). Given the sentence completion results, we expect that in comparison to the *was* conditions, in the *did* conditions readers will be more likely to entertain an upcoming structure where the wh-predicate finds a gap in a new binding domain. As a result, they will be more likely to retrieve the matrix subject as a BT-compliant antecedent. We expect an interaction in which only the *did* condition gives rise to a GMME [2,7]. **Results** Analyzing residualized reading times, at the critical gendered noun region (Figure 2; “saleswoman/man”), an interaction between Gender and Auxiliary was observed ($\beta=-61.63$, $SE=24.85$, $p<0.05$) as was a marginal effect of Gender ($\beta=21.86$, $SE=12.43$, $p=0.08$). Subset analysis revealed an effect of Gender only in the *did* condition ($\beta=53.34$, $SE=17.88$, $p<0.05$). At spillover region 2 (Figure 1, “California”) there was a significant interaction between Auxiliary and Gender ($\beta=-39.65$, $SE=17.53$, $p<0.05$). Subset analysis revealed a GMME only in the *did* condition ($\beta=24.44$, $SE=11.42$, $p<0.05$), not in *was* ($\beta=-15.90$, $SE=13.29$, $p>0.05$), suggesting that the mismatched *did* conditions were read more slowly than all other conditions.

Conclusions: One interpretation of the results is that the processor is sensitive to BT constraints like Principle B even when calculated over expected, but not yet verified, structures. Further investigation, however, is needed to test another possibility: that in *did* conditions, the processor accessed BT-non-compliant antecedents indiscriminately (see [8]) in the absence of more definitive evidence for the location of the gap (evidence that *is* available in the *was* conditions, which overwhelmingly trigger the expectation for a matrix/same domain gap). We are conducting a counterpart study using reflexives, where *was/did* make opposite predictions about binding domains, to address this possibility.

Table 1: was vs. did and binding domains

	Sentence fragment:	Possible continuation:	Binding domain?
(1)	How proud of him ₁ was the...	a. boy ₁ __?	same
		b. boy ₁ saying someone was __?	different
(2)	How proud of him ₁ did the...	a. boy ₁ feel/seem to be __?	same
		b. boy ₁ say someone was __?	different

Table 2: SPR Experiment stimuli

	Match/Mismatch
WAS	How impressed with him <u>was</u> the tall friendly salesman/saleswoman from California saying that Amanda's bosses were?
DID	How impressed with him <u>did</u> the tall friendly salesman/saleswoman from California say that Amanda's bosses were?

Figure 1. Word-by-word reading times

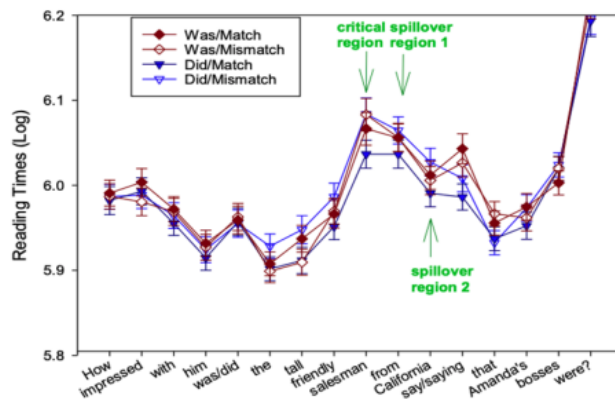
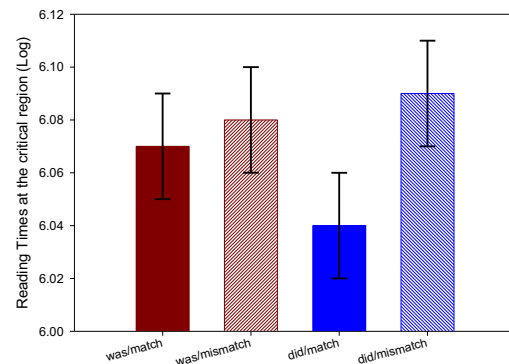


Figure 2. Reading Times at critical region



[1]Chomsky, N. (1981). *Lectures on government and binding*. [2]Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *JML* 48, 542-562. [3]Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *JML*, 56(3), 384-409. [4]Chow, W. Y., Lewis, S., & Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psych* 5, 630. [5]Kush, D. & Dillon, B. Disjoint is off the hook: Principle B constrains predictive resolution of cataphors. CUNY 2020. [6]Huang. (1993). Reconstruction and the structure of. *LI* 24. [7]Van Gompel, R. P. G., & Liversedge, S. P. (2003). The influence of morphological information on cataphoric pronoun assignment. *J. of Experimental Psych.* 29, 128–139.[8]Omaki, A., Ovans, Z., Yacovone, A., & Dillon, B. (2019). Rebels without a clause: Processing reflexives in fronted wh-predicates. *JML* 107, 80-94.

Classifier as a cue for structure building in head-final relative clause

Zirui Huang & Matthew Husband (University of Oxford)

Zirui.huang@ling-phil.ox.ac.uk

Previous studies have suggested a predictive mechanism for relative clause (RC) processing in languages that have a head-final RC structure, like Japanese (Yoshida et al., 2004) and Mandarin Chinese (Hsu, 2006; Wu, 2009). However, it still remains unknown what type of information the parser utilizes to anticipate the structure of an upcoming RC and how detailed such structure building is before receiving information from the head noun directly. To address this, we investigated how the semantic information provided by different classifiers (CL) in Mandarin Chinese (human, non-human, general) guides structure building of upcoming RCs.

Chinese “classifier + transitive verb” sequences are temporarily ambiguous between a subject gapped RC (1a) and a (null subject) object gapped RC construction (1b). Although the parser is bias to adopt a subject RC analysis, semantic cues of a CL may be used to guide which of these two RC structures is initially adopted. Non-human CLs in particular may guide the parser away from a subject RC analysis by indicating that the head noun is unlikely to be an eligible subject for a subject RC. We predicted that this should facilitate the analysis of a null subject RC. With human and general CLs, the parser may be more likely to assume a subject gap and expect a noun to fill the object position. This predicts reading disruption upon encountering an unexpected relativizer and head noun. In a series of studies, CL type was manipulated to examine whether the parser uses CL type to predict the gap site in a head-final RC.

Sentence completion: A sentence completion survey (N=439) was conducted online to investigate the parser’s bias for subject RC and null subject object RCs. The results suggest that the mismatch between a dislocated CL and following verb guides the parser to a RC structure (88.7%) and the RC type is influenced by the CL type. Human CLs produce an overwhelming preference for subject-gapped RC (92.2%). General CLs also elicit a subject-gapped preference (71.4%). Non-human CLs, however, produce more object-gapped RC (85.9%).

Eye-tracking: Verbs and head nouns were selected based on the responses in the completion study and used as stimuli in an eye-tracking while reading experiment (N=42). Using general CL as baseline, results of linear mixed effect model show reading facilitation with non-human CL at the relativizer region in first fixation (Est=-12.24 ms, $t=-2.399$, $p<0.05$), first pass (Est=-14.17 ms, $t=-2.545$, $p<0.05$), go past (Est=-39.38 ms, $t=-2.077$, $p<0.05$) and total fixation (Est=-48.62 ms, $t=-4.139$, $p<0.001$). Human CL show greater reading disruption compared with general CL in go pass reading (Est=66.30 ms, $t=3.499$, $p<0.01$) and total fixation time (Est=46.59 ms, $t=3.969$, $p<0.001$). These effects are largely recapitulated at the head noun region. In non-human CL condition, facilitation is significant in go past reading (Est=-58.27 ms, $t=-2.842$, $p<0.01$) and total fixation (Est=-81.35 ms, $t=-3.314$, $p<0.01$). For human CL, disruption is significant in first pass reading (Est=14.33 ms, $t=2.326$, $p<0.05$), go past reading (Est=86.64 ms, $t=4.310$, $p<0.001$) and total fixation (Est=58.67, $t=2.39$, $p<0.05$).

Self-paced reading: We extended the results using self-paced reading, keeping the head nouns as the same across different conditions by separately comparing non-human CL vs. general CL (N=43) and human CL vs. general CL (N=40). Both human and non-human conditions show reading disruptions at the verb (Est=35.08 ms, $t=2.898$, $p<0.01$; Est=30.37 ms, $t=2.892$, $p<0.01$), suggesting greater mismatch between the CLs and the verb. In human CL condition, disruptions continue in relativizer (Est=24.71 ms, $t=2.413$, $p<0.05$) and head noun (Est=37.16 ms, $t=2.75$, $p<0.01$) while in non-human CL condition, reading was facilitated at the relativizer (Est=-36.93 ms, $t=-3.916$, $p<0.001$) and the head noun (Est=-47.27 ms, $t=-4.941$, $p<0.001$).

Conclusion: The results indicate that the semantic properties of CLs can help parser to make structural predictions in head-final RC processing before accessing the head noun. In particular, non-human CLs guide the parser away from preferred subject-gapped RC structure, facilitating a null subject object-gapped analysis.

(1) a. 那个扔掉垃圾的小孩得到了表扬。

(subj RC)

That CL throw rubbish REL child receive PERF praise

That child who threw rubbish received praise.

b. 那个扔掉的娃娃变得脏兮兮的了。

(obj RC + null subj)

That CL throw REL doll become dirty PERF

That doll which (someone) threw away became dirty.

Sentence completion: Example stimuli:

那 { 个 / 名 / 张 } 扔掉 _____
That { General.CL / Human.CL / Nonhuman.CL } throw _____

Eye-tracking:

a. Human classifier condition:

那名捡到的孩子已经醒过来了。

That CL find REL child already awake PERF

The child that (someone) found is already awake.

b. General classifier condition:

那个捡到的硬币已经脏兮兮的了。

That CL find REL coin already dirty PERF

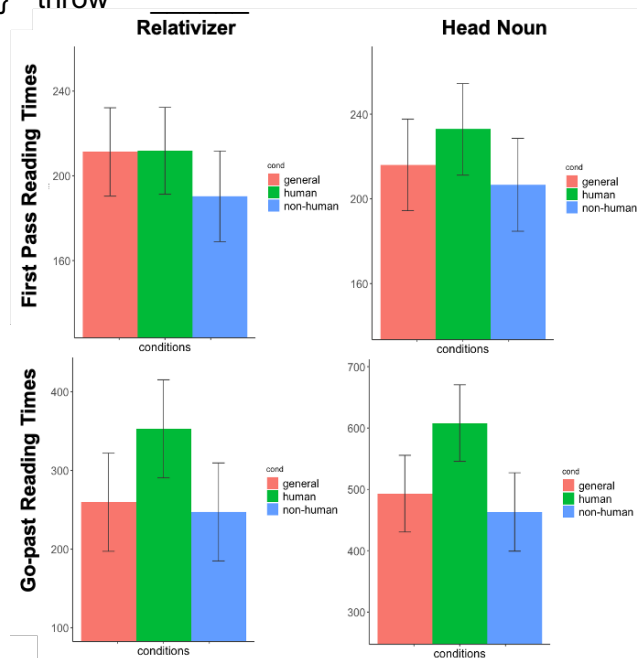
The coin that (someone) found is dirty.

c. Non-human classifier condition:

那张捡到的银行卡已经还给失主了。

That CL find REL card already return owner PERF

The credit card that (someone) found has already been returned to its owner



Self-paced reading:

Non-human vs. general classifier

a. Non-human classifier condition:

那条忽略的线索是破案的关键。

That CL ignore REL clue is solve case POSS. key

The clue that (someone) ignored is the key to solve the case.

b. General classifier condition:

那个忽略的线索是破案的关键。

That CL ignore REL clue is solve case POSS. key

The clue that (someone) ignored is the key to solve the case.

Human vs. general classifier

c. Non-human classifier condition:

那名忽略的证人是破案的关键。

That CL ignore REL passerby is solve case POSS. key

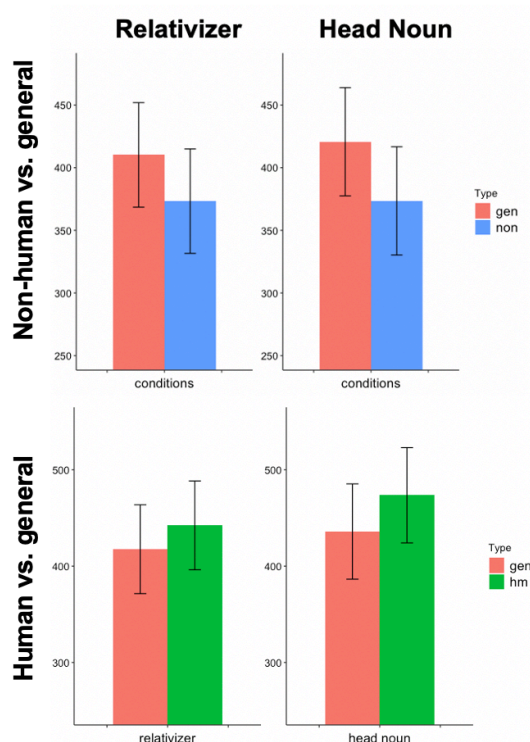
The passerby that (someone) ignored is the key to solve the case.

d. General classifier condition:

那个忽略的证人是破案的关键。

That CL ignore REL passerby is solve case POSS. key

The passerby that (someone) ignored is the key to solve the case.



ERPs reveal how semantic and syntactic processing unfolds across parafoveal and foveal vision in sentence comprehension

Chuchu Li (UC San Diego), Katherine Midgley & Phillip Holcomb (San Diego State University)
chl441@ucsd.edu

Sentence comprehension requires both semantic and syntactic processing, which elicit different patterns of neural activity. Previous ERP studies that investigated sentence comprehension usually adopted the RSVP paradigm that presents one word at a time, which showed a N400 for semantic anomaly and a left anterior negativity (LAN) and/or P600 for syntactic anomaly. However, in natural sentence reading upcoming words are available even before they are foveated, and at least semantic information seems to be processed for words in parafovea (i.e., parafoveal N400, which could mitigate the N400 when targets are foveated; Payne et al., 2019). The present study compared how semantic and syntactic processing unfolds across parafoveal and foveal vision in sentence comprehension by examining readers' EEG when unexpected content or function words were presented. Content words (e.g., dog, eat) have rich semantic information, while function words (e.g., in, her) carry less meaning but reveal grammatical relationship between content words. Thus, reading content words may involve more semantic processing while function words may elicit more syntactic processing (e.g., Brown et al., 1999). However, direct comparison between content versus function words generally involve some confounds (e.g., function words are typically shorter and have higher frequency), therefore in the present study critical comparisons were conducted within each class of word (i.e., unexpected vs. control words).

We tested 24 English monolinguals (*M* age=22; range 19-27). The critical stimuli included 120 sentences, each of which had three conditions: 1) the control condition with no errors, 2) the semantic violation condition where the critical content word was replaced by an unexpected one, and 3) the syntactic violation condition where the critical function word was replaced by an unexpected one (see Table 1). These sentences were evenly distributed in three lists in a Latin-square design. Thus, each list included 40 sentences in each of the three conditions plus 40 well-formed filler sentences. We adopted a modified visual RSVP flanker paradigm. Each sentence was presented in sequential three-word chunks, with the to-be-fixated word in the center of the display (foveal target), the upcoming next word to the right of fixation (parafoveal target), and the former central word to the left of fixation. At 400 ms intervals the three words were shifted leftward so that the old central word was on the left, the previous parafoveal word was now the central target word and a new word appeared to the right. To facilitate central word fixation, two yellow vertical bars were placed above and below the central word and the central foveal target was displayed in white letters while the two flanking words were displayed in a slightly dimmer grey (see Fig. 1). Horizontal eye-movement was closely monitored and all trials with horizontal eye-movement were removed. Each participant read a list of sentences silently while EEG was being recorded. At the end of each sentence, they judged if the sentence made sense (yes/no button press).

Unexpected content words elicited a right lateralized N400 when displayed in the parafovea, followed by a longer-lasting, widely distributed positivity starting around 300 ms once the target word was foveated (see Fig. 2). Unexpected function words elicited a left lateralized LAN-like component when presented in the parafovea, followed by a left lateralized, posteriorly distributed P600 when that word was presented in the fovea (see Fig. 3). As predicted, in sentence comprehension content versus function words elicit more semantic versus syntactic processing, respectively. Critically, our results suggest that the combination of negativities and positivities seen to critical words in typical word-by-word RSVP paradigms might mask what is actually a sequence of two overlapping stages in which fast, perhaps automatic processes, perform an initial semantic/syntactic assessment of the upcoming word when it is presented in the parafovea and is then followed by a more in depth attentionally mediated assessment once the word has been foveated (e.g., sentence level re-analysis or repairing).

Table 1. Examples of experimental sentences

Control	The old man was asleep in the chair when I came back.
Semantic Violation	The old man was asleep in the cherry when I came back
Syntactic Violation	The old man was asleep in of chair when I came back.

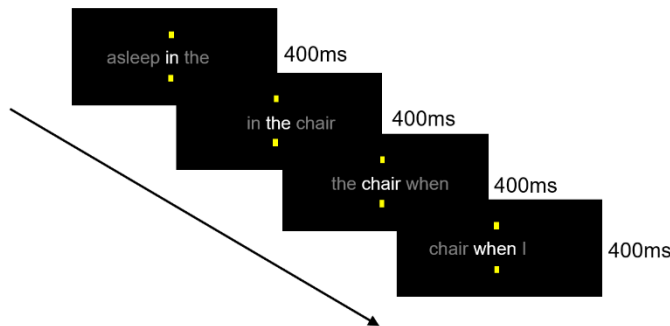


Fig. 1: An illustration of the modified visual RSVP hemi-field flanker paradigm adopted in the present study.

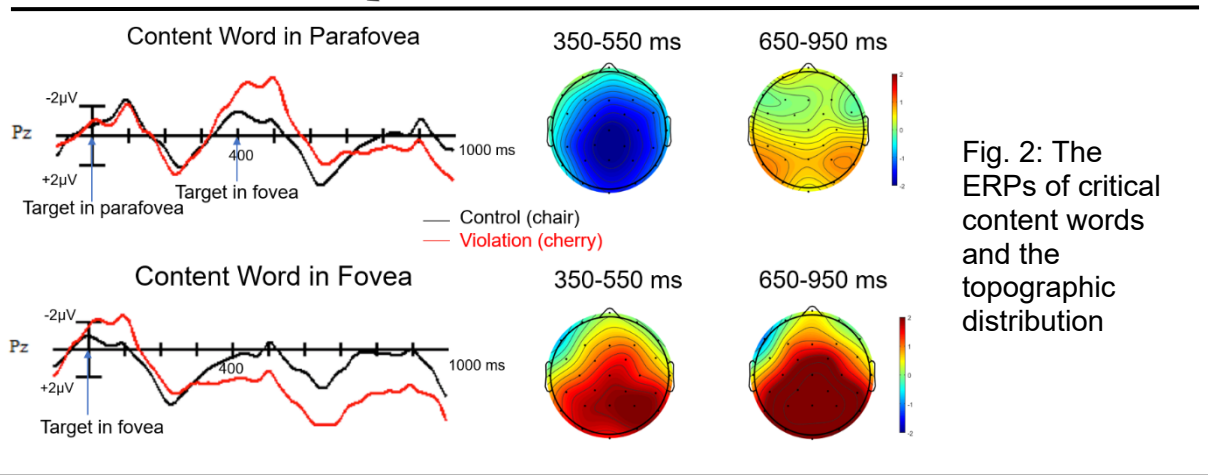


Fig. 2: The ERPs of critical content words and the topographic distribution

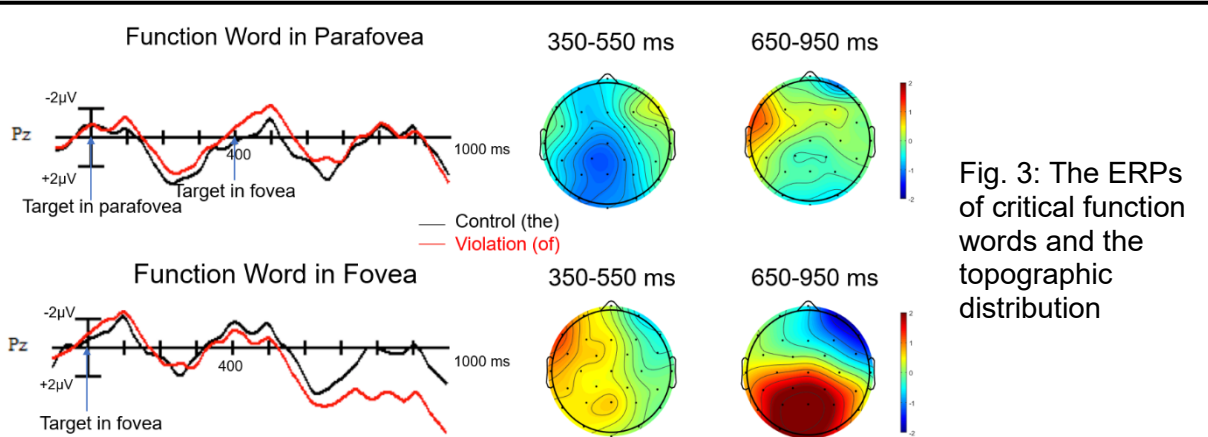


Fig. 3: The ERPs of critical function words and the topographic distribution

Reference

Brown, C. M., Hagoort, P., & Keurs, M. T. (1999). Electrophysiological signatures of visual lexical processing: Open-and closed-class words. *Journal of cognitive neuroscience*, 11(3), 261-281.

Payne, B. R., Stites, M. C., & Federmeier, K. D. (2019). Event-related brain potentials reveal how multiple aspects of semantic processing unfold across parafoveal and foveal vision during sentence reading. *Psychophysiology*, 56(10), e13432.

A noisy channel model of N400 and P600 effects in sentence processing

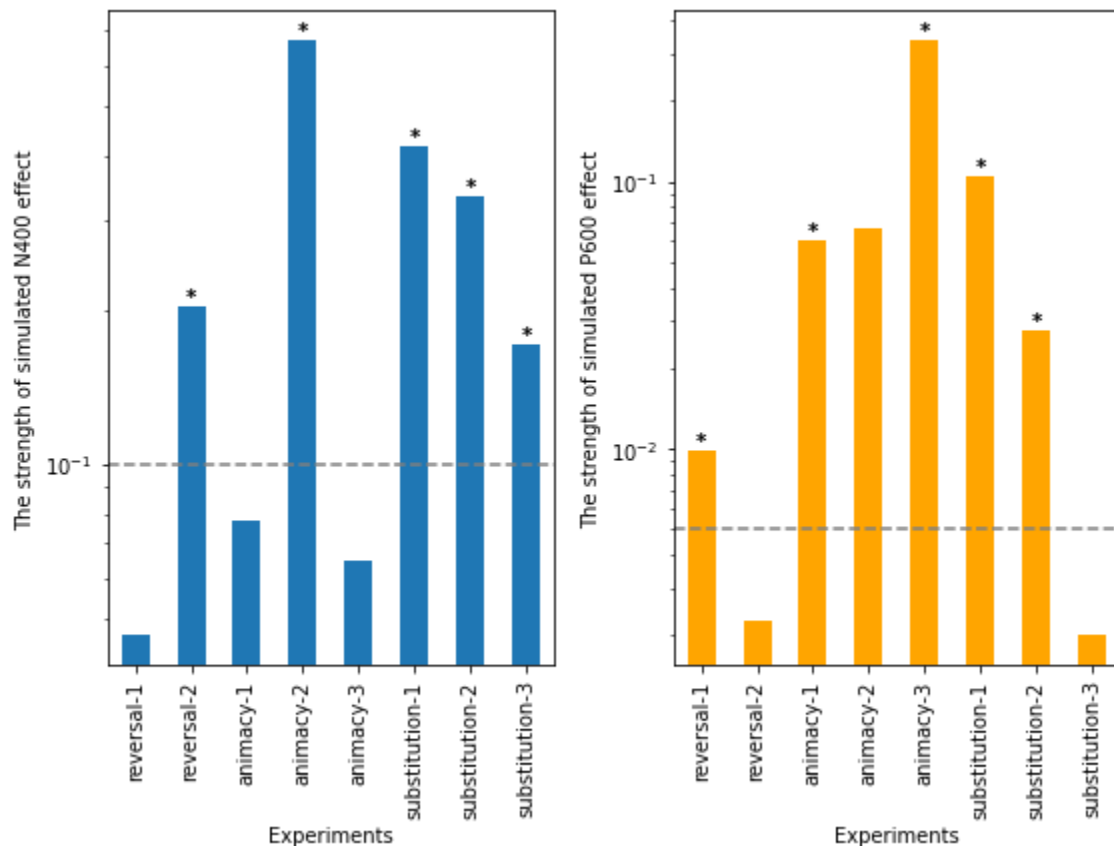
Jiaxuan Li & Allyson Ettinger (University of Chicago)

Introduction: N400 and P600 event-related potential (ERP) components have long been the object of study in psycholinguistics. Traditional accounts have associated N400 effects with semantic violations, and P600 effects with syntactic violations [1,2]. However, this picture is complicated by P600 effects—without N400 effects—in response to animacy [3,4] and thematic-role [5] violations (but only sometimes [6]), as well as biphasic N400/P600 effects for conventional semantic violations [5]. Building on explanations involving interplay of plausibility-driven and syntax-driven interpretations [3,7], we present a computational model that accounts for these complicating observations via a noisy channel modeling framework. Our model assumes early-stage sentence interpretations determined by noisy channel computation (influenced by plausibility), with these early interpretations driving the N400 amplitude. The P600 amplitude reflects reconciliation of the early interpretation with the true (syntax-driven) interpretation, and is modulated by the extent to which early interpretations deviate from the true input. Running this model on original experimental stimuli, we successfully simulate N400 and P600 effects from seven studies in this literature [3-6]. **Method:** We use original stimuli from psycholinguistic experiments featuring semantic / thematic violations, with empirical results varying between N400 effect only, P600 effect only, and biphasic N400/P600 effect (see Table 1). Our use of real experimental stimuli is of note because computational psycholinguistic models often use idealized inputs, while we account for idiosyncratic properties of the real stimuli. To estimate relevant properties of these stimuli (e.g., plausibility, semantic similarity), we draw on outputs of pre-trained models used in natural language processing (NLP). **Noisy channel model:** We implement a noisy channel model to estimate posterior probabilities of potential early interpretations (S_i) given presented input (S_p). These posterior probabilities are based on a) the prior probability of S_i , and b) the likelihood of seeing S_p as a distortion of S_i . For the prior $P(S_i)$, we aim to capture interpretation plausibility, which we approximate via sentence probability estimates from a large neural network pre-trained on word prediction (OpenAI GPT) [8]. We base the likelihood $P(S_p|S_i)$ on the Levenshtein edit distance between S_i and S_p , to capture stronger likelihood of smaller deviations from true input. For each stimulus item S_p , we compute posterior interpretation probabilities for the true input itself, and for one alternative (for anomalous items, a plausible alternative; for control items, an anomalous counterpart). The interpretation with the higher posterior probability is identified as the *early interpretation*. **N400 simulation:** N400 amplitude is approximated by the neural network probability of the target word, given prior context, within the selected early interpretation. **P600 simulation:** To capture reconciliation between interpretations, P600 amplitude is simulated as difference between representations of the early interpretation and the true input, obtained from a neural network pre-trained to detect semantic similarity (fine-tuned DistilBERT) [9]. **Results:** Simulated response amplitudes are averaged by condition, and effects are determined by amplitude differences between critical and control conditions. Results are shown in Fig 1. We see that the model successfully predicts N400 and P600 effects from seven of our eight target experiments. The one failure is a P600 effect appearing for animacy-2 [3]—but we believe that this can be attributed to limitations in the pre-trained neural networks (which show signs of particularly poor estimates on the stimuli in this experiment), rather than to fundamental limitations of our model. **Conclusions:** These results support an account of sentence processing involving early, plausibility-driven interpretation stages (informed by rational inference), reflected in the N400—followed by reconciliation with syntax-driven interpretations, reflected in the P600. Prior work has posited plausibility/syntax interplay [3,7], and other work has linked predictions of noisy channel models to patterns in comprehenders' final interpretations [10,11], and in the P600 [12]. However, to our knowledge this is the first fully-specified computational formalization of plausibility/syntax interplay, the first implemented noisy channel model for simulation of N400 and P600, and the first model of either type to carry out direct prediction of both N400 and P600 components, using real experimental stimuli, across this range of experiments.

Table 1. List of simulated experiments, with experimental manipulations and results.

ID	Manipulation	Violation type	Result	Source
reversal-1	role-reversal	Thematic role	P600	[5]
reversal-2	role-reversal	Thematic role	N400	[6]
animacy-1	Active/passive	Animacy	P600	[3]
animacy-2	Active/passive	Animacy	N400	[3]
animacy-3	Active/passive	Animacy	P600	[4]
substitution-1	word substitution	Lexical meaning	N400 & P600	[5]
substitution-2	word substitution	Lexical meaning	N400 & P600	[5]
substitution-3	word substitution	Lexical meaning	N400	[5]

Fig.1. Simulated N400 (left) and P600 effects (right) across experiments. * represents significant N400/P600 effect in the original human experiment. Dotted line represents a threshold (determined post-hoc) allowing for delineation between presence and absence of effect.



Reference

- [1] Kutas & Hillyard (1980). *Biological psychology*.
- [2] Hagoort, Brown, & Groothusen (1993). *Language and cognitive processes*.
- [3] Kim & Osterhout (2005). *Journal of memory and language*.
- [4] Kuperberg, Choi, Cohn, Paczynski, & Jackendoff (2010). *Journal of cognitive neuroscience*.
- [5] Chow, Smith, Lau & Phillips (2016). *Language, Cognition and Neuroscience*.
- [6] Ehrenhofer, Lau, & Phillips (2020). Forthcoming, (Submitted to *Neuropsychologia*)
- [7] Kuperberg (2007). *Brain research*.
- [8] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Preprint*.
- [9] Reimers, N., & Gurevych, I. (2019, November). In *EMNLP-IJCNLP*.
- [10] Levy (2008, October). In *Proceedings of the 2008 EMNLP*.
- [11] Gibson, Bergen & Piantadosi (2013). *Proceedings of the National Academy of Sciences*.
- [12] Ryskin, Stearns, Bergen, Eddy, Fedorenko & Gibson (2020). *bioRxiv 2020.02.08.930214*.

The benefits and costs of language prediction: Evidence from ERPs

Jiaxuan Li, Jinghua Ou & Ming Xiang (University of Chicago)

Introduction: Comprehenders actively anticipate upcoming material based on context, and the facilitation effect of prediction on expected words has been associated with an attenuated N400 [1]. However, it remains unclear what neural signature indexes the processing cost when an expectation is not fulfilled. A number of recent proposals have suggested that post-N400 positivities (PNP) with an anterior-frontal scalp distribution reflecting the cost of integrating an unexpected but still interpretable word [2]. Specifically, unexpected but plausible words have been found to elicit larger frontal PNPs relative to semantically anomalous continuations [3,4]. Moreover, context constraint plays a role in modulating the processing of unexpected plausible words, with highly constrained context eliciting larger frontal PNPs relative to less constrained context [5,6]. The anterior-frontal PNPs therefore have been interpreted as reflecting comprehenders' continuous effort to update from a previously expected semantic representation to a less expected but still interpretable one [6]. While contextual expectation of an event argument is typically used to examine the prediction cost in prior work, the current study makes a parallel comparison between two predictive contexts in Chinese: the verb-noun and the classifier-noun context, with both verbs and classifiers providing predictive cues for the noun phrases. The verb-object relation is based on a set of multidimensional features rooted in rich world knowledge, whereas the classifier-noun relation is often determined by a much narrower semantic dimension (e.g. shape). The current study aims at replicating the basic patterns of PNPs from previous studies, and further shedding light on the functional interpretation of this component.

Method: We constructed numeral-classifier-noun and verb-noun phrases in Chinese. Within each structure, there is a strongly and a weakly constraining context for the upcoming noun, with constraint defined as the max cloze probability of the possible continuation nouns, following the basic design in [6]. Under the high-constraint context, there are three levels of cloze probabilities for the upcoming noun, and under the low-constraint context, there are two levels (see Table 1). Cloze probabilities and constraints are matched between the classifier and the verb contexts, based on a separate noun-completion norming study. Twenty native Mandarin Chinese speakers participated in the study. There are 30 items per condition.

Result: Based on previous studies [3 - 6], our analyses focused on two comparisons: (i) a 3-way comparison based on the cloze probability of the noun, i.e. high cloze vs. low cloze vs. anomalous (zero cloze); (ii) and a 2-way comparison between the unexpected (low cloze) nouns in the high constraint context vs. those in the low constraint context. Fig. 1 shows grand-average ERP waveforms at critical electrodes. The current report focuses on the four regions (anterior to parietal, Fig. 2) around the midline electrodes. We analyzed the 300-500ms window from the onset of the critical noun as the N400 window, and the 600-1000ms window as the PNP window, using linear mixed-effects modeling. As expected, the N400 in the mid-frontal and mid-posterior regions is modulated by cloze probability (expectation) in both the classifier and verb conditions, with a larger N400 to the unexpected plausible noun than to the expected noun (CL: $ps < .05$; VB: $ps < .05$), and a larger N400 to the anomalous noun than to the unexpected one (CL: $ps < .05$; VB: $ps < .05$). In the PNP window, in the anterior and mid-frontal regions, a larger PNP effect is elicited by unexpected but plausible nouns than anomalous nouns in both classifier ($ps < .01$) and verb ($ps < .001$) conditions. Moreover, we found a larger PNP to the unexpected nouns in the high constraint context than in the low constraint context in the mid-frontal region. However, this effect is only present for the verb-noun structure ($p < .05$), and not in the classifier-noun structure ($p = 0.7$). These findings suggest that revising unfulfilled predictions in the verb-noun structure is more costly, likely due to the fact that predictions made based on verb information involve a richer set of semantic features, making it more difficult to inhibit the originally predicted representation and successfully shift to an alternative.

Conclusion: Examining two different structures in Chinese, we replicated previous findings that the anterior-frontally distributed PNPs index the cost of integrating unexpected but plausible words into context. More importantly, we also showed that the costs of revising unfulfilled predictions is modulated by the type of predictive cues.

Table 1. Experimental stimuli. Numbers in parenthesis indicate cloze probabilities of the corresponding noun. The HC classifier example here signals an object with a flat shape, the LC classifier example signals a more generic shape.

	High Constraint (HC)			Low Constraint (LC)	
	High Cloze (Exp)	Low Cloze (Unexp)	Anomalous (Anom)	Low Cloze (Unexp)	Anomalous (Anom)
Classifier-N (CL)	一扇门 (.52) one-CL door	一扇猪肉 (.02) one-CL pork	一扇水果 (.0) one-CL fruit	一块蛋糕 (.02) one-CL cake	一块水 (.0) one-CL water
Verb-N (VB)	激化矛盾 (.51) intensify conflict	激化能量 (.02) intensify energy	激化灯 (.0) intensify lamp	影响贸易 (.01) influence trade	影响时间 (.0) influence time

Fig. 1. Average ERP waveform from 200ms before to 1000ms after the onset of nouns following classifiers at Fz (1a), Cz (1b) Pz (1c) and verbs at Fz (2a), Cz (2b), Pz (2c).

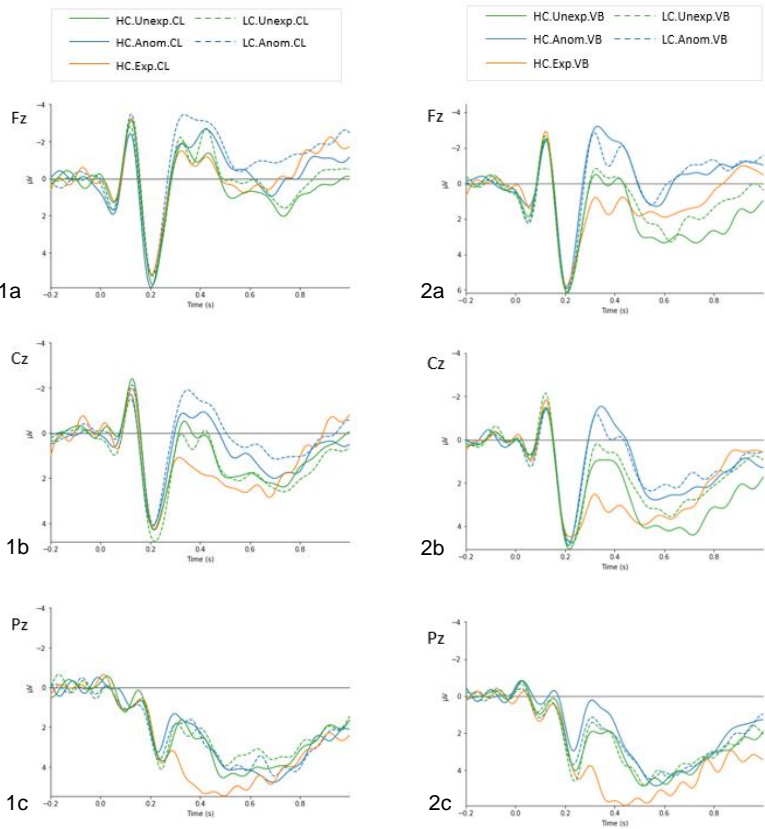
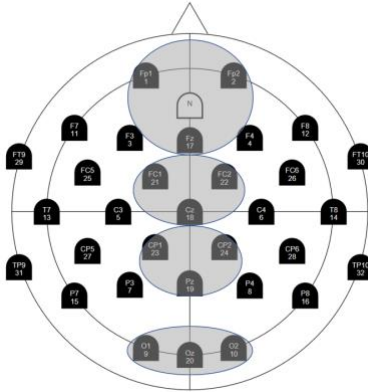


Fig. 2. Midline area channels included in the data analysis, with four regions from front to the back: Anterior, mid-frontal, mid-posterior, parietal.



Reference

[1] Kutas, M., & Federmeier, K. D. (2011). *Annual review of psychology*.
[2] Van Petten, C., & Luka, B. J. (2012). *International Journal of Psychophysiology*.
[3] DeLong, K. A., & Kutas, M. (2020). *Language, Cognition and Neuroscience*.
[4] DeLong, K. A., Quante, L., & Kutas, M. (2014). *Neuropsychologia*.
[5] Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). *Brain research*.
[6] Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). *Journal of Cognitive Neuroscience*.

Dissociating Effects of Predictability, Preview and Visual Contrast on Eye Movements and ERPs

Jon Burnsky¹, Franziska Kretschmar², Erika Mayer¹, Lisa Sanders¹ and Adrian Staub¹ (¹UMass Amherst, ²Leibniz Institute for the German Language, Germany & University of Cologne)

A predictable word receives shorter eye fixations in reading [1] and reduced N400 amplitudes in ERP experiments [2]. The effect on eye fixation durations appears to be dependent on valid parafoveal preview [3]; when a reader is provided with an invalid preview of the target word using the boundary paradigm [4] the predictability of the target no longer influences fixation durations. Nevertheless, predictability modulates the N400 in ERP experiments using the RSVP paradigm, in which there is no preview of upcoming words. We utilized a coregistration [5] paradigm where participants' eye movements and EEG are simultaneously recorded. Based on previous results, we predicted that the predictability effect on the N400 should persist with invalid preview, but the predictability effect on fixation durations should not. In a second experiment, we manipulated predictability and visual contrast, to further explore how low-level properties of the text may differentially influence eye movements and the N400 in normal reading. In [6] these variables demonstrated additive effects on eye fixations durations. We expected to replicate this pattern, while assessing whether contrast influences the amplitude or latency of the N400 [cf. 7].

In Experiment 1 participants ($N_{\text{subjects}}=33$) read sentences ($N_{\text{items}}=180$) distributed in a 2x2 design crossing the predictability of the target word (mean cloze = .93 vs mean cloze = .004) and target preview validity (Table 1). A linear mixed effects model (LMEM) of first fixation durations (Figure 5) revealed a significant interaction between predictability and preview validity, replicating [3]. Two sets of fixation-related potentials (FRPs) were created by time-locking the EEG to the onset of fixation on the target word, and the onset of the immediately preceding fixation (typically on the previous word) (Figures 1 and 2). These FRPs collapse across centro-parietal electrodes. The FRPs were analyzed using a standard N400-window (250-500ms) ANOVA; trial-level mixed effects models [8] revealed similar statistical patterns. Predictability reduced N400 amplitude on the target word FRP ($p < .001$), while preview did not have a significant effect ($p=.1$). There was a marginal interaction ($p=.06$). To assess differences in N400 latency, we also conducted ANOVAs in 50ms bins, as in [5], which revealed that the predictability effect began as expected at 350ms. This analysis also revealed an interaction ($p = .01$) in the 500 to 600ms interval; the predictability effect was larger with invalid preview, driven by a positivity in the invalid predictable condition. For the previous fixation FRP, there was a significant effect of preview ($p=.008$) and a significant interaction ($p=.007$), driven by a predictability effect in the valid preview conditions but not the invalid conditions. In sum, the reduced negativity associated with a predictable word appears earlier with valid preview than with invalid preview, but is present in both cases.

Experiment 2 ($N_{\text{subjects}}=25$) used the same procedure and same items, now crossing predictability with visual contrast (Table 2). The eye tracking data are shown in Figure 6; a LMEM confirmed [6] in showing significant and additive effects of predictability and contrast ($ps < .001$). FRPs are in Figures 3 and 4. Analyses of the target word FRP revealed that predictability significantly reduced N400 amplitude ($p=.001$), with stimulus quality showing no significant effect on amplitude or latency. A series of analyses of sequential 50ms bins revealed no evidence that the N400 was delayed in the faint text conditions [7]. There were no significant effects on the FRP for the previous fixation in the N400 window, though qualitatively the patterns for clear text are similar to the patterns for valid preview in Experiment 1 (Figures 1 and 3).

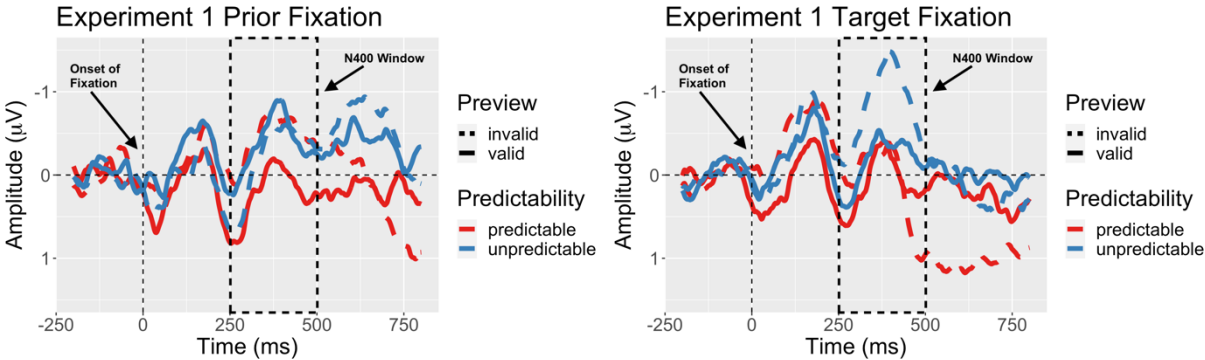
In sum, the usual predictability effect on eye movements in reading is eliminated when parafoveal preview is absent, while the predictability effect on the (foveal) N400 is not. Furthermore, the amplitude and latency of the N400 appears not to be influenced by visual contrast, while eye fixation durations are. These dissociations emphasize that distinct processing events determine eye fixation durations and N400 amplitude and latency. We are in need of a more explicit model of how each of these measures indexes specific stages of lexical processing.

	valid preview (mail → mail)	invalid preview (exit → mail)
predictable target	The package was sent through the {mail <u>mail</u> } two weeks ago.	The package was sent through the {exit <u>mail</u> } two weeks ago.
unpredictable target	If nobody claims the {mail <u>mail</u> } then it will be thrown away.	If nobody claims the {exit <u>mail</u> } then it will be thrown away.

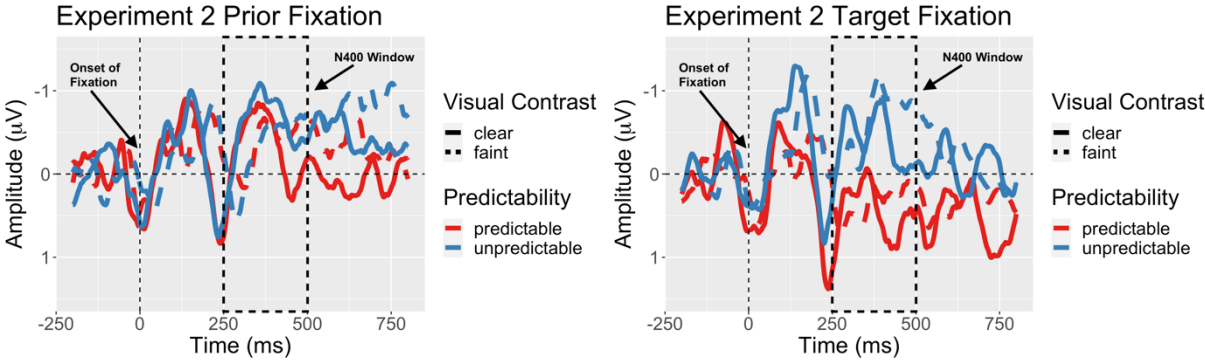
Table 1: Example stimuli from Experiment 1. Previews are to the left of the “|”. The display would show the preview until fixated, at which point the target word “mail” would be displayed.

	clear text	faint text
predictable target	The package was sent through the <u>mail</u> two weeks ago.	The package was sent through the <u>mail</u> two weeks ago.

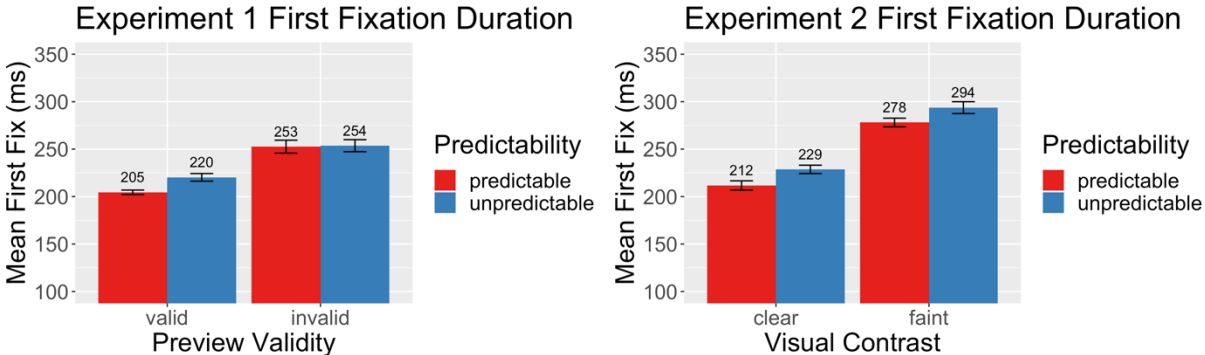
Table 2: Example stimuli from Experiment 2. Unpredictable contexts were used in addition.



Figures 1 (left) and 2 (right): Baseline-Corrected FRPs for the pre-target and target fixations.



Figures 3 (left) and 4 (right): Baseline-Corrected FRPs for the pre-target and target fixations.



Figures 5 (left) and 6 (right): First Fixation Durations on the critical target word by condition.

[1] Staub (2015), [2] Kutas & Federmeier (2011), [3] Staub & Goddard (2019), [4] Rayner (1975), [5] Kretschmar et al. (2015), [6] Staub (2020), [7] Holcomb (1993), [8] Alday (2019)

Modeling influences of coercion on N400 amplitudes as change in a probabilistic representation of meaning

Milena Rabovsky (University of Potsdam, Germany)

Coercion has been defined as ‘a semantic operation that converts an argument to the type that is expected by a function, where it would otherwise result in a type error’ [1, p. 425]. An example of complement coercion is given by the sentence ‘The journalist began the article’ where the predicate ‘began’ would require its complement to denote an event, but ‘the article’ instead denotes an entity. Thus, ‘began’ coerces ‘the article’ from an entity to an event involving this entity, allowing for the interpretation ‘The journalist began writing the article’. Influences of complement coercion on event related brain potentials (ERPs) have been investigated by presenting sentences such as ‘The journalist began/ wrote/ accomplished the article’ (i.e. ‘coerced’/ ‘non-coerced’/ incongruent) and comparing ERPs at the noun [2]. The authors observed larger N400s for ‘coerced’ and incongruent as compared to ‘non-coerced’ sentences. The goal of the current study was to investigate whether these observed influences of coercion on N400 amplitudes can be accounted for by the Sentence Gestalt (SG) model, a neural network model of sentence comprehension [3] that has previously been used to account for a broad range of N400 effects (Fig. 1; [4,5,6]).

The training environment of the SG model, which is based on a simple generative model (see [4] for details), was extended to include coercion like situations. Specifically, two additional verbs were included in the model’s vocabulary (‘begin’ and ‘finish’), which could be combined with all other verbs such as e.g., in ‘The man began/ finished reading the novel/ planting the rose/...’. For some sentences, such as the example sentence with the novel, complement coercion is possible, and the gerund was sometimes (with .2 probability) omitted. Ten independently initialized models were each trained on 800.000 example sentences produced by the simple generative model. For the simulation experiments, the ten trained models were each presented with 8 triplets of stimuli designed to mimic the ERP study reported above, e.g., ‘The man began/ read/ ate the novel’ (i.e., ‘coerced’/ ‘non-coerced’/ incongruent). The model’s N400 correlate, which is the magnitude of change in the model’s hidden SG layer induced by the current word (i.e., $Model\ N400 = |SG_t - SG_{t-1}|$), was compared at the noun.

The model’s N400 correlate was larger for ‘coerced’ and incongruent as compared to ‘non-coerced’ sentences over models and items ($ts > 6.6$, $ps < .001$; see Fig. 2), in line with the empirical data [2]. Note that this is the case despite the fact that the SG model does not assume a specific process such as ‘coercion’ to explain the interpretation of these sentences. Because the model does not assume fixed rules, no operation is required to prevent a presumed rule violation such as a type error. Instead, the model constantly estimates the probabilities of all relevant aspects of meaning involved in the described event based on the statistical structure of its environment, including aspects that are not explicitly mentioned in the sentence. It does not contain fixed lexical representations of words that would need to be converted into something else. Instead, each incoming word provides cues constraining the overall interpretation of the sentence. The model’s N400 correlate for sentences containing ‘coercion’ thus does not reflect any specific ‘coercion’ process of converting an argument into another type, but rather reflects the same process assumed to underlie N400 amplitudes in general from the model’s perspective, namely the amount of change in expected sentence meaning induced by the critical word. The amount of change was larger for ‘coerced’ as compared to ‘non-coerced’ sentences because the ‘coerced’ sentences were of lower constraint and lower cloze probability as was the case in the empirical study [2] (see also [6] and [7]). Thus, from this perspective, the available evidence reporting effects of complement coercion on ERPs (see also [8]) does not speak to the neurocognitive reality of this construct from compositional semantics.

References

- [1] Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.
- [2] Kuperberg, G.R., Choi, A., Cohn, N., Paczynski, M., & Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience*, 22, 2685–2701.
- [3] St. John, M.F., McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217–257.
- [4] Rabovsky, M., Hansen, S.S., & McClelland, J.L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2, 693-705.
- [5] Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, 143, 107466.
- [6] Rabovsky, M., & McClelland, J. L. (2020). Quasi-compositional mapping from form to meaning: a neural-network based approach to capturing neural responses during language comprehension. *Philosophical Transactions of the Royal Society B*. 375: 20190313.
- [7] Delogu, F., Crocker, M.W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: evidence from eye-movements and ERPs. *Cognition*, 161, 46–59.
- [8] Baggio, G., Choma, T., van Lambalgen, M., & Hagoort, P. (2009). Coercion and compositionality. *Journal of Cognitive Neuroscience*, 22, 2131-2140.

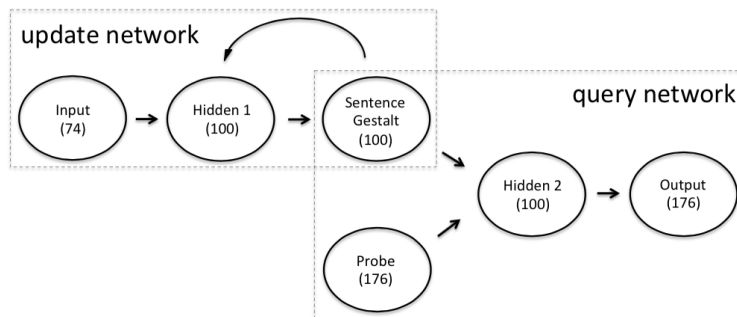


Figure 1. The Sentence Gestalt (SG) model architecture. Arrows represent all-to-all modifiable connections and ovals represent layers of units (with numbers of units in parentheses).

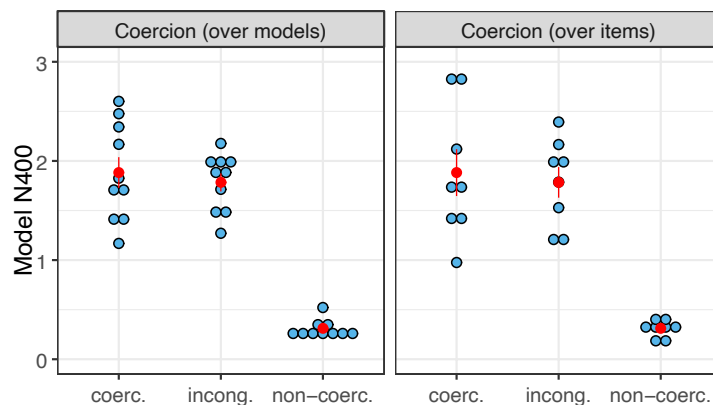


Figure 2. Influences of coercion on the SG model's N400 correlate. Blue dots represent results for models ($n = 10$, left) and items ($n = 8$, right); red dots represent condition means \pm standard error of the mean (SEM).

Neural correlates of expectation violations and discourse updating: The case of Bulgarian object agreement

Paul Compensis, Petra B. Schumacher (University of Cologne)

A core feature of language is to identify as quickly and unambiguously as possible *who did what to whom* and to keep track of participating referents as discourse unfolds. Thereby, the processing of argument structure is driven by contextual predictions as well as cue-based attention shifts. In order to facilitate predictions or highlight attention shifts, most languages use word order, case and/ or verb agreement, indicating to the listener which role a referent currently fulfils in a sentence. While order, case and subject-verb agreement received considerable attention in empirical linguistics, object agreement still requires experimental investigation.

In Bulgarian, objects in pre-verbal position are frequently marked with an object clitic (traditionally known as clitic doubling, CLD) [1]. Interestingly, these clitics can either serve as stand-alone pronouns in subject-clitic-verb (SCV) constructions (a) or as object agreement markers co-referring to an object NP in the same sentence (b). However, in line with the *subject-first preference* [2] we assume that in the case of two equally ranking referents an initial NP is always interpreted first as the subject of a sentence. In the case of CLD, the presence of the object agreement marker should enforce a reanalysis towards an object-initial interpretation. However, due to the normatively marked nature of CLD in Bulgarian, the validity of the object agreement marker as an attentional cue in role interpretation is a matter of debate. In order to test the online processing of CLD, we conducted an ERP study in which we contrasted SCV and CLD to reference mismatches (RFM, example c) and agreement violations (AGV, example d).

Previous ERP research found that cross-linguistically RFM and AGV typically engender an N400 followed by a late positivity (LPS) [3]. In general, the N400 component correlates with stimulus predictability in language processing [4] and particularly with expectation-based linking mechanisms with respect to referents [5]. The LPS is associated with reanalysis, also during referent shifts and discourse updating [5]. We assumed that CLD engenders a similar pattern due to its lower cue availability (reflected in the N400 pattern) and subsequent reanalysis (reflected in the LPS). However, these effects should be less pronounced in comparison to AGV and RFM.

In our ERP study, 20 participants read a context sentence introducing two referents of different gender and a target sentence (in either of the four conditions) presented as RSVP for 450 ms per word. Each target sentence started with a NP referring to one of the two referents from the context sentence. By manipulating grammatical gender of the clitic and the verbal ending, either the object or the subject agreement marker or both agreed with the gender of the first or second referent (or potentially with a third, non-specified referent), leading to the four conditions exemplified in (a-d). This allowed for testing expectation violations and discourse updating for both agreement types. 40 stimuli per condition were presented in segments and ERPs were measured time-locked to both agreement markers. After pre-processing, we calculated linear mixed-effect models with mean fitted values from 0 to 1000 ms in steps of 100 ms as dependent variable and CONDITION as fixed factor as well as two continuous factors SAGITTALITY and LATERALITY.

As predicted, the initial occurrence of the divergent clitic (in RFM and CLD) engendered an N400-LPS pattern at the position of the clitic (*ja/go*), indicating an expectation violation followed by an attempt to resolve the interpretation by searching for a new referent. At the subject agreement position (*napusna-lla*), a graded N400 effect (SCV < CLD < RFM < AGV) and a graded LPS effect (SCV < RFM/CLD < AGV) emerged for the non-canonical conditions. Thus, this study replicated previous findings concerning AGV and RFM and, in addition, showed that reanalysis towards an object-initial order by means of an object agreement marker (CLD) causes a smaller expectation violation than RFM and AGV, also reflecting some cue availability of CLD, but also causes discourse updating that is comparable to the establishment of reference to a non-specified referent (as in RFM).

Example stimuli for the four conditions (with clitic and agreement positions in bold):

Context sentence: Did you hear the news about Petar and Marija?

(a) Subject-Clitic-Verb (SCV)

Petar **ja** e napusna-l sled sporovete.

Petar.M **she.ACC** leave-PTCP.M after the argument.

'Petar left her after the argument.'

(b) Clitic doubling/ Object agreement (CLD)

Petar **go** e napusna-la sled sporovete.

Petar.M **he.ACC** leave-PTCP.F after the argument.

'She left (him) Petar after the argument.'

(c) Reference mismatch (RFM)

#Petar **go** e napusna-l sled sporovete.

Petar.M **he.ACC** leave-PTCP.M after the argument.

'Petar left him after the argument.'

(d) Agreement violation (AGV)

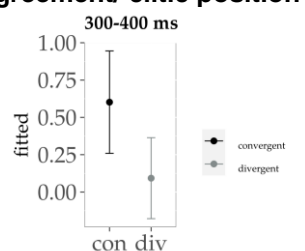
*Petar **ja** e napusna-la sled sporovete

Petar.M **she.ACC** leave-PTCP.F after the argument

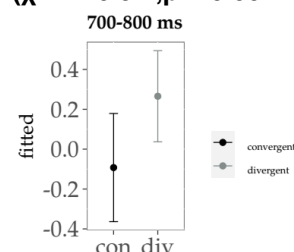
'Petar (she) left her after the argument.'

ERP effect plots in relevant time windows:

Object agreement/ clitic position

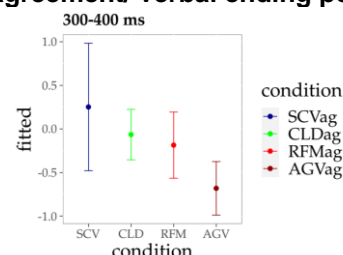


($\chi^2 = 10.32$, $p = 0.001^{**}$)

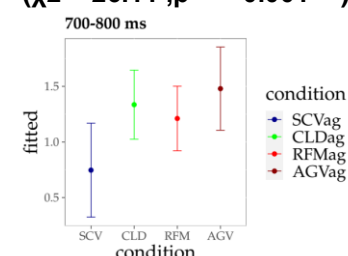


($\chi^2 = 6.34$, $p = 0.012^{*}$)

Subject agreement/ verbal ending position

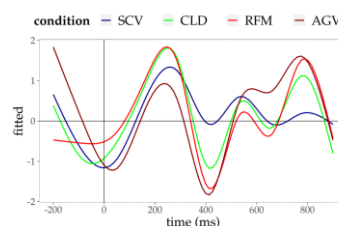
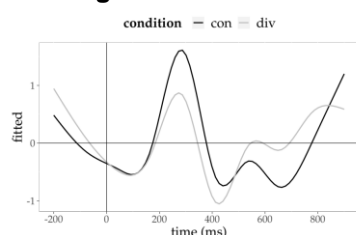


($\chi^2 = 26.11$, $p < 0.001^{***}$)



($\chi^2 = 24.8$, $p < 0.001^{***}$)

Grand-average ERPs at electrode Pz:



References:

- [1] Guentchéva, Z. (1994). *Thématisation de l'objet en bulgare*. Peter Lang.
- [2] Bickel, B., Witzlack-Makarevich, A., Choudhary, K. K., Schlesewsky, M., Bornkessel-Schlesewsky, I. (2015). The Neurophysiology of Language Processing Shapes the Evolution of Grammar: Evidence from Case Marking. *PLoS ONE*, 10(8): e0132819. <https://doi.org/10.1371/journal.pone.0132819>
- [3] Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of cognitive neuroscience*, 16(7), 1272–1288. <https://doi.org/10.1162/0898929041920487>
- [4] Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Frontiers in Psychology*, 10(298), 1–17. <https://doi.org/10.3389/fpsyg.2019.00298>
- [5] Hung, Y. C., & Schumacher, P. B. (2012). Topicality matters: position-specific demands on Chinese discourse processing. *Neuroscience letters*, 511(2), 59–64. <https://doi.org/10.1016/j.neulet.2012.01.013>

Feature Reactivation in Minimalist Parsing

Aniello De Santo (University of Utah)

Overview A top-down parser for Minimalist grammars [MGs; 9] can successfully predict a variety of off-line processing preferences, via metrics linking parsing behavior to memory load [6, 2, 4]. Given the close association between this model and modern minimalist syntax, it is important to extensively evaluate its empirical coverage. In this abstract we propose new metrics for the MG parser, that take into account the set of features triggering movement steps in a derivation — thus implementing a notion of memory reactivation. As a case study of how these metrics improve the empirical coverage of the MG approach, we successfully model the processing preferences for stacked relative clauses (RC) in [11], and a variety of previously modeled RC contrasts.

MG Parsing The MG parsing model systematically links syntactic structure to processing difficulty by connecting the stack states of a (deterministic) top-down parser [9] to memory burden. Memory usage [3, 7] is measured based on how long a node is kept in memory (**tenure**). Consider the MG derivation in Fig. 1. The index of a node n encodes the moment n was predicted and put in memory by the parser. The outdex encodes the moment n is confirmed and flushed out of memory. Tenure for n is measured as $outdex(n) - index(n)$, and can then be used to define a set of off-line metrics of processing difficulty (e.g., **max.** or **avg.** tenure across all nodes in the derivation [4]).

Implementing Feature Reactivation We want to make the parsing model sensitive to structural repetition. Inspired by previous literature on syntactic priming, we stipulate that if a moved element has been recently stored in memory, storing the next item of the same kind (e.g., triggered by a *wh*-feature) should be less costly (feature *reactivation*). Note that these items are not in memory at the same time, so this is different from interference effects. We implement this procedure by counting the number of parsing steps between movements of the same type. Consider the derivation in Fig. 2, with two NP movers associated to a feature f . Practically, reactivation for NP_2 is measured by subtracting from its index the outdex of the previous node associated to f (NP_1 ; so $w - y$). Finally, since reactivation is supposed to encode facilitatory effects induced by structural repetition, we operationalize it as: $R(m_i) := 1 - \frac{1}{i(m_i) - o(m_{i-1})}$. Additionally, we weight the tenure of a node by its reactivation value (*boost*, $BT := Tenure(m_i) * R(M_i)$), to investigate the interaction between reactivation and notions of storage previously employed by the MG parser. We then derive metrics that use reactivation and boost to compute processing costs over full derivations (e.g., **max.** R).

A Case Study We consider stacked RC constructions, in which a noun phrase (*the reporter*) is modified by two relative clauses. Zhang [11] explores the processing of stacked RCs in English (1) and Mandarin Chinese (2), in a 2×2 design crossing extraction type (subject or object) with the position of the RC (RC1 or RC2). She reports faster reading times when RC1 and RC2 are of the same type, than when they are of different types (i.e. $SS > OS$ and $OO > SO$). Crucially, none of the metrics used in the previous MG parsing literature is able to account for this effect. We model these contrasts as in (1) and (2), and we also consider a classical contrast between subject (SRC) and object (ORC) RCs both in English and Mandarin, which has been focus of much MG processing work in the past [4, 11, a.o.]. Since the parser is sensitive to detailed grammatical information, we consider two analyses for the RC construction: a promotion analysis [5], and a *wh*-movement analysis [1]. Our simulations show that the parser now successfully captures the facilitatory effect associated to consecutive processing of similar movement types (ORC-ORC; SRC-SRC), as well as the more classical SRC-ORC contrasts. We discuss how these results relate to the way different reactivation metrics are sensitive to differences between syntactic analyses. This extension to the computational model will clearly require extensive empirical evaluation. However, these results provide a valuable proof-of-concept in favor of a careful exploration of how ideas from the priming literature can be incorporated in formal models of structural processing.

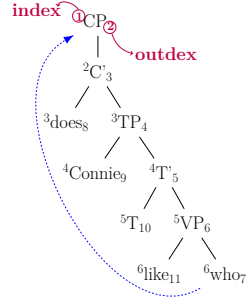


Figure 1: MG derivation tree with parse steps.

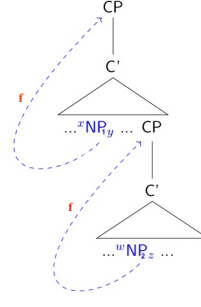


Figure 2: Example tree for memory reactivation.

(1) Test sentences for English Stacked Relative Clauses

- The horse that kicked the wolf on Tuesday that patted the lion just now went home **SS**
- The horse that the wolf kicked on Tuesday that patted the lion just now went home **OS**
- The horse that kicked the wolf on Tuesday that the lion patted just now went home **SO**
- The horse that the wolf kicked on Tuesday that the lion patted just now went home **OO**

(2) Example of test sentences for Mandarin Chinese Stacked Relative Clauses

- Nage zai xingqier tile xiaoma haojici de zai jintian zhuile daxiang
Dem on Tuesday kick-perf horse several-times de on today chase-perf elephant
de gongniu likaile jia
De bull leave-perf home
'The bull that kicked the horse for several times on Tuesday that chased the elephant earlier today left home.' **SS**
- Nage zai xingqier xiaoma tile haojici de zai jintian zhuile daxiang
dem on Tuesday horse kick-perf several-times de on today chase-perf elephant
de gongniu likaile jia
De bull leave-perf home
'The bull that the horse kicked for several times on Tuesday that chased the elephant earlier today left home.' **OS**

Language	Processing Contrast	$\langle \text{MaxR}', \text{AvgBT} \rangle$		$\langle \text{MaxBT}, \text{MaxR}'_R \rangle$	
		Promotion	Wh-movement	Promotion	Wh-movement
English	$OO < SO$	✓	✓	✓	✓
	$SS < OS$	✓	✗	✓	✓
Mandarin	$OO < SO$	✓	✓	✓	✓
	$SS < OS$	✓	✓	✗	✓
English	$SRC < ORC$	✓	✗	✓	✓
Mandarin	$ORC < SRC$	✓	✓	✗	✓

Table 1: Summary of results of ranked metrics by contrast and RC construction.

[1] Chomsky, N. (1977). On wh-movement. *Formal syntax*. [2] Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*. [3] Graf T., J. Monette, and C. Zhang. (2017). Relative clauses as a benchmark for Minimalist parsing. *Journal of Language Modelling*. [4] Kayne, R.S. (1994). The antisymmetry of syntax. *MIT Press*. [5] Kobele, G.M., S. Gerth, and J. Hale. (2013) Memory resource allocation in top-down minimalist parsing. *Formal Grammar*. [6] Rambow, O. and Joshi, A.K. (2015). A processing model for free word-order languages. *Perspectives on sentence processing*. [7] Reitter, D., Keller, F., and Moore, J. D.. (2011). A computational cognitive model of syntactic priming. *Cognitive science*. [8] Stabler, E. P. (2013). Two models of minimalist, incremental syntactic analysis. *Topics in cognitive science*. [9] Troyer M., O'Donnell, T.J., Fedorenko, E., and Gibson E. (2011). Storage and computation in syntax: Evidence from relative clause priming. *Proc. of the Cognitive Science Society*. [10] Zhang, C. (2017). *Stacked Relatives: Their Structure, Processing and Computation*. PhD thesis, Stony Brook U.

Can English Idioms Undergo the Dative Alternation? A Priming Investigation

Breanna Pratley, Philip J. Monahan (University of Toronto)

Theoretical Motivation. To date, though priming has been demonstrated to target syntax (Bock & Loebell, 1990; Pickering et al., 2002), priming experiments are equivocal to the level of structural representation targeted (Pickering & Ferreira, 2008). To better understand this issue, we examine competing syntactic representational hypotheses using priming and probe the level of abstract representation that priming targets. We use the English dative alternation as our test case, which has two structural options: Double Object (DO) and Prepositional Dative (PD). Idioms with verbs that should alternate are cited as being restricted to the DO (1a-1b; Richards, 2001). This restriction is often used as evidence to support theories in which the DO and PD are construed as categorically distinct (Harley, 1997; among others), and evidence against theories that analyse dative structures as derivationally related (Larson, 1988; among others); however, these same idioms appear to take on the PD form when the sentence involves heavy NP shift (1c). Bresnan and Nikitina (2007) take idioms in the PD form as evidence for derivational and probabilistic theories of the dative alternation. There is debate, however, about whether idioms like (1c) are truly PDs. An alternative hypothesis is that (1c) is a type of DO that has undergone *Rightward Dative Shift* (Figure 1; Bruening, 2010). Crucially, this construction is structurally a DO, but with the thematic goal projected to the right. This results in a surface order akin to the PD. The potential mismatch between surface word order and abstract structure makes idioms like (1c) a useful test case with which to understand whether priming targets a more abstract level of syntactic structure.

Current Experiment. To determine the structural representation of idioms like (1c) and investigate the depth of syntax that priming targets, we conducted a two-alternative forced-choice priming experiment. Primes were displayed in one of four conditions: Prepositional Dative, Double Object, Rightward Dative Shift, and a Control Condition (Table 1), and each trial included two test options: DO and PD. If idioms like (1c) are truly PD, then the results of the Rightward Dative Shift Condition should pattern like the results of the PD Condition. As such, if a Rightward Dative Shift prime (1c) results in fewer PD responses than a PD prime, idiomatic sentences in this form are not likely to have a PD structure. Our results suggest that these idioms are not structurally similar to PD, and thus cannot entirely undergo the dative alternation.

Methods. Native English-speaking participants ($n=40$) completed 144 trials. In each trial, they were shown a sentential prime, followed by a forced-choice picture description task. We created four lists in a Latin Square design. In each trial, participants read the prime aloud, then chose which of two sentences better described a drawing. Test sentences were presented in the lower portion of the screen, differed only in structure, and were counter-balanced for side of presentation. 48 trials tested the dative alternation, and 96 filler trials tested active/passive priming. Trials testing active/passive priming were included to ensure that the task was effective.

Results. Results were submitted to a linear mixed effects model with a logistic regression function (Jaeger 2008), including a fixed effect of condition, and a maximal random effects structure. Significant priming effects were found in the active/passive condition, ($\Delta=19\%$ between Active and Passive conditions), confirming task validity. Test trials after PD primes resulted in significantly more PD responses than after DO ($\Delta=8\%$, $\beta=0.36$, $SE=0.14$, $z=2.58$, $p<0.01$) or Control primes ($\Delta=6\%$, $\beta=-0.29$, $SE=0.14$, $z=-1.99$, $p<0.05$), see Figure 2. There was no difference between the Rightward Dative Shift condition and any other prime condition.

Implications. In our experiment, the PD response rate following a Rightward Dative Shift prime is not different from a PD prime; however, unlike PD primes, it is also not different from a DO prime. These results point to many influences in syntactic priming, including perhaps lexical overlap of *to* in both the Rightward Dative Shift and PD Conditions (Pickering & Branigan, 1998), and potential differences in semantics between conditions. If these idioms were truly PD, however, the rate of PD responses in the Rightward Dative Shift Condition should be different from the DO Condition. This suggests that, though it is unclear whether the structure in Figure 1 is responsible, idioms like (1c) are not true PD structures (cf. Bresnan & Nikitina 2007), which ultimately lends some support to theories which construe the dative alternation as distinct

structures, and interestingly suggests that syntactic priming may be sensitive to a more abstract level of structure.

References

- Bock, Kathryn, and Helga Loebell. 1990. "Framing Sentences." *Cognition*, vol. 35, no. 1, pp. 1-39., doi: 0010-0277/90
- Bresnan, Joan, and Tatiana Nikitina. 2007. "The gradience of the dative alternation." *Reality exploration and discovery: Pattern interaction in language and life*, ed. By Linda Uyeche and Lian Hee Wee. Stanford, CA: CSLI Publications.
- Bruening, Benjamin. 2010. "Double Object Constructions Disguised as Prepositional Datives." *Linguistic Inquiry* 41(2): 287-305.
- Harley, H. 1997. "If you have, you can give." In *Proceedings of the 15th West Coast Conference on Formal Linguistics*, ed. by Brian Agbayani and Sze-Wing Tang, 193-207. Stanford, CA: CSLI Publications.
- Jaeger, T. F. 2008. "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models." *Journal of Memory and Language*, 59(4), 434-446.
- Larson, R. 1988. "On the Double Object Construction." *Linguistic Inquiry*, vol. 19 no.3, pp. 335-391. www.jstor.org/stable/25164901
- Pickering, M. J., & Branigan, H. P. 1998. "The Representation of Verbs: Evidence from Syntactic Priming in Language Production." *Journal of Memory and Language*, 39(4), 633-651. doi:10.1006/jmla.1998.2592
- Pickering, M. J., Branigan, H.P. and McLean, J. F. 2002. "Constituent Structure is Formulated in One Stage." *Journal of Memory and Language*, 46, 586-605, doi:10.1006/jmla.2001.2824
- Pickering, M. J. and Ferreira, V. S. 2008. "Structural Priming: A Critical Review." *Psychological Bulletin*, 134(3), 427-459, doi: 10.1037/0033-2909.134.3.427
- Richards, N. 2001. "An Idiomatic Argument for Lexical Decomposition." *Linguistic Inquiry*, 32(1), 183-192, doi:10.1162/002438901554649

Examples and Figures

- (1) a. The lighting here gives me a headache.
 b. *The lighting here gives a headache to me.
 c. The lighting here gives a headache to everyone in the room.
 (Bresnan & Nikitina 2007)

Figure 1

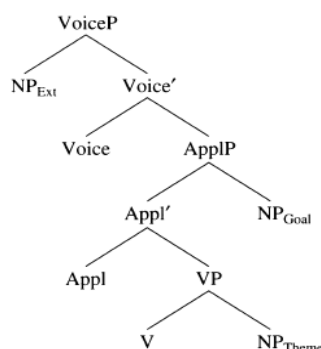


Table 1


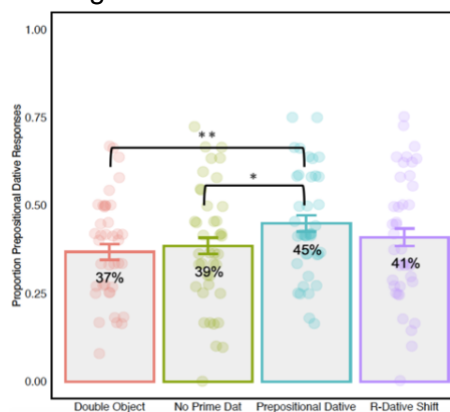
Prime Condition	Example Prime	Test Trial
Double Object	The conductor gave the quiet girl on the evening train the ticket	 <p>The man gave The man gave a</p>
Prepositional Dative	The conductor gave the ticket to the quiet girl on the evening train	
Rightward Dative Shift	The conductor gave the creeps to the quiet girl on the evening train	
Control	Fully flowery and intricately patterned	

Figure 2



		the child a cookie. cookie to the child.
--	--	--

The effect of representational complexity on working memory processes

Chi Dat (Daniel) Lam and Ming Xiang (University of Chicago)

Background. Working memory (WM) processes - encoding, maintenance and retrieval - are essential for sentence comprehension, especially for long-distance dependencies. It is an open question how representational complexity affects these processes. Some previous studies have argued that representational complexity increases encoding effort but decreases retrieval cost [1,2]. For example, it was found in [1] that reading time (RT) is longer when encoding a complex noun phrase (NP, e.g., *an alleged communist*) than a simpler one (e.g., *a communist*); but at a later retrieval site, retrieving the more complex NP antecedent elicits faster RTs. However, the effect of complexity on encoding and the trade-off between encoding and retrieval have only been observed in a limited set of constructions. The current study investigates whether the reported effects can be generalized by comparing coordinated NPs (e.g., *those judges and lawyers*), with simple NPs (e.g., *those lawyers*). Different from the findings in [1], our results showed that the encoding stage of the coordinated NPs, which are syntactically and semantically more complex, was facilitated (faster RTs); and their maintenance, but not retrieval, was also facilitated.

Experiment 1. We tested how complexity of an extracted NP affects the three WM processes in subject and object relative clauses (SRC and ORC). In English, it is known that ORCs pose more processing difficulty [3]. The experiment had a 2 (SRC/ORC) x 2 (complex/simple NP) design (Examples in (1)). 94 participants from Prolific performed a self-paced reading task with 32 experimental items and 32 fillers. Each sentence was followed by a comprehension question targeting the dependency. Raw RTs were first log transformed and residualized based on sentence position. We examined the **encoding region** (extracted NP *lawyers* and its spillover *who*), the **retrieval region** (RC verb), and the **maintenance region** (words between the encoding and retrieval sites). Bayesian statistical analyses using *brms* [4] were performed, with RTs on the previous word, NP type and RC type (both sum-coded) as fixed effects and a full random effect structure. In the **encoding region**, RTs on the extracted NP's final word *lawyers* showed an effect of NP type ($\beta=0.10$, 95% CrI[0.05, 0.15], complex < simple) (Fig 1). In the **maintenance region**, RTs on the adverb *harshly* showed an effect of NP type ($\beta=0.04$, [0.01,0.07]) and an NP x RC interaction ($\beta=-0.05$, [-0.09,-0.003]), driven by the fact that RTs for the complex NP conditions were faster in ORCs ($\beta=0.05$, [0.02,0.10], but not in SRCs. In the **retrieval region**, RTs on the verb *admitted* only showed an RC-type effect that ORCs are more difficult ($\beta=-0.11$, [-0.15,-0.07]).

Experiment 2 One difference between the RCs at *harshly* in Exp 1 is that for ORCs, there are two referents to be maintained as distinct representations, whereas for SRCs, the extracted NP is the only referent. In Exp 2 (n=75), we used the same design as Exp 1 with an additional embedding clause (*who John thinks*) so that additional referent(s) are present in both RC types. In the **encoding region**, we replicated the faster RTs on *lawyers* in the complex NP conditions, ($\beta=0.08$, [0.03,0.14]) (Fig 2). In the **maintenance region**, RTs on *thinks* were also faster in the complex NP conditions ($\beta=0.04$, [0.01,0.06]). RTs on *harshly* showed an RC type effect ($\beta=-0.09$, [-0.13,-0.05], ORC>SRC) and a marginal RC x NP interaction ($\beta=0.05$, [-0.001,0.11]), driven by faster RTs for complex NP condition only in SRCs ($\beta=0.05$, [0.01,0.09]). In the **retrieval region**, there was again only an RC type effect on *admitted* ($\beta=-0.07$, [-0.12,-0.02], ORC > SRC).

Discussion. Both experiments showed a speed-up for the more complex NP in the **encoding region**, contrary to the slowdown effect in [1]. Further analyses ruled out lexical priming from *judges* as the source of the speed-up, as semantic similarity between the two conjunct nouns did not predict RTs at *lawyers*. In the **retrieval region**, there was no facilitation due to complexity, again contrary to [1]. We are currently conducting a conceptual replication of [1] using the original adjective-noun structure. In the **maintenance region**, we hypothesize that richer features on complex NPs make it easier to maintain distinct representations of the extracted NP and another intervening referent. This facilitation was shown on *thinks* in Exp 2 and on *harshly* in ORCs in Exp 1 and SRCs in Exp 2, all of which require maintenance of two distinct referents. The facilitation effect of the complex NP diminishes, however, when the maintenance difficulty is overloaded with the addition of a third referent, as on *harshly* in ORCs in Exp 2.

Materials. (Encoding region in red, maintenance region in green, retrieval region in blue.)

(1) Experiment 1.

SRC, complex: It seems / that / those judges / and lawyers / who / harshly / reprimanded / Andy / today / admitted / the error.

ORC, complex: It seems / that / those judges / and lawyers / who / Andy / harshly / reprimanded / today / admitted / the error.

SRC, simple: It seems / that / those lawyers / who / harshly / reprimanded / Andy / today / admitted / the error.

ORC, simple: It seems / that / those lawyers / who / Andy / harshly / reprimanded / today / admitted / the error.

Comprehension question (complex): Was it Andy who reprimanded those judges and lawyers?

Comprehension question (simple): Was it Andy who reprimanded those lawyers?

(2) Experiment 2.

SRC, complex: It seems / that / those judges / and lawyers / who / John / thinks / harshly / reprimanded / Andy / today / admitted / the error.

ORC, complex: It seems / that / those judges / and lawyers / who / John / thinks / Andy / harshly / reprimanded / today / admitted / the error.

SRC, simple: It seems / that / those lawyers / who / John / thinks / harshly / reprimanded / Andy / today / admitted / the error.

ORC, simple: It seems / that / those lawyers / who / John / thinks / Andy / harshly / reprimanded / today / admitted / the error.

Comprehension question (complex): Was it Andy who John thinks reprimanded those judges and lawyers?

Comprehension question (simple): Was it Andy who John thinks reprimanded those lawyers?

Figures. (Log RTs were residualized with previous word's log RT and word position.)

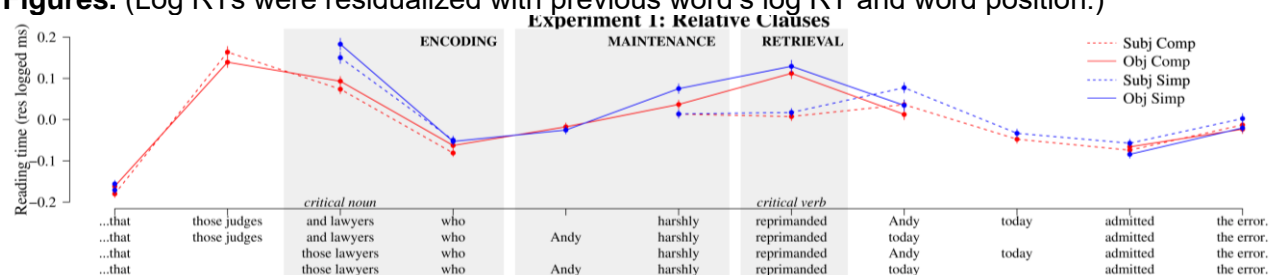


Fig 1. Residualized log RTs from Experiment 1. Error bars indicate +/- 1 standard error

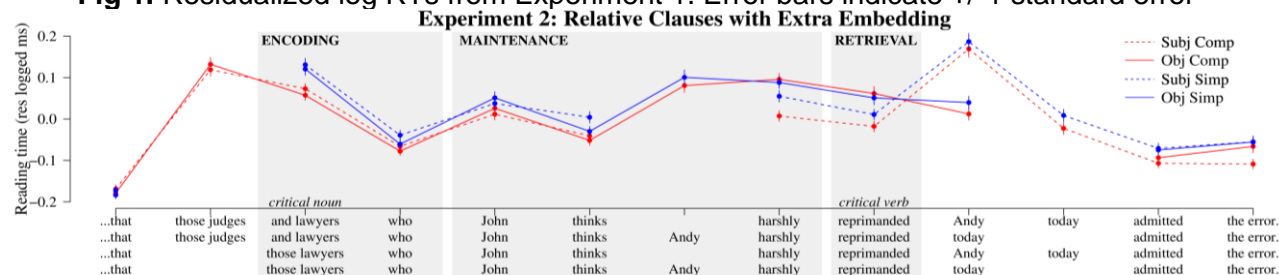


Fig 2. Residualized log RTs from Experiment 2. Error bars indicate +/- 1 standard error

References. [1] Hofmeister, 2011 [2] Hofmeister and Vasishth, 2014 [3] King and Just, 1991 [4] Bürkner, 2017

Null nouns can trigger intervention in Spanish relative clauses' comprehension

Marisol Murujosa (Universidad de Buenos Aires), Carolina Gattei (Universidad de Buenos Aires, Universidad Torcuato Di Tella & Pontificia Universidad Católica Argentina), Diego Shalom (Universidad de Buenos Aires & Universidad Torcuato Di Tella) & Yamila Sevilla (Universidad de Buenos Aires)

Introduction: The asymmetry in the comprehension of subject (S) and object (O) relative clauses (RC) is well documented in literature and seems to be present in a wide range of languages, for example, in Spanish (e.g. Betancort et al., 2009), in French (e.g. Cohen y Mehler, 1996), in Italian (e.g. Contemori & Belletti, 2010), in German (Schriefers, Friederici & Kuhn, 1995), in English (e.g. Gibson, 1998). Since Friedmann, Belletti & Rizzi (2009), but also cf. Grillo (2009), it has been argued, within the featural Relativized Minimality framework (fRM; Rizzi, 2004), that the advantage for SRCs can be explained as an effect of syntactic intervention. As both the subject NP and the object NP are lexically restricted, i.e. they share the [+N] syntactic feature, the subject NP functions as an intervener when the object NP moves to the left periphery, hindering the establishment of the syntactic dependency. Moreover, it has been claimed that when this element is not present, the comprehension of ORCs is facilitated, as it is the case of free RCs in Hebrew (Friedmann, Belletti & Rizzi, 2009). In Spanish, while headed RCs are headed by a fully realized noun, false free RCs are headed by a null noun (Panagiotidis, 2003), which is silent but present in the syntactic structure (Giollo & Muñoz Pérez, 2013). This study aims to answer one main question: can a null noun, with a [+N] syntactic feature, in the object RC head of false free RCs in Spanish hinder the establishment of the syntactic dependency and trigger intervention effects during comprehension? Following the fRM proposal, intervention effects should arise in both types of ORCs, but not in the case of SRCs. **Design:** 33 subjects participated in an auditory sentence comprehension task. They were asked to listen to a sentence; then were showed an image and were prompted to judge whether the image they saw faithfully reflected the content of the sentence heard or not (see Fig. 1). The stimuli (n=20) consisted of both, headed (1) and false free (2), RCs (Type of Antecedent). We manipulated the Type of RC in each case: SRCs (1a and 2a) and ORCs (1b and 2b). The images selected were counterbalanced to make the sentences either true or false. Response accuracy and response times (RTs) were measured during the task. **Results:** on average, participants answered 86% (SE=1.4%) of the total stimuli correctly; Figs. 2 and 3 show mean correct answers and standard error, and mean RTs and standard error (only RTs of correct answers were considered) according to condition respectively. Linear mixed-effect models were fitted for data analysis. Results showed a main effect of Type of RC: ORCs, both headed and false free, were harder to comprehend ($\beta=0.83$, SE=0.18, $z=4.54$, $p<.001$) and showed longer latencies ($\beta=-0.09$, SE= 0.02, $t=4.54$, $p<.001$). A main effect of the variable Type of Antecedent was not found, neither in accuracy ($\beta = -0.16$, SE = 0.17, $z = -0.90$, $p = 0.37$) nor in RTs ($\beta = 0.02$, SE = 0.02, $t = 1.26$, $p = 0.21$). No interaction between both factors (Type of RC x Type of Antecedent) was found, neither in accuracy ($\beta = -0.20$, SE = 0.17, $z = -1.18$, $p = 0.24$) nor in RTs ($\beta = 0.003$, SE = 0.02, $t = 0.45$, $p = 0.66$). **Discussion:** Our results confirm the predictions of the fRM account: in both RCs, headed and false free, intervention effects arose for ORCs (3b and 4b) but not for SRCs (3a and 4a). Although silent, the presence of a noun bearing the [+N] feature triggered the intervention effects predicted in the comprehension of ORCs. Our data are compatible with the results obtained in a similar study carried out in French (Bentea, Durrleman & Rizzi, 2016) where the comprehension of RCs headed by the complex pronominal forms *celui-celle* was evaluated. **Conclusions:** ORCs in Spanish, either headed by a fully realized noun or a silent one, are more difficult to comprehend than SRCs. Our investigation points to a structure-dependent account of the RCs comprehension asymmetries and highlights the importance of studying the comprehension of sentences in a wide array of languages with diverse grammatical properties, showcasing different syntactic configurations.

- (1) a. En la imagen aparece la maestra que le grita a la bruja.
In the image appears the teacher.NOM that CL.DAT yells at the.DAT witch.
'In the image appears the teacher that yells at the witch'.
b. En la imagen aparece la bruja a la que le grita la maestra.
In the image appears the witch the.DAT that CL.DAT yells at the teacher.NOM.
'In the image appears the witch that the teacher yells at'.
- (2) a. En la imagen aparece la que le grita a la bruja.
In the image appears the that CL.DAT yells at the.DAT witch.
'In the image appears the one that yells at the witch'.
b. En la imagen aparece a la que le grita la maestra.
In the image appears the.DAT that CL.DAT yells at the teacher.NOM.
'In the image appears the one that the teacher yells at'.

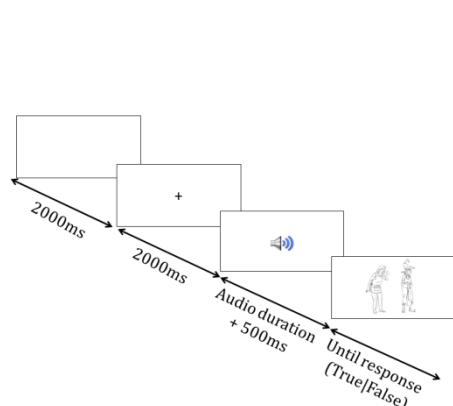


Fig. 1

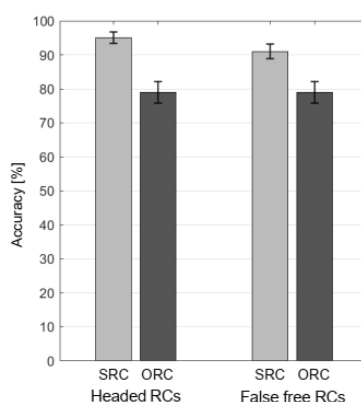


Fig. 2

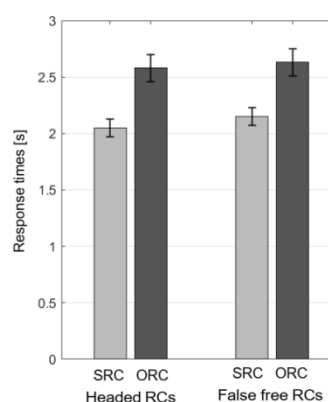


Fig. 3

- (3) a. En la imagen aparece la [_{RC} maestra_[+N+R] que le grita _[+N+R] a la bruja_[+N]].
b. En la imagen aparece la [_{RC} bruja_[+N+R] a la que le grita la maestra_[+N] _[+N+R]].
- (4) a. En la imagen aparece la [_{RC} N_{null}_[+N+R] que le grita _[+N+R] a la bruja_[+N]].
b. En la imagen aparece a la [_{RC} N_{null}_[+N+R] que le grita la maestra_[+N] _[+N+R]].

References

- BENTEA, A., DURRLEMAN, S., & RIZZI, L. (2016). Refining intervention: The acquisition of featural relations in object A-bar dependencies. *Lingua*, 169, 21-41. // BETANCORT, M., CARREIRAS, M., & STURT, P. (2009). The processing of subject and object relative clause in Spanish: An eye-tracking study. *Quarterly Journal of Experimental Psychology*, 62, 1915-1929. // COHEN, L., & MEHLER, J. (1996). Click monitoring revisited: An on-line study of sentence comprehension. *Memory and Cognition*, 24, 94-102. // CONTEMORI, C., & BELLETTI, A. (2014). Relatives and passive object relatives in italian-speaking children and adults: intervention in production and comprehension. *Appl. Psycholinguist.* 35, 1021-1053. // FRIEDMANN, N., BELLETTI, A., & RIZZI, L. (2009). Relativized relatives: types of intervention in the acquisition of A-bar dependencies. *Lingua*, 119, 67-88. // GIBSON, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1-76. // GIOLLO, N., & MUÑOZ PÉREZ, C. (2013). Sobre las construcciones relativas en español. In A. Marcoverchio, A. Ghio, & M. Cuñarro (Eds.). *En torno a la morfosintaxis del español* (pp. 49-59). Mendoza: Universidad Nacional de Cuyo. // GRILLO, N. (2009). Generalised minimality: feature impoverishment and comprehension deficits in agrammatism. *Lingua*, 119, 1426-1443. // PANAGIOTIDIS, P. (2003). Empty Nouns. *Natural Language & Linguistic Theory*, 21, 381-432 // RIZZI, L. (2004). Locality and the left periphery. In Belletti, A. (Ed.). *Structures and Beyond: The Cartography of Syntactic Structures* (vol. 3, pp. 223-251). Oxford-New York: OUP. // SCHRIEFERS, H., FRIEDERICI, A. D., & KÜHN, K. (1995). The processing of locally ambiguous relative clauses in German. *Journal of Memory and Language*, 34, 499-520.

Prosodic Phrasing in English and the Processing of Agreement Attraction
Adam J. Royer – UCLA Linguistics

Intro – English speakers occasionally produce erroneous subject-verb agreements when a subject NP has a singular head noun and a plural noun in some lower syntactic phrase (i.e. local noun) (Bock 1991, Bock et al. 2001). Evidence from production (Eberhard 2005) and comprehension (Badecker 2007, Wagers 2009) studies have conflicting accounts for the mechanisms at play in these errors (i.e. *Marking and Morphing* and *cues-based retrieval*). As of yet, however, neither account has incorporated prosody into our understanding of agreement despite what is known about prosody's role in sentence processing (Frazier 2006). This study bridges these areas of processing by investigating the role of phrasing in the processing of subject-verb agreement. Additionally, grammatical differences between participants were considered (“standardized” vs. “non-standardized” subject-verb agreements).

Methods – The experiment was a 2x2x2 design crossing the morphological number of the local noun and verb, and presence/absence of an intonation phrase break between the local noun and verb (e.g. “The key to the cabinets (L-H%) were placed...”). A ToBI trained linguist produced the 64 critical items and 64 distractor items. Participants (N = 106) listened to sentences and had 3sec to judge whether it sounded “acceptable” or “unacceptable” in a 2AFC task. Following this task was a short 2AFC task that gauged their sensitivity to the acceptability of default singular verb agreement (i.e. was-leveling). A dprime score was calculated for each participant from their responses in this survey.

Results – Both data were modeled using Bayesian mixed effects models. Ratings were modeled using a Bernoulli distribution and RTs with a shifted log-normal distribution. The fixed effects were the aforementioned factors with all interactions. Random effects included maximal intercepts and slopes for both participant and item. Rating data replicate findings that with a singular head noun and local noun, a plural verb drastically reduces acceptability ($\beta = -2.72$, CrI = -3.21, -2.24) but that with a local plural noun instead, acceptability increases ($\beta = 1.85$, CrI = 1.30, 2.41). The model for RTs shows that a mismatch in number of the head and local nouns resulted a slow-down in RT ($\beta = 0.104$, CrI = 0.028, 0.049) relative to number matched conditions. As for main effects of grammaticality and phrasing, there is weak evidence of a small to negligible effect that ungrammatical sentences resulted in slower RTs ($\beta = -0.044$, CrI = -0.102, 0.014) than grammatical sentences and that a prosodic break resulted in slower RTs ($\beta = -0.045$, CrI = -0.096, 0.006) than when no break was present. The model also shows that there is strong evidence that the slowdown in RT for mismatched number is diminished when there is a prosodic break ($\beta = -0.106$, CrI = -0.212, 0.001). In a three-way interaction between local noun, verb number, and response type, an effect of agreement attraction was found such that *unacceptable* responses were much slower in the presence of a plural local noun and plural verb ($\beta = 0.248$, CrI = 0.034, 0.461) than when the local noun was singular. The insertion of a prosaic break reduced the difference in RTs between response types in match and mismatch conditions, as compared to when no break was present ($\beta = -0.106$, CrI = -0.212, 0.001).

Discussion – The results show that a mismatch in morphological number results in a processing penalty for agreement, as shown by Staub (2009). This is across both grammatical and ungrammatical sentences. However, this effect only appears when the local noun and verb are prosodically phrased into the same intonational phrase. When phrased separately, the interference of the plural local noun is ameliorated. This is seen by comparing the difference in RTs for acceptable and unacceptable responses for agreement attraction sentences based on the presence or absence of a prosaic break. The RT for correct rejections (i.e. *unacceptable*) is much faster when a break is present. One explanation is that the local noun is less accessible as a source of agreement because of the prosodic hierarchical distance between it and the verb. I propose this mediates its interference in a similar way to syntactic depth.

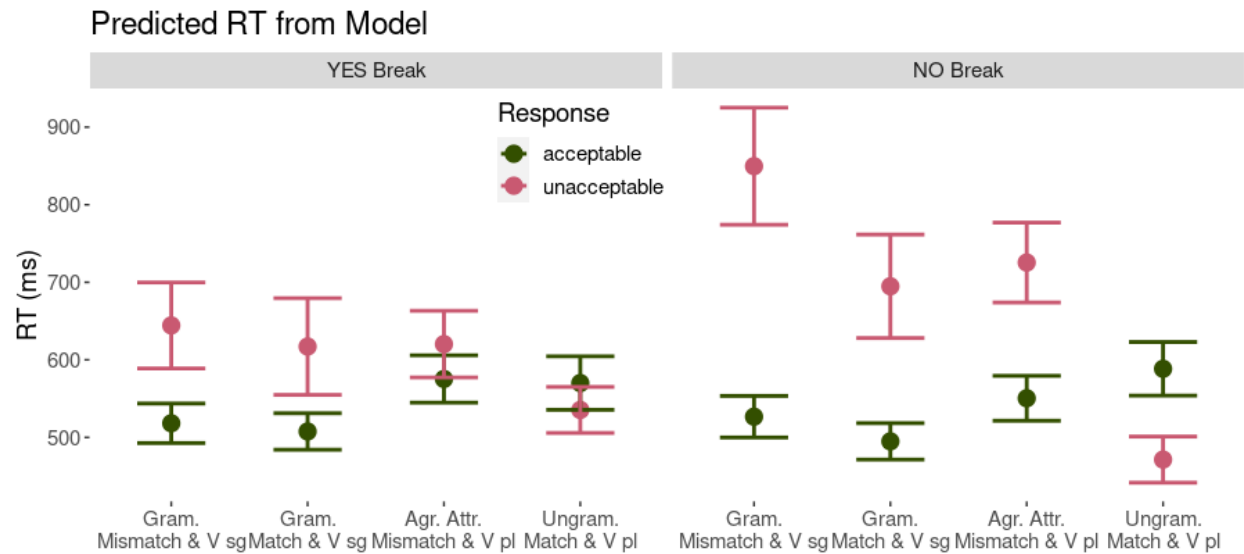


Fig 1. Mean estimated values and standard error bars. *Match* condition is when head noun and local noun are both singular, whereas *Mismatch* is when the local noun is plural. The *Grammatical* sentences have a singular verb and the *Agreement Attraction* and *Ungrammatical* conditions have a plural verb.

	Head-Local #		
Grammatical	Match	The actor in the film (%) was	popular with both young and old fans
	Mismatch	The actor in the films (%) was	
Ungrammatical	Match	The actor in the film (%) were	
	Mismatch	The actor in the films (%) were	

Table 1. Quartet of critical items.

References

- Bock, K. and Miller, C. A. (1991). "Broken agreement," *Cognitive psychology*, vol. 23, no. 1, pp. 45–93.
- Bock, K., Eberhard, K. M., Cutting, J. C. Meyer, A. S., and Schriefers, H. (2001) "Some attractions of verb agreement," *Cognitive psychology*, vol. 43, no. 2, pp. 83–128, 2001.
- Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making Syntax of Sense: Number Agreement in Sentence Production. *Psychological Review*, 112(3), 531-559.
- Frazier, L, Carlson, K., and Clifton, C. J. (2006) "Prosodic phrasing is central to language comprehension," *Trends in cognitive sciences*, vol. 10, no. 6, pp. 244–249, 06 2006.
- Schafer, A. J. (1998) "Prosodic parsing: The role of prosody in sentence comprehension," *Ph.D. dissertation*, UMass Amherst.
- Staub, A. (2009). "On the interpretation of the number attraction effect: Response time evidence," *Journal of Memory and Language*, vol. 60, pp. 308–327.
- Wagers, M., Lau, E. and Phillips, C. (2009). "Agreement attraction in comprehension: Representations and processes," *Journal of Memory and Language*, vol. 61, no. 2, pp. 206–237.

Prosody and eye movements on attachment in Brazilian Portuguese

Aline Fonseca (Federal University of Juiz de Fora), Andressa da Silva (Federal University of Juiz de Fora), Marcus Maia (Federal University of Rio de Janeiro)

This research explores how prosodic cues, such as contrastive pitch accent and prosodic boundary, can influence the attachment of adverbial prepositional phrases in ambiguous Brazilian Portuguese (BP) sentences like (1).

(1) O colega do Paulo revelou que a Camila fumou na varanda do sobrado.

(Paul's friend revealed that Camila smoked on the house balcony.)

a. *high attachment meaning*: Paul's friend revealed something to him on the house balcony.

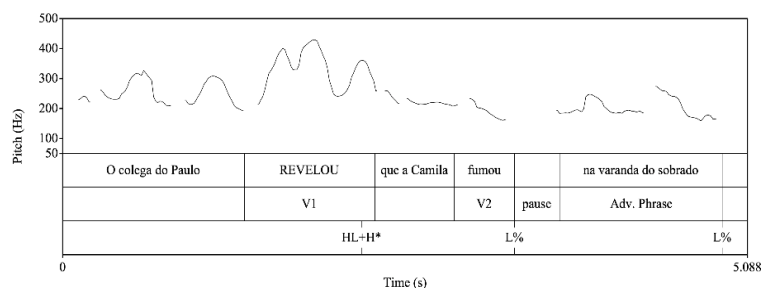
b. *low attachment meaning*: Camila smoked on the house balcony.

Both prosodic boundaries and pitch accents can affect attachment preferences in some syntactic structures. In English, Clifton *et al.* (2002) found that a prosodic boundary before the final adverbial phrase increased high attachments (HAs) to the first verb (e.g., *revealed*), while Carlson & Tyler (2018) showed that contrastive pitch accents (L+H*) on the first or second verb (e.g., *smoked*) drew attachment to the accented verb. For BP, Fonseca *et al.* (2019) found only 10% of HAs in a reading questionnaire, while in an auditory questionnaire there was evidence of the prosodic boundary effect on interpretation but not the accent effect. In the current study, the aim is to investigate if the accent on the first verb and the boundary before the adverbial phrase increase HA choices in a visual world paradigm experiment (Tanenhaus & Trueswell, 2006).

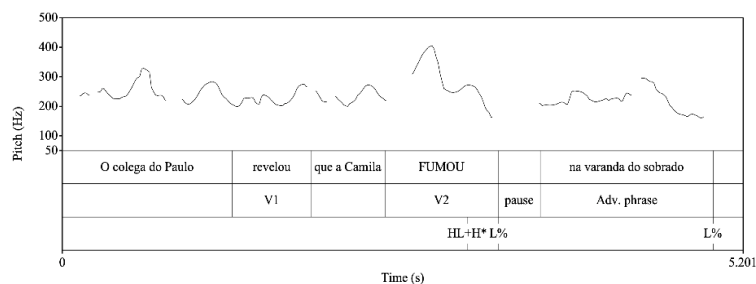
We crossed pitch accents on the first verb (*revealed*) vs. on the second verb (*smoked*) with a prosodic boundary (IP) before the adverbial phrase vs. none. The four conditions were named V1, V2, V1IP and V2IP. Pictures 1 and 2 show the pitch tracks of the two conditions with IP boundaries and accents. The accented verb has a HL+H* accent and it also has increased duration and intensity, mainly in the stressed syllable.

The experiment (N=28) was a spoken language comprehension task with the VWP which tested 24 sentences in four conditions. BP native speakers listened to the sentences while two pictures were being shown on the screen. One picture biased high attachment interpretation and the other one biased low attachment interpretation (see Pictures 3 and 4). After listening to the sentences and seeing the pictures, they had to answer a comprehension question like: *What happened on the balcony?* a) *Paul's friend revealed something there* or b) *Camila smoked there*. In V1 and V1IP conditions, we considered that the picture with high attachment bias was the target and the picture with low attachment bias was the control. In V2 and V2IP conditions, the picture with low attachment bias was the target and the picture with high attachment bias was the control.

We measured the participants' eye movements (total fixation duration/TFD and fixation count/FC) to both pictures on the screen, using an Eyelink 1000 eye tracker, while they were listening to the final part of the sentence (the ambiguous adverbial PP underlined in example 1). The means of TFD were higher to the target pictures than to the control pictures in all prosodic conditions (see Graph 1). In a linear regression model and a Tukey HSD post hoc test running in R Studio (R Core Team, 2020), we found out that participants looked more to the target pictures while they were listening to the final adverbial PP in all prosodic conditions. ($\beta = 94.73$, $SE = 36.273$, $df = 1158.752$, $t = 2.612$, $CI [23.64 \sim 165.83]$, $p = 0.009$). We also analyzed the interpretation choices and we found out that V1IP and V1 had more high attachment choices than the other two conditions (see Graph 2) (V1IP x V2IP conditions $\beta = -1.676$, $SE = 0.235$, $z = -7.138$, $CI [-2.161 \sim -1.218]$, $p < 0.001$). These results point out that listeners are sensitive to prosodic cues like contrastive accents and boundaries, and that they are able to use this prosodic information early in processing (Warren, 1996; Speer & Blodgett, 2006).



Picture 1: Example of Condition V1IP



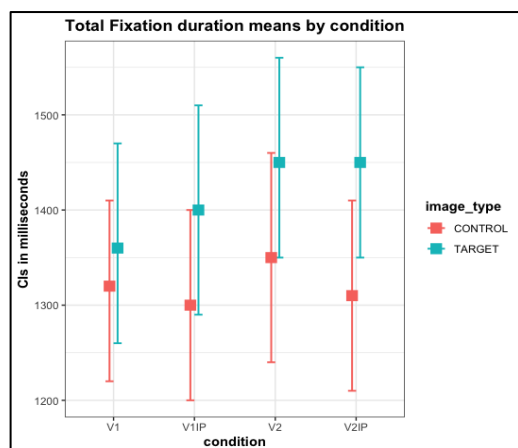
Picture 2: Example of Condition V2IP



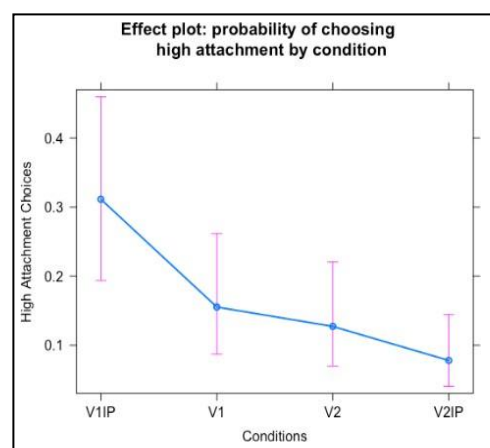
Picture 3: High Attachment Interpretation Bias



Picture 4: Low Attachment Interpretation Bias



Graph 1: TFD means



Graph 2: Effect plot of high attachment choices

References

- Carlson, K. & Tyler, J. C. (2018) Accents, not just prosodic boundaries, influence syntactic attachment. *Language and Speech*. 61, 246-276.
- Clifton, C. Jr., Carlson, K. & Frazier, L. (2002) Informative prosodic boundaries. *Language and Speech*, 45. 87-114.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Fonseca, A.; Carlson, K.; Silva, A (2019). Prosodic effects on attachment in Brazilian Portuguese. Poster presented at the CUNY Human Sentence Processing Conference, University of Colorado Boulder.
- Speer, S. & Blodgett, A. (2006) Prosody. In: Traxler, M. J. & Gernsbacher, M. A. (Ed.) *Handbook of Psycholinguistics*. 2nd Edition. Elsevier Academic Press. p. 505- 538.
- Tanenhaus, M. K. & Trueswell, J. C. (2006) Eye Movements and Spoken Language Comprehension. In: Traxler, M. J. & Gernsbacher, M. A. (Ed.) *Handbook of Psycholinguistics*. 2nd Edition. Elsevier Academic Press. p.863-900.
- Warren, P. (1996) Prosody and Parsing: An Introduction. *Language and Cognitive Processes*. 11:1-2. p. 1-16.

Two-dimensional parsing, the iambic-trochaic law, and the typology of rhythm

Michael Wagner, Alvaro Iturralde Zurita, and Sijia Zhang (McGill University)

Humans appear to be wired to perceive acoustic events rhythmically. English speakers tend to perceive alternating short and long sounds as sequences of binary groups with a final beat (iambs), but alternating soft and loud sounds as trochees (Bolton, 1894; Woodrow, 1909). This generalization (often called the ‘iambic-trochaic Law’ (ITL), following Hayes 1995), has been hypothesized to be a universal of auditory processing (Hay and Diehl, 2007). Kusumoto and Moreton (1997); Iversen et al. (2008), Bhatara et al. (2013), and Crowhurst and Teodocio Olivares (2014), however, found that the duration-side of the ITL fails to apply in Japanese, French, and Spanish, suggesting that rhythm perception is shaped by language experience. This has been attributed to cross-linguistic differences in word order (Iversen et al., 2008) or stress-systems (Bhatara et al., 2013). This prior work has an important limitation: If one parses a sequence of sounds (e.g. iterations of the syllables ‘ba’ and ‘ga’) into binary groups, there are not 2 but (at least) 4 potential percepts (e.g., BAgA, baGA, GAba, gaBA). Prior research usually asked in some way or other about the perceived foot (*Did you hear [X x] or [x X]?*). This task only narrows things down to 2 out of 4 possibilities (e.g. both BAgA and GAba are trochees). Crowhurst and Teodocio Olivares (2014) used a speech segmentation task which also only narrows things down to two possibilities (BAga/baGA for a ‘baga’ response; GAba/gaBA for a ‘gaba’ response).

Wagner (under review) argues that the ITL is a simple consequence of the cue distribution for the perceptual dimensions of grouping and prominence. Production data show that in words and phrases, prominent syllables are both louder and longer, but these two cues anti-correlate when encoding grouping: initial syllables are louder, final syllables longer. Using two tasks to fully determine the percept (*Which syllable is initial?*, *Which syllable is prominent?*), one can see that the ITL is simply a consequence of this cue distribution. The perception data can be predicted from the cue distribution seen in production, including the ITL effect: If a sound is sufficiently long, it will be perceived as final and stressed, and if it is sufficiently loud, as initial and stressed.

Our **first** contribution is to **replicate** the perception findings in Wagner (2020) (perception of syllable sequences e.g. *..bagaba...*, results in Fig. 1): Listeners make consistent prominence choices when intensity and duration correlate (consistent cues for prominence), and are closer to chance when they anti-correlate. They make more consistent grouping decisions when the cues anti-correlate (consistent cues for grouping since louder=initial; longer=final), and are closer to chance when they correlate. The foot decision, which can be reconstructed from the grouping and prominence decisions, shows the ITL pattern when only one cue is manipulated in an extreme way (trochees for an intensity difference; iambs for a duration difference), but shows little systematicity when both cues are manipulated. The choice between iamb and trochee is epiphenomenal, the choice of prominence and grouping highly systematic.

Our **second and central** contribution is to establish the beginnings of a **parsing typology** based on experiments in 4 more languages. The coefficients (log odds) for intensity and duration in logistic models of the prominence and grouping decisions (Fig. 2) show that cue interpretation is not universal. In Mandarin and Japanese, e.g., duration does not reliably cue grouping, and all languages differ significantly from each other in at least one coefficient. However, the differences between languages are small and non-significant for duration when looking at prominence, and for intensity when looking at grouping. So despite the observed differences between languages, could use the more invariable cue (duration for prominence; intensity for grouping) to bootstrap into signal, even if the respective other cue is language specific. We know that children show the ITL early on but then, depending on language, can ‘lose’ the duration-side. New work will be needed to establish whether this variance is due to grouping and/or prominence perception, since prior acquisition work used a single task that didn’t fully establish grouping and prominence. Existing quantitative measures of rhythm in the acquisition literature have been criticized as tapping phonotactics or phonemic differences rather than rhythm (see Arvaniti 2012 i.a.). The typology based on the cues for grouping and prominence perception jointly may provide a better quantitative map of cross-linguistic differences in what we intuitively call ‘rhythm.’

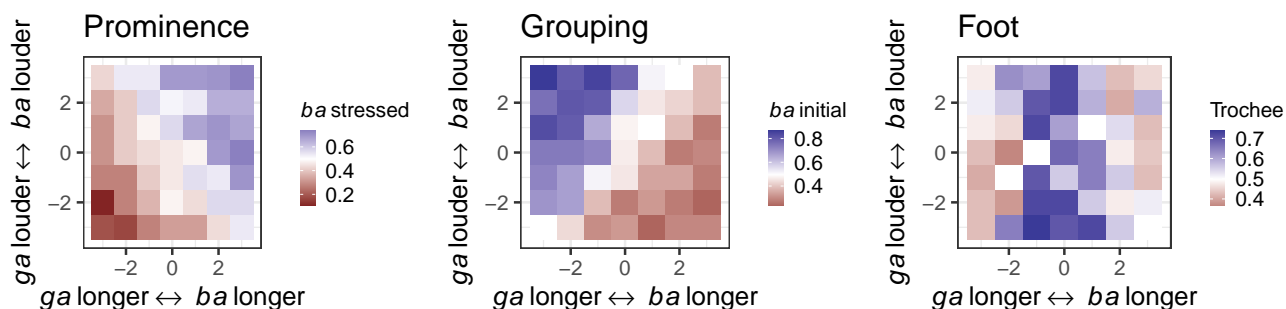


Figure 1: Listeners (50 adult native speakers of North American English) heard sequences of syllables *bagabaga...* or *gabagaba...*. The heatmaps show proportions of responses from the prominence task (*Which syllable was stressed (ba or ga)?*) and the grouping task (*Did you hear бага or гaba?*), plotted by relative duration (7 steps on x-axis) and intensity step (7 steps on y-axis); the foot decision (right) was reconstructed from these two responses.

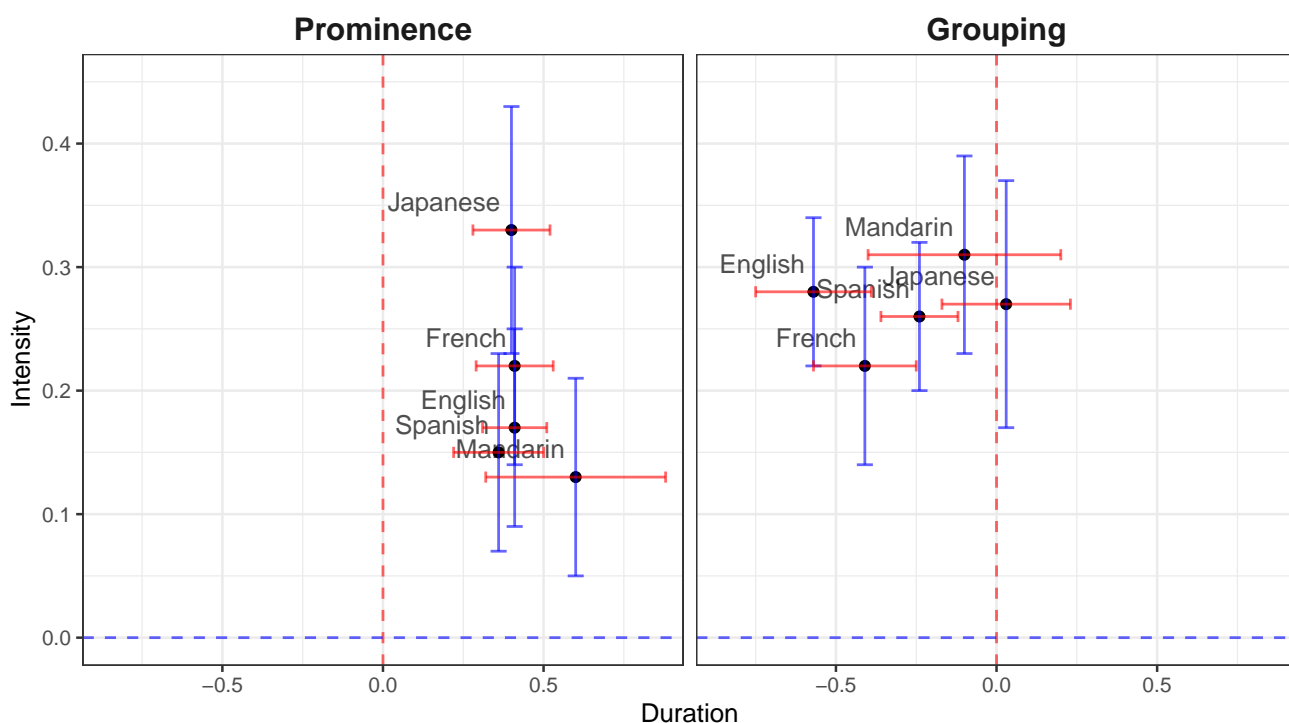


Figure 2: The parsing typology based on a small sample of 5 languages (NA English, French, Japanese, Mandarin, Mexican Spanish), plotting the coefficients from logistic MER models for the individual languages for each decision (50 listeners per language). Coefficients corresponds to the predicted change in log odds given a unit change in intensity/duration. Duration (x-axis) and intensity (y-axis) coefficients are shown for the prominence decision (left) and the grouping decision (right). Error bars show 2*se estimated by the logistic models. All error bars that don't cross the dashed zero line (red for duration, blue for intensity) came out significant (only two were not significant: duration in the grouping decision models in Mandarin and Japanese). Note that there is little variation on the horizontal (duration) for the prominence dimension (small differences, non-significant interaction duration*Language in all-language MER model), and little variation on the vertical (intensity) for the grouping decision (small differences, non-significant interaction duration*Language)

Case interference and phrase length effects in processing Turkish center-embeddings

Özge Bakay & Nazik Dinçtopal Deniz (Boğaziçi University)

Background: Doubly center-embeddings with relative clauses (2-CE-RCs) such as *The rat that the cat that the dog chased ate died* [1, p. 286] are reported to be extremely difficult to process despite their grammaticality. Several accounts have been proposed to explain their processing difficulty [e.g., 1,2,3]. The present study tests the predictions of (i) similarity-based interference [4] and (ii) prosodic phrase lengths [5]. (i) predicts that decreasing the similarity among the NPs, e.g., in their syntactic [e.g., 6] or phonological case markings [e.g., 7 but cf. 6], may ease the processing difficulty of CE structures. (ii) predicts that when the 2-CE-RCs have optimal and balanced phrase lengths, their processing is easier. An eye-tracking experiment was conducted to examine these predictions. **Materials:** The experiment employed Turkish 2-CE-RCs nested as a complement clause inside a matrix clause (see the examples in (1)). The experiment manipulated syntactic and phonological case interference (syntactic case indicates syntactic functions of NPs; phonological case indicates phonological similarity of cases irrespective of their syntactic functions), and prosodic phrase lengths. For case interference, in high syntactic-high phonological interference condition (HS-HP; 1a), all three subject NPs were marked with the genitive case, *-(n)in*. In low syntactic-high phonological interference condition (LS-HP; 1b), NP1 had the (null) nominative case to decrease the syntactic case similarity among the subject NPs, and the first object NP, NP4 was marked with genitive case to keep phonological case similarity high (at three). In low syntactic-low phonological interference condition (LS-LP; 1c), NP1 had the (null) nominative case, and the first object NP, NP4, was marked with the accusative case to decrease case similarity (at two). For phrase lengths, in conditions that encouraged a relatively balanced phrasing of Turkish 2-CE-RCs, viz., NP1||NP6||VP1, (ENC; 1a-c), NP1 and VP1 were each lengthened with two additional prosodic words (PWds), resulting in three PWds each [8]. In conditions that discouraged the optimal phrasing (DISC; 1a-c), NP6 was lengthened with four additional PWds, and NP1 and VP1 were one PWd each. Overall sentence length was the same across ENC and DISC conditions. **Procedure:** The participants' ($N = 44$) eye-movements were recorded as they read the sentences. A question followed each sentence to ensure comprehension. **Results:** The eye-tracking data, summarized in Table 1, were analyzed with mixed-effects linear/logistic regression models for the critical region (region 14: VP1) and the spillover region (region 15: matrix verb). Case interference and phrase lengths were fixed effects. The analyses on the critical region showed that HS-HP were harder to read than LS-HP (gaze duration (GD), regression path duration (RPD), rereading duration (RRD), total duration (TD) (t 's ≥ 2.99)) and LS-LP (GD, RPD, TD (t 's ≥ 3.92), probability of regression out (PRO) ($z = 1.95$)). There was an increased difficulty in reading DISC compared to ENC in the critical (first fixation duration (FFD), RPD, RRD, TD (t 's ≥ 3), PRO ($z = 2.51$)) and spillover region (RPD ($t = 2.52$)). In the critical region, the complex models with syntactic case interference and phrase lengths explained the data better than simpler models with a single predictor (GD, RPD, RRD, TD, PRO (χ^2 's (1) ≥ 6.22 , $p < .05$)). The follow-up analyses showed that HS-HP/DISC was the most difficult to read (RPD, RRD, TD (t 's ≥ 2.45)) and LS-HP/ENC was the easiest to read (RPD (t 's ≥ 2.31)). There were no effects of phonological case interference in either region. **Conclusion:** The results show that the processing difficulty of Turkish 2-CE-RCs can be alleviated with decreased syntactic case similarity (as in Japanese [6]) and with optimal and balanced phrase lengths (as in English [9]). This was the case in both early and late measures. Unlike syntactic case interference effects, phrase length effects persisted to a later region. This may suggest that the integration of prosodic phrase length information into the current structure may take longer [10] or may "alter a [parse] that was starting to take hold" [11, p. 119]. An end-of-sentence acceptability task to examine whether the two forces or only phrase lengths affect final decisions is underway. No effect of phonological case interference can be a true null effect [12] or due to the increased level of embeddings [7].

References: [1] Chomsky & Miller (1963). In Luce et al. (Eds.), *Handbook of Math. Psy.* [2] Bever (1970). In Hayes (Ed.), *Cog. and the Dev. of Lang.* [3] Gibson (2000). In Marantz et al. (Eds.), *Image, Lang., Brain.* [4] Lewis & Vasishth (2005). *Cog. Sci.*, 29(3). [5] Fodor (2013). In Montserrat et al. (Eds.), *Lang. Down the Garden Path: The Cog. & Bio. Basis for Ling. Str.* [6] Uehara & Bradley (1996). In Park & Kim (Eds.). *Lang., Info. & Comp.* [7] Nakayama et al. (2005). *Lang. Sci.*, 4. [8] Deniz & Fodor (2017). *Lang. & Speech* (60)4. [9] Fodor et al. (2017). In Almeida & Gleitman (Eds.), *On Con., Modules & Lang.: Cog. Sci. at its Core.* [10] Marcus & Hindle (1990). In Altmann (Ed.), *Cog. models of speech process: comput. and psy. persp.* [11] Fodor (2002). In *Proceed. of NELS* 32. [12] Obata et al. (2010). In *Proceed. of NELS* 41.

Materials: Brackets indicate clause boundaries. Case marking is in bold face. Colored words manipulate phrase lengths: green in ENC and red in DISC. || marks implicit prosodic boundaries predicted to be induced by phrase lengths.

1. a. **ENC/DISC, HS-HP:**

∅ [İşinin ehli marangoz-lar-**in** || [nakliyeciler-**in** [kiracı-**nın** **oldukça geniş** gri...
Pro expert carpenter-PL-GEN mover-PL-GEN renter-GEN extremely large gray
NP1-GEN NP2-GEN NP3-GEN

b. **ENC/DISC, LS-HP:**

∅ [İşinin ehli marangoz-lar-∅ || [nakliyeciler-**in** [kiracı-**nın** koltuğ-**un** **oldukça geniş**...
Pro expert carpenter-PL-NOM mover-PL-GEN renter-GEN sofa-GEN extremely large
NP1-NOM NP2-GEN NP3-GEN NP4-GEN

c. **ENC/DISC, LS-LP:**

∅ [İşinin ehli marangoz-lar-∅ || [nakliyeciler-**in** [kiracı-**nın** **oldukça geniş** gri...
Pro expert carpenter-PL-NOM mover-PL-GEN renter-GEN extremely large gray
NP1-NOM NP2-GEN NP3-GEN
koltuğ-u /minder-leri-ni **büyük özen-le** yerleştirdiğ-i] odaya taşıdıkları]
sofa-ACC/cushion-3POSS.PL-ACC great care-with place-FN-3SG room-DAT move-FN-3PL
VP3 VP2
dolabı || **dikkatli şekilde** kurdukları-nı/kurdu-lar] sandı-m.
clozet-ACC careful manner build-PAST-3PL-ACC/build-PAST-3PL think-PAST-1SG
VP1

'I know that the **expert** carpenters **carefully** built the closet that the movers moved to the room where the renter placed the **extremely large** gray sofa/sofa's cushions **with great care.**'

Table 1. Mean and standard error (SE) values for first fixation duration (FFD), gaze duration (GD), regression path duration (RPD), re-reading duration (RRD), total duration (TD) (in milliseconds) and probability of regression out (PRO) for the critical region (region 14) and the spillover region (region 15). ENC and DISC conditions are given in green and red, respectively.

		FFD	GD	RPD	RRD	TD	PRO
Critical Region	HS-HP, ENC	231 (6)	380 (14)	428 (19)	315 (31)	661 (31)	.10 (.02)
	LS-HP, ENC	233 (5)	318 (12)	363 (17)	253 (26)	565 (28)	.11 (.02)
	LS-LP, ENC	238 (7)	333 (12)	418 (24)	275 (30)	602 (30)	.15 (.03)
	HS-HP, DISC	239 (6)	409 (16)	554 (27)	434 (36)	793 (35)	.20 (.03)
	LS-HP, DISC	247 (8)	347 (14)	462 (23)	327 (32)	633 (31)	.21 (.03)
	LS-LP, DISC	239 (6)	324 (13)	417 (21)	352 (32)	648 (31)	.19 (.03)
Spillover Region	HS-HP, ENC	244 (7)	283 (10)	373 (27)	146 (18)	408 (19)	.15 (.03)
	LS-HP, ENC	238 (6)	267 (9)	332 (18)	180 (23)	439 (24)	.19 (.03)
	LS-LP, ENC	239 (7)	290 (11)	378 (24)	178 (26)	447 (25)	.19 (.04)
	HS-HP, DISC	247 (7)	280 (19)	434 (26)	210 (27)	463 (24)	.26 (.04)
	LS-HP, DISC	249 (7)	297 (11)	416 (27)	170 (22)	466 (26)	.26 (.04)
	LS-LP, DISC	251 (8)	278 (11)	416 (27)	192 (22)	465 (22)	.24 (.04)

Prosody drives eye movements from early on in semantic comprehension

Petra Augurzky, Ruth Kessler & Claudia Friedrich (University of Tübingen)

Background: Numerous studies have shown that sentence-level semantic comprehension may sometimes proceed in a non-incremental fashion. Such processing delays may not only relate to structural complexity and the avoidance of semantic revisions [1-3] but also to the visual presentation of sentence materials. In particular, word-by-word reading lacks relevant auditory cues that the parser can use for predicting upcoming sentence continuations [4]. However, in a former ERP study with spoken sentences, we did not find an effect of prosody on predictive processing. Though the manipulation of prosodic contours in that study was directly informative with respect to the semantic phenomenon under investigation (i.e. the processing of quantifier restriction), the amplitude of the N400 was exclusively modulated by the truth evaluation process.

Paradigm: The present study tested the effects of prosodic contours on semantic comprehension in sentences involving quantifier restriction. We applied the experimental design of [5] to a visual world paradigm, a method that has a long-established tradition in research on sentence-level prosody [6]. In each trial, participants viewed an array of four pictures (A-D; Fig.1 and 2), in which objects like triangles of different colors (e.g. blue or red) were presented inside and outside of a container form (e.g. a circle). After a short preview, participants heard naturally-produced sentences with the quantifier *alle* ('all') in one of two prosodic variants: Variant (1) involved a falling contour, signaling the end of the sentence on the adjective *blau* ('blue'), and variant (2) involved a continuation rise, signaling that a further restriction would follow [5]. Five seconds after the start of the audio signal, three of the pictures disappeared. Participants responded as fast as possible whether the remaining picture fitted with the sentence just heard by pressing buttons.

Conditions: Depending on the context pictures, sentences (1) and (2) involve different truth values at the adjective (*blau*, 'blue'): they are false for A,C and D but true for B. For (2), truth values may change for sentences related to pictures A and D at the preposition *innerhalb* ('inside-of') or *außerhalb* ('outside-of') in the relative clause. In these cases, the truth evaluation was shown to be postponed from the adjective to the preposition. We examined whether sentence-end prosody yields an immediate commitment on the adjective by signaling that no meaning shift would follow.

Eye Tracking and Hypotheses: We measured fixations during sentence listening on the four pictures for five seconds after the onset of the auditory signal and calculated empirical log transformations [7] of fixation proportions at the position of the adjective (*blau*, 'blue'). At this word, we expected more looks towards simple pictures that involved a local evaluation of the sentence as "true" (B), relative to all other pictures. If the prosodic manipulation is considered immediately, we expect that for sentence-end prosody (1), participants should equally reject false (C) and complex conditions (A,D), as prosody indicates that no meaning shifts would be possible. Thus, each of these three pictures should be equally rarely fixated. For sentence-continuation prosody (2), a similar pattern as in previous studies should arise: the looks towards the complex pictures (A,D) should be in-between the looks to simple pictures involving true (B) and false (C) truth evaluations.

Preliminary Results and Discussion: So far, we could not finish data acquisition due to COVID-19. Here, we report the results of 9 participants (see Fig.3). The overall task performance was high (94.9% correct responses). Similar to previous ERP studies, we found a processing bias towards true utterances: Participants fixated pictures associated with true judgments (B) more often than false ones (A,C,D) for sentence-end prosody (1) at the adjective. For sentence-continuation prosody (2), the fixation pattern was more variable at the adjective. We thus found an immediate interaction between prosodic information and the semantic truth evaluation, contrary to our previous ERP results. In our presentation, we will discuss this apparent discrepancy between the two methods. However, the current results generally need to be interpreted with caution, as they are only initial findings, and the total number of participants ($n=32$) still has to be measured. The current data are also the background for further studies on quantifier acquisition.

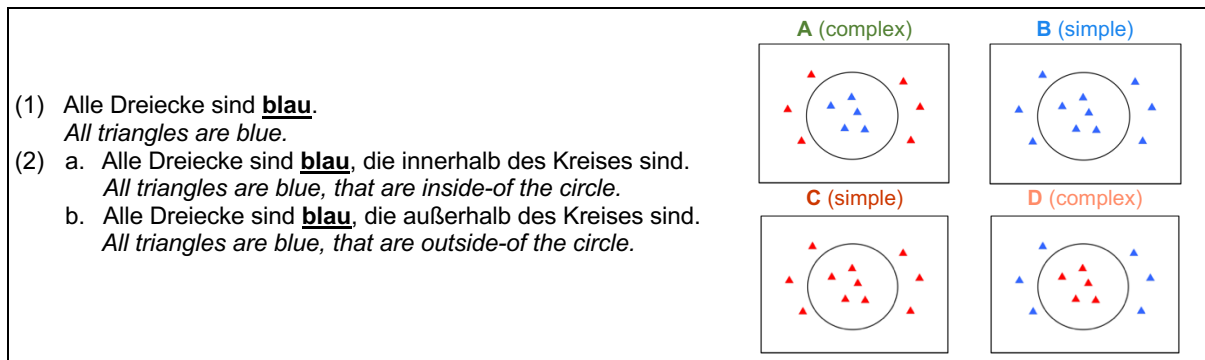


Fig. 1: Experimental sentences and the four pictures (A, B, C, D) that were used as targets for fixations in the experimental array. Figures were presented simultaneously, and their ordering on the screen was counterbalanced across items and participants.

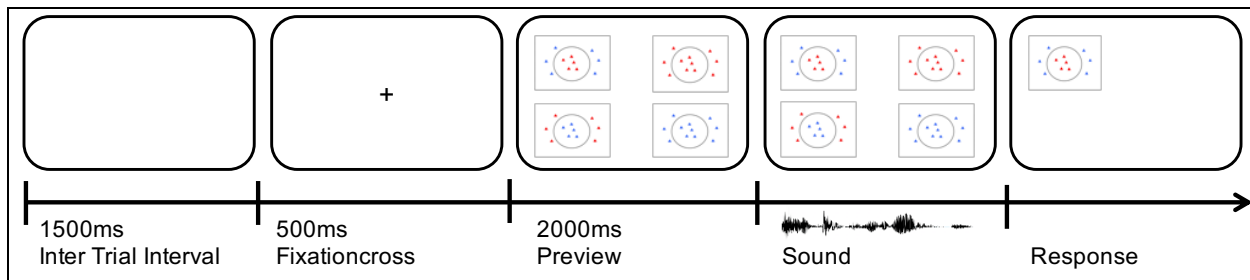


Fig. 2: Schematic overview of an experimental trial.

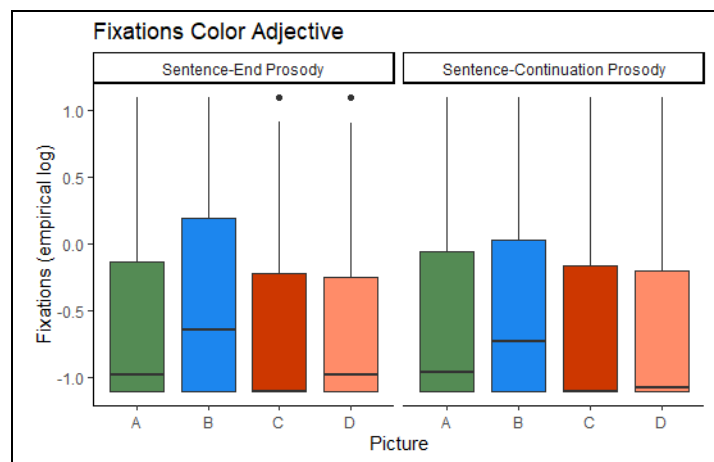


Fig. 3: Boxplot for empirical log transformed fixations to four pictures by prosodic variation in the time bin for the adjective. (Horizontal line = median; box = interquartile range; whisker = minimum/maximum; individual points = outliers)

References: [1] Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify online: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63 (2), 158-179. [2] Augurzy, P., Bott, O., Sternefeld, W., & Ulrich R. (2017). Are all the triangles blue? – ERP evidence for the incremental processing of German quantifier restriction. *Language and Cognition*, 9 (4), 603-636. [3] Augurzy, P., Schlotterbeck, F., & Ulrich, R. (2020): Most (but not all) quantifiers are interpreted immediately in visual context. *Language, Cognition and Neuroscience*. [4] Freunberger, D., & Nieuwland, M. S. (2016). Incremental comprehension of spoken quantifier sentences: evidence from brain potentials. *Brain Research*, 1646, 475-481. [5] Augurzy, P., & Ulrich, R. (2020). Prosodic cues in on-line semantic processing – ERP evidence on quantifier restriction. Poster presented virtually at CUNY 2020. [6] Watson, D. G., Gunlogson, C. A., & Tanenhaus, M. K. (2006). Online methods for the investigation of prosody. In S. Sudhoff, D. Lerner, R. Meyer, S. Pappert, P. Augurzy, I. Mleinek, N. Richter, J. Schlieer (eds.), *Methods in Empirical Prosody Research*. Walter de Gruyter: New York. 259-282. [7] Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4), 457-474.

Preferences for shorter dependencies in miniature language learning are modulated by the statistics of learners' L1

Yiyun Zhao (University of Arizona), Charles Torres (University of California, Irvine), Masha Fedzechkina (University of Arizona)

Human languages differ greatly in how they order words in sentences. These superficially different orders, however, result in short grammatical dependencies [1, 2]. Recent work using artificial languages provided a causal link between this bias in language learners and patterns in linguistic diversity: Adult native (L1) speakers of English confronted with a novel language that had unnecessarily long grammatical dependencies systematically restructured the language to reduce dependency lengths [3]. This work leaves open an important question: Are these preferences based on general cognitive principles or are they also influenced by the principles that are themselves learned from the statistics of the learners' L1? We tease apart these possibilities by comparing the strength of learners' preferences for shorter dependencies in a miniature language across L1 speakers of English and Mandarin. These L1s were chosen because they exhibit dependency length minimization (DLM) to different degrees ([1], Fig. 1).

Prediction: If learners whose L1 allows longer dependencies, exhibit DLM to a lesser degree in a structurally different miniature language, this behavior would suggest that learners' performance is subject to an abstract principle-based L1 transfer. If, however, the degree of DLM in a miniature language is the same across learners' L1s, it would argue that this bias is rooted in pre-L1 general cognitive biases ('UG' in the broad sense).

Method: 40 L1 speakers of English and Mandarin learned a novel miniature language consisting of simple transitive sentences over two 1-hour online sessions on consecutive days. Participants were exposed to a verb-final (50/50% SOV/OSV order) language (different from their verb-initial L1s) with obligatory case-marking on objects (never on subjects). Participants first learned novel nouns (*pilika*=CHEF) and then heard sentences using these nouns along with novel verbs. During training, participants heard utterances in a novel language paired with videos of actors performing simple two-participant actions ('chef kicks referee'), where both the subject and object were either long (i.e., modified by a postpositional phrase) or short (no modification). Balanced word order (SOV/OSV 50/50%) was maintained in all sentence types. Each session ended in a sentence production test: learners described previously unseen videos in the novel language, in which constituent length was manipulated by requiring PP-modification of either the subject, object, or neither of the constituents. Thus, the language allowed flexibility in constituent ordering, which had implications for DLM – ordering constituents long-before-short resulted in shorter dependencies in the verb-final miniature language.

Results: To assess whether learners exploited constituent order flexibility in the input to reduce dependency lengths, we analyzed average dependency lengths in the languages produced by individual participants in the final session of the experiment. Wilcoxon signed-rank test revealed that both English ($V=0$, $p<0.001$) and Mandarin ($V=24.5$, $p=0.025$) learners produced shorter dependencies compared to the input, suggesting an influence of the abstract DLM principle. However, Mandarin learners, whose L1 has on average longer dependencies than English, produced miniature languages with overall longer dependencies compared to English learners ($W=124$, $p=0.03$, Fig. 2), suggesting a clear influence of L1 statistics.

Conclusion: Learners' DLM preferences in the miniature language are influenced both by abstract pre-L1 and L1-driven biases. We find that both Mandarin and English learners follow the abstract DLM principle. This preference is, however, stronger in English speakers, reflecting the differences in the input statistics across the two L1s. Our work adds to the growing body of literature exploring L1 influences on miniature language learning [5, 6]. We show how by teasing apart pre-L1 processing biases and L1-driven cognitive biases, we can begin to better understand how these influences are captured in the miniature language learning paradigm.

References

1. Gildea, D. and D. Temperley, *Do grammars minimize dependency length?* Cognitive Sci, 2010. **34**(2): p. 286-310.
2. Futrell, R., K. Mahowald, and E. Gibson, *Large-scale evidence of dependency length minimization in 37 languages.* Proc Natl Acad Sci USA, 2015. **112**(33): p. 10336-10341.
3. Fedzechkina, M., B. Chu, and T.F. Jaeger, *Human information processing shapes language change.* Psychological Science, 2018.
4. Gildea, D. and T.F. Jaeger, *Human languages order information efficiently.* ArXiv e-prints, 2015.
5. Culbertson, J., Franck, J., Braquet, G., Barrera Navarro, M., & Arnon, I. (2020). A learning bias for word order harmony: Evidence from speakers of non-harmonic languages. *Cognition*, 204(November 2019), 104392.
6. Martin, A., & Culbertson, J. (2020). Revisiting the Suffixing Preference: Native-Language Affixation Patterns Influence Perception of Sequences. *Psychological Science*.

Figures:

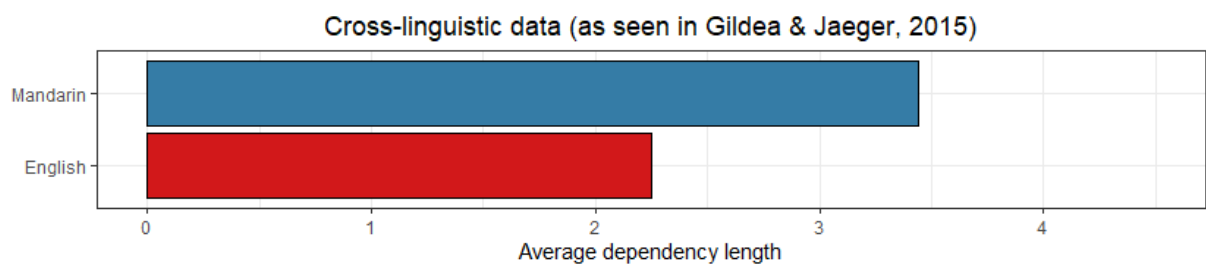


Figure 1: Average dependency length in English and Mandarin (adapted from Gildea & Jaeger, 2015).

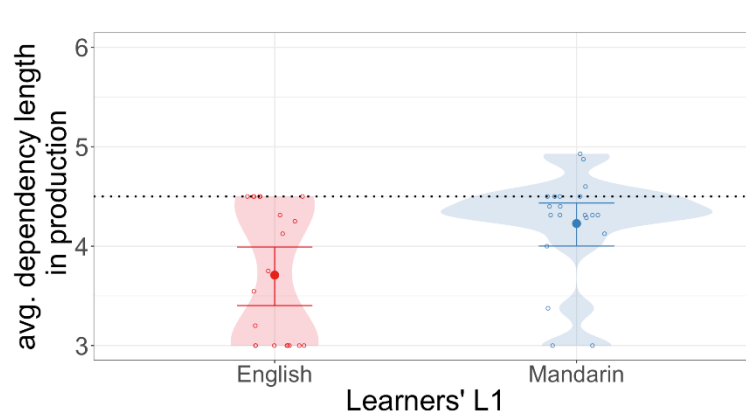


Figure 2: Average dependency length in learners' productions by L1 background. The dashed line represents average dependency length in the input miniature language. The dots represent individual learners' means. The error bars indicate 95% confidence intervals.

Children's acquisition of new/given markers in English, Hindi, Mandinka and Spanish

Vishakha Shukla, Madeleine Long, Vrinda Bhatia & Paula Rubio-Fernandez (University of Oslo)

paula.rubio-fernandez@ifikk.uio.no

Languages have different ways of marking new and given referents, and these markings can be obligatory or optional. **Here we studied four typologically diverse languages (English, Hindi, Mandinka and Spanish) to test the *Optionality Hypothesis*, according to which the acquisition of optional markers is protracted relative to obligatory ones.** In Hindi, the numeral 'one' ('ek') can be used optionally to introduce new referents [1-2]. Based on diachronic and semantic evidence [3-6], we hypothesized that 'ek' is in the process of grammaticalization into an indefinite article, and that due to its current stage of optionality, the acquisition of this marker would be protracted relative to the other languages. English and Spanish require indefinite articles for character introduction, but bare nouns are more permissible in English, making the use of articles less consistent. Mandinka lacks an article system and employs a default lexical morpheme for all nouns, thus being the most consistent. As such, we predicted the following order of acquisition of discourse introduction markers: Mandinka > Spanish > English > Hindi. Given that each of these languages has obligatory markers of givenness which are mastered early [7-9], we predicted no cross-linguistic developmental differences for givenness markers.

EXP 1 employed a narrative elicitation task with a series of 14 pictures featuring 1 or 2 animal characters carrying out actions (see Fig. 1; [10]). 20 children (aged 5) and 15 adult controls were tested in each language. Of interest was the way in which children marked new and given referents relative to adults. As these languages vary in markers, we created a coding system based on the most frequently used markers in each language (see Table 1): 'A' responses are appropriate for introducing new referents, and 'B' and 'C' for marking given. For new referents, an LMER model of Marking ($A=1$, $B \text{ \& } C=0$) with Age Group and Language (reference level: Hindi) as FE and maximal RE structure revealed a main effect of Spanish and English relative to Hindi ($p<.0001$), with more A responses in Spanish and English than in Hindi. There was also an Age x Spanish interaction relative to Hindi ($p=.021$), driven by a difference in A responses for Hindi-speaking children and adults ($p=.001$). Responses from Mandinka-speaking children and adults were uniform; thus, no effects were found relative to Hindi. As predicted, Hindi-speaking children and adults differed the most, followed by English-speaking children and adults, then Spanish-speaking children and adults, and finally Mandinka-speaking children and adults (see Fig. 2). For given referents, the same LMER model revealed no Age x Language differences in the way familiar referents were marked (all p 's $>.05$), confirming that by the age of 5 years, cross-linguistic developmental differences are not pronounced for obligatory givenness markers [7-9].

EXP 2 employed the same task, this time testing Hindi-speaking children and adults from outside of Delhi (Gorakhpur) to assess whether 'ek' is in the process of grammaticalization, or its use is simply a dialectal feature from Delhi. We tested 5-year-olds and adults from Gorakhpur, plus 10-year-olds from Gorakhpur and Delhi (because 10yos are known to have mastered even the harder discourse functions [7]). An LMER model of 'Ek' use for new referents ('Ek'=1, $B \text{ \& } C=0$) with Age Group (reference level: Adults) and Region (reference level: Delhi) as FE and maximal RE structure revealed a difference between 5-year-olds and adults ($p<.0001$), which disappeared by the age of 10, as 10-year-olds did not differ from adults ($p=.411$) (see Fig. 3). Crucially, while Delhi adults used 'ek' more frequently than Gorakhpur 5-year-olds ($p=.016$), there was no difference between Delhi adults and Gorakhpur 10-year-olds ($p=.287$). The same pattern holds when Gorakhpur Adults is the reference level. These results suggest that the use of 'ek' for new referents is not simply dialectal variation, as similar patterns emerged in both regions.

Overall, our findings show that discourse markers emerge earlier in languages that use them consistently, and that 'ek' seems to be undergoing grammaticalization into an indefinite article. Future work should further investigate the use of 'ek' to introduce new referents to shed light on the very process of language change and its implications for common ground marking and Theory of Mind development (for discussion, see [11]).

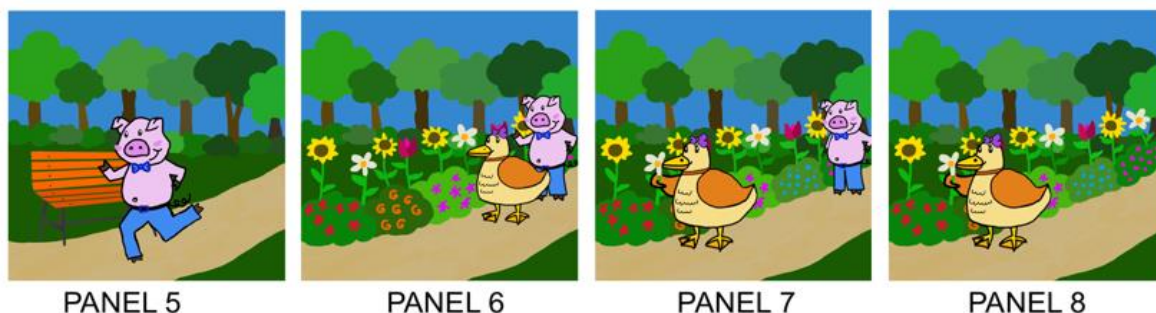


Figure 1. Panels from the narrative elicitation task used in the study, showing new and given characters. These materials were adapted from Long et al. (under review) [10].

Table 1. Coding of new and given markers in each of the languages.

Coding	Mandinka	Hindi	English	Spanish
A	Bare noun	Numeral 'one'	Indefinite	Indefinite
B	Demonstrative + noun	Bare noun	Definite	Definite
C	Pronoun	Pronoun (overt or not)	Pronoun	Pronoun (overt or not)

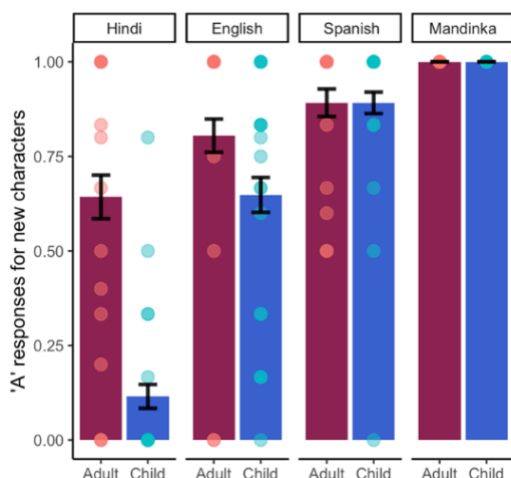


Figure 2. Mean proportions of 'A' markers for new characters across ages in each of the four languages. Error bars represent 95% confidence intervals and points reflect participant means.

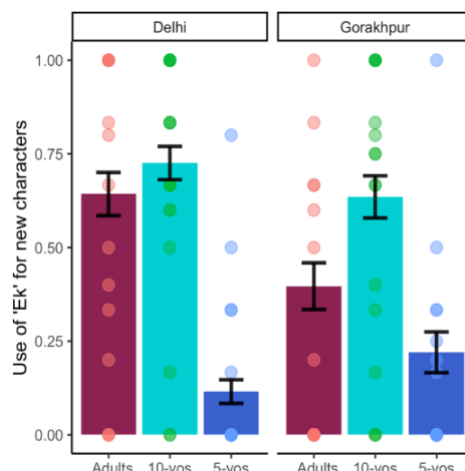


Figure 3. Proportion of 'ek' uses to introduce new discourse characters across three age groups and two regional varieties of Hindi. Error bars represent 95% confidence intervals and points reflect participant means.

References:

- [1] Kachru, 2006. *Hindi*. John Benjamins [2] Dayal, 2018. *Trends in Hindi Linguistics*. Mouton de Gruyter [3] Givón, 1981. *Folia Linguistica Historica* [4] Heine, 1997. *Cognitive Foundations of Grammar*. OUP [5] Chierchia, 1998. *Natural Language Semantics* [6] Heine & Kuteva, 2006. *The changing languages of Europe*. OUP [7] Hickmann et al., 2015. *The Acquisition of Reference*. John Benjamins [8] Vion & Colas, 1999. *Journal of Psycholinguistic Research* [9] Wong & Johnston, 2004. *Journal of Child Language* [10] Long, Rohde, Oraa Ali & Rubio-Fernandez (under review), Speaker-internal biases in referential choice remain stable over the adult lifespan [11] Rubio-Fernandez, 2020. *Synthese, SI The Cultural Evolution of Human Social Cognition*.

The role of language context in the acquisition of novel words

Anna Alberski, Kathryn Schuler (University of Pennsylvania)

Researchers have developed a robust understanding of how mutual exclusivity is used by word learners to make predictions about possible referents (e.g. Markman & Wachtel, 1988). Yet, most studies have focused on acquiring words in isolation, despite children's input consisting of words embedded in rich linguistic contexts (Hoff-Ginsberg, 1990). We propose that such linguistic contexts play an equally important role in acquiring word meanings. While a large body of work has investigated children's use of language context to acquire verb meanings (e.g. Gleitman, 1990), considerably less attention has been given to the role of language context in the acquisition of noun meanings. We know that adults use verb information to predict upcoming familiar nouns (Altmann & Kamide 1999), but it is not clear whether, or how, such linguistic information is used to acquire the meanings of novel words.

In the present experiment, we ask whether adults can use verb information to predict upcoming **novel** nouns during sentence processing, just as they can for familiar nouns (as in Altmann & Kamide 1999). Further, we ask how their use of this language context cue compares to their use of mutual exclusivity alone. On each of 24 trials, participants saw two images—one novel and one known—and were asked to select one (e.g. “Mary wants to eat the wug. Click on the wug!”). Crucially, while mutual exclusivity was always informative, the language context (here, the verb “eat”) was only informative when one of the available referents was edible (and uninformative when both were; see Figure 1). In half of the trials, the correct referent was novel (e.g. unfamiliar fruit with the label “wug”), and in the other half, the correct referent was known (e.g. bananas).

Figure 2 shows a mixed-effect analysis of participant reaction times (log RT and untransformed raw RT). Each model included **word type** and **language context** as simple coded fixed effects and by-participant random intercepts. Overall, participants took longer to select the target referent when the language context was uninformative ($\chi^2(1)=55.42$, $p<0.001$; $\beta=350.39$, $SE=44.62$, $t=7.85$) and took longer to select the target referent when the target word was novel ($\chi^2(1)=8.38$, $p<0.001$; $\beta=143.79$, $SE=44.66$, $t=3.22$). An interaction between word type (known, novel) and language context (informative, uninformative) suggests that the main effect of word type depended on language context ($\chi^2(1)=13.36$, $p<0.001$): participants took significantly longer to select novel targets, but only when the language context was uninformative ($\beta=327.22$, $SE=89.23$, $t=3.67$, see Figure 2).

Our results suggest that learners can use verb information to predict upcoming nouns equally well, regardless of whether the nouns are novel or known. Further, learners may be able to use rich language contexts to predict the meanings of upcoming novel words, even **before** these words are heard. This process could have a facilitative effect on word-learning, whereby meanings are first predicted and then reinforced upon hearing the novel word. In ongoing work, our lab is conducting several follow-up experiments to ask whether (and at what age) children can similarly make use of language context to acquire novel word meanings, and whether children can leverage both the language context and mutual exclusivity in sophisticated ways to acquire novel words in noisy environments. Our findings will not only have important implications for theories of word learning, but will also emphasize the important role that language itself plays in children's early vocabulary development.

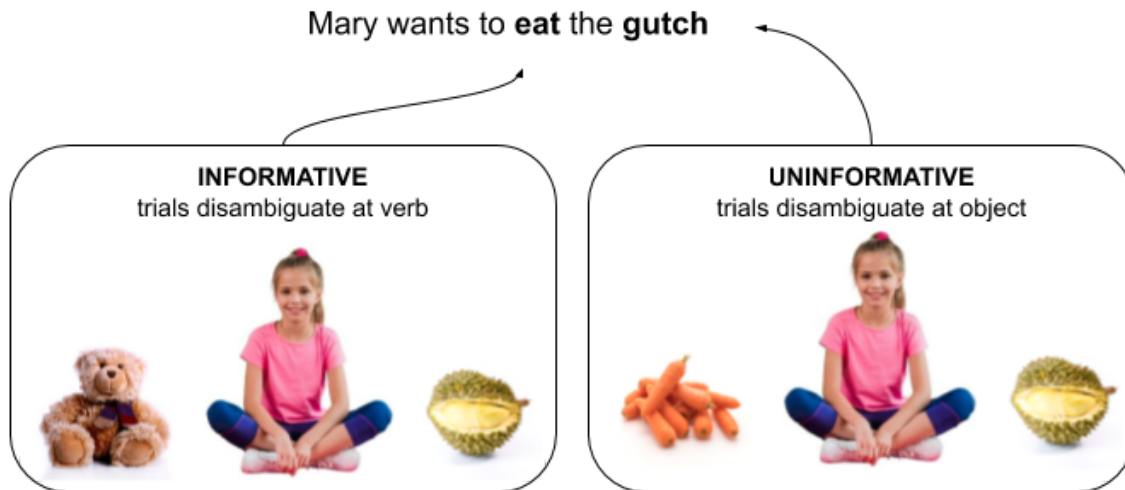


Figure 1. Two sample trials. Participants hear the sentence “Mary wants to eat the” followed by known (e.g. carrots) or novel (e.g. gutch) nouns. The verb “eat” is informative when one referent is edible and uninformative when both are.

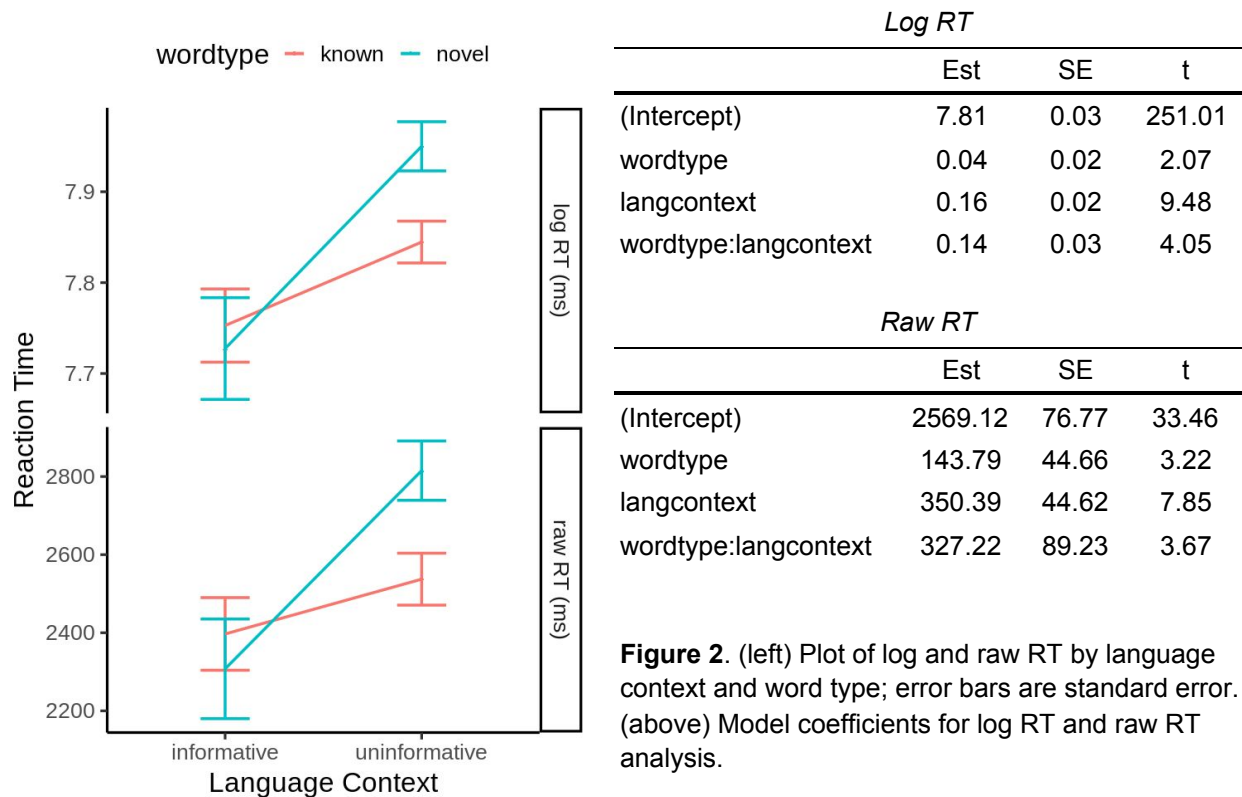


Figure 2. (left) Plot of log and raw RT by language context and word type; error bars are standard error. (above) Model coefficients for log RT and raw RT analysis.

References. Altmann & Kamide (1999) Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*. Gleitman (1990) The structural sources of verb meanings. *Language Acquisition* Hoff-Ginsberg (1990) Maternal speech and the child’s development of syntax: A further look. *Journal of child language*. Markman & Wachtel (1988) Children’s use of mutual exclusivity to constrain the meaning of words. *Cognitive psychology*.

Effects of lifetime and fact knowledge in language comprehension

Daniela Palleschi^{1,2,3}, Pia Knoeferle^{1,2,3}

¹Humboldt-Universität zu Berlin, Department of German Studies and Linguistics; ²Berlin School of Mind and Brain;

³Einstein Center for Neurosciences Berlin; daniela.palleschi@hu-berlin.de

Background Various forms of knowledge can rapidly affect language comprehension, such as who-does-what-to-whom (Kamide, Scheepers, & Altmann, 2003), what can be done with objects (Chambers, Tanenhaus, & Magnuson, 2004), and knowledge like the color of Dutch trains (Hagoort, Hald, Bastiaansen, & Peterson, 2004). Effects of world knowledge (e.g., the color of Dutch trains) resemble effects of lexical semantics (Hagoort et al., 2004) in EEG studies, each eliciting an N400 effect. Knowledge of a referent's lifetime (dead/alive) is, by contrast, integrated with temporal morphology in reading only at sentence end (Chen & Husband, 2018). The extent to which findings for rapid integration of world knowledge extend to biographical knowledge of individuals (e.g., alive or dead; biographical facts) has yet to be more fully explored.

Present Study The current study examines how specific biographical knowledge stored in long-term memory, prompted by a picture of a famous cultural figure, is integrated during processing of two types of information: temporal phrases (in relation to lifetime) and biographical information (in relation to biographical knowledge). The study thus informs theories of sentence processing about the integration of long-term knowledge of specific individuals: their lifetime (dead/alive) and biographical facts (e.g., starring in a certain film).

Procedure In an internet-based self-paced reading study (run on Ibex Farm), native German speakers ($N = 160$, aged 18-31) were presented pictures of famous cultural figures, half living and half dead ($N = 24$). After indicating whether they were familiar with the cultural figure, participants were presented with a fictional statement from the cultural figure (ex. 1) in which the cultural figure mentions in which year some accomplishment of theirs occurred (e.g., appearing in a film). Participants indicated whether the sentence was true given the picture preceding it. Critical items contained two two-level factors: *life-time congruence* (match vs. mismatch; at *year sentence region*) and *fact congruence* (match vs. mismatch; at *fact sentence region*), resulting in 4 conditions (full match, life-time mismatch, fact mismatch, or double mismatch). If long-term knowledge of lifetime and biographical facts of cultural figures are each rapidly integrated with language processing, we predicted (i) processing costs would be elicited from the year and fact regions in conditions containing the respective violations. We further predicted (ii) stronger effects for fact (e.g., song) than life-time (year) mismatches motivated by stronger effects for referential than non-referential relations in psycholinguistic research.

Results Post-trial responses (Fig. 1) indicated a main effect of life-time mismatch in accuracies (i.e., life-time mismatch + double mismatch vs. full match + fact mismatch): sentences containing lifetime violations received significantly higher accuracies than those that did not ($z = -15.3$, $p < .001$). Trials which received an incorrect response were excluded from reading time analyses. A main effect of *life-time mismatch* was found in the regions 'fact' ($p < .05$), fact+1 ($p < .001$), fact+2 ($p < .001$), name ($p < .001$), and the final region ($p < .001$). No main effect of *fact mismatch* was found, nor an interaction effect. Upon visual inspection of the reading times by condition (Fig. 2), this main effect of *life-time mismatch* resulted in shorter reading times for both the *life mismatch* and *double mismatch* conditions. Results are summarised in Table 1.

Summary and Conclusion The longer reading times for lifetime matches than mismatches (only) go against our expectations of (i) violations eliciting longer reading times, and (ii) a larger effect for fact mismatches than lifetime mismatches. It is possible the shorter reading times for life mismatches reflect explicit detection of the violation during comprehension, leading to 'speeding-up' in later regions, as participants had enough information to make the post-trial binary decision. The lack of a main effect of fact mismatch, despite a 73% rejection rate for the fact mismatch condition, could be attributed to later integration of specific biographical knowledge, which is more varied (singers release many songs) than lifetime knowledge (someone is either dead or alive). Predicting specific biographical knowledge effects during comprehension then seems to involve also modeling the multi-faceted and more or less variable nature of experience-based knowledge.

Example sentence

(ex1)	„Im Jahr	<u>2016</u> / <u>1968</u>	habe ich	das Lied	<u>„Formation“</u> / <u>„Hey Jude“</u>
gloss:	<i>In the year</i>	<i>2016_{match} / 1968_{mismatch}</i>	<i>I (have-aux)</i>	<i>the song</i>	<i>„Formation“_{match} / „Hey Jude“_{mismatch}</i>
region:	(pre-year)	(year)	(year+1)	(year+2)	(fact)
	aufgenommen,“	das verkündete	Beyoncé	gegenüber der Presse.	
gloss:	<i>released,</i>	<i>announced</i>	<i>Beyoncé</i>	<i>to the press</i>	
region:	(fact+1)	(fact+2)	(name)	(final)	
translation:	“In the year <u>2016</u> / <u>1968</u> , I released the song ‘ <u>Formation</u> / <u>Hey Jude</u> ,’ Beyoncé told the press.				

Results

	fact			fact+1			fact+2			name			final			Responses	
	t=	p<	d=	t=	p<	d=	t=	p<	d=	t=	p<	d=	t=	p<	d=	z=	p<
Life	3.7	.05	.07	7.8	.001	.11	4.1	.001	.06	4.8	.001	.07	10	.001	.15	-15.3	.001
Fact																	
Inter.																	

Table 1: t-values (df = 147), p-values, and Cohen’s d for reading times per region, and post-trial responses. Reading time p-values are Bonferroni corrected for multiple comparisons (multiplied by eight; once for each region analysed). Critical regions where effects were insignificant (i.e., $t < 2$, $p > .05$; regions ‘year’, ‘year+1’ and ‘year+2’) are omitted for visual simplicity. Responses: z-scores are reported (rather than t-values) and Cohen’s d is omitted as it is not suitable for binomial data

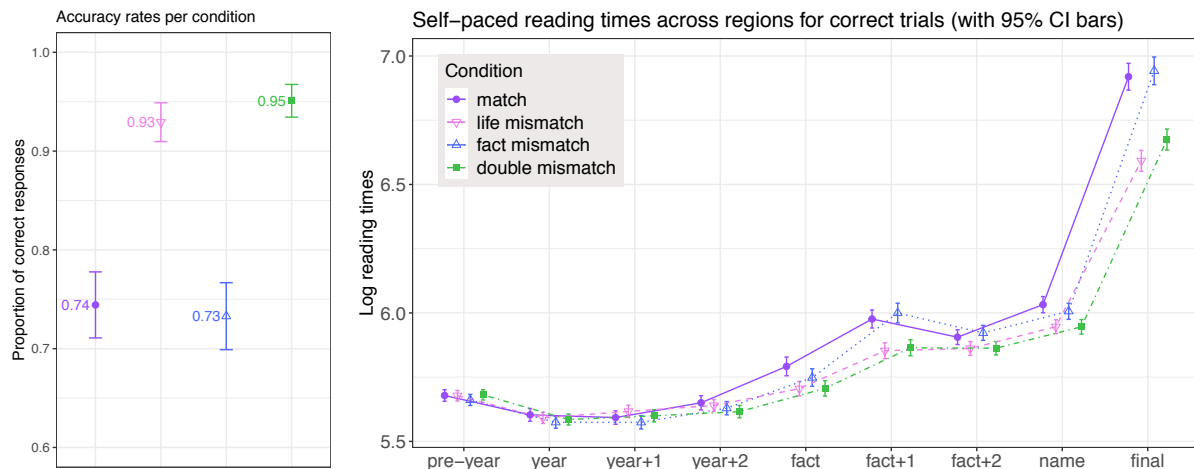


Figure 1 (left): mean accuracies per condition (with 95% confidence intervals); conditions correspond to legend from Figure 2.

N.B., ‘accuracy’ corresponds to proportion of acceptances for the full match condition, and rejections for all other conditions

Figure 2 (right): mean log-transformed self-paced reading times across sentence regions.

References

- Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3), 687.
- Chen, S. Y., & Husband, E. M. (2018). Contradictory (forward) lifetime effects and the non-future tense in Mandarin Chinese. *Proceedings of the Linguistic Society of America*, 1–14.
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, 304, 438–442.
- Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of Syntactic and Semantic Information in Predictive Processing: Cross-Linguistic Evidence from German and English. *Journal of Psycholinguistic Research*, 32(1), 37–55.

Effect of referent lifetime in the processing of verbal morphology: a self-paced reading study

Daniela Palleschi^{1,2,3}, Camilo Rodríguez Ronderos^{1,4}, Pia Knoeferle^{1,2,3}

¹Humboldt-Universität zu Berlin, ²Einstein Center for Neurosciences Berlin, ³Berlin School of Mind and Brain, ⁴University of Oslo
daniela.palleschi@hu-berlin.de

Background Theories of sentence processing, as well as studies, suggest multiple linguistic and non-linguistic sources of information are integrated during comprehension (e.g., Altmann & Kamide, 2007; Nieuwland & Martin, 2012). Information about a referent's lifetime (dead or alive), for example, has been shown to be integrated with temporal morphology in a phenomenon known as *the Lifetime Effect*, eliciting processing costs and lower ratings when the Present Simple or Present Perfect are used to refer to dead referents (Chen & Husband, 2018; Palleschi et al., 2020). Building on these findings, the current study contrasts the English Present Perfect with the Past Simple in the context of the Lifetime Effect. The former links completed past events to the present through 'current relevance' or 'future possibility' (Klein, 1992; ex. 1a), whereas the Past Simple requires a link to a completed past time frame, and has been described as anaphoric (e.g., Partee, 1973; ex. 1b). When no explicit mention of a time frame is mentioned, the temporal context may be inferred to be the lifetime of a referent, invoking *The Lifetime Effect* (dead = past, living = present; Musan, 1997; ex. 2). Thus, when the Past Simple is used with a living individual in the absence of a completed past time frame, the utterance is left 'hanging in the air' due to the missing past temporal antecedent. Meanwhile, the use of the Present Perfect to describe a dead referent violates the 'current relevance' requirement (Klein, 1992). The current study thus involves the integration of lifetime knowledge with temporal morphology, further refining the types of information considered immediately available in theories of language processing, and provides a first glimpse into the processing of the Present Perfect Lifetime Effect contrasted with the Past Simple.

Present Study In a cumulative self-paced reading experiment, the Present Perfect and Past Simple were presented in sentences describing accomplishments of dead and living cultural figures, with no temporal references given. The lifetime of the cultural referents therefore provided the frame of temporal reference, with the dead and living being congruent with the Present Perfect and Past Simple, respectively. Verbs ($n=10$) were counterbalanced across conditions. Differences between the dead and living conditions within each verb tense would be evidence of the integration of lifetime context in the processing of temporal morphology.

Procedure In an online cumulative self-paced reading experiment, native British English speakers ($n = 160$, 111 female, aged 18-31) read sentences (20 critical and 30 filler items) describing the occupation and life status of a cultural figure (ex. 3a/b), followed by a critical sentence containing either the present perfect (PP; ex. 4a) or past simple (PS; ex. 4b). A post-trial binary naturalness judgement task followed. Lower proportions of 'yes' responses and longer reading times from the verb region onward were expected for the *dead-PP* and *living-PS* conditions compared to their congruent lifetime counterparts, respectively. Stronger effects were expected for violations containing the PP, following Roberts & Liszka (2013). Linear mixed-effects regression models were fitted to the log reading times from the verb region onward. A generalised linear mixed model was run on the binary response data.

Results Conditions were contrast coded using sliding contrasts. Of interest, the *dead-PP* elicited significantly longer reading times than the *living-PP* in the 'adjective' region and the two penultimate sentence regions, while the *living-PS* elicited longer reading times in the 'object-NP' and sentence-final region (Fig. 1). The effect (*Cohen's d*) was larger for the PP violations. In addition, the *dead-PP* and *living-PS* both elicited higher rejection rates (Fig. 2).

Conclusion The effect found in the present perfect conditions indicate that violations of the Present Perfect Lifetime Effect elicit processing costs, indicating difficulties integrating the Present Perfect in the context of a completed lifetime. Meanwhile, the past simple effect provides initial support for processing delays elicited by sentences left 'hanging in the air' by a lack of a completed past temporal antecedent in the living condition. For violations of both the Present Perfect Lifetime Effect and the Past Simple anaphora, the ratings indicated explicit awareness of the violations. The reading results indicate that lifetime contexts of well-known cultural figures are integrated with temporal morphology, further informing processing theories regarding the types of information available during comprehension. To what extent long-term knowledge of the cultural figures contributed to these effects will be explored in future studies.

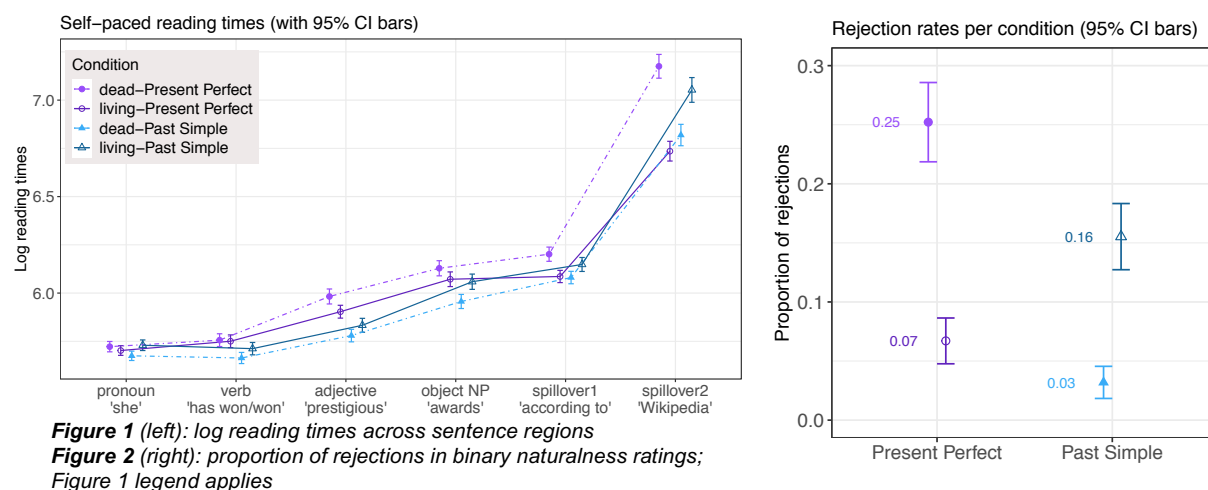
Example sentences

1a.	John <u>has seen</u> his sister twice <u>since last year</u> /*last year.	Present Perfect
1b.	John <u>saw</u> his sister twice <u>last year</u> /*since last year.	Past Simple
2a.	Angela Merkel <u>has accomplished</u> / ?? <u>accomplished</u> a lot.	Living
2b.	Abraham Lincoln <u>accomplished</u> / ?? <u>has accomplished</u> a lot.	Dead
3a.	Beyoncé <u>is</u> an American performer. She <u>lives</u> in California.	Living
3b.	Whitney Houston <u>was</u> an American performer. She <u>died</u> in California.	Dead
4a.	She <u>has performed</u> in many arenas, according to Wikipedia.	Present Perfect
4b.	She <u>performed</u> in many arenas, according to Wikipedia.	Past Simple

Results

	Tense			Lifetime			Interaction			Dead-livingPP			living-deadPS		
	$t_{138} =$	$p <$	$d =$	$t_{138} =$	$p <$	$d =$	$t_{138} =$	$p <$	$d =$	$t_{138} =$	$p <$	$d =$	$t_{138} =$	$p <$	$d =$
verb	5.95	.001	.08												
adj	7.3	.001	.11				7.3	.01	.05	3.7	.01	.15			
Obj-NP	4.2	.001	.07				3.7	.01	.06				3.5	.01	.15
Spill1							4.7	.001	.08	4.6	.001	.22			
Spill2							2.5	.001	.15	7.5	.001	.44	3.1	.05	.18
Rating							-22	.001	N/A	-6.8	.001	N/A	-6.7	.001	N/A

Table 1: t -values, p -values, and Cohen's d for reading times per region and ratings. Reading time p -values are Bonferroni corrected for multiple comparisons (multiplied by five; once for each region analysed). Insignificant effects are omitted for visual simplicity. Ratings: z -scores are reported (rather than t -values) and Cohen's d was not calculated as it is not suitable for binomial data



References

- Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57, 502–518.
- Chen, S. Y., & Husband, E. M. (2018). Contradictory (forward) lifetime effects and the non-future tense in Mandarin Chinese. *Proceedings of the Linguistic Society of America*, 3(1), 6-1.
- Klein, W. (1992). The present perfect puzzle. *Language*, 68(3), 525–552.
- Musan, R. (1997). Tense, Predicates, and Lifetime Effects. *Natural Language Semantics*, 5(3), 271–301.
- Nieuwland, M. S., & Martin, A. E. (2012). If the real world were irrelevant, so to speak: The role of propositional truth-value in counterfactual sentence comprehension. *Cognition*, 122(1), 102–109. <https://doi.org/10.1016/j.cognition.2011.09.001>
- Palleschi, D., Ronderos, C. R., & Knoeferle, P. (2020). Effects of linguistic context and world knowledge on the processing of tense and aspect: evidence from eye-tracking. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.
- Partee, B. H. (1973). Some Structural Analogies between Tenses and Pronouns in English. *The Journal of Philosophy*, 70(18), 601–609.

A protracted developmental trajectory for English-learning children's detection of consonant mispronunciations in newly learned words

Carolyn Quam (Portland State University), Daniel Swingley (University of Pennsylvania)

Conventionally, children are said to learn the consonants and vowels of their language in infancy. Though refinement of these categories extends throughout childhood, toddlers are expected to know their language's phonological distinctions and to encode and differentiate words using those sounds. This account is supported by demonstrations of native-language category formation, mispronunciations hindering word recognition, and minimal-pair learning. However, there are some wrinkles. Word recognition is blocked at 11 months when stressed syllables (e.g., Swingley, 2005), but not unstressed syllables (Segal et al., 2020) are mispronounced. Toddlers are poor at learning novel neighbors (Stager & Werker, 1997; Swingley & Aslin, 2007) and even 30-month-olds don't spontaneously consider novel neighbors to be new words (Swingley, 2016). To what extent, then, do toddlers really have mature phonology? We addressed this question in a series of word-learning experiments with children, at 19, 24, and 30 months, and adults, included as a developmental endpoint. All children were monolingual native English speakers.

Taught a novel word, adults' and 30-month-olds' recognition is impaired when the stressed vowel is altered, but not when a distinct pitch contour is used (Quam & Swingley, 2010; also Ma et al., 2017; Singh et al., 2014). Here, we evaluated interpretation of consonants—which are widely argued to play a leading role in word differentiation—vs. pitch. We taught a novel word, “deebo,” in an English narrated, animated story followed by ostensive labeling. The word was always pronounced with a consistent intonation contour: rise-fall or low-falling. A second novel object was present but never labeled. Then, recognition was tested in a language-guided looking task. The two objects appeared on the screen, and a spoken sentence presented the correct pronunciation (CP) of the target word, “deebo,” or a version with a one-feature consonant mispronunciation (“consonant MP”: “teebo”). In 18-month-olds and adults, we also tested interpretations of a version with the pitch-contour mispronounced (“contour MP”) from rise-fall to low-falling or vice-versa, as in Quam and Swingley (2010; for 24- and 30-month-olds, pitch interpretations are reported elsewhere). We measured whether participants looked less at the *deebo* for either MP.

Adults (N=18) were each tested with both MPs (**Fig. 1**, left). They showed the expected reduction in recognition for consonant MPs relative to correct pronunciations, $t(17) = 4.62$, $p < .001$, and no reduction for contour MPs. Nineteen-month-olds (N=43) were tested with either consonant or contour MPs (**Fig. 1**, right). Recognition performance was above chance, indicating word learning, but neither MP reduced looking, with a (ns) trend toward better performance on consonant MPs. Thus there was no evidence that 19-month-olds weighted phonologically contrastive consonantal-feature variation more heavily than pitch-contour variation. Given this result, we tested 24- and 30-month-olds on the consonant MP (total N=34; **Fig. 2**). Children readily learned the words, but in contrast to 30-month-olds' substantial vowel-MP effects in the same procedure (Quam & Swingley, 2010), here children's recognition was not measurably impaired by a consonantal MP at either age. By comparison, experiments testing highly familiar words nearly always reveal recognition decrements for consonantal MPs (Von Holzen & Bergmann, 2018).

These results suggest that well into the second year, newly learned words may not be represented with intact phonological features. Children can be trained to attend to phonologically relevant lexical distinctions in fully novel words (e.g., Werker's *Switch* procedure), but processing phonologically divergent variants as distinct words does not necessarily follow from this ability. Further research could also investigate whether cue-weighting differences between children and adults could explain children's weaker sensitivity to consonant MPs.

References

Ma, W., Zhou, P., Singh, L., & Gao, L. (2017). Spoken word recognition in young tone language learners: Age-dependent effects of segmental and suprasegmental variation. *Cognition*, 159, 139-155.

- Quam, C., & Swingle, D. (2010). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *JML*, 62, 135–150.
- Segal, O., Keren-Portnoy, T., & Vihman, M. (2020). Robust effects of stress on early lexical representation. *Infancy*, 25, 500–521.
- Singh, L., Hui, T. J., Chan, C., & Golinkoff, R. M. (2014). Influences of vowel and tone variation on emergent word knowledge: A cross-linguistic investigation. *Devel. Sci.*, 17, 94–109.
- Stager, C. L., and Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640), 381–382.
- Swingle, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Devel. Science*, 8, 432–443.
- Swingle, D. (2016). Two-year-olds interpret novel phonological neighbors as familiar words. *Developmental Psychology*, 52(7), 1011–1023.
- Swingle, D., & Aslin, R.N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, 54, 99–132.
- Von Holzen, K., & Bergmann, C. (2018). A meta-analysis of infants' mispronunciation sensitivity development. In Proceedings of the 40th Annual CogSci Conference (pp. 1157–1162).

Figures

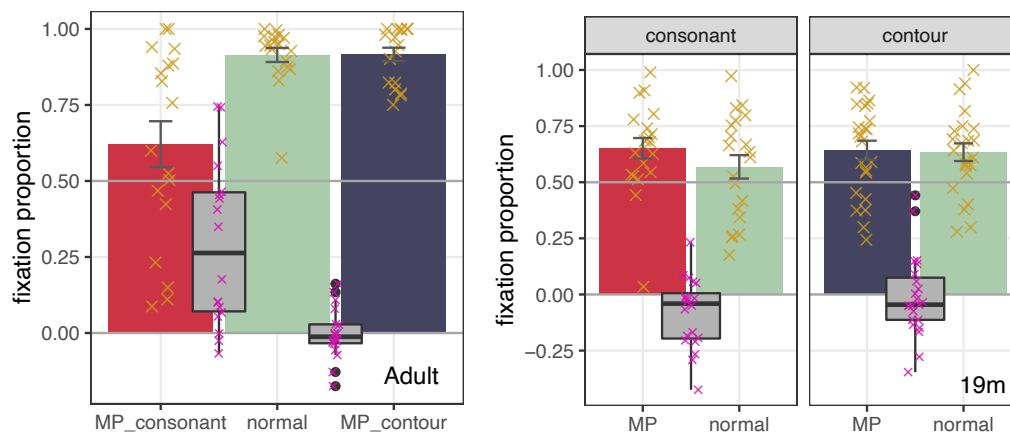


Figure 1. Adults' (left) and 19-month-olds' (right) fixation of the *deebo* object in response to the trained pronunciation ("normal") and the two MPs. The horizontal line indicates chance fixation, or 50%. Each adult was tested with both MPs, while each child was tested with either the consonant or pitch-contour MP. Box plots indicate within-subject difference scores between correct-pronunciation and MP trials.

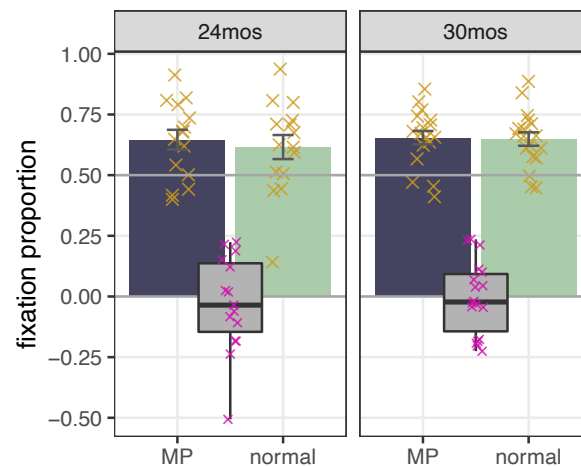


Figure 2. 24- and 30-month-olds' *deebo* fixation in response to the trained pronunciation and the consonant MP. Boxplots in gray show within-subjects difference scores.

34th Annual CUNY Conference on Human Sentence Processing

Friday Morning March 5, 2021

Hour	Session	Time	Title	Authors
Hour 1	1	9:00	Modeling effects of incremental memory and prediction pressures on phoneme learning from speech	Cory Shain and Micha Elsner
Hour 1	1	9:00	Analyzing complex human sentence processing dynamics with CDRNNs	Cory Shain and William Schuler
Hour 1	1	9:00	The Effect of Context on Typing Time: Evidence from 100,000 TypeRacers	Robert Chen, Roger Levy and Tiwalayo Eisape
Hour 1	1	9:00	BERT, a deep-learning language model, processes NPI licensing without suffering from NPI illusion	Unsub Shin and Sanghoun Song
Hour 1	1	9:00	Cross-situational Word Learning from Naturalistic Headcam Data	Wai Keen Vong, Emin Orhan and Brenden Lake
Hour 1	1	9:00	Artificial Language Models Do Not Learn Syntax-Semantics Mappings	Xiaohan Guo, Bryor Sneffjella and Idan Blank
Hour 1	2	9:00	EARLY LEXICAL COMPREHENSION AND GENDER AGREEMENT IN ITALIAN TODDLERS.	Giulia Mornati, Valentina Riva, Elena Vismara, Massimo Molteni and Chiara Cantiani
Hour 1	2	9:00	Spreading jam with a butter knight: Near-homophones and phonological pre-activation	Kari Schwink and Jeffrey Green
Hour 1	2	9:00	Planning ahead: Interpreters predict source language in consecutive interpreting	Nan Zhao, Xiaocong Chen and Zhenguang Cai
Hour 1	2	9:00	Perception of disfluencies in non-native speech	Rajalakshmi Satarai Madhavan and Martin Corley
Hour 1	2	9:00	Evidence for two-stage accounts of prediction	Ruth Corps, Charlotte Brooke and Martin Pickering
Hour 1	3	9:00	Turning the young parser into the adult parser: Working memory matters	Jiawei Shi and Peng Zhou
Hour 1	3	9:00	The Meaning and Processing of Conditionals – German ‘wenn’ (if) vs. ‘nur wenn’ (only if)	Mathias Barthel and Mingya Liu
Hour 1	3	9:00	Adults process Number and Gender head-subject mismatches differently during the	Nicoletta Biondo, Vincenzo Moscati, Luigi Rizzi and Adriana Belletti

Hour	Session	Time	Title	Authors
			online comprehension of object-relative clauses (as children do, offline).	
Hour 1	3	9:00	Developmental effects in the real-time use of morphosyntactic cues: Evidence from Tagalog	Rowena Garcia, Gabriela Garrido Rodriguez and Evan Kidd
Hour 1	3	9:00	Selective Modulation of Syntactic Processing by Anodal tDCS over the Left Inferior Frontal Region	Shinri Ohta
Hour 1	3	9:00	What is the upper limit of working memory? Evidence from Chinese recursive possessive structure	Sihan Zhang, Shuqi Ni, Shuyang Liu and Fuyun Wu
Hour 1	4	9:00	Reliance on semantic and structural heuristics across the lifespan	Anastasiya Lopukhina, Anna Laurinavichyute and Svetlana Maljutina
Hour 1	4	9:00	Keep calm and move on: Reduced processing advantage of an early-arriving morphological cue in comprehension of Korean suffixal passive construction	Chanyoung Lee and Gyu-Ho Shin
Hour 1	4	9:00	Processing noncanonical sentences: online and offline effects on misinterpretation errors	Markus Bader and Michael Meng
Hour 1	4	9:00	Online representations of implausible non-canonical sentences are more than good-enough	Michael Cutter, Kevin Paterson and Ruth Filik
Hour 1	4	9:00	Age invariance in syntactic prediction during self-paced reading	Michael Cutter, Kevin Paterson and Ruth Filik
Hour 1	4	9:00	Agreement attraction in grammatical sentences arises only in the good-enough processing mode	Anna Laurinavichyute, Titus von der Malsburg
Hour 1	5	9:00	The distributional learning of recursive structures	Daoxin Li, Lydia Grohe, Petra Schulz and Charles Yang
Hour 1	5	9:00	The effects of input typicality (or variability) on the acquisition of argument structure constructions	Eunkyung Yi and Jia Kang
Hour 1	5	9:00	Predictive effects of number-marked verbs and copulas in Czech 2-year-olds	Filip Smolík and Veronika Bláhová
Hour 1	5	9:00	The acceptability of null subjects	Juliana Gerard

Hour	Session	Time	Title	Authors
Hour 1	5	9:00	Second language acquisition and language processing: Grammatical gender in Norwegian	Yulia Rodina, Valantis Fyndanis, Bjørn Lundquist, Nina Hagen Kaldhol, Eirik Tengedal and Emel Tjørker-Van der Heiden
Hour 1	6	9:00	Pronouns attract in number but (much) less so in person. Evidence from Romanian.	Adina Camelia Bleotu and Brian Dillon
Hour 1	6	9:00	Dynamics of referent demotion and promotion: Consequences for pronoun interpretation	Jina Song and Elsi Kaiser
Hour 1	6	9:00	Antecedent prominence and the Chinese reflexive ziji	Jun Lyu and Elsi Kaiser
Hour 1	6	9:00	Anaphora resolution in causal coherence relations in Chinese	Jun Lyu and Elsi Kaiser
Hour 1	6	9:00	Investigating perspective-sensitivity during the resolution of Korean anaphors	Sarah Hye-Yeon Lee and Elsi Kaiser
Hour 1	6	9:00	Interpretation of null pronouns in Mandarin Chinese does not follow a Bayesian model	Suet Ying Lam and Heeju Hwang
Hour 2	7	10:00	Source monitoring and false information endorsement in native and foreign language: an online study with Russian-English bilinguals	Aleksandra Dolgoarshinnaia and Beatriz Martín
Hour 2	7	10:00	ERP decoding shows bilinguals represent the language of a code-switch after lexical processing	Anthony Yacovone, Moshe Poliak, Harita Koya and Jesse Snedeker
Hour 2	7	10:00	The use of pronoun interpretation biases in unbalanced Spanish-English bilinguals: the role of language experience.	Carla Contemori and Alma Armendariz
Hour 2	7	10:00	Changing pronoun interpretations across-languages: discourse priming in Spanish-English bilingual speakers	Carla Contemori and Natalia Irene Minjarez Oppenheimer
Hour 2	7	10:00	Similarity-Based Interference in Native and Non-Native Comprehension	Ian Cunnings and Hiroki Fujita
Hour 2	7	10:00	How do structural predictions operate between languages for multilinguals? Evidence from cross-language structural priming in comprehension	Xuemei Chen and Robert Hartsuiker
Hour 2	8	10:00	What to expect when you are expecting an antecedent: processing cataphora in Dutch	Anna Giskes and Dave Kush

Hour	Session	Time	Title	Authors
Hour 2	8	10:00	The COMP-trace effect and sentence planning: Evidence from L2	Boyoung Kim and Grant Goodall
Hour 2	8	10:00	Prominence guides incremental interpretation: Lessons from obviation in Ojibwe	Christopher Hammerly, Brian Dillon and Adrian Staub
Hour 2	8	10:00	Interference and Filler-Gap Dependencies in Native and Non-Native Comprehension	Ian Cunnings and Hiroki Fujita
Hour 2	8	10:00	Inference in the processing of complement control: an eye-tracking study on lexically determined long-distance dependencies	Iria de Dios Flores, Juan Carlos Acuña-Fariña, Simona Mancini and Manuel Carreiras
Hour 2	8	10:00	Processing embedded clauses in Korean: silent element or a dependency formation?	Nayoun Kim, Keir Moulton and Daphna Heller
Hour 2	9	10:00	Does negation influence the choice of sentence continuations? Evidence from a four-choice cloze task	Elena Albu, Carolin Dudschig, Tessa Warren and Barbara Kaup
Hour 2	9	10:00	Contrary to expectations: Is negation more difficult than affirmation?	Elena Albu, Oksana Tsaregorodtseva and Barbara Kaup
Hour 2	9	10:00	Negation cancels discourse-level processing differences: Evidence from reading times in concession and result relations	Ludivine Crible
Hour 2	9	10:00	Verifying negative sentences - How context influences which strategy is used	Shenshen Wang, Chao Sun and Richard Breheny
Hour 2	9	10:00	Processing polar questions in contexts with varying epistemic biases in English	Vinicius Macuch Silva and E Jamieson
Hour 2	9	10:00	Uniformity and variability in the understanding of expletive negation across languages	Yanwei Jin and Jean-Pierre Koenig
Hour 2	10	10:00	Testing the influence of the listener's perspective in the epistemic step.	Blanche Gonzales de Linares and Napoleon Katsos
Hour 2	10	10:00	The costs and benefits of different metaphoric structures: evidence from pupillometry	Camilo Rodriguez Ronderos, Ernesto Guerra and Pia Knoeferle
Hour 2	10	10:00	Ageing and communication in face-threatening contexts	Madeleine Long, Sarah MacPherson and Paula Rubio-Fernandez

Hour	Session	Time	Title	Authors
Hour 2	10	10:00	The social benefits of being a non-native speaker	Martin Ho Kwan Ip and Anna Papafragou
Hour 2	10	10:00	Viewing the Metaphor Interference Effect in context	Shaokang Jin and Richard Breheny
Hour 2	10	10:00	How many response options in a TVJT? It depends	Yuhan Zhang, Giuseppe Ricciardi and Kathryn Davidson
Hour 2	11	10:00	Are there segmental and tonal effects on syntactic encoding? Evidence from structural priming in Mandarin	Chi Zhang, Sarah Bernolet and Robert Hartsuiker
Hour 2	11	10:00	The dynamic prominence status of Patient in Mandarin sentence production	Fang Yang, Martin Pickering and Holly Branigan
Hour 2	11	10:00	Morphological boost in structural priming: Evidence from Czech	Maroš Filip and Filip Smolik
Hour 2	11	10:00	Syntactic Rule Frequency as a Measure of Syntactic Complexity: Insights from Primary Progressive Aphasia	Neguine Rezaii, Kyle Mahowald, Rachel Ryskin, Bradford Dickerson and Edward Gibson
Hour 2	11	10:00	Does deciding what to say involve deciding how to say it?	Ruth Corps, Holly Abercrombie, Alix Dobbie, Luke Raben and Martin Pickering
Hour 2	11	10:00	Early preparation during question-answering: Speakers prepare content but not form	Ruth Corps, Laura Lindsay and Martin Pickering
Hour 2	12	10:00	Source of processing costs of indirect anaphors - self-paced reading and ERP data	Magdalena Repp and Petra B. Schumacher
Hour 2	12	10:00	What reaction times can reveal behind acceptability judgments	Eunkyung Yi and Sang-Hee Park
Hour 2	12	10:00	Limits on failure to notice word transpositions during sentence reading	Kuan-Jung Huang and Adrian Staub
Hour 2	12	10:00	Learning speaker-specific 'stylistic' preferences	Nitzan Trainin and Einat Shetreet
Hour 2	12	10:00	Variability in the agreement attraction effect	Sanghee Kim and Ming Xiang
Hour 2	12	10:00	Bayesian surprise predicts incremental processing of grammatical functions	Thomas Hörberg and Florian Jaeger

Modeling effects of incremental memory and prediction pressures on phoneme learning from speech

Cory Shain and Micha Elsner, Ohio State

What learning signals enable infants to discover linguistic patterns from a noisy, information-rich perceptual stream? Some theories of language acquisition invoke *memory pressures* to explain infant learning, arguing that linguistic representations constitute efficient compression codes whose discovery might optimize long-term storage demands [24, 29] and/or working memory demands during real-time speech processing [3]. This view is supported by experimental [18] and modeling [9, 33] evidence, but other work has questioned the efficiency of human mental codes [23] and the utility of memory pressures for language learning [27]. An alternative class of theories invokes *prediction pressures* as a learning signal [31, 15, 1], since knowledge of linguistic regularities might make speech more predictable. Recent work has argued that incremental language models [16, 30] acquire language-like representations from a prediction objective [22] and covary with measures of human processing [13, 36]. This discussion mirrors related discussion about the relative importance of memory and prediction in theories of adult sentence processing [20, 11] and broad neuronal-level learning [2, 26, 5, 17, 34], and, as in those fields [21], memory and prediction pressures may play complementary roles in infant language learning.

In this study, we develop a broad-coverage unsupervised neural network model (Fig 2) to examine possible influences of memory and prediction pressures on infant phoneme learning from speech. Cochlear output is submitted to a hierarchical multiscale recurrent neural network (HM-RNN) [6] speech processor. Each layer of the network processes representations from the layer below, dividing them into discrete segments; at predicted segment boundaries, the layer both (1) emits its segment label (hidden state) to the layer above, and (2) flushes its working memory and refreshes it with top-down guidance from the layer above. In this way, the encoder generates a sparse, hierarchical speech representation over multiple timescales. We apply a novel incremental objective function that at any point in time attempts to reconstruct B segments into the past and predict F segments into the future from the layer below, applied only at incoming segment boundaries. Learning is driven only by these objectives, without supervision for boundary locations or segment labels. Our model implements several independently supported cognitive constraints: incrementality [35]; hierarchically organized [14, 25], feature-rich [7, 28] segmental [32, 19] representations; interactive top-down and bottom-up information flow [38, 10]; modeling of its own sequence of latent representations [12]; and local error signals that are plausibly supported by human working memory [4, 8].

We use the model to study phoneme learning from English and Xitsonga speech in the Zerospeech 2015 dataset [37], (1) experimentally manipulating $B \in \{0, 5, 25, 50\}$ (strength of memory pressure) and $F \in \{0, 1, 5, 10\}$ (strength of prediction pressure), along with number of layers $L \in \{2, 3, 4\}$, and (2) evaluating the impact of these manipulations on three measures of phoneme induction quality: (i) alignment between modeled and human-annotated phoneme boundaries, and (ii) phoneme and (iii) phonological feature (e.g. $[\pm\text{voice}]$) classification accuracy from a linear probe of the first layer's hidden state (the most phoneme-like in tuning experiments). Evaluations on a held-out test set (Fig 1) show statistically significant benefits of working memory pressures (better performance when $B > 0$), prediction pressures (better performance when $F > 0$), and depth (better performance when $L > 2$), with a general peak in performance when $B \approx 25$ and $F \approx 5$. The optimality of these values is intriguing because they correspond respectively to 250ms and 50ms intervals, which fall within estimates of the storage duration in humans of the unanalyzed acoustic traces [8] that are needed to compute the objectives. Performance patterns are largely consistent across languages and metrics, supporting a language-general, complementary influence of memory and prediction pressures on overall phoneme learning. We also compare against an architecturally matched untrained baseline (Baseline U) and against an architecturally matched cross-language baseline (Baseline X, i.e. English training and Xitsonga evaluation, or *vice versa*). Baseline U measures architectural inductive bias (how much phoneme knowledge can be derived from processor design, without learning), while Baseline X measures domain inductive bias (how much phoneme knowledge can be derived from knowledge of some form of human speech, without exposure to the target language). Both kinds of biases might plausibly be innately specified, and comparisons indicate that the full model systematically outperforms them only in the presence of both memory and prediction pressures. Memory and prediction pressures thus modulate not only absolute acquisition performance, but also the utility of language experience *vis-a-vis* plausible inductive biases. Our study therefore suggests that both memory-driven and prediction-driven learning signals may be available to infants during early phoneme acquisition.

Selected References

- [1] Apfelbaum, K. S. and McMurray, B. *Cognitive science*, 2017.
- [2] Atneave, F. *Psychological review*, 1954.
- [3] Baddeley, A., Gathercole, S., and Papagno, C. *Psychological Review*, 1 1998.
- [4] Baddeley, A. D., Thomson, N., and Buchanan, M. *Journal of Verbal Learning and Verbal Behavior*, 1975.
- [5] Bialek, W., Nemenman, I., and Tishby, N. *Neural computation*, 2001.
- [6] Chung, J., Ahn, S., and Bengio, Y. In *International Conference on Learning Representations 2017*, 2017.
- [7] Clements, G. N. *Phonology*, 1985.
- [8] Cowan, N. *Psychological bulletin*, 1984.
- [9] Elman, J. L. *Cognition*, 1993.
- [10] Feldman, N., Griffiths, T., and Morgan, J. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2009.
- [11] Ferreira, F. and Chantavarin, S. *Current directions in psychological science*, 2018.
- [12] Friston, K. *Nature reviews neuroscience*, 2010.
- [13] Goodkind, A. and Bicknell, K. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 2018.
- [14] Hasson, U., Chen, J., and Honey, C. J. *Trends in cognitive sciences*, 2015.
- [15] Johnson, M. A., Turk-Browne, N. B., and Goldberg, A. E. *Behavioral and Brain Sciences*, 2013.
- [16] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. *arXiv preprint arXiv:1602.02410*, 2016.
- [17] Keller, G. B. and Mscis-Flogel, T. D. *Neuron*, 2018.
- [18] Kersten, A. W. and Earles, J. L. *Journal of Memory and Language*, 2001.
- [19] Kooijman, V., Junge, C., Johnson, E. K., Hagoort, P., and Cutler, A. *Frontiers in psychology*, 2013.
- [20] Levy, R. *Cognition*, 2008.
- [21] Levy, R., Fedorenko, E., and Gibson, E. *Journal of Memory and Language*, 2013.
- [22] Linzen, T., Dupoux, E., and Goldberg, Y. *Transactions of the Association for Computational Linguistics*, 2016.
- [23] McMurray, B., Tanenhaus, M. K., and Aslin, R. N. *Cognition*, 2002.
- [24] Newport, E. *Cognitive Science*, 1990.
- [25] Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E., Feldstein, N. A., McKhann, G. M. V., Schevon, C., Flinker, A., and Mesgarani, N. *bioRxiv*, 2020.
- [26] Olshausen, B. A. and Field, D. J. *Nature*, 1996.
- [27] Perfors, A. *Journal of Memory and Language*, 2012.
- [28] Pierrehumbert, J. B. *Frequency and the emergence of linguistic structure*, 2001.
- [29] Pinker, S. *Science*, 1991.
- [30] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. *OpenAI Blog*, 2019.
- [31] Rohde, D. L. T. and Plaut, D. C. *Cognition*, 1999.
- [32] Sanders, L. D. and Neville, H. J. *Cognitive Brain Research*, 2003.
- [33] Shain, C. and Elsner, M. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
- [34] Singer, Y., Teramoto, Y., Willmore, B. D. B., Schnupp, J. W. H., King, A. J., and Harper, N. S. *eLife*, 2018.
- [35] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. E. *Science*, 1995.
- [36] van Schijndel, M. and Linzen, T. In *EMNLP 2018*, 2018.
- [37] Versteegh, M., Thiollère, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., and Dupoux, E. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [38] Warren, R. M. *Science*, 1970.

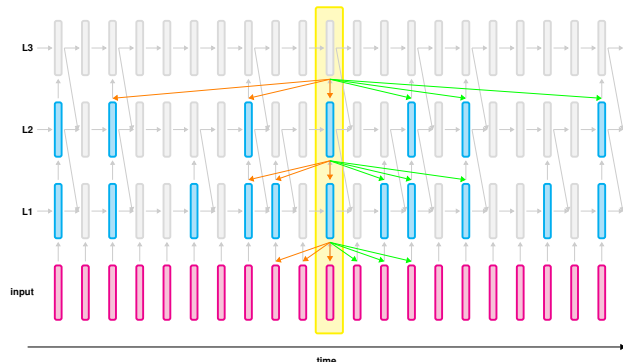


Figure 2: Model. Cyan indicates boundaries, grey arrows show encoder information flow, and orange and green arrows respectively show backward and forward decoder information flow.

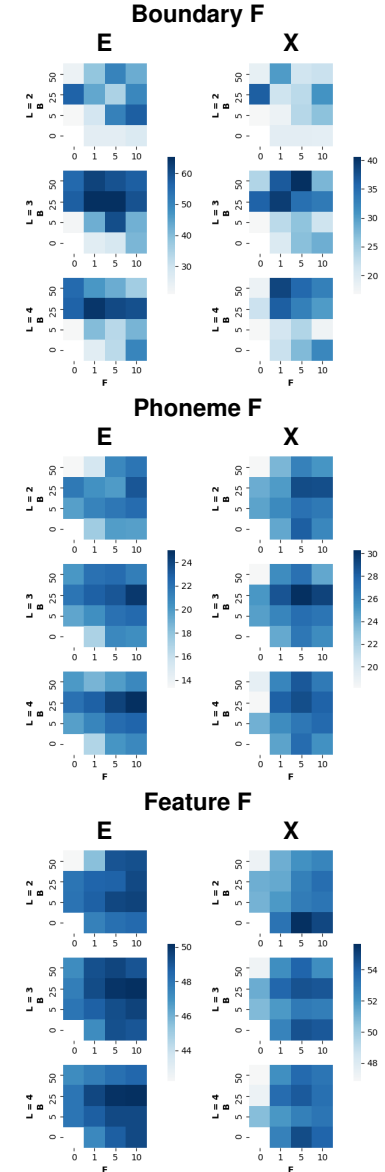


Figure 1: Acquisition performance. Phoneme boundary, label, and feature learning scores respectively (darker is better) in English (E) and Xitsonga (X) as modulated by memory pressure (y -axes), prediction pressure (x -axes), and depth (top to bottom by column, shallowest to deepest). All three variables improve phoneme acquisition.

Analyzing complex human sentence processing dynamics with CDRNNs

The empirical predictions of theories of incremental sentence processing are often cached out in word-level features [7, 5, 9, 8, 12], but experimental measures from human participants reflect the dynamics of real-time cognition, which may be complex, non-linear, and time-varying [1]. Thus, a central challenge in psycholinguistic theory evaluation is to specify a sufficiently expressive linking function between theory-driven word features and measures of human sentence processing, and prior work has argued that widely-used linear time series models may be inadequate, especially for naturalistic designs [1, 2, 14]. The recently proposed technique of continuous-time deconvolutional regression (CDR) relaxes assumed *instantaneity* of effects [14], instead directly estimating continuous-time *impulse response functions* (IRFs) that describe the temporal extent of a predictor's influence on the response (cf. e.g. spillover regressors [10], which ignore clock time). Empirically, CDR learns plausible effect estimates that generalize significantly better than linear models across experimental modalities [15]. However, CDR retains simplifying assumptions that may not hold of human cognition: the parametric form of the IRF must be specified in advance, the IRF is fixed over time (stationary), effects are strictly linear and additive, and the response variance is assumed to be constant (homoskedastic). Any of these constraints may be violated in practice when analyzing the outputs of a complex system like the human mind, with potential impacts on the resulting model.

In this study, we reimplement the CDR impulse response and error distribution using a time-varying deep neural network applied to the predictor sequence (CDRNN). The IRF in CDRNN is therefore a joint (potentially non-linear and interactive) function of all the predictors and time, which can be arbitrarily queried with respect to any collection of feature values, yielding a highly flexible model that relaxes all of the aforementioned simplifying assumptions. We use CDRNN both to (1) reanalyze prior claims based on CDR analysis and (2) shed new exploratory light on the dynamics of human sentence processing. In particular, we focus on the CDR-based claim from [13] that participants' word predictions are syntax-sensitive (significant effects over 5-gram surprisal of probabilistic context-free grammar or PCFG surprisal) based on activity in language-selective voxels during naturalistic listening in an fMRI experiment. To obtain this result, the authors assumed a parametric stationary *hemodynamic response function* (HRF) based on the double-gamma canonical HRF [3] and tied the parameters of the HRF across predictors within each brain region. This design improves on the standard approach of assuming the canonical HRF, instead discovering the HRF from the data [11] and allowing it to vary parametrically by region [6]. However, the HRF is known to be non-stationary, since the vascular response saturates over time [4]. Furthermore, processing effects may coordinate non-linearly [16] and non-additively, especially correlated measures such as different variants of word surprisal.

Nonetheless, CDRNN also shows that PCFG surprisal significantly improves generalization error against a 5-gram baseline ($p < 0.0001^{***}$), validating the prior result obtained using (non-neural) CDR. *Post hoc* analyses show an estimated HRF that independently replicates known features of the HRF, including initial dip, peak response, and undershoot components (Fig 1a), despite the fact that (unlike [13]), no such *a priori* knowledge was provided. The exploratory insights afforded by CDRNN go beyond those obtainable using CDR. For example, similar *post hoc* visualizations (Fig 1b) show a PCFG surprisal response that only ramps up at values larger than its mean of 1.45 standard units (cf. [16]), suggesting that the system may be calibrated to the expected information gain per word. In addition, we find a coordination between 5-gram and PCFG surprisal near the HRF peak (Fig 1c): substantial increases in activity occur only when both variables are large, suggesting a unitary predictive mechanism that exploits both string-level and structural features (and thus incurs high error when both predictive cues are poor), rather than separate mechanisms that track each information source independently. By contrast, PCFG surprisal and unigram surprisal show a different kind of interplay (Fig 1d): there is a large jump in response to PCFG surprisal for highly frequent words (low unigram surprisal) compared to highly infrequent words (high unigram surprisal), possibly because the most frequent words in English tend to be function words that play an outsized role in syntactic structure building (e.g. prepositional phrase attachment decisions). Development of statistical methods for testing non-linearities in the CDRNN-estimated IRF is ongoing, but even in their absence, these results show CDRNN to be a valuable tool for exploratory data analysis, providing detailed insights about how predictors coordinate (potentially non-linearly) over time in order to determine the response. These insights can support both theoretical innovation and the design and testing of standard statistical models.

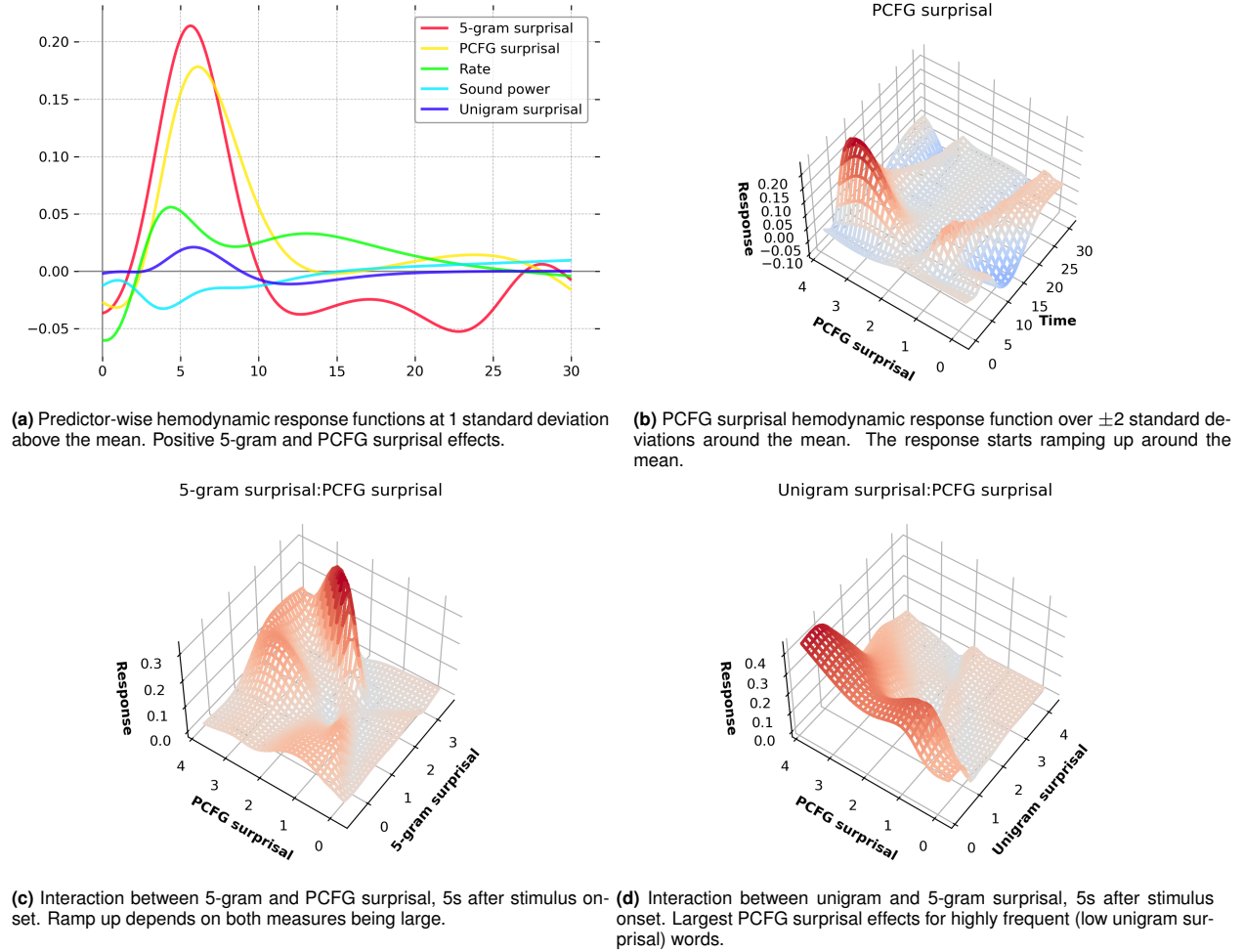


Figure 1: CDRNN estimates from naturalistic fMRI. All predictor values are in standard units.

References

- [1] Baayen, H., Vasishth, S., Kliegl, R., and Bates, D. *Journal of Memory and Language*, 2017.
- [2] Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. 2018.
- [3] Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. *Journal of Neuroscience*, 1996.
- [4] Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. *NeuroImage*, 2000.
- [5] Gibson, E. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 2000.
- [6] Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. *NeuroImage*, 2004.
- [7] Hawkins, J. A. *A performance theory of order and constituency*, 1994.
- [8] Levy, R. *Cognition*, 2008.
- [9] Lewis, R. L. and Vasishth, S. *Cognitive Science*, 2005.
- [10] Mitchell, D. C. *New methods in reading comprehension research*, 1984.
- [11] Pedregosa, F., Eickenberg, M., Ciuciu, P., Gramfort, A., and Thirion, B. *NeuroImage*, 2014.
- [12] Rasmussen, N. E. and Schuler, W. *Cognitive Science*, 2018.
- [13] Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. *Neuropsychologia*, 2020.
- [14] Shain, C. and Schuler, W. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [15] Shain, C. and Schuler, W. *PsyArXiv*, 2019.
- [16] Smith, N. J. and Levy, R. *Cognition*, 2013.

The Effect of Context on Typing Time: Evidence from 100,000 TypeRacers

Robert Chen, Roger Levy, Tiwalayo Eisape (MIT)

robertcc@mit.edu

Context effects in human spoken language are well-documented and play a central role in language production. However, the role of context in written language production is far less well understood, even though a considerable proportion of the language produced by many people today is written. Here we use computational language models (LMs) to quantify the effect of context-based predictability on typing time in a subset of the data available on TypeRacer.com.

TypeRacer is a viral online typing game where players race against themselves, friends, or strangers in groups of up to 10 to complete a short text as quickly as possible (Fig. 1). With races in 50 languages and a wide variety of text genres, TypeRacer is an openly accessible, massive, and untapped dataset of typing times that contains data from over 100,000 users — across 35,000 distinct texts, and 70,000,000 races.

We take a random sample of 100 users from TypeRacer and a random sample of up to 100 races from each of those users, resulting in a total of 317,000 measures of word typing times ($\mu=49.1$ words per race, $\sigma=20.4$). Of our sample of 100 users, 4 do not list their location, 41 are in the United States, followed by 9 in Canada, 5 in India, 4 in the United Kingdom, and 37 from other countries. 92 do not list their age, and the ages of the remaining range from 13 to 28 ($\mu=19.1$, $\sigma=4.8$). 77 do not list their keyboard layout, 22 use Qwerty, and 1 uses Colemak.

We use LMs trained on the WikiText-2 dataset (Merity et al., 2016) to estimate in-context probability for the words in our dataset. The models we compare on this task are as follows. **Forward full surprisal**: An LSTM language model trained with a standard autoregressive language modeling objective. **Backward full surprisal**: A variant of **Forward full surprisal** trained to predict the text in WikiText-2 in reverse. **Forward bigram surprisal**: A variant of **Forward full surprisal** where, during training and inference, context is limited to only the previous word. **Backward bigram surprisal**: A variant of **Forward bigram surprisal** trained to estimate bigram probability in reverse. **Unigram surprisal**: negative log-frequency estimates based on data from the COCA corpus (Davies, 2010).

We use generalized additive models (GAMs; Wood 2006) to determine the functional form of the relationship between each of our context-based predictability estimates and typing time (Fig. 2). Furthermore, to capture both fixed and participant level effects we use a “two-stage” approach (Gelman, 2005) in which we fit a linear mixed-effects model with the above effects (plus word length and random by-word intercepts/slopes for all surprisal effects) for each participant individually, and then analyze the distribution of fitted coefficients (Fig. 3).

We find the same general shape of effects of word properties and context-based predictability on typing time as has been documented for word duration in spoken language production, but we also find key differences in the detailed patterns of sensitivity. Firstly, the results of our GAM analysis show that, in the predictor ranges where most of the data lie, typing times are roughly linear in word length and log-probabilities, with the notable exception of unigram surprisal (word frequency), which has a nonlinearity in the 10–15 bit range for which we do not yet have an explanation (Fig. 2). In our second analysis, the median and distribution of surprisal coefficients show that the effect of predictability based on left context is a stronger determinant of typing time than predictability based on right context (Fig. 3). This contrasts results in spoken language production that show the opposite effect (Bell et al., 2009). Furthermore, we find predictability based on local context (bigram surprisals) is a stronger determinant of typing time than predictability based on global context. Notably, word frequency (unigram surprisal) has no predictive value for typing time once surprisal effects are taken into account (Fig. 3).

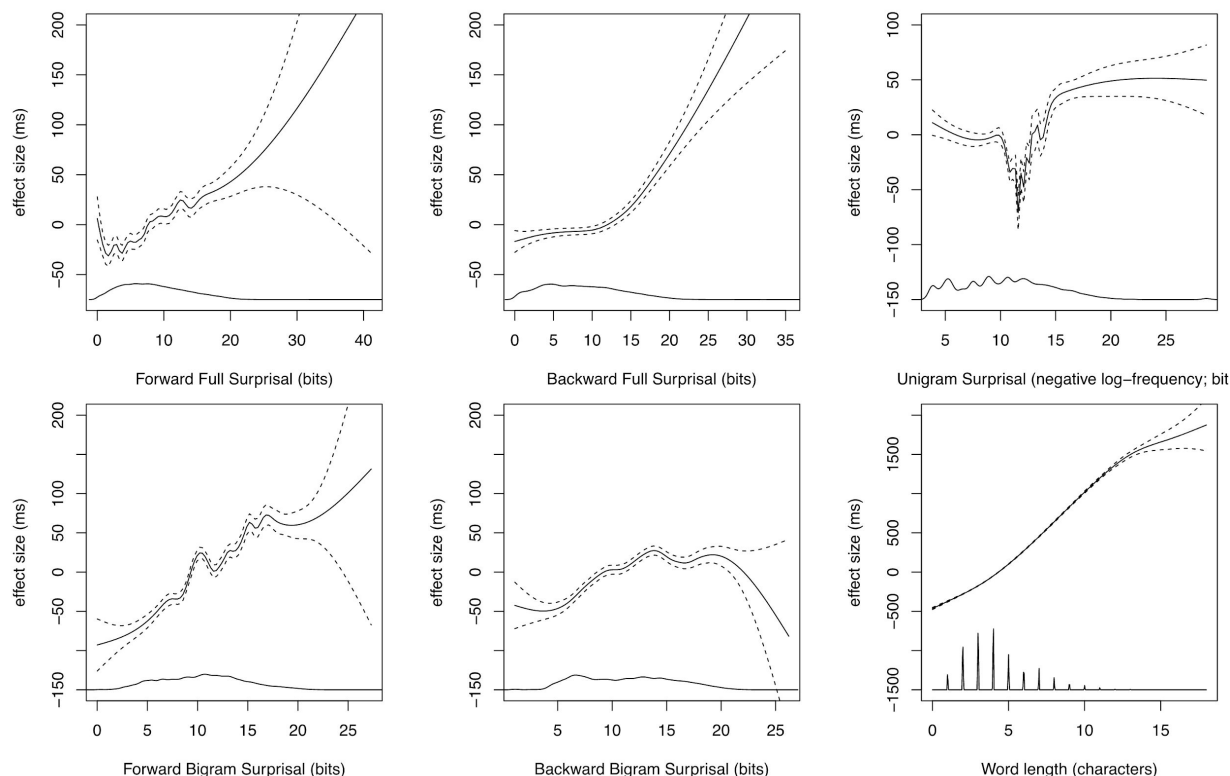


Figure 2: Relationship between various estimates of contextual probability (and frequency) and typing speed slowdown. Regression lines from fitted GAM models are shown as solid lines. Dashed lines indicate 95% confidence intervals but do not take into account the repeated-measures structure of the data or uncertainty in GAM hyperparameter values, and hence should be taken with a grain of salt. The marginal density of each predictor is shown at the bottom of each plot.

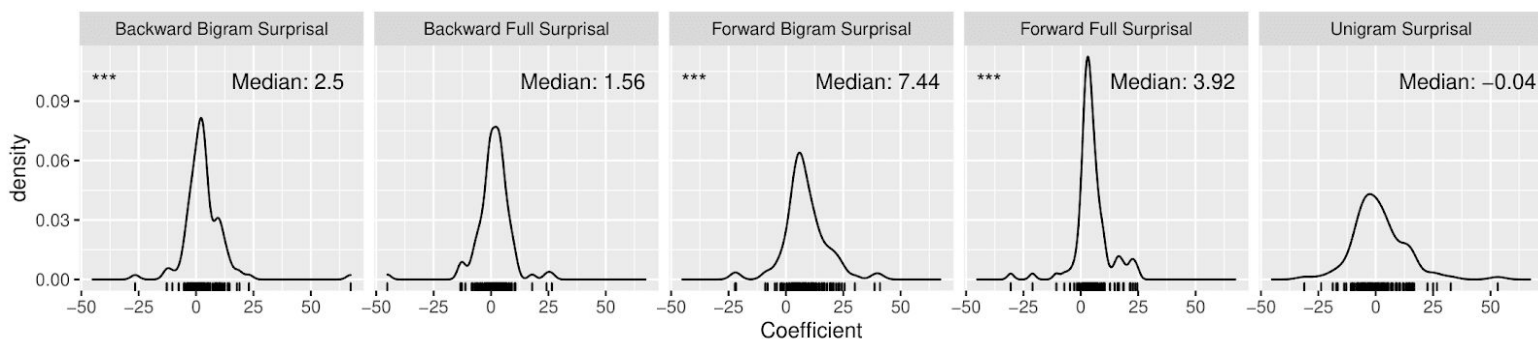


Figure 3: Median and distribution across participants of estimated predictor coefficients. Predictors marked with *** have mean above 0 at $p < 0.001$ (t-test); other predictors have mean not significantly different from 0. For word length, all participants have estimated coefficient above 0, with median 153ms/character (not shown).

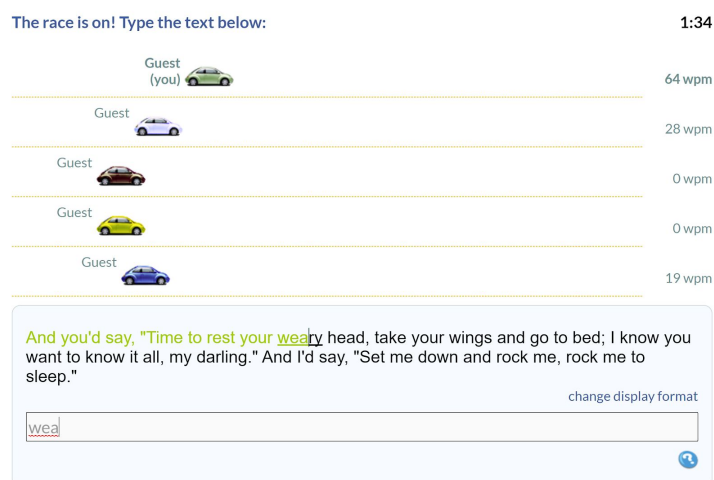


Figure 1: An in-progress race on TypeRacer.com. Racers are given up to 12 seconds to read the race prompt before the race starts.

References:

1. Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92-111.
2. Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, 25(4), 447-464.
3. Gelman, A. (2005). Two-stage regression and multilevel modeling: a commentary. *Political Analysis*, 13(4), 459-461.
4. Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
5. Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.

BERT, a deep-learning language model, learns NPI licensing but does not suffer from NPI illusion

Unsub Shin and Sanghoun Song (Korea University)

Recent development in computational models made it easier and more accurate to simulate human behavior in sentence processing (Wilcox et al., 2020; Merkx & Frank, 2020). Present study investigated whether a deep Transformer model BERT (Devlin et al., 2019) processes long-distance dependency and grammatical illusion in the same way as human language processors do. More specifically, we examined how BERT processes NPI licensing and NPI illusion.

Negative Polarity Items (NPIs) such as *ever* constitute a grammatical sentence only when it is c-commanded by a word or licenser that provides a negative context such as *no* or *few* e.g., *No! *Some! *The prisoner has ever talked to the priest* (Ladusaw, 1980). Research showed human processors are good at detecting whether an NPI and its licenser make a legitimate structural relationship. It was also shown they may mistakenly accept a potential licenser not occurring in a c-commanding position, for example in an embedded clause such as **The man [that no woman liked] has ever been to the party*. This phenomenon is called NPI illusion (Vasishth et al. 2008). Recent studies suggested BERT can capture some semantic features and structural information (Hewitt & Manning, 2019) but it is not fully resolved whether BERT can also learn the linguistic mechanism underlying NPI processing (Warstadt & Bowman 2020).

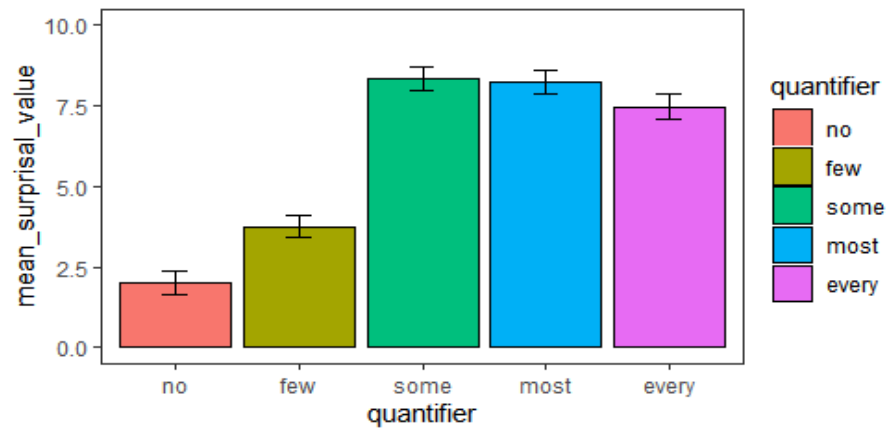
We conducted two experiments with BERT: We first investigated whether BERT can discern between semantically licit (negative) and illicit (positive) licensors of the NPI *ever* by testing five different quantifiers, *no*, *few*, *some*, *most* and *every* (Experiment 1). We used 150 sentence stimuli adapted from Xiang et al.'s (2009) (Table 1). Second, we examined whether BERT is susceptible to NPI illusion like humans by varying syntactic positions of the potential licensors (Experiment 2). We tested another set of 150 sentence stimuli in which a potential licenser *no* occurs in an embedded clause, violating the c-commanding condition of NPI licensing, and compare it with the condition where *no* occurs in the legitimate matrix clause. We also tested stimuli with no negative word as a control, i.e. *the*. In both experiments, we evaluated model performance by computing lexical surprisal values (Smith & Levy, 2013) from the output softmax layer, i.e. higher surprisal as a sign of increased processing difficulty.

The results of Experiment 1 (Figure 1) using Dunn's pairwise comparison shows that BERT captures the difference between strong NPI licenser *no* and weak NPI licenser *few* ($z = 3.45$, $p < .006$) and between negative quantifier *no* and non-negative quantifier *most* ($z = 11.74$, $p < .001$). The results of Experiment 2 (Figure 2) reveal that BERT discriminates between the licit and illicit position of NPI licensors ($z = 14.82$, $p < .001$). A much higher surprisal score for the embedded position indicates that the model successfully detects a structural violation. The fact that it is slightly higher than the surprisal for the no-licenser condition ($z = 2.07$, $p < .115$) further supports that *no* in the embedded clause is never considered a licenser for *ever* in the matrix clause. Overall, the results show that BERT successfully encoded the semantic feature of NPI licensors and structural c-command constraints while it was hardly led into NPI illusion as opposed to human language processors. We conducted post hoc analyses using sequential LSTM-RNN (Jozefowicz et al. 2016), which will be discussed in the paper as well.

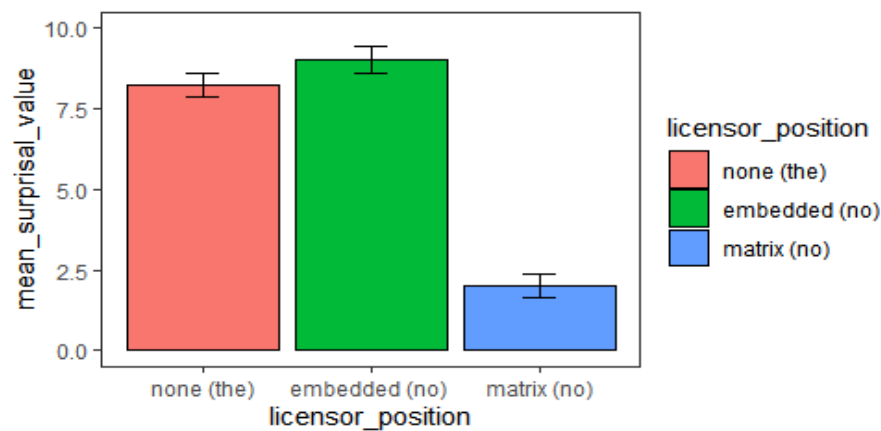
The results of this study suggests that a deep learning language model BERT is fully capable of extracting semantic and syntactic features or constraints required for processing long-distance dependencies such as NPI licensing. However, the fact that BERT is immune to NPI illusion may also suggest that the mechanisms or algorithms BERT relies on in language processing may fundamentally differ from those which humans rely on, e.g. cue-based retrieval, feature-matching, similarity-based analogical reasoning, etc. The current results do not exclude the possibility BERT depends on some surface-related naïve heuristics as well (McCoy et al. 2019). This is, to our knowledge, the first study that investigated BERT's capability in NPI processing and compared its performance between its legitimate licensing and the illusion phenomenon.

Table 1. Example sentence stimuli

Licensor Position	Sentence examples for experiments
Matrix clause	{ <i>no/few/some/most/every</i> } bears [that the competent trainers have treated kindly at all times] have <u>ever</u> gotten out of control.
Embedded clause	The bears [that { <i>no/the</i> } competent trainers have treated kindly at all times] have <u>ever</u> gotten out of control.



<Figure 1> Experiment 1: Surprisal of five potential licensors in the matrix clause



<Figure 1> Experiment 2: Licensing interactions of the negative quantifier *no* and targeted NPI *ever*.

Reference

- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). ACL vol.1, 4171–4186
- Hewitt, J. & Manning, C. D. (2019). ACL, vol.1, 4129–4138
- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). arXiv:1602.02410.
- McCoy, R. T., Pavlick, E., & Linzen, T. (2019). ACL, vol.1, 3428–3448
- Merkx, D. & Frank, S. L. (2020). arXiv:2005.09471.
- Smith, N. J. & Levy, R. (2013). *Cognition* 128(3), 302–319
- Vasishth, S., Brüssow, S., Lewis, R. & Drenhaus, H. (2008). *Cognitive Science* 32, 685-712.
- Warstadt, A. & Bowman, S. (2020). Proceedings of the 42nd Annual Conference of the CSS.
- Wilcox, E. Gauthier, J. Hu, J. Qian, P. & Levy, R. (2020) Proceedings of the 42nd Annual CSS.
- Xiang, M., Dillon, B., & Phillips, C. (2009). *Brain and Language* 108(1), 40-55.

Cross-situational Word Learning from Naturalistic Headcam Data

Wai Keen Vong, Emin Orhan and Brenden Lake (New York University)

One of the challenges of word learning is the problem of reference. When a child hears a word like “ball”, how are they able to figure out which referents in the world this word refers to? One proposed learning mechanism for resolving this is through cross-situational learning: rather than learning from an ambiguous single instance, children can aggregate information across multiple ambiguous co-occurrences of the word “ball” to correctly determine the underlying referent. While the topic of cross-situational word learning has received significant attention, both empirically (Yu & Smith, 2007), and in the development of various computational models (Frank, Goodman & Tenenbaum, 2009; Stevens et al., 2017), many well-known models require the visual referents to be preprocessed as discrete entities so it is unclear how these models could scale to explain cross-situational word learning from naturalistic data.

Recently, researchers have begun to combine convolutional neural networks with egocentric headcam data, and have shown that such models can learn useful visual representations (Bambach et al., 2018; Orhan, Gupta & Lake, 2020; Tsutsui et al., 2020). One limitation of these approaches is that they are trained using supervised feedback on category labels. In contrast, cross-situational learning provides no direct supervision akin to supervised feedback during the learning process, but only a form of weak supervision based on which words co-occur with which referents in the scene.

Inspired by these challenges, we recently developed a computational model to perform cross-situational word learning using the SAYCam dataset, a large-scale longitudinal egocentric headcam dataset (Sullivan et al., 2020). We trained our model using data from a single child, by creating a dataset of roughly 35000 image-utterance pairs extracted from roughly 60 hours of raw video with transcribed utterance data. Our model architecture is a multimodal neural network model, which embeds images using a pre-trained convolutional neural network trained only from the raw visual data from the same child (Orhan et al., 2020), and separately embeds utterances using a language encoder consisting of either a single embedding layer or an LSTM. The model is trained via a contrastive loss, learning to pair images with their corresponding utterances in the embedding space, which allows the model to learn word-referent mappings.

The model was evaluated on a separate dataset of frames extracted from 22 common visual categories in SAYCam, by presenting the model with a target word and four images (one target, and three foils), and asking the model to select the correct target referent. Our results show that the model is above chance in selecting the correct referent for more than half of these categories. We also show that qualitatively, the model can also localize the referents in a given scene (as shown below). To our knowledge, this is the first model that successfully captures cross-situational word learning using longitudinal egocentric data from a single child, and demonstrates that such learning is possible from raw inputs using recent advances in deep learning.

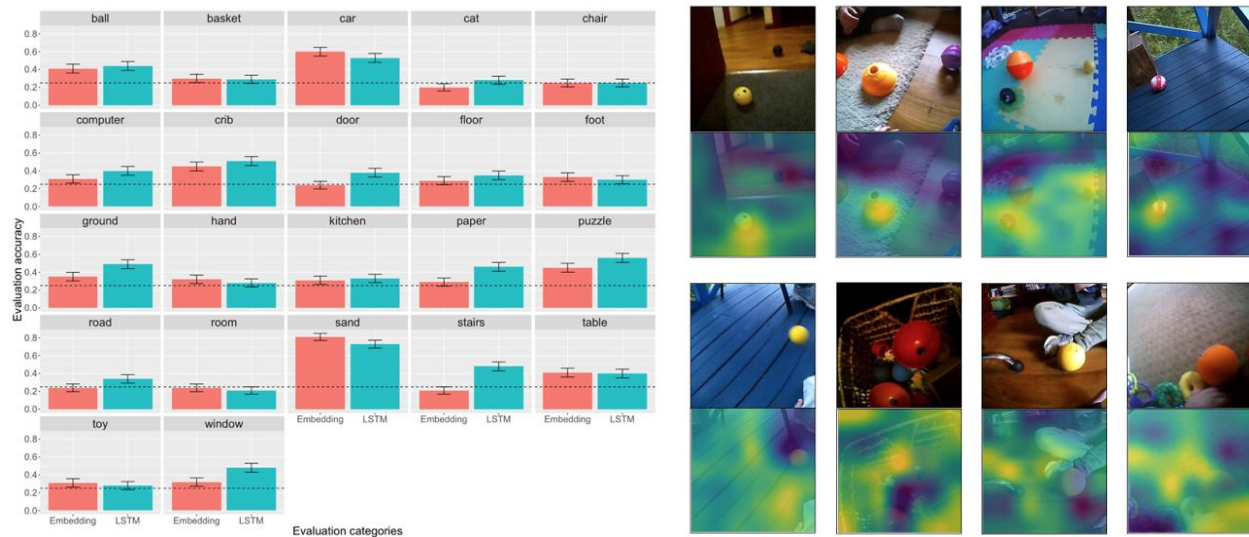


Figure 1. The left figure shows the evaluation performance of the two models for the 22 visual categories in SAYCam, with error bars as standard error and the dotted line representing chance. The right figure shows examples of localization of referents using attention maps extracted from our model, with the top row showing successes for the word “ball”, and the bottom showing some failures.

References

- Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, 31, 1201-1210.
- Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578-585.
- Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33.
- Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, 41, 638-676.
- Sullivan, J., Mei, M., Perfors, A., Wojcik, E. H., & Frank, M. C. (2020). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective.
- Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A Computational Model of Early Word Learning from the Infant's Point of View. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414-420.

Do Artificial Language Models Learn Syntax–Semantics Mappings?

Xiaohan (Hannah) Guo, Bryor Snefjella, Idan A. Blank, (UCLA)

Background. Artificial neural networks (ANNs) have recently emerged as successful models of language processing [1,2]. Specifically, these models implicitly learn a surprising amount of syntactic knowledge [e.g., 3,4]. However, ANNs have been criticized for having no semantic knowledge [e.g., 5,6]. Such criticisms often conflate several issues: grounding linguistic meaning in non-linguistic experience; having common sense / world knowledge; and representing semantic relations, such as event structure. Here, we focus on the latter and test whether ANNs implicitly represent “who did what to whom”. Because event semantics might be internally represented even if not evident in ANNs’ output (e.g., next word prediction; cf. [7]), we study the hidden representations of these networks.

Materials. We borrow our design from fMRI studies in [8,9]. We use a set of “base” items, and edit each item to create several distinct versions (“conditions”). In Experiment 1 (**Table 1**), base items are simple transitive sentences, and are edited to create 4 conditions, differing from the base in: (A) only lexical items (using synonyms), but not syntax or global meaning; (B) only syntax (active vs. passive), but not words or global meaning; (C) only global meaning (switching agent and patient), but not syntax or words (the critical condition); and (D) all 3 aspects (control). In Experiment 2 (**Table 2**), conditions differ from the base in: (A) one synonymous word, not affecting global meaning; (B) one non-synonymous word, changing global meaning; (C) syntax (active vs. passive / direct- vs. prepositional-object) but not meaning; (D) both syntax and meaning (switching agent and patient); or (E) all aspects (control).

Procedure. We evaluated two representative state-of-the-art transformer architectures, BERT [10] and GPT2 [11]. For each sentence, we extracted unit activations from the last hidden (non-embedding) layer (results hold in other layers); the last sentence token was used (BERT: [SEP]; GPT2: ‘.’; results hold for all-token averages). For each item, we computed cosine similarities between activations for the “base” sentence and each other version (condition). Similarities were Fisher-transformed to improve normality. We compared conditions in terms of similarities to the “base” via a non-parametric, repeated-measures ANOVA based on restricted permutation of residuals [12-13] (results hold under two other permutation regimes). Specifically, pairs of conditions were compared via Tukey tests within this ANOVA model.

Results and discussion. See **Figure 1**. In Experiment 1, two sentences with the same words and syntax but different meaning (switching agent and patient; “base” vs. condition C) were *more* similar to each other than pairs that had the same meaning but differed in either words (“base” vs. A) or syntax (“base” vs. B). Thus, ANNs represent sentences with different event structures as more similar than sentences with the same event structure. In Experiment 2, sentence pairs that differed in both syntax and meaning (“base” vs. D) were *no less* similar than pairs that differed only in syntax but not meaning (“base” vs. C). Thus, a difference in syntax influenced ANNs’ representations to a similar extent regardless of whether it led to a change in meaning or not (in contrast, a word changing to a non-synonym had a larger influence than it changing to a synonym, as expected). Overall, the ANNs we studied might be severely limited as models of human language processing: at least in terms of the overall, distributed pattern of activations across hidden units, ANNs fail to represent sentence semantics, even in a test of “bare” event structure divorced from world knowledge or grounding.

Table 1. Experiment 1 sample materials (94 sets for BERT; 92 sets for GPT2)

Base	Different words	Different syntax	Different meaning	Different all
The teacher praised the thinker	(A) The educator lauded the theorist	(B) The thinker was praised by the teacher	(C) The thinker praised the teacher	(D) The educator was lauded by the theorist

Table 2. Experiment 2 sample materials (113 sets for BERT; 106 sets for GPT2)

Base	Different word		Different syntax		Different all
	Mean same	Mean different	Mean same	Mean different	
Anna invited the composer	(A) Anna invited the songwriter	(B) Anna invited the translator	(C) The composer was invited by Anna	(D) Anna was invited by the composer	(E) Anna was invited by the translator

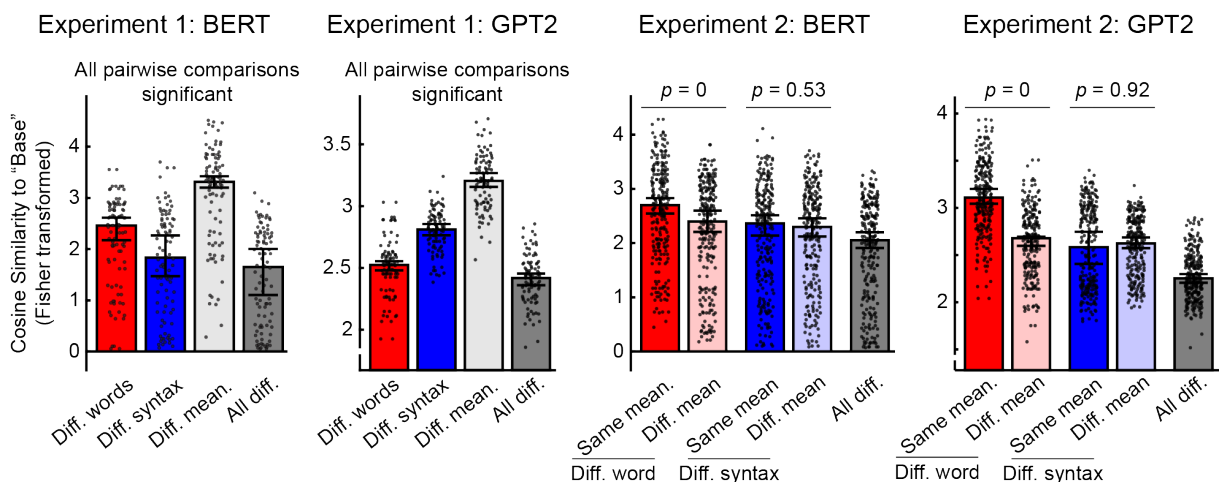


Figure 1. Similarities between sentence pairs. In Experiment 1, note that sentence pairs with *different* meanings (light gray) are more similar to each other, compared to sentence pairs with the *same* meaning but different words (synonyms; red) or syntax (blue). In Experiment 2, note that pairs with different syntax and *different* meanings (light blue) are no less similar to each other compared to pairs with different syntax but the *same* meaning (dark blue). Bars show medians. Error bars show 95% confidence intervals. Dots show individual items.

References. [1] Schrimpf et al. (2020). *bioRxiv preprint*. [2] Hu et al. (2020). *ACL*. [3] Rogers et al. (2020). *ACL*. [4] Manning et al. (2020). *PNAS*. [5] Bender & Koller (2020). *ACL*. [6] Marcus (2020). *ArXiv preprint*. [7] Ettinger (2020). *TACL*. [8] Fedorenko et al. (2020). *Cognition*. [9] Dapretto & Bookheimer (1999). *Neuron*. [10] Devlin et al. (2018). *ArXiv preprint*. [11] Radford et al. (2019). OpenAI blog. [12] Kherad-Pajouh & Renaud (2015). *Statistical Papers*. [13] Anderson & Braak (2003). *JSCS*.

EARLY LEXICAL COMPREHENSION AND GENDER AGREEMENT IN ITALIAN TODDLERS.

Giulia Mornati (University of Milano-Bicocca), Valentina Riva, Elena Vismara, Massimo Molteni, Chiara Cantiani (Scientific Institute IRCCS E. Medea, Bosisio Parini, LC, Italy)

Experimental evidence of lexical comprehension in children younger than one year of age is limited^{1,2}. To date, studies applying online techniques on Italian toddlers have shown successful lexical comprehension after age 15 months³. Italian is a gender marked language in which determiners agree in gender and number with the following noun: being aware of such agreement relationship is crucial because it could facilitate the processing of words, allowing children to predict what they are going to listen to next. The processing of gender features - specifically related to the characteristics of each language^{4,5} - has been sporadically investigated in children under 2 years of age⁵. Therefore, the aim of this study is to investigate early lexical comprehension and the role of determiners in the processing of Italian, in children aged 12 and 20 months.

The Looking While Listening (LWL) procedure⁶ is an online paradigm allowing to analyze comprehension in real time, by recording children's eye-movements in relation to an auditory stimulus. In each trial, two pictures (a target and a distractor) appeared on a monitor while sentences, including determiner and noun, were auditory presented (*Where is the_{FEM} ball_{FEM}?*). Two conditions were created: a same-gender condition, in which a target and a distractor share the same grammatical gender (e.g., *dog_{MASC}* vs. *boy_{MASC}*) and determiner was uninformative; and a different-gender condition in which nouns have different grammatical gender (*dog_{MASC}* vs. *girl_{FEM}*) and determiner was informative. Children's looking patterns were recorded by an eyetracker (Tobii X50) for the whole duration of the trials (5s). Children were divided into two groups based on age: 12-months (N=17) and 20-months (N=15). Separately for each age-group, we conducted three cluster-based permutation analyses: one for each experimental condition comparing the average looking proportion toward the target to chance level (0,5), and one comparing the looking proportions between conditions.

In the 20-month group, looking proportions to the target for the same-gender condition were significantly different from chance level from the middle to the end of the trial (1120-2220ms, $p < .001$, blue line Fig.1). For the different-gender condition, looking proportions to the target were significantly different from chance level (i) already just after hearing the informative determiner (560-1160ms, $p = .012$, first red line in Fig.1) and (ii) when they heard the full name of the target (1260-1920ms, $p = .005$, second red line in Fig.1). The direct comparison between conditions confirmed this pattern, as shown by the significant clusters represented with dashed lines in Fig.1 (260-980ms, $p = .018$, 1400m-1740ms, $p = .047$, and 1760-2220ms, $p = .041$). Moreover, in the 12-month group looking proportions to the target were significantly different from the chance level for the different-gender condition just after hearing the informative determiner (380-700ms, $p = .042$ —red line in Figure 2). Results in this age-group were however less robust, as confirmed by the absence of significant cluster in the direct comparison between conditions.

In conclusion, this study extends the results found in literature⁵. Already at 12 months of age, and with an improvement seen at 20 months of age, Italian toddlers seem to be able to extract and use the grammatical gender carried by determiners, to make predictions about the following target noun. The results found in the 12-month group with the informative determiner may suggest that infants at this age have access to the grammatical traits. Alternatively, we can hypothesize that 12-month-olds rely on other cues (i.e. the probability of occurrence between the determiner and the noun) that are not relevant when the determiner is uninformative. When infants have no cues, they can rely only on the meaning of the noun, and this process could be slower and not detectable before the trial's end. Future studies are needed to further understand these aspects.

20-MONTH GROUP

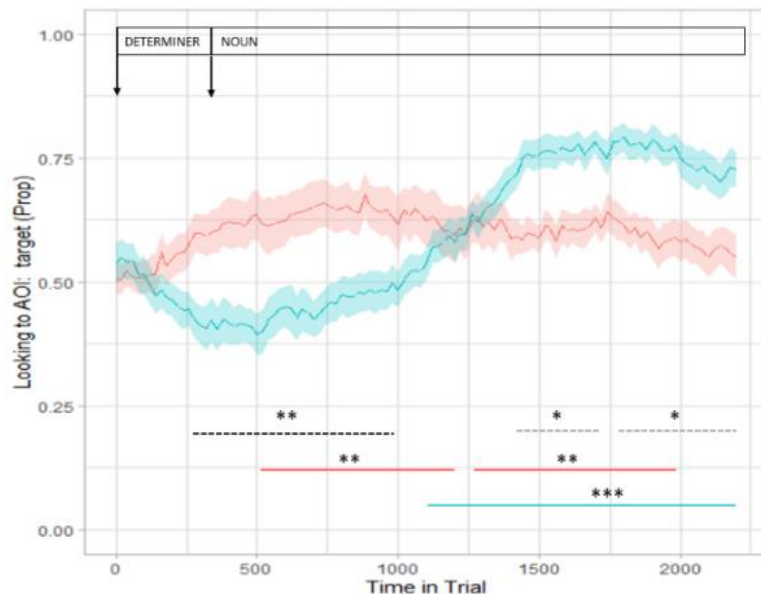
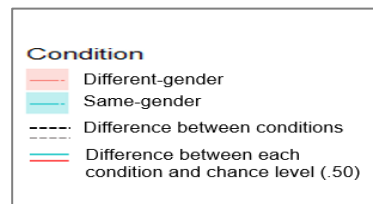


Fig. 1 Proportion towards the target picture from the determiner onset till the end of the trial for the same-gender condition (blue line) and the different-gender condition (red line) in toddlers aged 20 months. In the different gender condition, the looks to the target significantly increased (above the chance level of 0.5) in two time-windows from 560ms to 1160 and from 1260ms to 1920ms (red lines). In the same-gender condition, the looks to the target significantly raised above the chance level from 1120 to 2220ms (blue line). Moreover, toddlers behaved differently according to the conditions: they detected the target picture faster when the determiner was informative (different-gender condition) than when it was uninformative (same-gender condition), as shown by the dashed black line (from 260ms to 980ms). However, in the same-gender condition, toddlers were able to detect the target picture when they heard its name (dashed grey lines, from 1400 to 1740ms and from 1760-2220 respectively).



12-MONTH GROUP

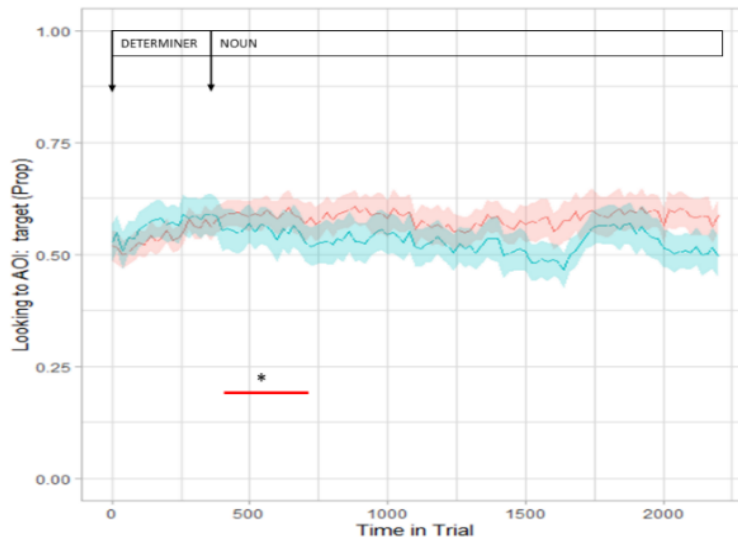


Fig2. Proportion towards the target picture from the determiner onset till the end of the trial for the same-gender condition (blue line) and the different-gender condition (red line) in infants aged 12 months. Despite there was no difference between the two conditions, infants significantly increased the looks to the target picture above the chance level (0.5) just after hearing the determiner (red line from 360ms to 700ms).

BIBLIOGRAPHY

1. Bergelson, E. & Swingle, D. At 6-9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci.* **109**, 3253–3258 (2012).
2. Friedrich, M. & Friederici, A. D. Phonotactic knowledge and lexical-semantic processing in one-year-olds: Brain responses to words and nonsense words in picture contexts. *J. Cogn. Neurosci.* **17**, 1785–1802 (2005).
3. Suttora, C. *et al.* Relationships between structural and acoustic properties of maternal talk and children's early word recognition. *First Lang.* **37**, 612–629 (2017).
4. Johnson, E. K. Grammatical Gender and Early Word Recognition in Dutch. *Proc. 29th Annu. Bost. Univ. Conf. Lang. Dev.* 320–330 (2005).
5. Lew-Williams, C. & Fernald, A. Young Children Learning Spanish Make Rapid Use of Grammatical Gender in Spoken Word Recognition. *Psychol. Sci.* **18**, 193–198 (2007).
6. Fernald, A., Zangl, R., Portillo, A. L. & Marchman, V. A. Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. in *Developmental psycholinguistics: On-line methods in children's language processing* 97–135 (2008). doi:10.1075/lald.44.06fer

Spreading jam with a butter knight: Near-homophones and phonological pre-activation

Kari Schwink & Jeffrey J. Green (University of Illinois at Urbana-Champaign)

Lexical pre-activation in predictive contexts is a well-established effect, but additional evidence is required to determine the level of detail in comprehenders' predictions. The current ERP study uses near-homophones to ask whether comprehenders pre-activate phonological representations in their predictions. Preliminary results ($N=14$) do not support the hypothesis that phonological information can be pre-activated, even in highly predictive contexts.

In previous studies, phonological pre-activation has been approached using the allomorphy of the English indefinite article *a/an*, since the correct form of the article is determined by the phonology of the following word. Most notably, [1] found an N400 effect when predictive contexts were followed by an unexpected determiner. Under the view that the N400 reflects lexical activation processes [4], the increased N400 amplitude on unanticipated articles was interpreted as an indicator that phonological features may be preactivated in sentence processing. However, a recent large-scale study [5] failed to replicate these findings. Our ERP experiment investigates the role of phonological prediction in sentence processing by comparing the amplitude of the N400 response to semantically implausible near-homophones of predicted sentence completions versus predicted and unrelated completions. A sample item set is given in Table 1. Materials were normed for predictability and for semantic similarity between target words in an item set in separate online tasks ($N=30$ each). In order to reduce potential shallow processing of the near-homophones, participants were asked after one-third of items whether a specific phrase was seen in the previous sentence (e.g. "Did the sentence you just read include 'along the sandy beach'?").

We hypothesized that if comprehender predictions include phonological detail, spreading activation to phonological neighbors of an anticipated continuation would attenuate the amplitude of the N400 effect in the near-homophone condition relative to the unrelated condition. At first glance, preliminary results appear to support this prediction. Both the near-homophone and unrelated conditions show a significantly larger N400 than the predicted condition ($p<0.01$), and the N400 is significantly reduced for the near-homophones relative to unrelated words ($p<0.01$). This is illustrated in Figure 1. The near-homophone condition also showed greater positivity in the 500–800ms time window relative to the other two conditions ($p<0.01$). One possible explanation for these results is that the predicted word was pre-activated sufficiently to allow spreading activation to the phonologically-related near-homophone, causing the reduced N400 seen. The P600 effect may reflect participants noticing that the word they saw is very similar to the highly predicted word, which may have triggered a reanalysis or monitoring process [2, 3]. This would suggest that at least in very highly constraining contexts, prediction of a word may include pre-activation of its phonological features, inducing spreading activation to phonologically similar words. However, an alternative explanation is that the reduced N400 was simply caused by component overlap with the P600. This would mean that there was no real facilitation in accessing the near-homophone relative to the unrelated item, and that the P600 reflecting reanalysis or monitoring began early enough to reduce the amplitude of the N400 in the near-homophone condition. This possibility is supported by a high correlation between the difference in size of the N400 and P600 effects in the unrelated and near-homophone conditions ($R=0.81$, $p<0.001$). A reduced N400 was only seen in the near-homophone condition for participants who also had a relatively large P600 effect.

Our results do not, therefore, provide convincing evidence for phonological pre-activation in lexical prediction, and instead demonstrate that when participants read an unexpected word that is related phonologically to a highly predicted word, they may reanalyze or monitor their interpretation of the sentence, leading to a post-N400 P600 effect. However, more data is needed in order to draw firm conclusions about the source of the reduced N400 seen.

Table 1: Sample item set

Pre-critical region	
Sandra looked out at the ocean as she walked along the sandy...	
Condition	Critical word and end of sentence
Predicted:	... beach that day.
Near-homophone	... beef that day.
Unrelated:	... lime that day.

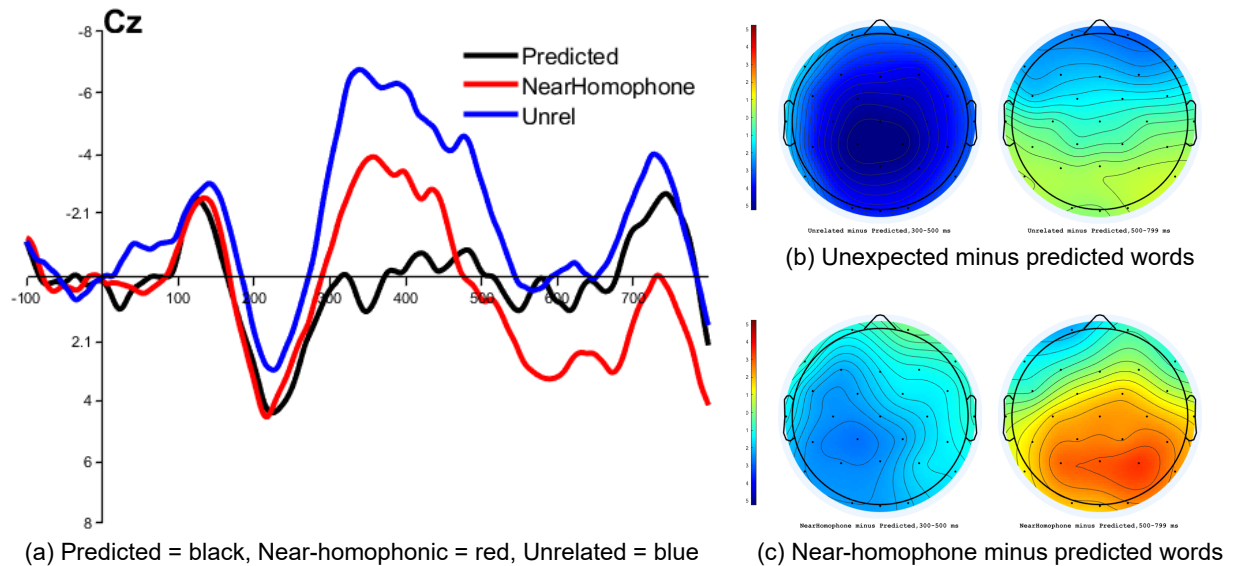


Figure 1: Experiment results, N=14. (a) gives the ERP at the Cz electrode for the critical word. (b) and (c) are scalp maps showing the effect of the unexpected (b) and near-homophone (c) conditions relative to the predicted condition in the N400 and P600 time windows (300–500ms and 500–800ms).

References

- [1] DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). *Nature Neuroscience*. [2] Kaan, E., & Swaab, T. Y. (2003). *Journal of Cognitive Neuroscience*. [3] Kolk, H., & Chwilla, D. (2007). *Brain and Language*. [4] Lau, E. F., Namyst, A., ... Delgado, T. (2016). *Collabra*. [5] Nieuwland, M. S., Politzer-Ahles, S., ... Huettig, F. (2018). *eLife*.

Planning ahead: Interpreters predict source language in consecutive interpreting

Nan Zhao (Baptist University of Hong Kong), Xiaocong Chen (The Hong Kong Polytechnic University), Zhenguang G. Cai (Chinese University of Hong Kong)

People predict upcoming linguistic content in reading and listening (Pickering & Gambi, 2018). In particular, it has been hypothesized that interpreters anticipate upcoming words and syntax in both source language (SL) and target language (TL) to facilitate timely interpreting delivery (Amos & Pickering, 2020; Chernov, 1994). In three experiments (E1a, E1b and E2), we asked whether interpreters predict lexico-semantic content in SL comprehension in consecutive interpreting to a greater extent than in regular language comprehension and whether such enhanced prediction (if any) is constrained by cognitive resources.

E1a and E1b examined whether interpreters make more lexico-semantic predictions when they read a sentence to later interpret than to later repeat (Macizo & Bajo, 2006). E1a (52 participants, 48 target items and 64 fillers) had a design of 2 (predictability: predictable vs. unpredictable) x 2 (task: repetition vs. interpreting; blocked). Based on results of a cloze test, we manipulated a critical word (e.g., eyes) to be predictable or unpredictable in a sentence (*Without the sunglasses/hat, the sun will hurt your eyes on the beach*). Participants were Chinese-English bilinguals with interpreting training/experience. In an online experiment on Gorilla, participants self-paced read an English sentence word by word to either repeat it (as a form of regular language comprehension) or to interpret it into Mandarin (as a form of SL comprehension). E1b (50 participants, 72 target items and 24 fillers) had the same design and was intended to replicate E1a using more and refined items.

LME analyses showed that participants read the critical word and the following regions more quickly in the predictable than unpredictable condition (in C-1, C, C+1 and C+2 in E1a and in C-1, C, and C+2 in E1b; see Fig 1). More importantly, there was an interaction between predictability and task (in C+1 in E1a and C and C+1 in E1b) such that the prediction effect was stronger when participants read a sentence to later interpret than to repeat.

E2 (64 participants, 72 target items, 24 fillers) further examined whether the enhanced prediction in interpreting is constrained by cognitive resources. It had a design of 2 (predictability: predictable vs. unpredictable) x 2 (task: repetition vs. interpreting; blocked) x 2 (load: low vs. high). In the low-load condition, participants read one sentence and then repeated/interpreted it (as in Expt 1a and 1b). In the high-load condition, we added a 5-word sentence before the original sentence. Participants read the first sentence, kept it in memory, read the second (target) sentence, before they repeated/interpreted both sentences. As shown in Fig 2, we replicated the finding in Expt 1a and 1b: The prediction effect was stronger in reading to interpret than in reading to repeat (in C and C+1). More importantly, there was also a three-way interaction (in C), with enhanced prediction in reading to interpret in the low- but not high-load condition.

In all, the results suggest that interpreters are more predictive of lexico-semantic content in SL comprehension in interpreting than in regular language comprehension, giving support to the hypothesis that interpreters use an anticipatory strategy to maximize interpreting timeliness. Also, prediction in interpreting seems to require cognitive resources.

References

- Amos, R. M., & Pickering, M. J. (2020). A theory of prediction in simultaneous interpreting. *Bilingualism: Language and Cognition*.
- Macizo, P., & Bajo, M. T. (2006). Reading for repetition and reading for translation: Do they involve the same processes? *Cognition*, 99(1), 1-34.
- Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10), 1002–1044.
- Chernov, G. V. (1994). Message redundancy and message anticipation in simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation* (pp. 139–153). John Benjamins.

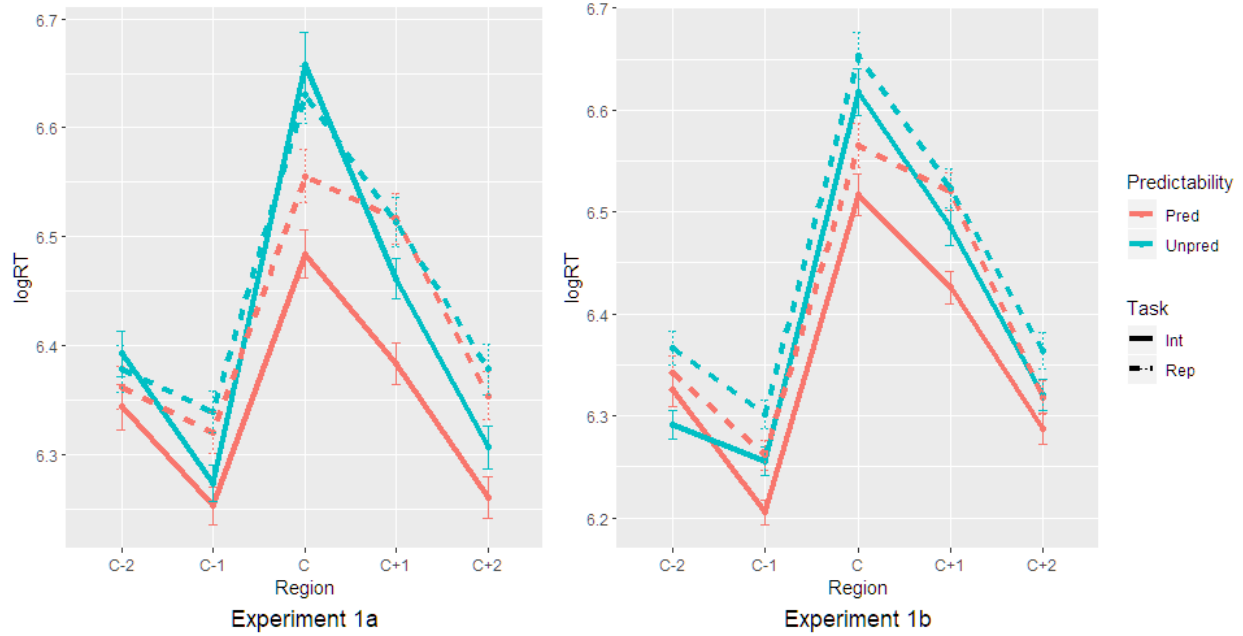


Fig 1. Log RTs for the critical word and surrounding words in self-paced reading in Experiment 1a (left panel) and Experiment 1b (right panel). For regions, *Without the sunglasses/hat, the sun will hurt* (C-2) *your* (C-1) *eyes* (C) *on* (C+1) *the* (C+2) *beach*; same for **Fig 2**.

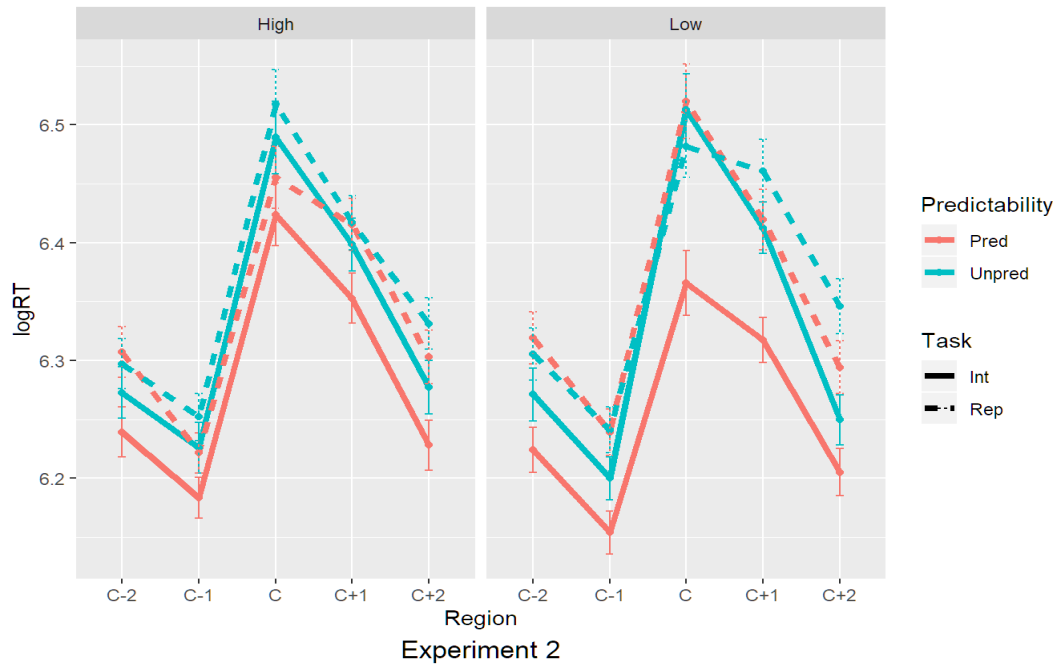


Fig 2. Log RTs for the critical word and surrounding words in self-paced reading in Experiment 2 for the high-load (left panel) and low-load condition (right panel).

Perception of disfluencies in non-native speech

Rajalakshmi Satarai Madhavan (University of Göttingen) and Martin Corley (University of Edinburgh)

Background:

Disfluencies are usually defined as the false starts, hesitations, and filled pauses that occur in speech (Corley & Stewart, 2008). They are common in spontaneous speech, and can occur due to difficulties in lexical access (Arnold, Losongco, Wasow, & Ginstrom, 2000). Listeners use disfluencies to predict upcoming words: For example, listeners look more towards a low frequency object (LFO: objects with names that occur rarely during speech) when preceded by disfluency (Arnold, Fagnano & Tanenhaus, 2003; Arnold, Kam & Tanenhaus, 2007). Research about perception of disfluencies in non-native speech shows that disfluencies do not influence native listeners' predictions in the same way (Bosker, Quené, Sanders & De Jong, 2014). However, there has been no investigation to date into whether these differences stem from difficulties in comprehending non-standard accents, or from taking the speaker's perspective and attributing any disfluencies to general difficulties in formulation. The aim of this study was to distinguish these two views.

Methodology and procedure:

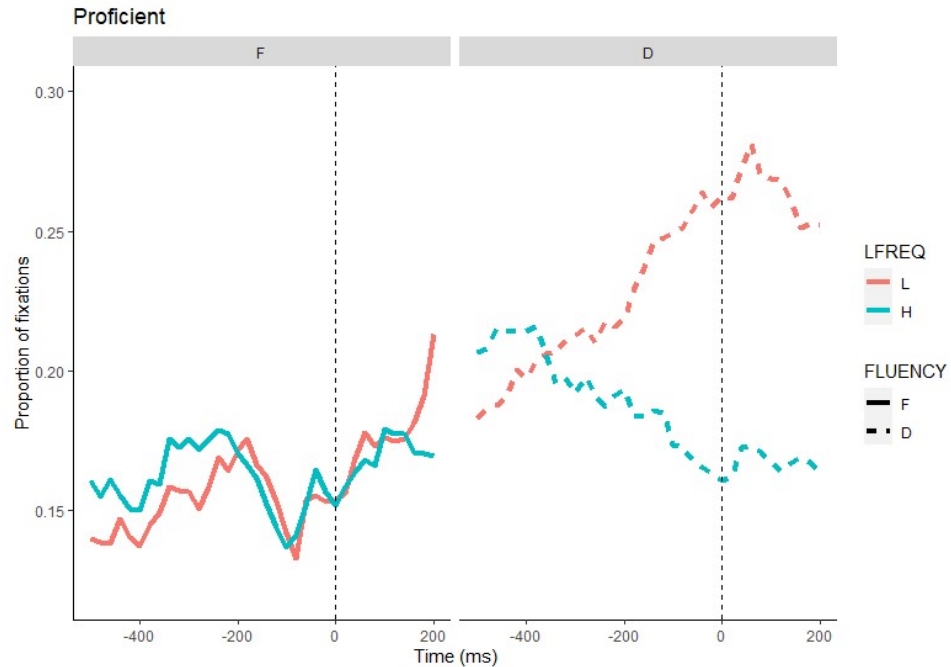
Sixty participants performed an eye-tracking study where they were randomly assigned to either a 'proficient' or 'non-proficient' non-native speaker (30 participants per condition). In both conditions the same speaker of Indian English introduced himself differently, in a brief audio recording, so as to appear 'proficient' ("I enjoy reading historical fiction") or 'nonproficient' ("I only learnt English for a short amount of time"). The only difference between conditions was the content of the introductory stories; the accent of the speaker stayed the same, and identical recordings of experimental items, using the accent and pronunciation of Indian English, were used in both conditions. Participants were presented with pictures of a high-frequency (e.g., egg) and a low-frequency (e.g., wheelbarrow) object. The speaker then gave either fluent [*Click on the...*] or disfluent [*Click on thee uh...*] instructions to click on one of the objects in the visual array. After the task was complete, participants were given a questionnaire to assess their exposure to non-native accents.

Results:

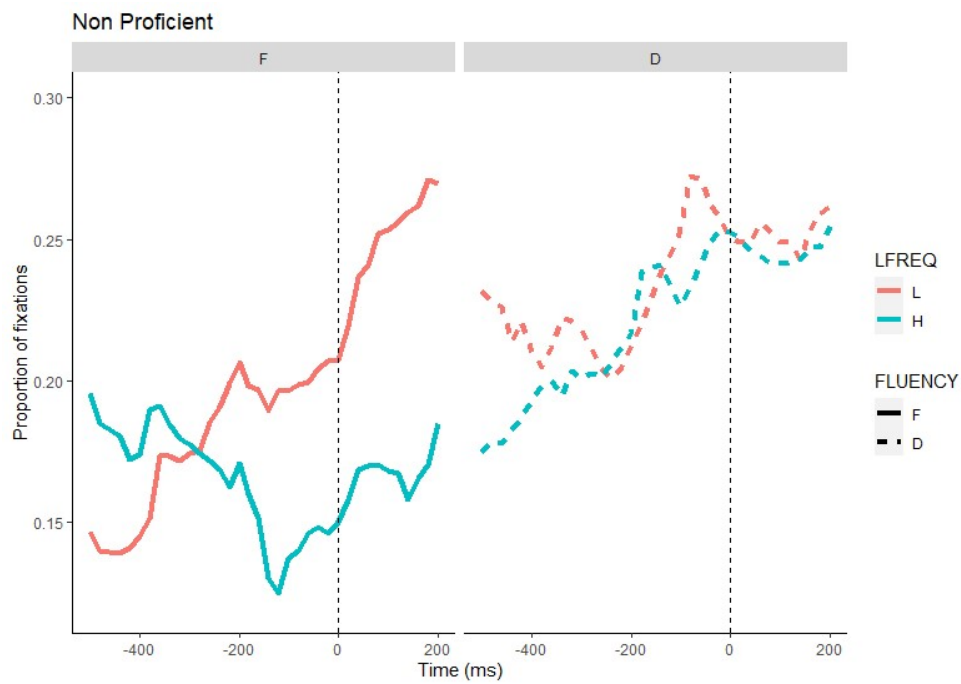
Two linear mixed models were run, one for Fluent and one for Disfluent trials, predicting the proportions of looks toward the LFO in the visual array with predictors of proficiency and linear and quadratic time terms. Following Bosker et al. (2014), the time window for analysis ran from the start of the sentence to the onset of the target word. There were no effects of any predictors in the fluent trials. However, for disfluent trials, both time terms, and the interactions between proficiency condition and time terms were significant. Participants in the proficient speaker condition looked more towards the LFOs in the disfluent trials. The analysis of the questionnaire answers showed no differences between participants in exposure to non-native speech in daily life.

Discussion:

When they encounter non-native-sounding speech, listeners engage in perspective taking. In the present study, they anticipated the low-frequency referent when the supposedly 'proficient' speaker was disfluent, while there was no bias toward the low frequency referent when they were listening to the supposedly 'nonproficient' speaker. Thus, it appears that listeners are able modulate their assumptions about non-native speakers' disfluencies, perhaps inferring that a less proficient speaker is more likely to be disfluent for reasons other than retrieving a low-frequency object name.



(a)



(b)

Figure 1 shows the proportion of looks toward both high and low frequency objects in the (a) proficient and (b) non-proficient condition. The vertical black dashed lines show the time of the target onset. The graphs with the full lines show fluent trials, and those with the dashed lines show disfluent trials. The pink lines and blue show proportion of fixations toward high frequency objects and low frequency objects respectively.

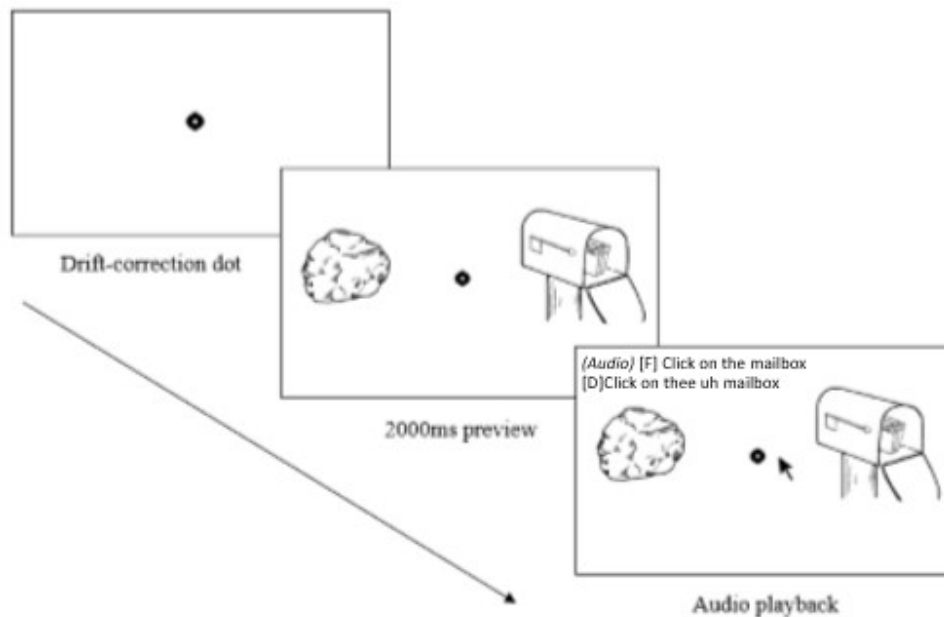


Figure 2 shows the timeline of one fluent or disfluent trial.

Key References

- Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1), 28-55.
- Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal thee, um, new information. *Journal of psycholinguistic research*, 32(1), 25-36.
- Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5), 914.
- Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of memory and language*, 75, 104-116.
- Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4), 589-602.

Evidence for a two-stage account of prediction

Ruth Corps^{1,2}, Charlotte Brooke², & Martin Pickering²

¹ Psychology of Language Department, Max Planck Institute for Psycholinguistics,
Ruth.Corps@mpi.nl

² Department of Psychology, University of Edinburgh

Comprehenders often predict what they are going to hear. For example, they will preferentially look at edible objects immediately after hearing *The boy will eat...*, and thus predict that the speaker is about to mention such an object [1]. But what exactly do comprehenders predict? And more importantly, what information do they use to make these predictions? Do they immediately make the best (most appropriate) predictions they can, or do such predictions take time and resources?

Comprehenders may immediately predict *appropriately* (from the speaker's perspective), because their predictions will tend to correspond to what the speaker actually says – a one-stage account. But considering perspective is effortful [2], and so comprehenders may initially predict on the basis of automatic *associations* [3], or on the basis of what they would (*egocentrically*) say if they were speaking [4] – two different two-stage accounts.

We tested among these alternatives in three experiments using the visual-world paradigm, in which participants listened to sentences (N=28; e.g., *I would like to wear...*), while viewing four objects on-screen. We manipulated the gender of the speaker (as indexed by their voice and face; [5]), the participants, and the characters in the sentences. In particular, participants heard a male or a female speaker producing sentences about gender-stereotyped objects (as assessed in a pre-test; N=80). One target (a dress) and one distractor (a hairdryer) were stereotypically female; the other target (a tie) and one distractor (a hairdryer) were stereotypically female; the other target (a tie) and distractor (a drill) were stereotypically male. To make different perspectives salient, sentences began with *I* in Experiment 1 (speaker's perspective), *You* in Experiment 2 (participant's perspective), and the name *James* or *Kate* in Experiment 3 (character's perspective). We fitted Bayesian generalized linear mixed effects models to binomial fixations in 50 ms time bins from 1000 ms before to 1500 ms after critical verb onset (*wear*).

In Experiment 1, participants (N=24, 12 males) fixated targets more than distractors from 450 ms after verb onset ($ps < .05$), before the target was mentioned, suggesting they predicted associatively. Participants also fixated appropriate targets (which matched the speaker's gender) more than inappropriate targets (which matched their own gender) from 600 ms ($ps < .05$; see Figure 2) and there was no point at which they predicted egocentrically. This appropriateness effect occurred later than the associative effect. For example, a male participant listening to a female speaker initially fixated the dress and the tie (over the hairdryer and the drill), and then homed in on the dress (over the tie).

We found similar effects in Experiment 2, in which sentences used the pronoun *You* rather than *I*, so that appropriate prediction was not tied to the speaker's perspective. Participants (N=32, 16 males) predicted associatively from 300 ms after verb onset ($ps < .05$) and appropriately from their own perspective from 1000 ms ($ps < .05$). Note that this appropriate effect was later in Experiment 2 than Experiment 1, perhaps because there is some ambiguity as to who *You* refers to [6]. But importantly, participants again predicted appropriately later than they predicted associatively, providing further evidence for a two-stage account.

In Experiment 3, participants (N=32, 16 males) listened to sentences referring to a male (James) or a female (Kate) character. Participants predicted associatively from 300 ms after verb onset ($ps < .05$). They also predicted appropriately (looking at the target that matched the character's gender) from 450 ms, again later than the associative effect.

We conclude that comprehenders predict in two different ways – associatively, by drawing on information associated with the verb, and appropriately, by drawing on relevant contextual information. We show how these findings are compatible with initial resource-free prediction-by-association, followed by slower resource-intensive prediction-by-production [7].

References

- [1] Altmann, G. T. M. & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247-264.
- [2] Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46, 551-556.
- [3] Neely, J. H. (1977). Semantic priming and retrieval from lexical memory. Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106, 226-254.
- [4] Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11, 32-38.
- [5] Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of cognitive neuroscience*, 20, 580-591.
- [6] Brunyé, T. T., Ditman, T., Mahoney, C. R., Augustyn, J. S., & Taylor, H. (2009). When you and I share perspectives: Pronouns modulate perspective taking during narrative comprehension. *Psychological Science*, 20, 27-32.
- [7] Pickering, M. J. & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144, 1002-1044.

Figure 1. Examples of stimuli used in the three experiments

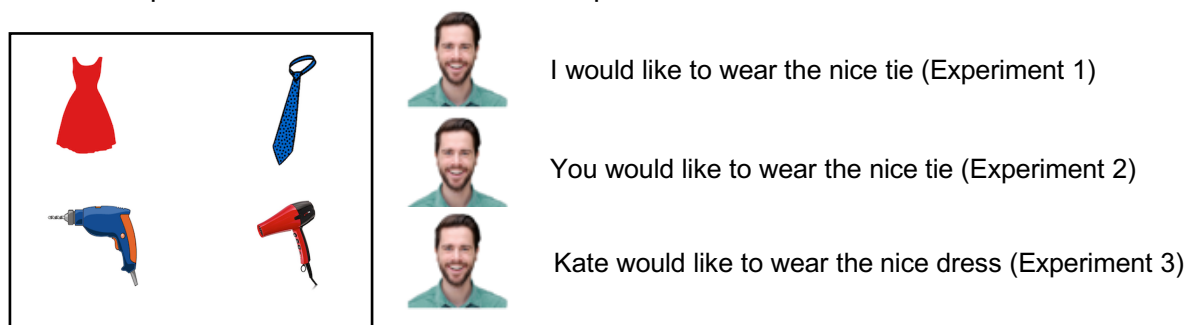
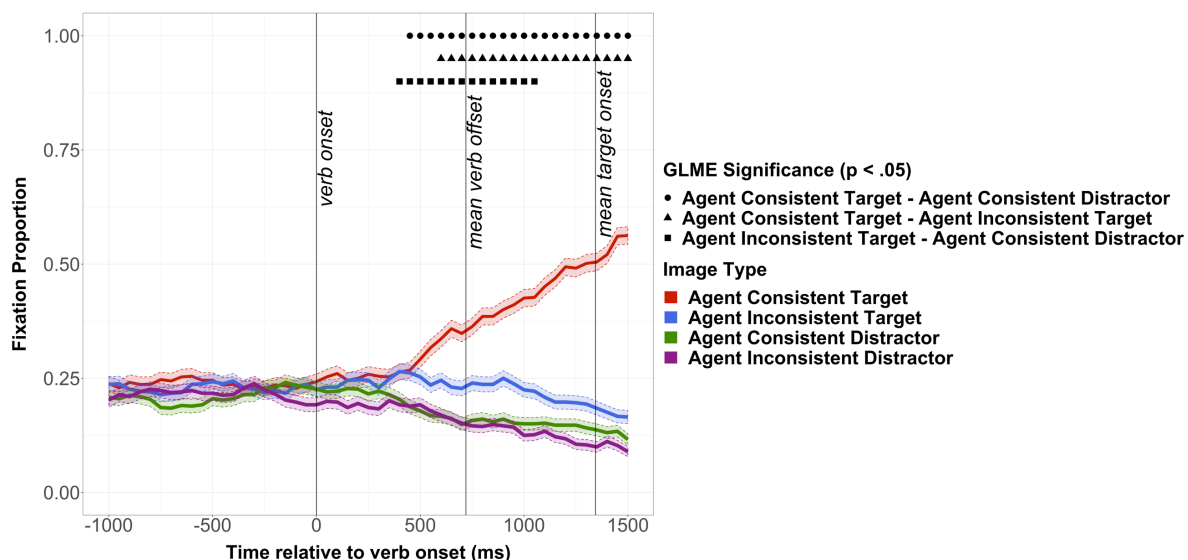


Figure 2. Eye-tracking results for Experiment 1. Shapes at the top of the graph show significant differences ($p < .05$) for the time bin on the x-axis between the critical pairs of pictures.



Turning the young parser into the adult parser: Working memory matters

Jiawei Shi and Peng Zhou (Tsinghua University)

Children exhibit difficulties in processing structural ambiguity due to their failure to revise their initial misinterpretation (Trueswell et al., 1999). This difficulty is often attributed to their non-adult cognitive attributes, one of which is their limited working memory capacity. Our study investigates whether children become more adult-like in processing structural ambiguity when the working memory burden associated with reanalysis is alleviated. The rationale is based on several adult working memory models, such as the one by Lewis et al. (2006), which proposes that when the ambiguous word is adjacent to the disambiguation point, the linear distance between them is minimized, and so is the working memory burden with reanalysis. The present study aims to explore whether the same rationale can be applied to child sentence processing.

Using the visual world paradigm, the eye movement data of 25 Mandarin-speaking four-year-olds, 25 five-year-olds and 30 adults were collected. The participants were presented with 8 target and 8 control items in random order, each containing a spoken sentence and a picture (see Fig.1). The target sentences (see (1)) had the following structure: “NP1 + Modal + V + NP2 + *DE* + NP3”. The morpheme *DE* is a possessive marker, so “NP2 + *DE* + NP3” indicated a possessive relation in which NP2 (*xiaogou* “dog”) was the possessor and NP3 (*piqiu* “ball”) was the possessee. The verb *ti* ‘kick’ could take either NP2 or NP3 as a plausible complement. If the parser incrementally processed the sentence, it might initially analyze “NP1 + Modal + Verb + NP2”, as in (2), as a complete sentence before encountering the disambiguating point *DE* which is adjacent to the ambiguous NP2. Upon hearing *DE*, the parser had to reanalyze NP2 as the modifier of the actual object NP3 (*piqiu* ‘ball’). By contrast, the control sentences (see (3)) followed the structure of the target sentences up until the point of disambiguation, but crucially did not involve a garden path. If the participants were able to revise their initial interpretation, when hearing *DE* in the target sentences, they should be expected to: 1) switch their looks from the dog to the dog’s ball; 2) exhibit more looks to the dog’s ball and fewer looks to the dog than when hearing the adverb *yixia* “once” in the controls.

Fig.2 and Fig.3 show the average fixation proportions on two critical areas: *Target_Mod* (the dog) and *Target_Obj* (the dog’s ball). As shown in both figures, all the three age groups showed similar eye gaze patterns. They initially looked more at the dog and then switched their looks to the dog’s ball when hearing *DE* (Fig.2). Besides, they exhibited more looks to the dog’s ball and fewer looks to the dog when hearing *DE* in the targets than when hearing *yixia* in the controls (Fig.3). However, 4-year-olds showed an overall delay in exhibiting the relevant pattern than the older groups. The observed eye gaze patterns were then confirmed by statistical modelling.

The findings suggest that 4-year-olds could revise their initial representation, though not as effective as 5-year-olds and adults, when the working memory burden associated with reanalysis was reduced to minimum. The findings also provide a good example of how adult processing models can inform us about child sentence processing, as well as calling for a fine-grained model of child sentence processing that specifies how each cognitive component contributes to the development of the young parser.

- (1) Xiaomao yaoqu ti xiaogou DE piqui
Cat will kick dog DE ball
"The cat is going to kick the dog's ball."
- (2) Xiaomao yaoqu ti xiaogou
Cat will kick dog
"The cat is going to kick the dog."
- (3) Xiaomao yaoqu ti xiaogou yixia
Cat will kick dog once
"The cat is going to kick the dog once."

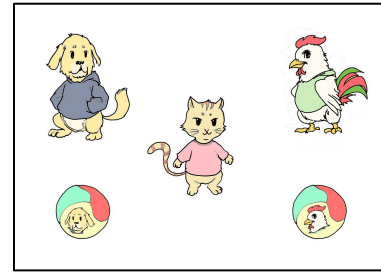


Fig.1 Example visual stimulus

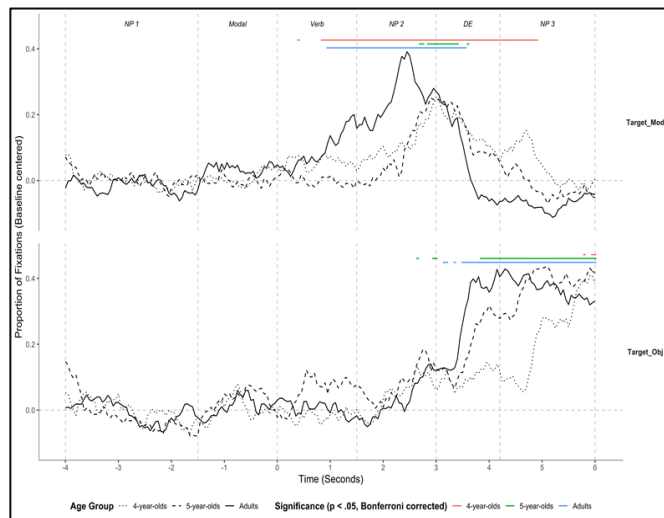


Fig.2 Average fixation proportions in the *Target_Mod* area (upper panel) and in the *Target_Obj* area (lower panel) by the 4-year-olds (dotted line), the 5-year-olds (dashed line) and the adults (solid line). The illustrated proportions are baseline centered (subtracting the mean fixation proportion in that area before the verb). The colored lines indicate a significantly higher fixation proportion than the baseline in this area during this temporal bin; the red line represents the 4-year-olds, the green line the 5-year-olds and the blue line the adults.

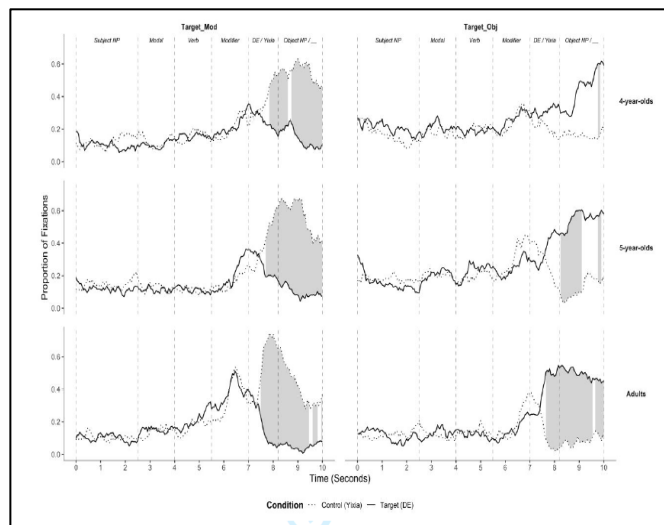


Fig.3 Average fixation proportions in the *Target_Mod* area (e.g. the dog, left column) and in the *Target_Obj* area (e.g. the dog's ball, right column) by the 4-year-olds (upper panel), the 5-year-olds (middle panel), and the adults (lower panel). The gray areas indicate significant differences between the target and control baseline conditions on the basis of the adjusted p values ($p < .05$).

Selected references

- Lewis, R. L., Vasishth, S., & van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10, 447–454.
- Trueswell, J.C., Sekerina, I., Hill, N.M. & Logrip, L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition*, 73, 89-134.

The Meaning and Processing of Conditionals – German ‘wenn’ (if) vs. ‘nur wenn’ (only if)

Mathias Barthel & Mingya Liu (Humboldt University Berlin, Germany)

This paper focuses on the semantics, pragmatics and processing of the lexically related German conditional connectives (CCs) ‘wenn’ (if) and ‘nur wenn’ (only if). In logic, *if* is treated as a binary truth-functional CC of material implication ($p \rightarrow q$). However, the interpretation of conditionals in natural language is subject to semantic and/or pragmatic modulation [1-3]. The modulating role of CCs for a conditional’s interpretation hitherto remains unclear.

Logically, modus ponens (MP) should be valid for all conditional sentences, irrespective of their CC (If $p \rightarrow q$; p . // q .). Based on the semantics of ‘only’ proposed in [4], ‘nur-wenn’ sentences should also entail the affirmation of the consequent inference (AC) (Only if $p \rightarrow q$; q . // p .), making ‘nur wenn’ a promising candidate for a natural language bi-conditional CC. The bi-conditional status of ‘nur wenn’ is doubted by [5], however. In a series of three experiments (E1-3), we contrasted the meaning and interpretation processes of the respective CCs.

In E1 ($N_{subj} = 24$, $N_{items} = 108$), participants read short scenarios including a conditional (If p , q .) with ‘wenn’ or ‘nur wenn’ and a second sentence containing the confirmed or negated antecedent proposition (p / not- p). Participants completed a final sentence fragment by either affirming or negating the consequent proposition (q / not- q ; see (1)). After confirmed antecedents, <1% of completions in ‘wenn’ but 11% in ‘nur wenn’ contained a negated consequent. After negated antecedents, however, 15% of completions in ‘wenn’ but <1% in ‘nur wenn’ contained a negated consequent, suggesting that neither of the CCs was treated as bi-conditional, with AC being questionable for ‘wenn’ and MP being questionable for ‘nur wenn’.

In E2 ($N_{subj} = 48$, $N_{items} = 48$, $N_{fillers} = 48$), participants were presented with a conditional sentence containing ‘wenn’ or ‘nur wenn’ and a second sentence containing either the confirmed or the negated antecedent proposition. In a final sentence, participants were asked to rate the truth of the consequent on a 5-point Likert scale (see (2)). A Bayesian ordinal mixed model with CC and antecedent plus their interaction revealed the bi-conditional interpretation to be most prominent overall, with mean ratings for both CCs above 4.6 after confirmed antecedents and below 1.6 after negated antecedents. However, after confirmed antecedents, acceptance rates were decisively lower for ‘nur wenn’ than for ‘wenn’ ($BF_{10} = 499$), suggesting that in ‘nur wenn’, less p -cases have been interpreted to be q -cases than in ‘wenn’. After negated antecedents, on the other hand, ratings for ‘wenn’ were decisively higher than for ‘nur wenn’ ($BF_{10} > 2000$), suggesting that in ‘wenn’, less not- p -cases have been interpreted to be not- q -cases than in ‘nur wenn’ (Fig. 1). Analyses of rating latencies support these results, with faster decisions for ‘wenn’ after confirmed than after negated antecedents and for ‘nur wenn’ after negated than after confirmed antecedents (Fig. 2). These results again cast doubt on the strict bi-conditionality of ‘nur wenn’ (or ‘wenn’, as expected).

To compare the CCs’ online interpretation, participants in E3 ($N_{subj} = 24$, $N_{items} = 108$, $N_{fillers} = 24$) did a self-paced reading task on scenarios containing a conditional sentence with either ‘wenn’ or ‘nur wenn’ and a follow-up sentence which, in critical trials, always contained the negated antecedent. A final sentence contained either the confirmed or the negated consequent (see (3)). A Bayesian mixed effects regression model (Fig. 3) with CC and consequent plus their interaction revealed that reading times for the positive quantifier in the final sentence (indicating the confirmed consequent) were statistically equivalent between CCs, but the negative quantifier was read decisively faster in ‘nur wenn’ than in ‘wenn’, suggesting that the meaning ‘not- $p \rightarrow$ not- q ’ is activated more strongly by ‘nur wenn p , q ’ than by ‘wenn p , q ’ conditionals.

In conclusion, neither ‘wenn’ nor ‘nur wenn’ are interpreted as strictly bi-conditional connectives. While for ‘wenn’, all p -cases are interpreted to be q -cases, only some not- p -cases are not- q -cases. For ‘nur wenn’, on the other hand, all not- p -cases are interpreted to be not- q -cases and only some p -cases are q -cases. This finding contradicts common conceptions of the meaning of *only if* and calls for adequate formal analyses of the meaning contributions of CCs.

- (1) S1: Kristian las die Zeitung und dachte sich: (K. read the newspaper and thought:)
 S2: **Wenn/Nur wenn** die Artikel interessant sind, schneide ich einen aus. (If/Only if the articles are interesting, I'll cut one out.)
 S3: Wie sich zeigte, waren die Artikel **(nicht)** interessant. (As it turned out, the articles were (not) interesting.)
 S4: Von denen schnitt er (Of these he cut)
- (2) S1: **Wenn/Nur wenn** heute gutes Wetter ist, geht Kai Eis essen. (If/Only if the weather is good, Kai will go have ice cream.)
 S2: Heute ist **(kein)** gutes Wetter. (The weather is (not) good today.)
 S3: Geht Kai Eis essen? (Is Kai going to have ice cream?)
- (3) S1: Kristian las die Zeitung und dachte sich: (K. read the newspaper and thought:)
 S2: **Wenn/Nur wenn** die Artikel interessant sind, schneide ich einen aus. (If/Only if the articles are interesting, I will cut one out.)
 S3: Wie sich zeigte, waren die Artikel **nicht** interessant. (As it turned out, the articles were not interesting.)
 S4: Von denen schnitt er **einen / keinen** aus und las weiter. (Of these he cut one / none out and continued to read.)

Figure 1. Rating results in E2.

Mean Ratings (CI) in Experiment 2

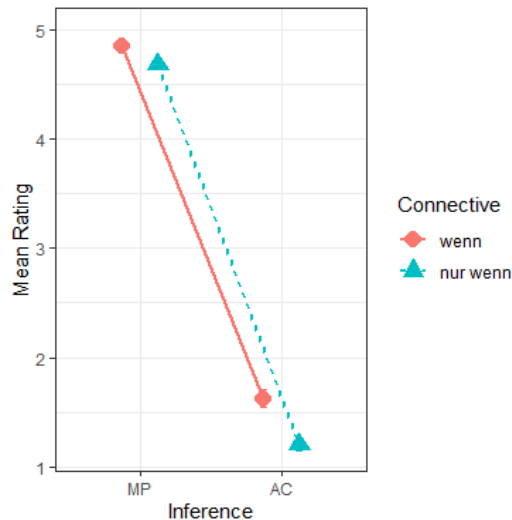


Figure 2. Rating latencies in E2.

Mean RTs (CI) in ms in Experiment 2

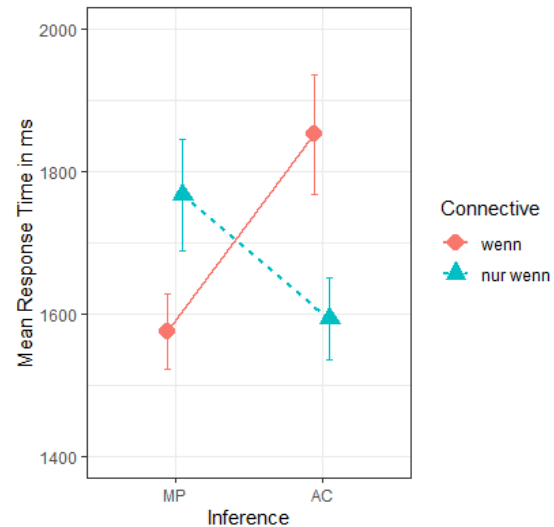
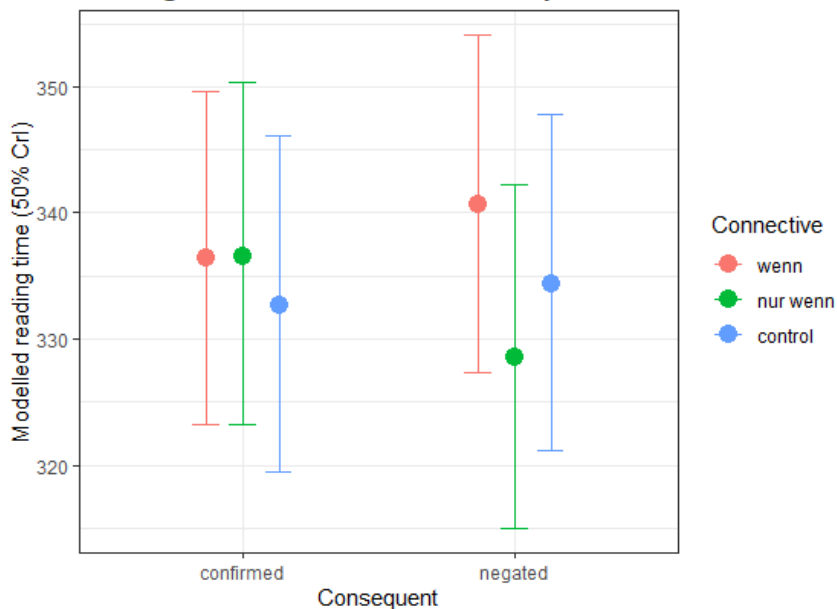


Figure 3. Reading times for critical word in E3.

Reading Times for determiner in Experiment 3



References:

- [1] Evans & Over (2004). If.
- [2] Johnson-Laird & Byrne (2002). Conditionals: A theory of meaning, pragmatics, and inference.
- [3] von Fintel (2011). Conditionals.
- [4] Horn (2002). Assertoric inertia and NPI-licensing.
- [5] Herburger (2015). Only if: If only we understood it.

Adults process Number and Gender head-subject mismatches differently during the online comprehension of object-relative clauses (as children do, offline).

Nicoletta Biondo (University of Siena), Vincenzo Moscati (University of Siena), Luigi Rizzi (Collège de France) & Adriana Belletti (University of Siena)

Children selectively struggle with the comprehension of embedded relative clauses, cross-linguistically. Children comprehend subject-relative clauses (SRC) as in (1) more accurately than object-relative clauses (ORC) as in (2) [1]. Moreover, in Italian [2], ORC comprehension improves when the two NPs (the head of the RC *the waiter* and the subject of the RC *the boy*) mismatch in Number features, as in (3), while there is no improvement when the two NPs mismatch in Gender features, as in (4). Number and Gender behave differently because of their different morphosyntactic status in Italian: Number plays an active role (i.e., triggers movement) while Gender does not, as theorized in the featural Relativized Minimality approach [3-5,1].

(1) *Il cameriere che saluta il ragazzo lavora qui.* (The waiter that is greeting the boy works here.)

(2) *Il cameriere che il ragazzo saluta lavora qui.* (The waiter that the boy is greeting works here)

(3) *Il cameriere che i ragazzi salutano lavora qui.* (The waiter that the boys are greeting ...)

(4) *Il cameriere che la ragazza saluta lavora qui.* (The waiter that the girl is greeting ...)

Adults are not expected to show low accuracy in the comprehension of these sentences since they are all grammatical in Italian. In this self-paced reading study, we investigate whether Italian speaking adults show a selective facilitation effect for Number (compared to Gender) mismatches during online sentence comprehension (as children do “offline”).

Several studies show that SRC are easier to process than ORC, and that the dissimilarity between the head (e.g., NP *the waiter*) and the subject of the RC (e.g., pronoun *he*) can make ORC easier to process (e.g., [6]). Still, there is limited evidence of the SRC/ORC asymmetry in Italian adults [7,8], and no study has directly compared SRC and ORC with different instances of head-subject (Gender, Number) morphosyntactic mis/match, as we do in this study (see Table 1). ORCs should trigger longer reading times (RTs) compared to SRC (2 vs 1); ORC Number mismatches are expected to show a facilitation effect (faster RTs) compared to ORC All-match (2.c vs 2.a), while ORC Gender mismatches are not expected to show a similar facilitation effect (2.b vs 2.a).

Forty-six Italian native speakers accessed Ibex Farm to read 102 experimental sentences plus 60 fillers, constituent-by-constituent, and to answer comprehension questions. RTs data of the two critical words (RC verb, main clause verb) were analyzed through a 2-stage analysis [9,10]. RTs were log-transformed and regressed against word length and trial position. The residual log RTs then entered a parsimonious [11] linear mixed-effect model analysis. The fixed-effect factors were coded as repeated contrasts: *Clause* (SRC -0.5; ORC 0.5), *Gender* and *Number* (match -0.5; mismatch 0.5). Figure 1 shows average RTs; Table 2 shows the output of the analysis.

Both the RC verb and the main verb showed longer RTs in the ORC compared to the SRC condition, in line with previous studies. We also found a Number mismatch facilitation effect on the ORC verb while the same effect did not reach significance for Gender [1-5]. Our findings show that children and adults appear to be subject to the same syntactic constraints, offline and online respectively, with morphological information analyzed during syntactic processing in a selective way depending on the nature of the morphosyntactic feature involved.

We also found that the main verb of both SRCs and ORCs showed smaller RTs for Number and Gender compared to All-match. This non-selective mismatch effect may mirror both a late (spill-over) processing of the RC verb and the processing of the long-distance subject-verb relation (*The waiter... works*) [12]. To disentangle these processes, we are designing a follow-up study testing sentences where the head of the RC is the object, so that the word following the RC verb is not the main verb of the sentence (e.g., *John is watching the waiter that the boy is greeting during the lunch-break at the restaurant*).

References: [1] Friedmann et al. (2009), *Lingua*, 119(1), 67–88. [2] Adani et al. (2010), *Lingua*, 120(9), 2148–2166. [3] Rizzi (1990). MIT Press. [4] Rizzi (2004). In A. Belletti (Ed.), *Structures and Beyond: the Cartography of Syntactic Structures*, Vol. 3 (pp. 223-251). Oxford University Press. [5] Belletti et al. (2012). *Lingua*, 122(10), 1053–1069. [6] Gordon et al. (2001), *Journal of experimental psychology: learning, memory, and cognition*, 27(6), 1411. [7] Di Domenico & Di Matteo (2009). *The Journal of General Psychology: Experimental, Psychological, and Comparative Psychology*, 136(4), 387-406. [8] Guasti, Vernice & Frank (2018). *Languages*, 3(3), 24. [9] Hofmeister (2011), *Language and cognitive processes*, 26(3), 376-405. [10] Villata et al. (2018), *Frontiers in psychology*, 9, 2. [11] Bates et al. (2015). *arXiv preprint arXiv:1506.04967*. [12] Staub et al. (2017), *Cognitive Science*, 41, 1353-1376.

Table 1. Experimental conditions

SRC	All-match	(1.a) <i>Il professore che chiama lo studente apre la porta dell'aula.</i> (The professor _(sg,m) that calls the student _(sg,m) opens the door of the class)
	Gender (mismatch)	(1.b) <i>Il professore che chiama la studentessa apre la porta dell'aula.</i> (The professor _(sg,m) that calls the student _(sg,f) opens the door of the class)
	Number (mismatch)	(1.c) <i>Il professore che chiama gli studenti apre la porta dell'aula.</i> (The professor _(sg,m) that calls the students _(pl,m) opens the door of the class)
ORC	All-match	(2.a) <i>Il professore che lo studente chiama apre la porta dell'aula.</i> (The professor _(sg,m) that the student _(sg,m) calls opens the door of the class)
	Gender (mismatch)	(2.b) <i>Il professore che la studentessa chiama apre la porta dell'aula.</i> (The professor _(sg,m) that the student _(sg,f) calls opens the door of the class)
	Number (mismatch)	(2.c) <i>Il professore che gli studenti chiamano apre la porta dell'aula.</i> (The professor _(sg,m) that the students _(pl,m) call opens the door of the class)

Figure 1. Average RTs and standard errors.

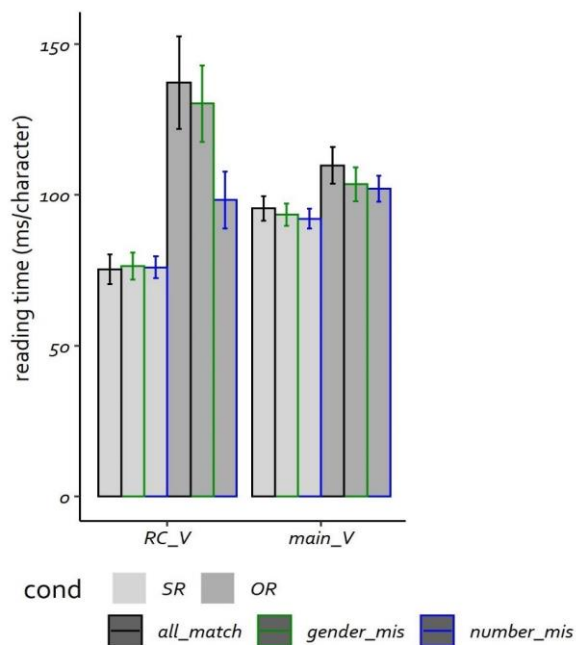


Table 2. Output of the statistical data analysis.

	Estimate	SE	t
Verb of the relative clause (RC_V)			
Clause	0.36	0.05	7.48
Gender	-0.02	0.02	-1.17
Number	-0.04	0.02	-2.75
Clause:Gender	-0.03	0.03	-1.04
Clause:Number	-0.11	0.04	-2.99
▪ Interaction model (SR clause)			
Number	0.01	0.02	0.78
▪ Interaction model (OR clause)			
Number	-0.08	0.03	-2.91
Verb of the main clause (main_V)			
Clause	0.06	0.02	4.04
Gender	-0.04	0.01	-2.49
Number	-0.03	0.02	-2.16
Clause:Gender	-0.02	0.03	-0.51
Clause:Number	-0.02	0.03	-0.47

Developmental effects in the real-time use of morphosyntactic cues: Evidence from Tagalog

Rowena Garcia (Max Planck Institute for Psycholinguistics), Gabriela Garrido Rodriguez (MPI Psycholinguistics, University of Melbourne, ARC Center of Excellence for the Dynamics of Language) & Evan Kidd (MPI Psycholinguistics, Australian National University, ARC CoEDL)

How children acquire and use different cues to rapidly process language is a matter of intense debate. On the one hand, *early abstraction* accounts predict that children process sentences using early emerging (or innate) adult-like linguistic generalizations (Özge, Kuntay, & Snedeker, 2019; Phillips & Ehrenhofer, 2015; Snedeker, 2013). In contrast, *experience-based* accounts assume a greater role of children's input, predicting that both acquisition and parsing decisions are input-driven (Chang, Dell, & Bock, 2006; MacDonald, 2013). In this research, we tested the predictions of these accounts in Tagalog (Austronesian), an understudied verb-initial language that uses pre-nominal morphosyntactic markers to assign thematic roles (i.e., voice-marking on the verb and a prenominal marker).

In Tagalog, the agent voice *-um-* indicates that the *ang*-marked noun is the agent [Table 1a, b], while the patient voice *-in-* marks the *ang*-phrase as the patient [Table 1c, d]. Post-verb word order is relatively flexible. Evidence from child-directed speech shows that the patient voice is overall more frequent, as well as the agent-initial order (Garcia, Roeser, & Höhle, 2019). Given this distribution, experience-based accounts predict that children would learn the patient voice mapping before that of the agent voice, as children have more exposure to the former than the latter, facilitating the rapid implementation of online parsing decisions. In contrast, early abstraction accounts do not predict a voice difference.

To test these predictions, we conducted an eye-tracking experiment with 32 adults (controls) and 151 children (fifty-three 5-year-olds, forty-nine 7-year-olds, forty-nine 9-year-olds), who saw a picture depicting a transitive action between two animals. After 1500ms of silence, they heard an audio-recorded sentence [Table 1a-d] that corresponded to the picture. They were told to pay attention because there would be questions about what they had seen and heard. There were 32 experimental items (8 per sentence condition) and 32 fillers. Our independent variables were voice and the order of the thematic roles; and the dependent variable was the proportion of fixations to the agent in the picture. Our analyses determined whether participants looked at the referent of the upcoming noun before it is mentioned (Noun1 region), based on the voice-marking on the verb and the noun marker that they had previously encountered.

A permutation analysis revealed that the ability to use morphosyntactic markers to assign thematic roles develops with age. The 5-year-olds showed divergence in the looks to the agent between agent-initial and patient-initial conditions only after the noun onset (Figure 1). However, similar to adults, 7- and 9-year-old children showed predictive use of the morphosyntactic markers in the patient voice. Thus, in Figure 2 (bottom panel), 7-year-olds looked more to the agent during the pre-noun regions when the sentence was agent-initial than when it was patient-initial (significant regions are shaded grey). However, in the agent voice we only found divergence after noun onset.

Our results showed that children's online use of morphosyntactic markers develops with age, with adult-like online predictive processing only beginning to emerge at 7 years. Furthermore, we found that the real-time use of the markers is modulated by voice—with the patient voice being used more efficiently than the agent voice. We interpret this to reflect the participants' sensitivity to the distributional properties of the language in line with experience-based accounts.

Table 1. Sample stimuli sentences

(a) Agent voice agent-initial	H<um>uhuli <AV>capture	noong Martes last Tuesday	ang SBJ	malusog healthy	na LIN	unggoy monkey	ng NSBJ	baka cow
'The healthy monkey was capturing a cow last Tuesday.'								
(b) Agent voice patient-initial	H<um>uhuli <AV> capture	noong Martes last Tuesday	ng NSBJ	malusog healthy	na LIN	baka cow	ang SBJ	unggoy monkey
'The monkey was capturing a healthy cow last Tuesday.'								
(c) Patient voice agent-initial	H<in>uhuli <PV> capture	noong Martes last Tuesday	ng NSBJ	malusog healthy	na LIN	unggoy monkey	ang SBJ	baka cow
'The/A healthy monkey was capturing the cow last Tuesday.'								
(d) Patient voice patient-initial	H<in>uhuli <PV> capture	noong Martes last Tuesday	ang SBJ	malusog healthy	na LIN	baka cow	ng NSBJ	unggoy monkey
'The/A monkey was capturing the healthy cow last Tuesday.'								

Note. The vertical lines show the division between the sentence regions namely, verb + temporal adverb, first noun marker + adjective, first noun, second noun marker + second noun. Abbreviations: AV (agent voice), PV (patient voice), SBJ (subject), NSBJ (non-subject), LIN (linker).

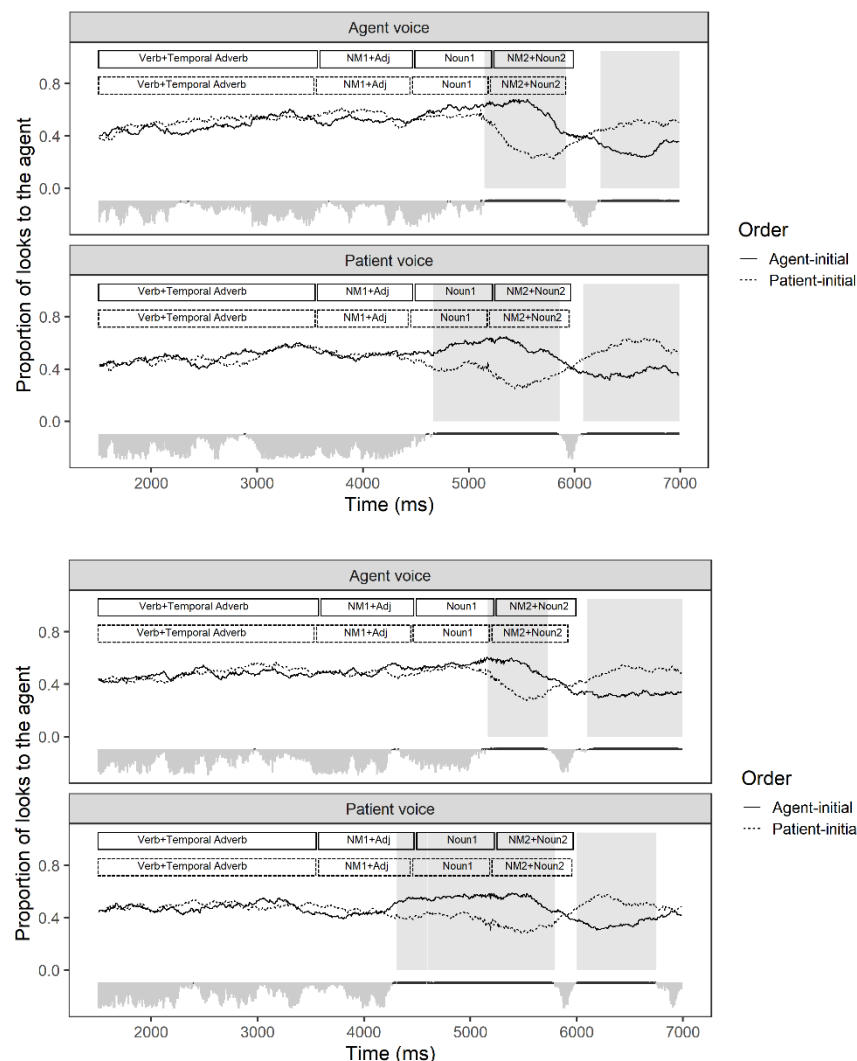


Figure 1. Five-year-olds' average proportion of looks to the agent. The sentence regions are indicated by the rectangles (NM1=1st noun marker; Adj= adjective; NM2=2nd noun marker). The small grey/black bars around - 0.01 indicate the *p* values for each time bin. The large grey bars indicate the time bins which were found to be significant in the permutation analysis.

Figure 2. Seven-year-olds' average proportion of looks to the agent from verb onset until the end of the trial.

Selective Modulation of Syntactic Processing by Anodal tDCS over the Left Inferior Frontal Region

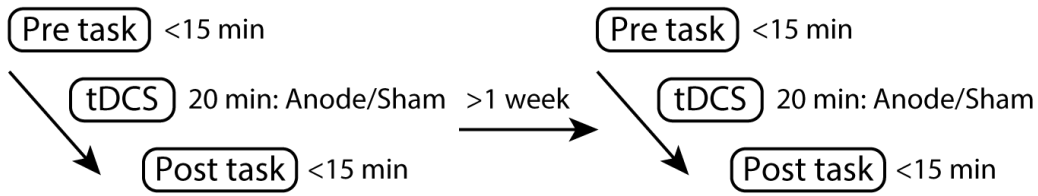
Shinri Ohta (Kyushu University)
ohta@lit.kyushu-u.ac.jp

Previous neuroimaging studies have demonstrated that the left inferior frontal gyrus (IFG) is critical for syntactic processing. To test the causal relationship between the left IFG activation and syntactic processing, we examined whether anodal (i.e. excitatory) transcranial direct current stimulation (tDCS), a non-invasive brain stimulation technique applicable in humans, over the left IFG facilitates syntactic processing. We hypothesize that behavioral performance of sentences with additional syntactic loads (e.g. passive sentences) is improved by the anodal tDCS.

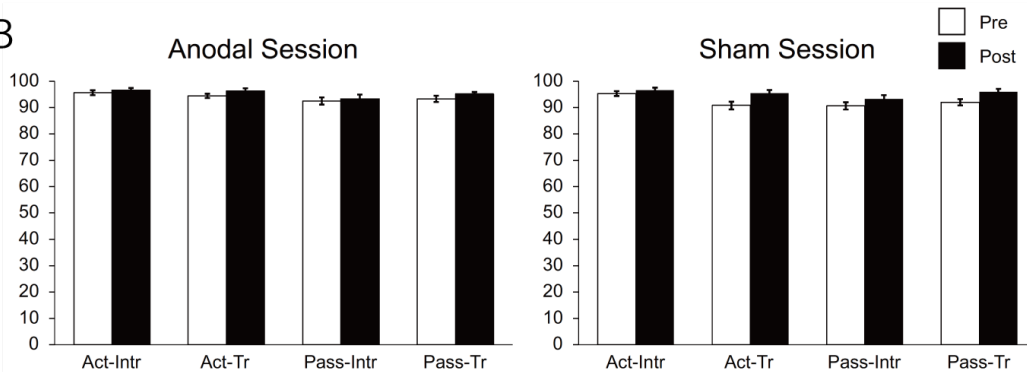
We recruited 20 right-handed native speakers of Japanese (10 males, mean \pm SD = 22.5 \pm 0.8 years), who had no history of neurological or psychiatric diseases. The same participants were tested for both anodal stimulation session and sham session (Fig. 1A). We used 30 Japanese sentences for each of active intransitive (e.g., *Taro-to Hanako-ga aruita*, *Taro and Hanako walked*), active transitive (*Taro-ga Hanako-o tataita*, *Taro hit Hanako*), passive intransitive (*Hanako-ga Taro-ni arukareta*, *Hanako was adversely affected by Taro's walking*), and passive transitive sentences (*Hanako-ga Taro-ni tatakareta*, *Hanako was hit by Taro*) (total 120 stimuli). To examine the effect of active/passive voice as well as that of transitivity, we used these four sentence types. Note that the passive intransitive sentences, the so-called indirect passive, are grammatical in Japanese. Each sentence consisted of two noun phrases and one verb, immediately followed by a question consisted of a subject and a verb (e.g., *Taro-ga aruita?*, *Did Taro walk?*). In the present experiment, we used a sentence comprehension task, in which the participants were instructed to judge whether the meaning of the sentence matched with the question by pressing one of two buttons. We used a single-blinded sham-controlled design. Stimulation was delivered using DC-Stimulator Plus (NeuroConn GmbH, Germany). The anode and cathode electrodes were placed over F5 and F6 according to the International 10-20 EEG system, which were right above the left and right IFG, respectively. For anodal tDCS, stimulation was given for 20 minutes (1 mA, 5 cm * 7 cm saline-soaked sponge electrodes). Sham stimulation, which controls for the placebo effect, ramped up to 1 mA over 10 s, remained at that level for 30 s, ramped back down over 10 s. In the sham session, the participants felt the initial ramp up event, which is the most noticeable in tDCS, without receiving an effective stimulation in the anodal tDCS. Before and after the anodal and sham stimulations, the participants performed the sentence comprehension task (Pre and Post task).

The participants showed high accuracies (> 90%) and short reaction times to comprehension questions (RTs, <1600 ms) for all of the four conditions (Fig. 1B, 1C). A three-way repeated-measures analysis of variance (rANOVA) (Stimulation*Condition*Pre/Post) for the accuracies showed significant main effects of Condition ($F(3,57)=11$, $p<.0001$) and Pre/Post ($F(1,19)=8.4$, $p=.009$), while the main effect of Stimulation and interactions were not significant ($p>.18$). The rANOVA for the RTs also showed significant main effects of Condition ($F(3,57)=42$, $p<.0001$) and Pre/Post ($F(1,19)=21$, $p=.0002$), as well as the interaction of these factors ($F(3,57)=3.7$, $p=.002$). These results suggest that the active intransitive condition was easiest, while the passive conditions were more demanding. The significant main effect of the Pre/Post also shows the learning effect. To consider the random variabilities of participants and stimuli, we further analyzed the RTs by using a linear mixed-effect model (lme4 and lmerTest packages on R). We found that the model with the effect of Stimulation was significantly better than the simpler model without such effect ($\chi^2(3)=38$, $p<.0001$), suggesting the effect of anodal tDCS. Moreover, the anodal stimulation over the left IFG significantly decreased the RTs of the passive sentences ($p=.002$, Fig 1D). In the present tDCS study, we demonstrated that the anodal tDCS over the left IFG facilitated the processing of syntactically more demanding passive sentences, suggesting the causal relationship between the left IFG activation and syntactic processing.

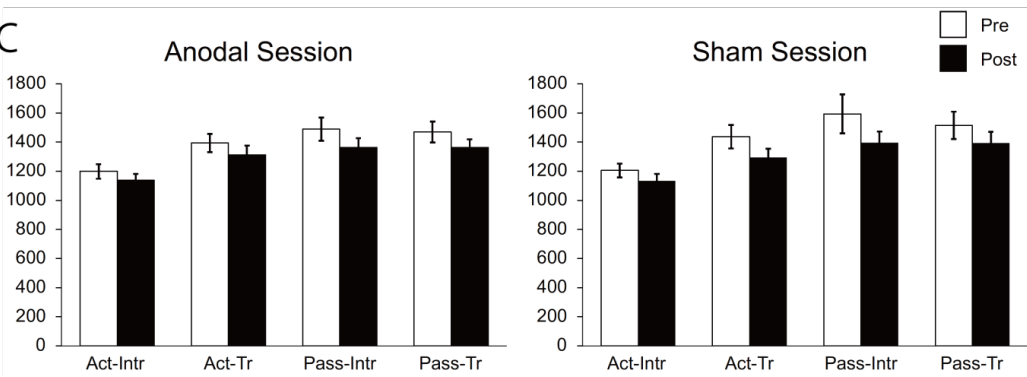
A



B



C



D

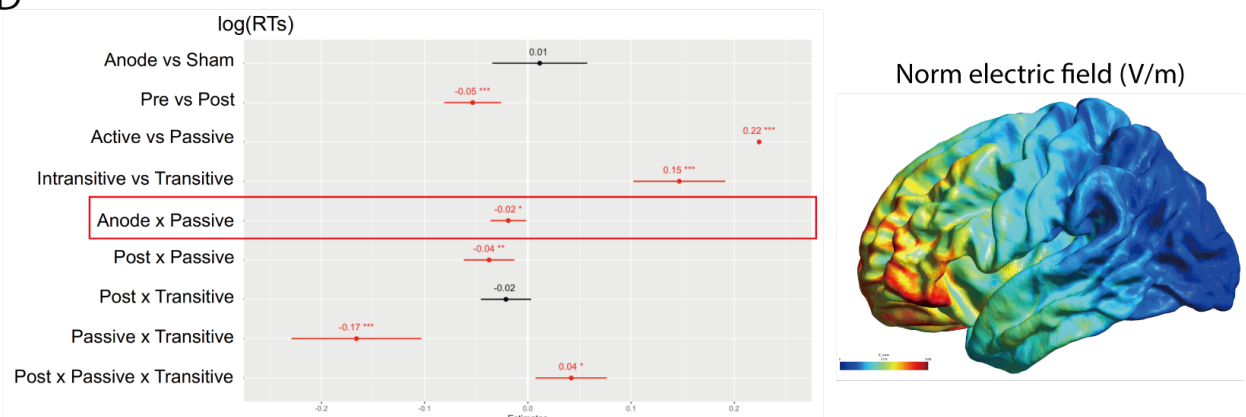


Figure. (A) Schematic illustration of the tDCS procedures, (B) accuracies to the comprehension questions, (C) reaction times to the comprehension questions, and (D) the LME results and estimated electric fields during anodal tDCS.

What is the upper limit of working memory? Evidence from Chinese recursive possessive structure

Zihan Zhang, Shuqi Ni, Shuyang Liu and Fuyun Wu (Shanghai Jiao Tong University)

In Mandarin possessive structure ($N_{\text{possessor}}$ de N_{possesum}), the possessive marker DE can be dropped when the possessor bears an inalienable relationship with the possesum, as in *wo (de) mama* ‘my mom’. Importantly, a possessive phrase can be nested within another possessive phrase in succession, forming possessive chains. Thus, possessive chains provide a good test case for probing into memory capacity, a factor that is known to affect real-time parsing. Existing work has yet to give a definitive answer regarding the upper limit of short-term memory capacity. While Miller (1956) posits that short-term memory capacity is “7±2”, Cowan (2001) proposes that it is limited to “4±1”. Using the recursive possessive chain structure in Mandarin, we set out to explore native Chinese speakers’ upper limit of working memory capacity by manipulating the presence versus absence of the possessive marker “DE” between the six possessor nouns.

In **Experiment 1** (N=80), we ran a grammaticality acceptability task using a 5-point scale (1 = least acceptable, 5 = most acceptable) on *wenjuanxing* (www.wjx.cn). In the experimental stimuli, totaling 20 sets, the sentential subject consisted of 7 noun phrases (NPs), where the first NP was always the first-person pronoun *wo* ‘I’, followed by six inalienable kinship terms, with the 7th NP being the possesum (see ex.(1)). Between the NP2 and the NP6, we manipulated the presence or absence of “DE”, yielding 5 conditions (1a-e). To prevent participants from developing test-taking strategies, we created 10 versions of tests, each having 10 experimental sentences, with 2 from each condition. In each version, 10 experimental stimuli were intermixed with 20 filler sentences of various structures, and then pseudo-randomized. Fig. 1 shows participants’ mean ratings by conditions. When using (c) as the baseline, we found that (a) was rated significantly higher than (c) ($\beta = 0.32$, SE = 0.15, $t = 2.23$, $p = 0.026$), and (c) was rated significantly higher than (e) ($\beta = -0.38$, SE = 0.15, $t = -2.59$, $p = 0.0098$). When using (d) as the baseline, we found (d) was rated significantly lower than (a) ($\beta = 0.52$, SE = 0.15, $t = 3.66$, $p = 0.0003$) and (b) ($\beta = 0.31$, SE = 0.15, $t = 2.14$, $p = 0.03$). But no differences were found between (c) and (d), nor between (d) and (e). These patterns suggest that starting from the 4th and 5th consecutive nouns (i.e., c & d), the acceptability gets drastically decreased.

To control potential effects of “similarity-based interference” (Gordon et al. 2006) presented in Experiment 1 due to kinship terms in a row, we ran **Experiment 2** (N=50) by alternating kinship terms with descriptive NPs (see ex.(2a-e)). We used self-paced reading with a stop-making-sense task, following Boland et al. (1989). Participants took the online test on Gorilla, followed by an offline paper-&-pen test, in which they not only rated the grammaticality of experimental sentences (a version different from the online test) on a 5-point scale – as in Experiment 1, but identified their sensitive points after which the sentences started to become incomprehensible. We found the condition (2d) (i.e., 5 consecutive nouns) had the highest percentage of ‘stop-making-sense’ button-press (29%). Furthermore, regarding the offline GJ data, the results basically replicate Experiment 1 with a much clearer pattern (Fig. 2). When (2c) was set as the baseline, (2c) was rated significantly lower than (2a) ($\beta = 1.04$, SE = 0.15, $t = 7.12$, $p < 0.0001$) and (2b) ($\beta = 0.73$, SE = 0.15, $t = 5.00$, $p < 0.0001$), but was rated higher than (2d) ($\beta = -0.75$, SE = 0.15, $t = -5.13$, $p < 0.0001$) and (2e) ($\beta = -1.04$, SE = 0.15, $t = -7.12$, $p < 0.0001$). When (2d) was set as the baseline, (2d) was rated significantly lower than (2a), (2b) and (2c) ($ps < 0.0001$). These patterns suggest that participants’ acceptability ratings decreased significantly between (2c) (i.e., 4 consecutive nouns) and (2d) (i.e., 5 consecutive nouns).

Taken together, our results showed that the upper limit of processing Chinese possessive chains is four consecutive nouns. Our study supports Cowan’s (2001) hypothesis, providing novel evidence for precise quantification of human working memory capacity that underlies language processing.

References

- [1] Miller, G.(1956). *The Psychological Review*.
 [2] Miller, G.& Chomsky, N.(1963). *Handbook of Mathematical Psychology*.
 [3] Cowan, N.(2001). *Behavioral and Brain Sciences*.
 [4] Gordon,P., Hendrick, R., Johnson, M., &Lee, Y.(2006). *JEP: LMC*.
 [5] Boland, Tanenhaus, Carlson, &. Garnsey (1989). *Journal of Psycholinguistic Research*.
 [6] Comrie, B. Possessive chains and possessor camouflage.
 [7] Lu, B. (1983). The conditions of infinite recursion and finite segmentation. *Hanyu Xuexi*.

(1) Sample stimulus set in English gloss (shown in Chinese characters)of Experiment 1

condition	NP1	DE1	NP2	DE2	NP3	DE3	NP4	DE4	NP5	DE5	NP6	DE6	NP7	Predicate
a	My		son	DE	uncle _m	DE	daughter	DE	aunt _m	DE	elder brother	DE	father	won the prize.
b	My		son		uncle _m	DE	daughter	DE	aunt _m	DE	elder brother	DE	father	won the prize.
c	My		son		uncle _m		daughter	DE	aunt _m	DE	elder brother	DE	father	won the prize.
d	My		son		uncle _m		daughter		aunt _m	DE	elder brother	DE	father	won the prize.
e	My		son		uncle _m		daughter		aunt _m		elder brother	DE	father	won the prize.

Note: The subscript 'm' means 'on the maternal side', 'p' means 'on the paternal side'.

(2) Sample stimulus set in English gloss (shown in Chinese characters) in the online version of Experiment 2, using self-paced reading with a stop-making-sense task

condition	context	NP1	NP2	NP3	NP4	NP5	NP6	DE7	NP7	Predicate
a	Last night, news program said that	our	son	/	/	/	/	DE	colleague	has brilliant achievements in war
b		our	son	classmate	/	/	/	DE	colleague	has brilliant achievements in war
c		our	son	classmate	uncle	/	/	DE	colleague	has brilliant achievements in war
d		our	son	classmate	uncle	comrade	/	DE	colleague	has brilliant achievements in war
e		our	son	classmate	uncle	comrade	elder brother	DE	colleague	has brilliant achievements in war

Note: The slash '/' means that the slot is not filled in with any lexical content

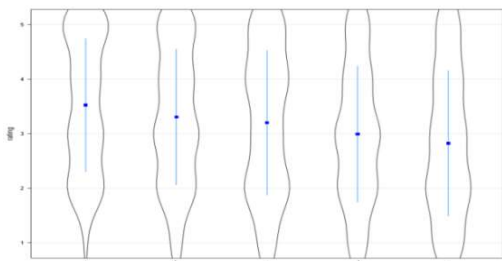


Fig. 1 Mean ratings of GJ in Exp. 1

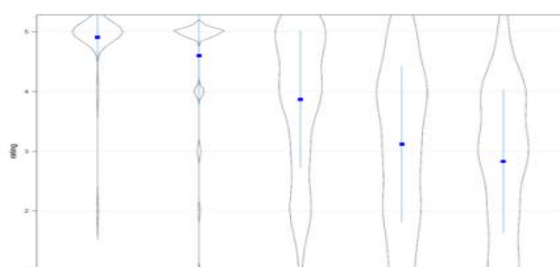


Fig. 2 Mean ratings of GJ in the offline task of Exp. 2

Reliance on semantic and structural heuristics across the lifespan

Anastasiya Lopukhina (HSE University, Russia), Anna Laurinavichyute (HSE University, Russia; University of Potsdam, Germany), Svetlana Malyutina (HSE University, Russia)

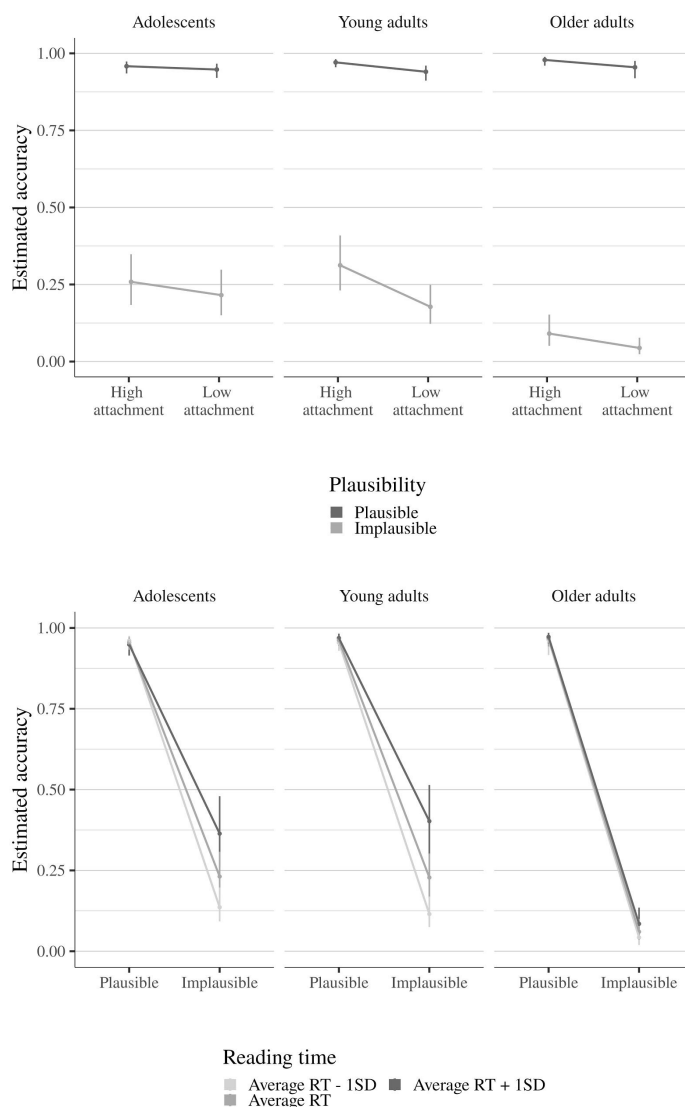
People sometimes misinterpret the sentences that they read. One possible reason suggested in the literature is a race between grammar-driven incremental bottom-up processing and “fast and frugal” top-down heuristic processing that serves to support fast-paced communication but sometimes results in incorrect representations. Heuristics can be semantic, relying on world knowledge and semantic relations between words [1], or structural, relying on structural economy [2]. According to the online equilibrium hypothesis of the good-enough processing theory [3], heuristic-based representations are computed faster than full syntactically-based representations. However, empirical studies have rarely evaluated this assumption directly, by analyzing the relationship between the accuracy of responses to comprehension questions (as an indicator of sentence representation accuracy) and reading speed.

Scattered experimental evidence preliminarily suggests that reliance on heuristics may change from greater reliance on syntactic information in younger people to greater reliance on semantic information in older people. Several studies showed that 7-to-12-year-old children relied on syntactic information and structural heuristics while disregarding semantic plausibility information [4]. At the same time, older adults were shown to rely more on semantic than syntactic information [5].

To test whether reliance on semantic and structural heuristics changes with age and whether heuristic processing is indeed faster than algorithmic processing, we tested three groups of Russian-speaking participants: 137 adolescents (87 female; age range 13-17 years, $M=15$), 135 young adults (99 female; age range 20–40 years, $M=25$), and 77 older adults (57 female; age range 55–91 years, $M=64$). The participants read 56 high- vs. low-attachment sentences that were marked by case inflection, and all stimuli sentences were therefore completely unambiguous (Russian speakers show bias to high-attachment interpretations even in unambiguous sentences, see [6]). The sentences were either semantically plausible or implausible, i.e., the syntactic structure either matched or contradicted the typical semantic relations, see Example 1 (all materials are available online <https://osf.io/4f2px/>). Sentences were presented in a non-cumulative self-paced reading paradigm and were followed by a two-alternative comprehension question targeting the attachment site of the relative clause.

To assess the reliance on heuristics, we analyzed question response accuracies using Bayesian mixed-effects logistic regression, see the model structure below. As expected, we found that young adults made more errors in the dispreferred implausible and low-attachment conditions. Older adults had lower accuracy than young adults across the board and showed a greater decrease in accuracy in implausible sentences, thus demonstrating increased reliance on semantic heuristics. Adolescents did not differ from young adults in overall accuracy, but had similar accuracy in high- and low-attachment conditions, thus demonstrating the lack of reliance on the structural heuristic of high attachment. We found that when participants read sentences faster, their accuracy decreased. However, specifically in implausible sentences, faster reading times were associated with an additional decrease in accuracy indicating that semantic heuristic processing was faster than incremental bottom-up processing.

To summarize, we showed heuristic mechanisms appear already in adolescence and then keep maturing across the adult lifespan, via emerging reliance on structural heuristics in adulthood and increasing reliance on semantic heuristics in older age. We also for the first time showed that heuristic processing is indeed faster than incremental processing, as predicted by the good-enough processing model.



Example 1:

High attachment, plausible

Rimma dressed the child-ACC of the writer-GEN, who was babbling-ACC incomprehensibly.

Question: Who was babbling incomprehensibly? **Child** / Writer

Low attachment, plausible

Rimma dressed the child-ACC of the writer-GEN, who published-GEN a popular novel.

Question: Who published an interesting novel? Child / **Writer**

High attachment, implausible

Rimma dressed the child-ACC of the writer-GEN, who published-ACC a popular novel.

Question: Who published an interesting novel? **Child** / Writer

Low attachment, implausible

Rimma dressed the child-ACC of the writer-GEN, who was babbling-GEN incomprehensibly.

Question: Who was babbling incomprehensibly? Child / **Writer**

The model structure:

$$\text{accuracy} \sim \text{age} * (\text{plausibility} + \text{attachment}) + \text{RT} * (\text{age} + \text{plausibility} + \text{attachment}) + \text{plausibility} : \text{RT} : \text{age} + (1 + \text{age} * (\text{plausibility} + \text{attachment}) + \text{RT} * (\text{age} + \text{plausibility} + \text{attachment}) + \text{plausibility} : \text{RT} : \text{age} \parallel \text{ItemID}) + (1 + \text{plausibility} + \text{attachment} \parallel \text{ParticipantID}).$$

- [1] Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1), 11-15.
- [2] Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4), 355-370.
- [3] Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, 69(5), 1013-1040.
- [4] Joseph, H. S. S.L., Liversedge, S.P., Blythe, H.I., White, S.J., Gathercole, S.E., & Rayner, K. (2008). Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, 61(5), 708-723.
- [5] Beese, C., Werkle-Bergner, M., Lindenberger, U., Friederici, A.D., & Meyer, L. (2019). Adult age differences in the benefit of syntactic and semantic constraints for sentence processing. *Psychology and Aging*, 34(1), 43-55.
- [6] Chernova, D., & Chernigovskaya, T. V. (2015). Syntactic Ambiguity Resolution in Sentence Processing: New Evidence from a Morphologically Rich Language. In *EAPCogSci*.

Keep calm and move on: Reduced processing advantage of an early-arriving morphological cue in comprehension of Korean suffixal passive construction

Chanyoung Lee (Yonsei University) & Gyu-Ho Shin (Palacký University Olomouc)

'Good-enough' processing argues that a linguistic processor favours a simpler and less effortful analysis available.^{[1][2]} The processor seeks to achieve cognitive equilibrium in online processing at the earliest opportunities and remain in this state as long as possible; these properties lead the processor to prefer heuristic processing over algorithmic processing.^[3] A core force that establishes heuristics, involving morpho-syntactic typicality and semantic-pragmatic plausibility,^[2] comes from frequency in use.^{[4][5]} Against this background, we investigate how sentence processing is modulated by heuristics and the assumed early-arriving morphological cue benefit in parsing^{[6][7]} during sentence comprehension. Korean, an SOV language, provides an intriguing testbed for this issue because scrambling of sentential components is permitted (albeit infrequent compared to the canonical counterpart) with the propositional meaning intact (yet inviting particular discourse effects).^[8] We focus on suffixal passives (Table 1) engaging in the unusual form-function mapping of case-marking: the NOM indicating a theme (but usually indicating an agent) and the DAT indicating an agent (but usually indicating a recipient), with passive morphology serving as a key disambiguation point for these form-function pairings.^{[8][9]}

Methods. Forty native speakers of Korean (mean age = 23.6; $SD = 4.05$) participated in two tasks sequentially in web-based platforms: self-paced reading (SPR; a non-cumulative moving-window paradigm) and acceptability judgment (AJ; a 6-point Likert scale from zero to five). Sixteen sentences (one half for the verb-final (VF) pattern; the other half for the verb-initial (VI) pattern), together with fillers, split into two sub-lists and were randomly assigned to participants. Sentences for the AJ were adapted from those for the SPR (Table 2) by reducing R1, R5, and R6. The data from each task (outliers excluded → AJ: Z-transformed; SPR: log-transformed) were fitted to separate linear mixed-effects models (AJ: canonicity as a fixed effect & participant / sentence as random effects; SPR: canonicity as a fixed effect & participant / word-in-region as random effects).

Prediction. (AJ) The VI pattern should be rated less acceptable than the VF pattern due to the infrequent word order with no relevant context. (SPR) If the position of passive morphology affects comprehension more strongly than heuristics, RTs for the VI pattern should be shorter than those for the VF pattern. This is because passive morphology in the VI pattern guides the whole interpretation from R2 whereas the same morphology in the VF pattern necessarily requires revision of the previous interpretation at R4. In contrast, if the opposite happens, we should expect RTs for the VI pattern to be longer than those for the VF pattern. This is due to continuous online disequilibrium incurred by the VI pattern— infrequent word order and weak plausibility, along with the unusual form-function associations of case-marking—relative to the VF pattern.

Results. (AJ; Fig 1) Participants rated the VI pattern significantly less acceptable than the VF pattern. Given the no-context setting, their judgment may have been affected by canonicity and plausibility of the sentences. (SPR; Fig 2) RTs for the VI pattern were numerically longer than those for the VF pattern in all regions (with statistical significance in R3/5/6), indicating that the VI pattern incurred more processing cost than the VF pattern. This is ascribable to (i) infrequent word order with no proper context and (ii) cumulative computation cost for integrating the unusual case-marking information (requiring realignment of the form-function mapping; R3/4) into the entire construction (R5/6) to arrive at a complete interpretation. The VF pattern involves the same revision/integration process, but the pattern is frequent and context-neutral within this construction type, so participants may have handled the processing challenge efficiently when encountering passive morphology in its typical location—a sentence-final position.

Together, our findings suggest that the extent to which a processor benefits from an early-arriving morphological cue may be limited to heuristic processing which is subject to morpho-syntactic typicality and semantic-pragmatic plausibility. This aligns nice with how good-enough processing occurs during sentence comprehension, continuously seeking online cognitive equilibrium.

Table 1. Korean suffixal passive construction

Pattern	Composition	How does PSV work in comprehension?	Frequency in use (within the construction)
Verb-final (canonical)	N-NOM + N-DAT + V-PSV	Requires revision of the initial interpretation	Frequent
Verb-initial (scrambled)	V-PSV + N-NOM + N-DAT	Guides the following interpretation	Infrequent

Note. The passive morphology consists of four allomorphs: *-i-*, *-hi-*, *-li-*, and *-ki-*.

Table 2. Scheme of stimuli (SPR)

	R1	R2	R3	R4	R5	R6
Verb-final (canonical)	<i>I heard that</i>	N-NOM	N-DAT	V-PSV	<i>yesterday</i>	<i>night</i>
Verb-initial (scrambled)		V-PSV	N-NOM	N-DAT		

Note. English translations in R1, R5, and R6 are only for the readers' sake; all test sentences were presented in Korean.

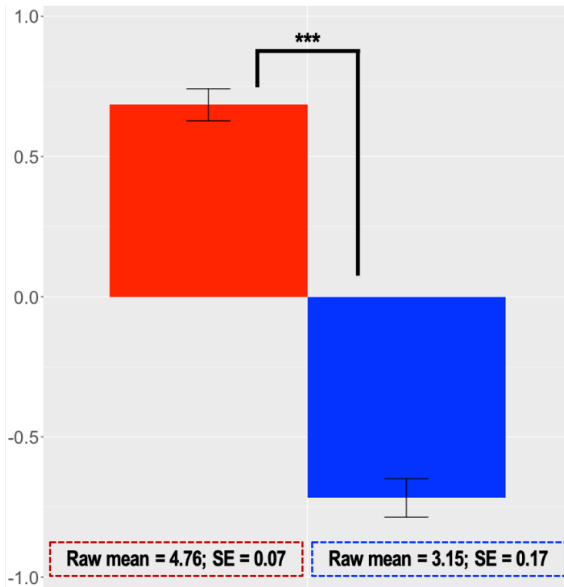


Figure 1. Result: AJ. X-axis: pattern; Y-axis: rating (1000 ms ≤ response time for each value ≤ 10000 ms (data loss: 4.37%) → Z-transformation); red: verb-final; blue: verb-initial. *** < .001.

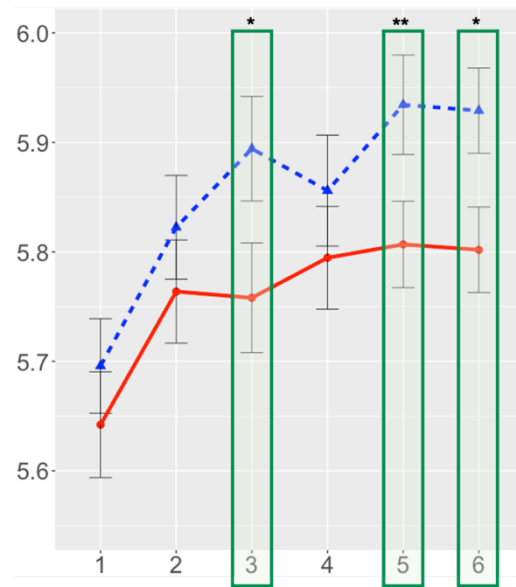


Figure 2. Result: SPR. X-axis: region; Y-axis: RT (3SD cut-off (data loss: 4.32%) → log-transformation); red: verb-final; blue: verb-initial. * < .05; ** < .01.

Abbreviations: DAT = dative marker; N = noun; NOM = nominative case marker; PSV = passive suffix; V = verb

References: [1] Christianson (2016) *Quarterly Journal of Experimental Psychology*, 69(5), 817–828. [2] Ferreira (2003) *Cognitive Psychology*, 47, 164–203. [3] Karimi & Ferreira (2016) *Quarterly Journal of Experimental Psychology*, 69(5), 1013–1040. [4] Ambridge et al (2015) *Journal of Child Language*, 42(2), 239–273. [5] Goldberg (2019) *Explain me this: Creativity, competition, and the partial productivity of constructions*. [6] Pozzan & Trueswell (2015) *Cognitive Psychology*, 80, 73–108. [7] Choi & Trueswell (2010) *Journal of Experimental Child Psychology*, 106(1), 41–61. [8] Sohn (1999) *The Korean language*. [9] Choo & Kwak (2008) *Using Korean*.

Processing noncanonical sentences: online and offline effects on misinterpretation errors

Markus Bader (University of Frankfurt), Michael Meng (Merseburg University of Applied Sciences)

Sentences with noncanonical argument order (e.g., patient/theme-before-agent instead of the more common agent-before-patient/theme order) have provided a longstanding challenge for theories of human sentence comprehension. Experimental studies on noncanonical sentences have been dominated by two issues. First, does discourse context facilitate the online processing of noncanonical sentences? Second, what is the source of offline misinterpretation errors observed by Ferreira (2003) and others?

Here, we report the results of two experiments that examined whether factors that modulate online processing difficulty also affect comprehenders' final interpretation. Both experiments investigated noncanonical object-before-subject (OS) sentences in German, using self-paced reading to assess online processing difficulty and offline comprehension questions presented without delay to probe the content of the final interpretation. Besides varying word order (SO versus OS), we varied the type of NP serving as object and the question probing offline comprehension.

In Experiment 1, 62 participants read 24 three-sentence texts using a non-cumulative word-by-word moving-window display. The object of the target sentence was either a definite NP (*den Verteidiger*) or a demonstrative NP (*diesen Verteidiger*, see Table 1). In a corpus study (see Bader, 2020), a rate of 76% OS order for demonstrative objects contrasted with a rate of 29% OS order for definite objects. Demonstrative objects were therefore hypothesized to reduce online processing difficulty, as compared to definite objects. The subject and object of the target sentences were both given in the preceding context sentences. *Order* and *Object Type* were within-sentence factors. The additional between-sentence factor *Question Type* varied whether the comprehension question asked for the subject/agent or the object/patient of the preceding clause. As in English, a subject question has SO order whereas an object question has OS order. Figure 1 and 2 show the results. A reading time disadvantage was found on the initial NP for definite objects but not for demonstrative objects. On the sentence-final verb, however, reading slowed down for OS sentences regardless of object type. For offline comprehension, the object manipulation had a marginal effect, whereas the question type manipulation led to a robust effect. Accuracy was high when target sentence and question had both SO order, but was reduced to varying extents in the other conditions.

Experiment 2 tested 32 participants using the same presentation method. All 24 target sentences were either SO or OS sentences with a demonstrative object. Question Type was now a within-sentence factor. The question asked again for either the agent or patient. However, instead of containing two arguments as in Experiment 1 (subject and object), all questions had a single argument (a subject). To this end, all target sentences now contained an optionally transitive verb. Intransitive active questions asked for the agent of the target sentences; passive clauses without a by-phrase asked for the patient. Reading times revealed a similar OS disadvantage as Experiment 1, but, as shown in Figure 3, accuracy was quite high across all conditions in Experiment 2.

In sum, small effects of the object manipulation contrast with large effects of the question type manipulation. Importantly, even a most favorable discourse context together with a preferred referential expression did not prevent misinterpretations when comprehension was probed by a two-argument question. When probed by a one-argument question, in contrast, answer accuracy was generally high, which strongly argues that the target sentences were parsed correctly. Since the crucial difference between Experiment 1 and 2 was whether a two- or a one-argument question was used to probe comprehension, we hypothesize that misinterpretation errors reflect difficulties of extracting retrieval cues for querying the target sentence representation held in working memory.

Table 1: A complete stimulus item for Experiment 1 and Experiment 2. Target sentences with definite objects were only included in Experiment 1.

Context:	[C1]	Schon vor dem Spiel war der Ton ziemlich rau gewesen. 'Already before the game, the atmosphere had been rather charged.'
	[C2]	Der eingewechselte Stürmer hatte nämlich einen Verteidiger des Gegners mehrfach beleidigt. 'The new striker had insulted a defender of the opposing team several times.'
Target:	SO	Der Stürmer hat den/diesen Verteidiger dann auch ziemlich rüde gefault. 'The striker then fouled the/this defender very badly.'
	OS	Den/Diesen Verteidiger hat der Stürmer dann auch ziemlich rüde gefault. 'The/This defender, the striker then fouled very badly.'
Question:	Exp1	Wer hat jemanden gefault? / Wen hat jemand gefault? / Stürmer – Verteidiger 'Who fouled someone? / Who did someone foul? / striker – defender'
	Exp2	Wer hat gefault? / Wer wurde gefault? / Stürmer – Verteidiger 'Who fouled? / Who was fouled? / striker – defender'

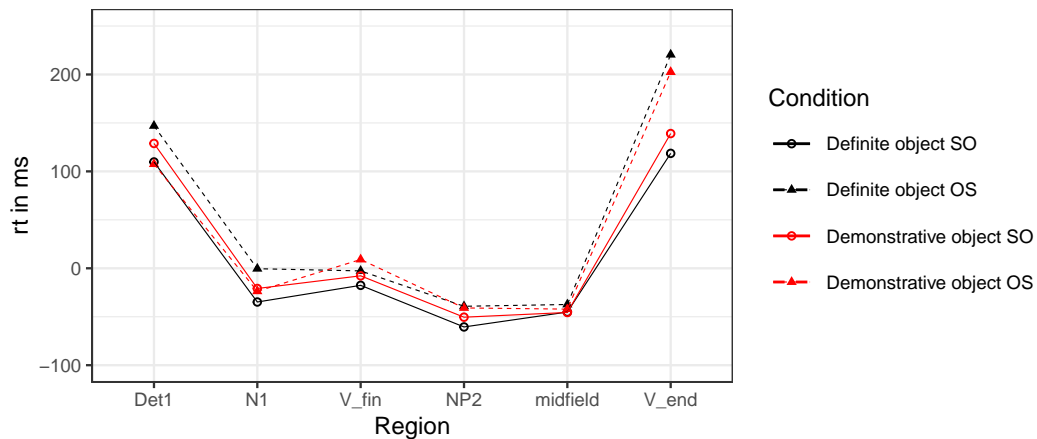


Figure 1: Mean residual reading times in Experiment 1.

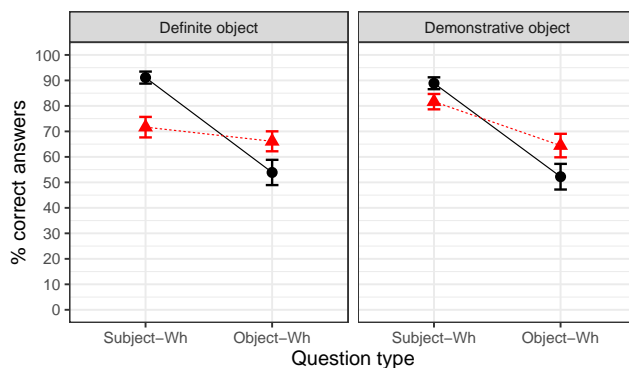


Fig. 2: Answer accuracy in Experiment 1

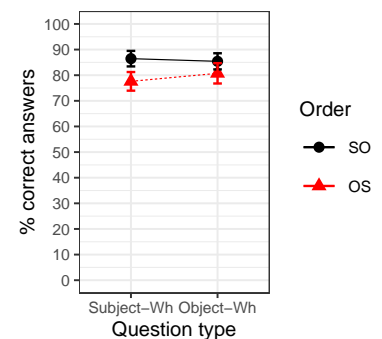


Fig. 3: Answer accuracy in Experiment 2

References

- Bader, M. (2020). Objects in the German prefield: A view from language production. In Woods, R. and Wolfe, S., editors, *Rethinking verb second*, pages 15–39. Oxford University Press, Oxford.
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2):164–203.

Online representations of implausible non-canonical sentences are more than good-enough

Michael G. Cutter (University of Nottingham), Kevin B. Paterson (University of Leicester), Ruth Filik (University of Nottingham).

Given an implausible non-canonical sentence such as ‘*The dog was bitten by the man*’ people often state that the dog did the biting and the man was bitten, despite the opposite being true [1, 2]. This does not occur for the sentence in canonical form (*The man bit the dog*), suggesting that when an algorithmic parse is more complex readers may form ‘good-enough’ representations of sentences based on word order and pragmatic heuristics rather than an algorithmic parse. However, recent work suggests these findings may be attributable to task demands, with the paradigm used in [1,2] causing participants to query their sentence representation in a way that leads to misinterpretation [3,4]. We therefore tested whether we could find evidence for “good-enough” processing under conditions that did not impose explicit demands on comprehension.

We presented readers with two-sentence texts. The first sentence was implausible and shown in canonical (e.g. *It was the peasant that executed the king*) or non-canonical form (e.g. *It was the king that was executed by the peasant*). The following sentence was either 1) Algorithmically Consistent, such that it was plausible given a correct interpretation of the first sentence, but implausible given a good-enough interpretation of the first sentence (e.g. *Afterwards, the peasant rode back to the countryside*; the peasant is dead in a good-enough representation), or 2) Good-Enough Consistent, where the opposite was true (e.g. *Afterwards, the king rode back to his castle*; the king is dead in an algorithmic representation). If a good-enough representation is assigned to non-canonical sentences, we would predict an interaction between first sentence canonicity and follow-up sentence type. Specifically, reading times at the underlined region in the example sentences would be longer for the Algorithmically Consistent follow-up after a non-canonical than canonical sentence, and shorter for the Good-Enough Consistent follow-up after a non-canonical than canonical sentence. Given that prior work suggests older adults depend more on good-enough processing [5] we also compared effects for older vs. young adults.

We presented 44 items, normed for plausibility, with 80 filler items, to 120 participants (60 aged 18-25 years; 60 aged 65+ years) in non-cumulative phrase-by-phrase self-paced reading (see Fig. 1) using Gorilla.sc, a browser-based research platform [6]. We analysed log-transformed reading times for a target region at which implausibility emerged and a post-target region (see Fig. 1). Canonicity, Follow-Up Type, Age Group, and their interactions were set as predictor variables in a Bayesian mixed model. This showed no interaction of Canonicity and Follow-Up Type as a two-way interaction (see Fig. 2), or part of a three-way interaction with Age Group. A Bayes Factor analysis using default Cauchy priors favoured a null Canonicity * Follow-Up Type interaction (Target: $BF_{10}=0.04$; Post-Target: $BF_{10}=0.05$); however, there was strong evidence of longer reading times for Good-Enough Consistent vs. Algorithmically Consistent follow-ups in the Post-Target region ($BF_{10} > 1000$), suggesting that reading of the follow-up sentence was affected by its compatibility with the correct interpretation of the first sentence, with no evidence of misinterpretation. Young adults read faster than older adults ($BF_{10} > 1000$).

The results offer no evidence that participants formed “good-enough” representations of our sentences, rather than performing a full algorithmic parse. We argue that in the absence of specific task demands, participants do not arrive at a semantically incorrect interpretation of non-canonical sentences, consistent with the arguments put forward by [3].

C-AC: It was the peasant | that executed | the king. | Afterwards, | the peasant | rode back to | the countryside.
 NC-AC: It was the king | that was executed by | the peasant. | Afterwards, | the peasant | rode back to | the countryside.
 C-GEC: It was the peasant | that executed | the king. | Afterwards, | the king | rode back to | his castle.
 NC-GEC: It was the king | that was executed by | the peasant. | Afterwards, | the king | rode back to | his castle.

Figure 1. An example of an item in each of our four conditions. “|” symbols represent the gaps between self-paced reading regions. “C-” and “NC-” represent canonical and non-canonical initial sentences, respectively. “-AC” represents a second sentence which is only plausible with an algorithmic parse of the first sentence, while “-GEC” represents a second sentence which is only plausible with a good-enough interpretation of the first sentence. The implausibility of the second sentence was always located in the third region of this sentence (e.g. *rode back to*), with a final sentence wrap-up region following this (e.g. *the countryside/his castle*).

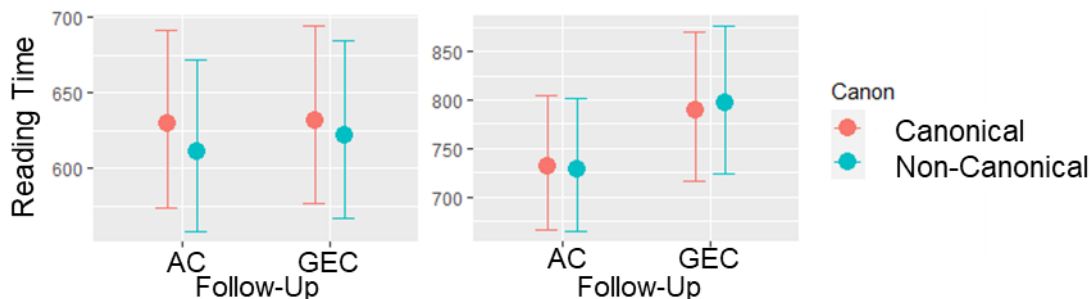


Figure 2. Predicted reading times from our Bayesian mixed models for our target region (left; *rode back to*) and post-target region (right; *the countryside/his castle*). AC represents the Algorithmically Consistent follow-up sentences, and GEC the Good-Enough Consistent follow-up sentences.

References: [1] Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47, 164-203. [https://doi.org/10.1016/S0010-0285\(03\)00005-7](https://doi.org/10.1016/S0010-0285(03)00005-7) [2] Christianson, K., Luke, S. G., & Ferreira, F. (2010). Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 538-544 <https://doi.org/10.1037/a0018027> [3] Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44, 1286-1311. <https://doi.org/10.1037/xlm0000519> [4] Meng, M., & Bader, M. (2021). Does comprehension (sometimes) go wrong for noncanonical sentences? *Quarterly Journal of Experimental Psychology*, 74, 1-28. <https://doi.org/10.1177/1747021820947940> [5] Christianson, K., Williams, C. C., Zacks, R. T., Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*, 42, 205-238. https://doi.org/10.1207/s15326950dp4202_6 [6] Anwyl-Irvine, A. L., Massoné, J., Flitton, A., Kirkham, N. & Evershed, J. K. (2020). Gorilla in our midst: An online behavioural experiment builder. *Behavior Research Methods*, 52, 388-407. <https://doi.org/10.3758/s13428-019-01237-x>

Age invariance in syntactic prediction during self-paced reading

Michael G. Cutter (University of Nottingham), Kevin B. Paterson (University of Leicester), Ruth Filik (University of Nottingham).

A great deal of controversy exists as to whether older adults are more, less, or equally as likely as young adults to make predictions about upcoming linguistic information during reading [1]. Many studies examining linguistic prediction in ageing have focussed upon lexical prediction of specific target words in sentences. In the current study we examined whether older adults use reliable linguistic cues to make syntactic– rather than lexical– predictions to a similar extent to young adults.

We presented readers with sentences in which an upcoming noun-phrase coordination structure was made predictable or left unpredictable through the presence or absence of the word *either* (e.g. *Josh will order either a large pizza or tasty calzone at the restaurant*). Prior work shows faster reading at *or* + the second noun phrase (e.g. *or tasty calzone*) when *either* is present earlier in the sentence in eye movements by young adults [2] and self-paced reading by older adults [3]. However, whether this effect is equivalent in the two age groups is unclear. Furthermore, [3] found a cost of the presence of *either* in a pre-target region (e.g. *a large pizza*) using self-paced reading with older adults, while [2] found no such cost for young adults during eye-tracking. As such, a secondary interest in the current study was to determine if this effect in older (and not young) adults was a form of prediction ‘cost’ due to cognitive ageing, or whether a similar effect is present in young adults in self-paced reading.

Sixty young adults (18-25 years) and 60 older adults (65+ years) read 32 sentences, half with *either* and half without *either*, in non-cumulative phrase-by-phrase self-paced reading. These items were presented alongside 88 filler items. This task was administered online using Gorilla.sc, a browser-based platform for remote data collection [4]. Sentences were presented in four regions (see Fig. 1). We examined effects in both target and pre- target regions. We analysed log-transformed reading times using Bayesian mixed models with Age Group and the presence of *either* as predictor variables, and a two-way interaction between these variables (see Fig. 2 for conditional means). At the target region, older adults read more slowly ($b = 0.40$, $\text{CrI}[0.27, 0.52]$, $p(b > 0 = 1)$), and there was a facilitative main effect of the presence of *either* ($b = 0.06$, $\text{CrI}[0.04, 0.09]$, $p(b > 0 = 1)$), but no interaction between these factors ($b = 0.00$, $\text{CrI}[-0.05, 0.05]$, $p(b > 0 = 0.51)$). To further determine whether there were age differences in our effects we calculated Bayes Factors comparing a model including an interaction between age group and the presence of *either* with a model in which only main effects were present. The Bayes factor favoured the non-interactive model ($\text{BF}_{10} = 0.068$), suggesting that syntactic prediction is age invariant. In the pre-target region, older adults read more slowly ($b = 0.32$, $\text{CrI}[0.21, 0.44]$, $p(b > 0 = 1)$), and there was a cost of the presence of *either* ($b = -0.06$, $\text{CrI}[-0.08, -0.03]$, $p(b > 0 = 0)$) but no interaction ($b = 0.01$, $\text{CrI}[-0.04, 0.06]$, $p(b > 0 = 0.61)$; $\text{BF}_{10} = 0.035$), which might suggest a cost of making a prediction in the pre-target region, for both age groups.

We conclude that there are no differences between younger and older adults in the use of *either* to make syntactic predictions during self-paced reading. This was true for both the benefit of having made the prediction upon reading the target region, and any earlier cost associated with the presence of *either*. We suggest that efforts should be made to further investigate syntactic prediction in ageing, to determine whether a clearer pattern of results emerges across paradigms than has typically been the case for lexical prediction.

Predictable: Josh will order either | a large pizza | or tasty calzone | at the restaurant.
 Unpredictable: Josh will order | a large pizza | or tasty calzone | at the restaurant.

Figure 1. An example of an item in each condition, with “|” symbols representing the demarcation of regions in the self-paced reading study. The target region always consisted of the word *or* and the following noun phrase, while the pre-target region consisted of the first noun phrase of the co-ordination structure.

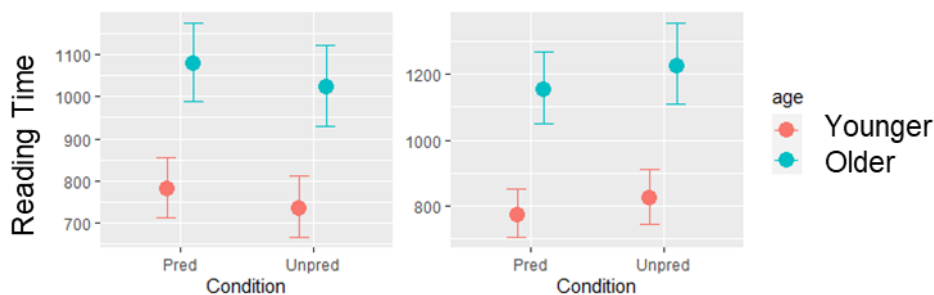


Figure 2. Predicted reading times from our Bayesian mixed models for our pre-target region (left; *a large pizza*) and target region (right; *or tasty calzone*). *Pred* represents the sentences in which *either* appeared as a predictive cue, while *Unpred* represents sentences in which this cue was absent.

References: [1] Payne, B. R., & Silcox, J. W. (2019). Aging, context processing, and comprehension. *Psychology of Learning and Motivation*, 71, 215-264. <https://doi.org/10.1016/bs.plm.2019.07.001> [2] Staub, A., & Clifton, C., Jr. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 425–436. <https://doi.org/10.1037/0278-7393.32.2.425> [3] Warren, T., Dickey, M. W., & Lei, C. M. (2016). Structural prediction in aphasia: Evidence from either. *Journal of Neurolinguistics*, 39, 38-48. <https://doi.org/10.1016/j.jneuroling.2016.01.001> [4] Anwyl-Irvine, A. L., Massoné, J., Flitton, A., Kirkham, N. & Evershed, J. K. (2020). Gorilla in our midst: An online behavioural experiment builder. *Behavior Research Methods*, 52, 388-407. <https://doi.org/10.3758/s13428-019-01237-x>

Agreement attraction in grammatical sentences arises only in the good-enough processing mode

Anna Laurinavichyute, Titus von der Malsburg (University of Potsdam)

In comprehension, agreement attraction errors are known to facilitate the processing of ungrammatical sentences, such as *The key to the cabinets are rusty*^[1]. There is only scarce evidence suggesting that agreement attraction can also increase processing difficulty in grammatical sentences. The Marking & Morphing account [1] predicts a slowdown at the verb in sentences such as *The key to the cabinets is rusty*, due to erroneous representation of the subject number (the ungrammaticality illusion). The majority of studies haven't found any evidence for this effect, and most of those that did had design confounds. However, recently evidence in favor of the predicted effect has begun to accumulate: [2] reported the expected effect in grammaticality judgments, and [3] found an illusion of ungrammaticality in reading times in three self-paced reading experiments. It seems that the illusion is subject to some unknown constraints, and there is no explanation of why [3] detected the illusion absent in other studies. We hypothesized that the crucial factor might be good-enough processing: [3] presented participants with a single experimental sentence preceded by three simple training sentences without comprehension questions. We suggest that the training phase might have encouraged superficial processing of the experimental sentence. The superficial processing, in turn, may have allowed the illusion of ungrammaticality to appear. We test whether increasing the depth of processing would make the illusion of ungrammaticality disappear.

Methods. Participants were presented with the materials from Experiment 3 by [3], which were not changed in any way (see 1). Instead, we manipulated the training sentences that preceded the experimental sentence to induce deeper processing: we used three new, more complex training sentences, each of them accompanied by a difficult comprehension question (see 2). The original experiment had data from 3,559 participants. We aimed to collect at least as much data as in the original experiment and acquired data from 3,702 individuals. For the analysis, we used Bayesian LMMs.

Results. No main effects or interactions were detected at the verb or on the region following the verb. We pooled the data from the original Experiment 3 and the new experiment to test for an interaction between the illusion of ungrammaticality and the depth of processing. An interaction was found at the critical verb *n*, words *n*+1 and *n*+2. At the verb and word *n*+1, the interaction was driven by the the illusion of ungrammaticality in the superficial processing condition (the verb: 59ms, 95%-CrI:[15, 103]ms; the following region: 34ms, 95%-CrI:[9, 59]ms). At word *n*+2, nested comparisons showed the opposite effect: a slowdown in conditions with a number-matching interfering noun in the deep processing condition (-25ms, 95%-CrI:[-49, -2] ms), predicted by the cue-based retrieval accounts.

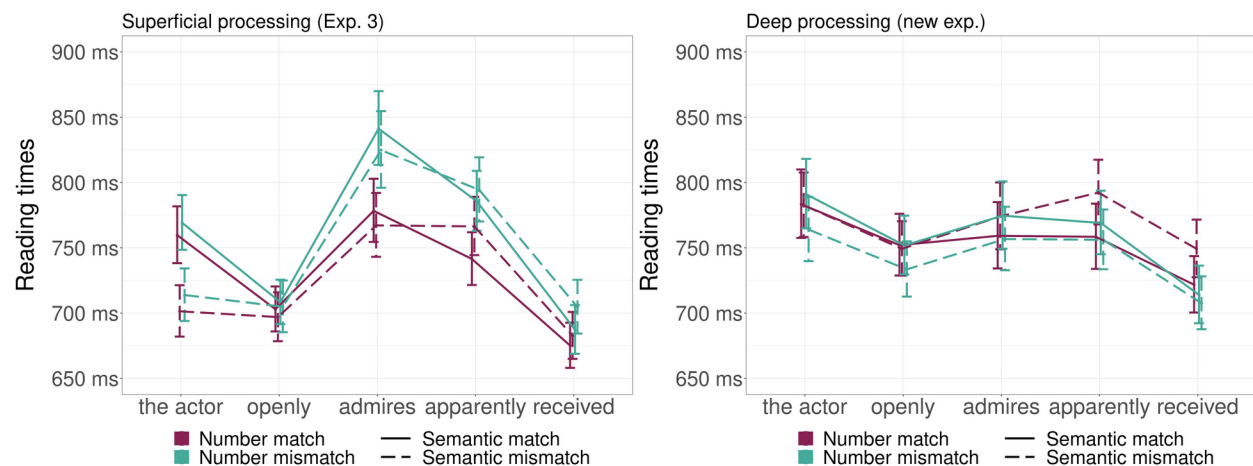
Discussion. Our results demonstrate that the illusion of ungrammaticality can be switched off when participants engage in deep processing. This finding sheds light on why the illusion was so rarely observed in previous studies and consistently found in Expts. 1 through 3 by [3]: superficial processing mode is difficult to achieve when using a repeated measures design, where experimental sentences are followed by comprehension questions. From the theoretical perspective, our findings are difficult to reconcile with the Marking & Morphing account: although it predicts the illusion, the postulated cause is not the misidentification of the subject noun or falsely assembled syntactic structure. Therefore, deeper processing and potentially more accurate memory encoding should not influence the rate of agreement attraction according to Marking & Morphing. Our findings are more compatible with a simple heuristic tracking the instances of plural features, a heuristic that might be initiated when deep parsing is not the main priority. On a broader level, our findings add to the surprisingly sparse causal evidence supporting the existence of different processing modes (the only demonstration so far being the case of global ambiguity resolution [4,5]).

Example experimental item:

- (1) a. The singer that the actor openly **admires** apparently ...
 b. The singers that the actor openly **admires** apparently ...
 c. The play that the actor openly **admires** apparently ...
 d. The plays that the actor openly **admires** apparently ...
 ...received some harsh criticism.

New practice sentences (response options were presented in random order):

- (2) 1. The priest who had privately advised the lawyer of the art dealer, is accused of withholding information.
 Who was accused? — The priest/The lawyer/The art dealer/The art dealers/I'm not sure.
2. The personal assistant who the bodyguard of the delegate does not trust attracts great public attention.
 Who attracted public attention? — The personal assistant/The bodyguard/The delegate/The bodyguards/I'm not sure.
3. The philanthropist who had greeted the secretary of the director, later participated in the fundraising committee.
 Who took part in the committee? — The philanthropist/The secretary/The director/The secretaries/I'm not sure.



Geometric mean reading times across conditions. Number match and Number mismatch refers to the match/mismatch between the interfering noun and the verb (since all experimental stimuli are grammatical, the subject noun always fully matches the verb).

- [1] Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychological review*, 112(3), 531.
- [2] Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology*, 110, 70-104.
- [3] Laurinavichyute, A., & von-der-Malsburg, T. (2019). Agreement attraction effects in the comprehension of grammatical sentences. Poster presented at CUNY, Boulder.
- [4] Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36(1), 201-216.
- [5] Logačev, P., & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, 40(2), 266-298.

The Distributional Learning of Recursive Structures

Daoxin Li (University of Pennsylvania), Lydia Grohe (Goethe University Frankfurt), Petra Schulz (Goethe University Frankfurt), Charles Yang (University of Pennsylvania)

Problem Although the ability for recursive embedding may be universally available, languages differ regarding depth, structure, and syntactic domains [1]. As the Appendix illustrates, English allows infinite stacking of the prenominal genitive -s (1a), but in German, this option is restricted to only one level, and to a narrow set of items (1b-c) [2]. For post-nominal PP *of*-genitives, *von* ‘of’ can embed infinitely in German (2a) while *of* in English is more limited (2b-c). In Chinese, genitives can stack freely with the possessive marker *de* (3a) but are restricted to one level when the marker is omitted (3b-c). What learning mechanism enables children’s early acquisition of these recursive structures [3]?

Proposal We propose that productivity is a prerequisite for recursion. In the more familiar case of English determiners [4], productivity is defined as the interchangeability of *a* and *the* in combination with nouns. For genitive structures, we take productivity as the interchangeability of structural position. For a structure such as **X’s-Y** or **Y-of-X** to be recursive, the child needs first to discover the interchangeability of the **X** and **Y** positions: that the possessum can productively appear in the possessor position. This view of recursion enables us to apply distributional learning models such as the Tolerance/Sufficiency Principle [TSP; 5]: a rule defined over **N** lexical items productively generalizes iff $e \leq N / \ln N$ where **e** is the cardinality of the subset not attested under the rule. Under the TSP, **N** pertains to the child learner’s modest, and likely high-frequency, vocabulary [6-8]. The recursion of a genitive structure (**X’s-Y** or **Y-of-X**) is licensed if a sufficiently large proportion—à la the TSP—of nouns attested in the **Y** position in the input is also attested in the **X** position in the input.

Method Our analyses combined automatic search with manual inspection and were comparable for three languages (Table 1); the English results are reported in detail. We targeted a 5.5-million-word input corpus and focused on the nouns established to be representative of 3-year-old children [9]. For the **X’s-Y** sequences in the input, 59 head nouns appeared in the **Y** position. 46 also appeared in the **X** position, clearing the TSP threshold (45; $59 / \ln 59 = 14$): **X’s-Y** is thus productive. For the **Y-of-X** sequences in the input, 43 head nouns appeared in the **Y** position but only 28 also appeared in the **X** position, falling below the TSP threshold (32, $43 / \ln 43 = 11$). Thus **Y-of-X** does not productively generalize, while subregularities within the attested nouns in the **Y** position may be derived by further applications of the TSP [5].

Conclusion Productivity, as a necessary condition for recursion, can be acquired from level-1 input data for specific syntactic domains, given that the child can recognize the relevant syntactic (e.g., noun) and semantic categories (e.g., possessor/possessum). Explicit evidence for deep embedding [10] is not necessary.

Appendix

English allows free embedding with *-s*, but not with *of*:

- (1) a. the neighbor's lawyer's briefcase's price
- b. the price *of* the briefcase
- c. ?the price *of* the briefcase *of* the lawyer
- d. ?*the price *of* the briefcase *of* the lawyer *of* the neighbor

German allows free embedding with *von* ('of'), but not with *-s*:

- (2) a. das Buch *von* dem Nachbarn *von* dem Mann ('the book of the neighbor of the man')
- b. Vaters Buch ('father's book'), *Manns Buch ('man's book')
- c. *das Manns Nachbars Buch ('the man's neighbor's book')

Chinese allows recursive genitive with *de*, but one level without [11]:

- (3) a. nage ren *de* linju *de* shu ('that man's neighbor's book')
- b. nage ren linju ('that man's neighbor')
- c. *nage ren linju shu ('that man's neighbor's book')

Table 1. Distributional analysis of recursive and non-recursive possessive structures with the Tolerance/Sufficiency Principle

Language	Chinese*		English		German*	
Structure	X de Y	X Y	X's Y	Y of X	X's Y	Y von X
N in Y	41	27	59	43	34	40
N in X & Y	35	15	46	28	5	34
TSP Threshold	30	19	45	32	24	29
Productive?	Yes	No	Yes	No	No	Yes

*The Chinese and German input corpora contain 1.7 million words and 3.5 million words, respectively. The Chinese analysis made use of the vocabulary previously established to be representative of three year olds [8]. No such vocabulary list is available for German, so we used the set of the most frequent nouns of comparable cardinality, 50 in this case, found in the input.

References

- [1] Pérez-Leroux et al. (2018). *Language*, 94, 332-359. [2] Weiss (2008). In *Microvariation and syntactic doubling*. Emerald. [3] Giblin et al. (2019). *Proc. BUCLD 43*, 270-286. [4] Yang (2013). *PNAS*, 110(16), 6324-6327. [5] Yang (2016). *The price of linguistic productivity*. MIT Press. [6] Hart & Risley (1995). *Meaningful differences in the everyday experience of young American children*. Brookes. [7] Szagun et al. (2006). *First Language*, 26(3), 259–280. [8] Hao et al. (2008). *Behavior Research Methods*, 40(3), 728- 733. [9] Carlson et al. (2014). *Journal of Memory and Language*, 75, 159-180. [10] Roeper (2011). *Biolinguistics*, 5, 57-86. [11] Li & Thompson (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

The effects of input typicality (or variability) on the acquisition of argument structure constructions

Eunkyung Yi (Ewha Womans University) and Jia Kang (University of Hawaii at Manoa)

Partial productivity of argument structure constructions poses a major challenge to language learners (Pinker 1989). It is acknowledged that learners can generalize over individual sentences like *John pulled the drawer open* and add an abstract form-and-meaning representation or *construction* to their mental grammar such as $[NP_x \text{ verb } NP_y \text{ RP}_z]$ meaning 'X makes Y become Z (by verb-ing)' (Goldberg 1995). Then, they can use it productively with other verbs and result phrases such as *Tom pushed the door shut*. At the same time, however, they are expected not to produce odd sentences such as **John scolded Tom unhappy* to mean 'John makes Tom unhappy by scolding him.' It is not clear yet how learners eventually learn and use an abstract construction while avoiding such errors. The present study investigates the nature of language input that can facilitate the generalization and production of an argument structure construction called the resultative construction by varying the typicality (or variability) of verbs and result-phrases, respectively.

We conducted two experiments on two groups of subjects (L1 Korean), i.e. advanced and high-intermediate English learners based on test scores such as TOEFL. The experiments consisted of a reading phase and a test phase while no explicit teaching was involved. In Experiment 1, we tested one of the most influential proposals in the acquisition of argument structure constructions that the acquisition is driven by the most typical and frequent verb of the construction, e.g. *give* for the ditransitive construction and *make* for the resultative construction (Boas 2011). Two sets of nine stimuli were prepared to test the proposal (Table 1). One set contains nine resultative sentences with three different verbs (*pull*, *rub* and *kick*) paired with three different result phrases, respectively; the other set used the same verb *make* across the nine stimuli. Each resultative sentence is preceded and followed by a context sentence to help readers capture the meaning of the resultative sentences in a natural way (an example in (1)). In the reading phase, participants read the stimuli, each followed by a comprehension question; in the test phase, they were presented with a short video clip (snapshots of an example video in (2)) and asked to describe the event occurring in the video most preferably in a single sentence. We annotated the production data as to whether they used the resultative construction in describing the event or not.

In Experiment 2, we further investigated the role of variability (or typicality) in verbs and result phrases by manipulating the number of verbs and result phrases in the input. We tested whether the variability (or typicality) of the verbs or that of result phrases is more effective in facilitating the acquisition of the resultative construction. We prepared a third set of nine stimuli where one result phrase is paired with three different verbs and compared it with the first set in Experiment 1 (Table 1). Namely, in the reading phase, participants were exposed to 3 verbs x 9 RPs in the verb-centered condition and to 9 verbs x 3 RPs in the result-centered condition. All other settings were kept constant across the two experiments.

The results showed that subjects produced significantly fewer resultative sentences in the *make*-only exposure condition than in the resultative exposure condition ($b=-1.355$, $p<.05$) in Experiment 1 and they also produced fewer resultatives in the result-centered condition than in the verb-centered condition ($b=-1.27$, $p<.05$) in Experiment 2. In both experiments, advanced learners produced more resultative sentences than high-intermediate learners. Our results disconfirm the previous contention that the *make*-construction plays the key role in the acquisition of the resultative construction (Exp 1) and support that subjects tend to make verb-centered generalization in learning resultatives (Exp 2). Our study provides empirical evidence on the effect of typicality/variability on the acquisition of the resultative construction and also suggests that exposure to a small set of different verbs with some repetition is crucial in facilitating the argument structure acquisition.

Example stimuli:

- (1) Learning phase (reading & comprehension): a stimulus consisting of three sentences
*The detective suspected the woman concealed the jewelry in the drawer. **He pulled the drawer open** to see what was inside. He found the diamond watch in there.*
- (2) Test phase (production): snapshots of a video stimulus (about 15 sec.) with captions



Table 1. Verb & result-phrase pairings in stimuli for each condition

Condition 1 in both Experiments 1 & 2		Condition 2 in Experiment 1		Condition 2 in Experiment 2	
Verb	Result phrase	Verb	Result phrase	Verb	Result phrase
pull	open	make	open	pull	open
	shut		shut	push	
	loose		loose	break	
rub	clean		clean	rub	clean
	dry		dry	wipe	
	smooth		smooth	sweep	
kick	dead		dead	kick	dead
	high		high	knock	
	unconscious		unconscious	shot	

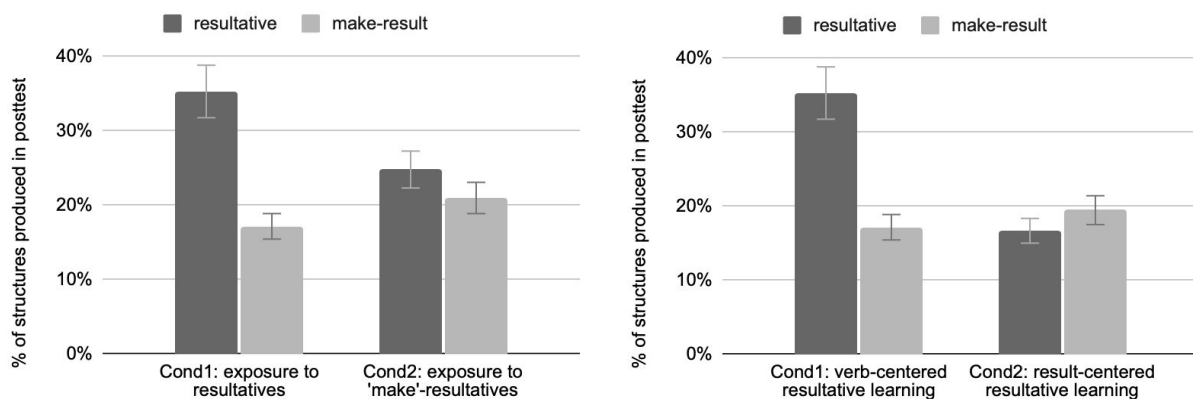


Figure 1. Percentages of resultatives in Experiments 1 & 2 (advanced+high-intermediate)

References

- Boas, Hans C. (2011). *Linguistics* 49, 1271-1303.
- Goldberg, Adele E. (1995). *Constructions: a construction grammar approach to argument structure*. UCP.
- Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. MIT Press.

Predictive effects of number-marked verbs and copulas in Czech 2-year-olds

Filip Smolík, Veronika Bláhová (Institute of Psychology, Czech Academy of Sciences, Prague)

There are conflicting findings regarding the early comprehension of grammatical number marking in verbs. While some studies found limited comprehension (e. g. Johnson, de Villiers, & Seymour, 2005; Smolík & Bláhová, 2017), other others found that children comprehend verb number marking (Brandt-Kobeles & Höhle, 2010), and even demonstrated that 2,5-year-olds can use the number of English copula to predict the upcoming nouns (Lukyanenko & Fisher, 2016). Predictive processing of sentences based on grammatical morphemes has also been shown for grammatical gender, but the findings could not distinguish between the facilitation of lexical search and active anticipation of upcoming words.

Two preferential studies reported here examined the predictive effects of number marking in the comprehension of Czech copulas and lexical verbs. In both studies, children saw pairs of pictures, one showing a single instance of an object, the other showing a group of two to four instances of another object. While watching the pair, children heard a phrase referring to one of the pictures. The phrase ended with a noun labeling the picture, which was preceded by a short introduction. In the informative condition, the introduction contained a copula (E1) or a lexical verb (E2) agreeing in number with the sentence-final noun.

Experiment 1

Podívej, tady je/jsou na obrázku kniha/míče.

Look there is/are in the picture book/balls.

Experiment 2

Podívej, tady skáče/skáčou na obrázku kůň/žáby

Look there jumps/jump in the picture horse/frogs.

In the uninformative condition, the phrase only contained non-agreeing attention-getters. Each experiment included 16 experimental trials and 16 uninformative control trials. Four yoked pairs of pictures were used, each 4 times in each condition.

A total of 40 27-month-olds participated in Experiment 1, and 40 30-month-olds in Experiment 2. Children's faces were recorded and their gaze direction coded, focusing on the effect of the number-marked copula/verb on the proportion of looks towards the target picture. Differences were tested using random permutation analysis. In addition to the experiment, receptive grammar and vocabulary tasks were given to children.

In Experiment 1, children looked towards the target picture as early as 1000 ms after the copula onset, demonstrating predictive looks towards target due to copula. In Experiment 2, similar effect of main verbs was found, but only in children who were above the median according to the grammar and vocabulary tasks. Overall, the results confirm that Czech 2-year-olds understand number marking of verbs and its meaning, and can use it for predictive processing. Contrary to existing research, the findings were made in a language with often ambiguous morphological marking and flexible word order, confirming that predictive processing is not limited to fixed sequences of units. Predictive effects of number were observed for lexical verbs, suggesting that the comprehension of number is not lexically specific.

Figure 1. Proportions of looks towards target for all participants in Exp. 1. The vertical shaded bar shows the range of onsets of the target nouns.

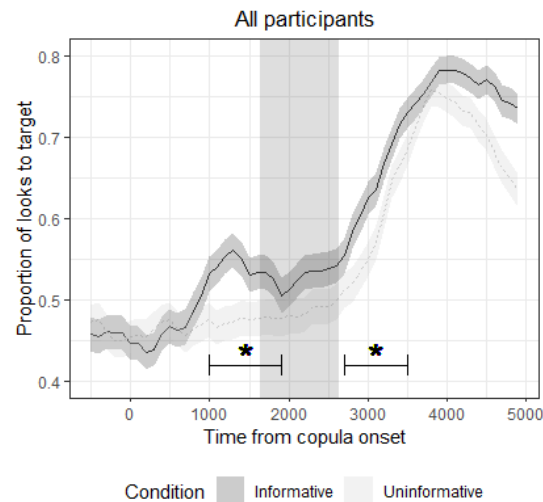
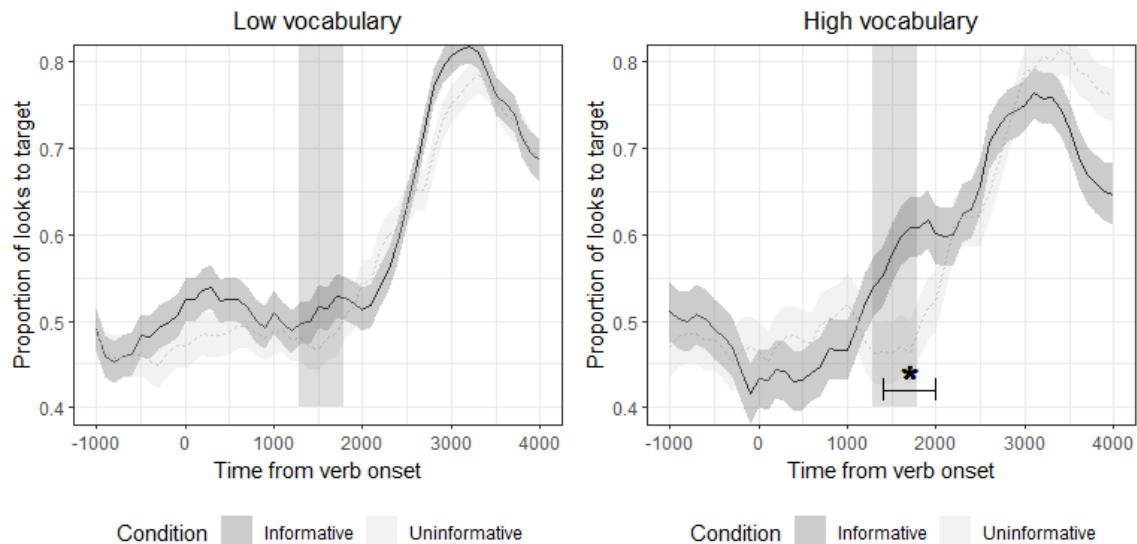


Figure 2. Proportions of looks towards target in Experiment 2, separately for children with vocabulary below/at the median and above median in Experiment 2. The vertical shaded bar shows the range of onsets of the target nouns.



References

- Brandt-Kobe, O.-C., & Höhle, B. (2010). What asymmetries within comprehension reveal about asymmetries between comprehension and production: The case of verb inflection in language acquisition. *Lingua*, 120, 1910–1925.
- Johnson, V. A., de Villiers, J. G., & Seymour, H. N. (2005). Agreement without understanding? The case of third person singular /s/. *First Language*, 25, 317–330.
- Lukyanenko, C., & Fisher, C. (2016). Where are the cookies? Two-and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition*, 146, 349–370.
- Smolík, F., & Bláhová, V. (2017). Comprehension of verb number morphemes in Czech children: Singular and plural show different relations to age and vocabulary. *First Language*, 37(1), 42–57.

The acceptability of null subjects

Juliana Gerard (Ulster University)

Languages vary in their subject requirements: some languages permit the subject to be dropped in declarative clauses, as in (1), while others require an overt subject.

- (1) a. \emptyset plays with blocks.
b. \emptyset plays with exciting new blocks.

Two-year-olds often produce sentences without a subject, even in overt subject languages. These *null subject* sentences may be due to a non-adult grammar which *permits* null subjects¹⁻⁴, or to a processing bottleneck, which *causes* the subject to be dropped despite the adult grammar⁵⁻¹².

If null subjects are due to a non-adult grammar, then sentences with null subjects will be grammatical before the adult grammar is acquired, and only considered ungrammatical after the grammar changes. But **if null subjects are due to a processing bottleneck, then they are more likely to be accepted in contexts with a high processing load - at any age.**

We test this processing prediction with 85 adults in a speeded acceptability judgment task (1-7 rating scale). Participants saw sentences with a *null* (1) or *overt* (2) subject, and with an *inflected* (1;2) or *bare* (3) verb (within-subjects).

- (2) a. The child plays with blocks.
b. The child plays with exciting new blocks.
(3) a. {The children/ \emptyset } play with blocks.
b. {The children/ \emptyset } play with exciting new blocks.

We manipulated the availability of processing resources in two ways:

VP length: VP length is varied from 3-5 words (within-subjects). Since null subjects are produced more often with longer VPs⁵⁻⁹, as in (1b and 3b), **null subject sentences should be more acceptable with a longer (1b) than shorter (1a) VP.**

Timing: sentence presentation time is varied from 1200ms (N=30), 2000ms (N=25), or no limit (N=30). If null subjects are due to limited processing resources⁸⁻¹⁰, then **greater acceptability for null subjects is predicted under stricter time limits.**

Results are presented in Fig.1, with z-scored ratings. A mixed effects model (Table 1) revealed a significant three-way interaction between subject form, verb form, and VP length:

- as predicted, null subject sentences are less acceptable than with an overt subject
- **the difference between null and overt subjects is greater with an inflected form (*plays*; bottom 3 figures) than with a bare form (*play*; top 3 figures)**
- **within the bare forms, null subjects are more acceptable with short than long VPs**, particularly in the timed conditions (A vs B) - an unexpected finding on both accounts

In addition, overt subject sentences are more acceptable with inflection (bottom white bars in Fig.1) than without inflection (top white bars), a further unexpected finding given that overt subjects are grammatical in general.

While a grammatical account predicts no effect of VP length for comprehension, the processing account predicts the *reverse* of the observed effect: greater acceptability for a longer VP. However, the bare forms *are* grammatical if interpreted as an imperative rather than as a declarative⁴. An imperative is possible regardless of VP length, but the null version of the short VP in (3a) is more likely as an imperative than the null version of the long VP in (3b). This explains the greater acceptability for null subjects with a bare verb than with an inflected verb.

The imperative form thus interferes with judgments under a processing load – i.e. the timed conditions. If children's null subjects involve similar interference from imperatives in English, then individual differences in null subjects may be predicted by imperatives in the linguistic input. Cross-linguistic differences are also predicted based on verb form, for acquisition and processing.

Fig.1. z-scored ratings: null subjects are more acceptable with a bare verb (top row), and more acceptable with a short VP (**A**) than a long VP (**B**) in the timed conditions (1200ms & 2000ms)

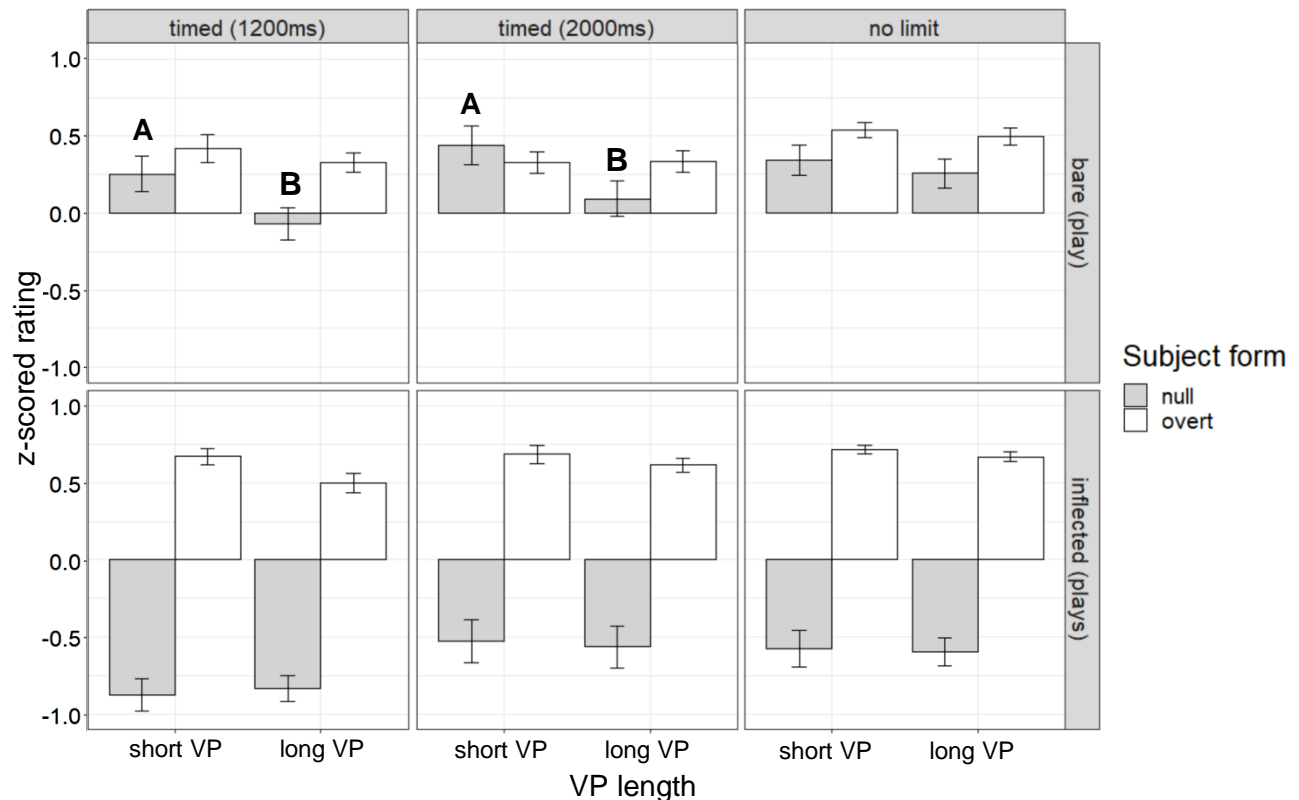


Table 1. Model with fixed effects subject form, verb form, VP length (within-subjects), and timing (between-subjects), and random effects subject and item; coding is effects coding

Fixed effects	Estimate	SE	t-value	p-value
Intercept	-0.66	0.04	-14.72	<.001
subject form (null/overt)	1.26	0.05	26.74	<.001
verb form (bare/inflected)	0.75	0.05	13.84	<.001
timing (no limit - 1200ms)	-0.40	0.11	-3.51	<.001
timing (no limit - 2000ms)	0.24	0.12	2.07	0.04
subject form (null/overt) : verb form (bare/inflected)	-0.96	0.07	-14.36	<.001
subject form (null/overt) : VP length (modifier/no modifier)	0.11	0.07	1.70	0.09
verb form (bare/inflected) : VP length (modifier/no modifier)	0.29	0.08	3.74	<.001
subject form (null/overt) : timing (no limit - 1200ms)	0.22	0.13	1.69	0.09
subject form (null/overt) : timing (no limit - 2000ms)	-0.21	0.14	-1.55	0.12
subject form (null/overt) : verb form (bare/inflected) : VP length (modifier/no modifier)	-0.31	0.09	-3.32	<.001

References

¹Hyams 1992 in *Theoretical Issues in Language Acquisition*, ²Hyams & Wexler 1993 *LI*, ³Hyams 2011 in *Handbook of Generative Approaches to Language Acquisition*, ⁴Orfitelli & Hyams 2012 *LI*, ⁵Bloom 1989 in *Papers and Reports on Child Language Development*, ⁶Bloom 1990 *LI*, ⁷Bloom 1993 *LI*, ⁸Valian 1991 *Cognition*, ⁹Valian, Hoeffner, & Aubry 1996 *Developmental*

Psychology, ¹⁰Valian & Aubry 2005 *JCL*, ¹¹Chen, Valian, & Chodorow 2016 *BUCLD41*, ¹²Valian 2016 in *Oxford Handbook of Developmental Linguistics*.

Second language acquisition and language processing: Grammatical gender in Norwegian

The present study investigates a recent proposal that the effects of L1 in L2 gender production and predictive processing are fine-grained and that the degree of the overlap between the gender systems in the L1 and L2 (rather than the presence vs. absence of gender in the L1) determines the extent to which grammatical gender production and predictive processing in the L2 is nativelike (Hopp & Lemmerth, 2018; Dussias, Valdés Kroff, Guzzardo Tamargo & Gerfen, 2013). To address a granular perspective on the effects of lexical and structural similarities and differences between gender systems in SLA, we extend the scope of research to previously unstudied language pairs L1 Greek/L2 Norwegian and L1 Russian/L2 Norwegian, which exhibit a varying degree of overlap in gender properties. Although Norwegian, Greek and Russian categorize nouns into one of the three gender classes (masculine, feminine or neuter), they differ in lexical congruency, i.e. whether individual nouns are assigned the same (e.g., Russian: *jabloko*(N) 'apple'; Norwegian: *eple*(N) 'apple') or different gender (e.g., Russian: *dom*(M) 'house'; Norwegian: *hus*(N) 'house'). At the syntactic level, there is an overlap between Norwegian and Greek, which both mark gender on indefinite articles, while Russian does not. Speakers of L1 Turkish, a genderless language, are also included for comparison.

The study includes two experimental tasks. Experiment 1 was the noun-naming task which elicited indefinite noun phrases in Norwegian. Experiment 2 was an eye-tracking Visual World Paradigm experiment with a two-picture design. The auditory stimuli were phrases like *Jeg tenker på en/et avbildet NOUN* 'I am thinking of a(M/N) depicted NOUN'. The participants were 66 late L2 learners of Norwegian: L1 Greek ($n=23$, age 27-64), L1 Russian ($n=23$, age 28-64), and L1 Turkish ($n=20$, age 32-65). We also included a control group of L1 Norwegian speakers ($n=19$, age 25-55) in Experiment 2. The production and eye-tracking experiments had the same stimuli, which were 64 depicted nouns/objects: congruent neuter (16), incongruent neuter (16), congruent masculine (16) and incongruent masculine (16). Feminine gender was not tested, because it is disappearing from the dialects of Norwegian examined in the study. The materials were identical for the Greek and Turkish groups. However, it was impossible to match the nouns for gender and lexical congruency across all languages, therefore 20 out of 64 nouns were different in the experiments with L1 Russian speakers.

In Experiment 1, gender assignment was near target-like with the masculines and at approximately 65% accuracy rate with the neuters across all groups (Table 1). Thus, all participant groups, including the speakers of genderless Turkish, performed equally well. To check for language proficiency effects, the L2 participants were divided into an advanced-proficiency and intermediate-proficiency groups based on their gender assignment scores in Experiment 1 which also matched their general proficiency in Norwegian. Experiment 2 revealed a striking asymmetry between L1 Greek and L1 Russian vs. L1 Turkish (Figure 1). L1 Greek and L1 Russian showed nativelike gender processing, yet, only at advanced proficiency levels. L1 Turkish failed to use gender predictively even at advanced proficiency levels. This difference was significant and robust. Thus, while all L2 learners showed similar knowledge of gender in production, only those who have gender in their L1 could access and use this knowledge during online gender processing in the L2. We also found no difference between the two gendered languages, i.e. no effect of syntactic similarity or lexical congruency. These results suggest that predictive gender processing in the L2 is determined by the presence vs. absence of gender in the L1, rather than by the degree of the overlap between the gender systems in the L1 and L2 (cf. Hopp & Lemmerth, 2018; Dussias et al., 2013). Furthermore, this effect was moderated by learner proficiency, but not by lexical congruency, as no congruency effects emerged in any of the groups. This result may be due to the fact that the overlap between the Norwegian and Greek or Russian gender systems is not sufficient to modulate predictive gender processing at intermediate proficiency levels and perhaps for lexical congruency to have an effect.

Table 1. Results of the Experiment 1: Noun naming and gender assignment.

	Greek	Russian	Turkish
A. Nouns named	83%	83%	92%
B. Gender accuracy, all named nouns	63%	62%	71%
C. Gender accuracy, all correctly named nouns	74%	74%	77%
D. Gender accuracy, all correctly named nouns, Masculine vs. Neuter	M: 87% N: 63%	M: 82% N: 66%	M: 85% N: 70%

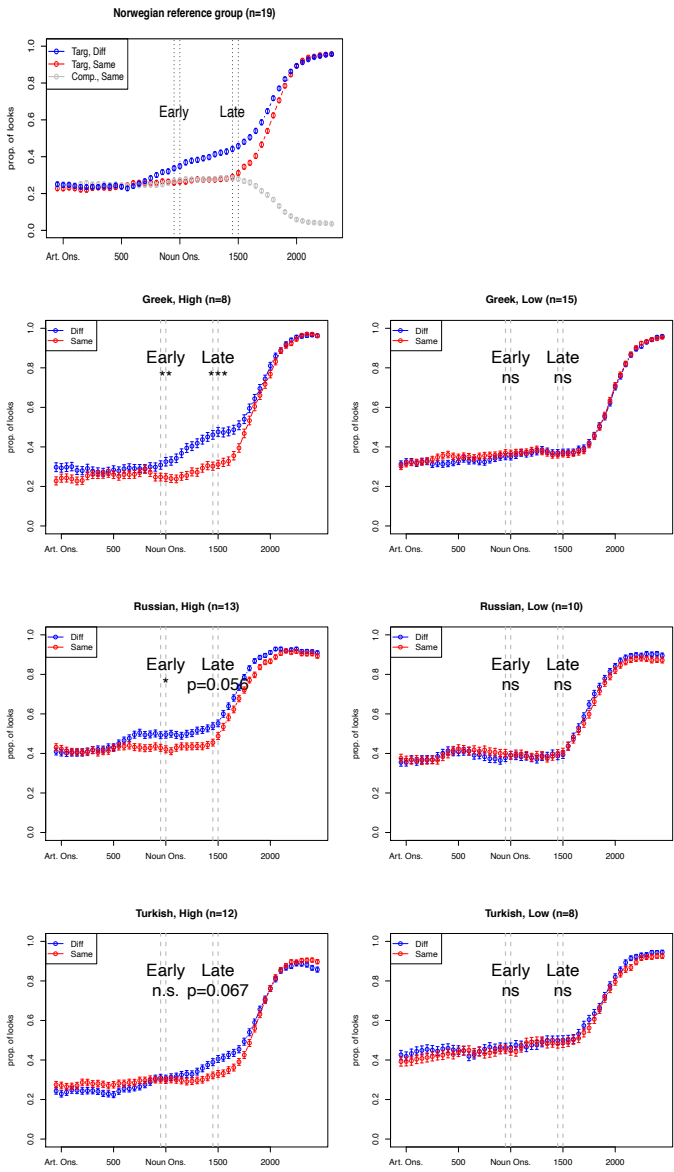


Figure 1. The eye-tracking gender prediction results of the Norwegian control group and the three L2 groups (proportion of looks to target per 50ms time slot)

References

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition* 35, 353-387. Hopp, H., & Lemmerth, N. (2018). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition*, 40(1), 171-199.

Pronouns attract in number but (much) less so in person. Evidence from Romanian.

Adina Camelia Bleotu (University of Bucharest), Brian Dillon (UMass Amherst)

Agreement attraction happens when a verb erroneously agrees with an intervening *distractor* instead of the **target** (*The **key** to the cabinets are on the table) [1]. Attraction has been widely observed in number and gender features [e.g., 2], it is less clear whether agreement attraction can also occur with person features, like 1st (*I, we*) and 2nd person (*you*). Previous research [3] concluded on the basis of a self-paced reading task that (1st, 2nd and 3rd person) pronouns in Russian lead to a person agreement attraction effect (though small in size) but did not examine the size of the effect in comparison to number. The current study investigates number and person attraction comparatively. We provide evidence on the basis of two 2 forced-choice experiments in Romanian that person features cause less attraction than number features.

Existing theories of agreement attraction do not explicitly consider person features. However, cue-based retrieval theories of agreement attraction [4, 6, 7] may suggest that interference should not be limited to particular features; this would imply that person features should create attraction through interference just like number or gender features. On the other hand, so-called ‘representational’ theories of agreement attraction [8] do not so clearly predict attraction with person features. Unlike number, 1st and 2nd person can neither percolate to the head noun, nor contribute to (the person of) the resulting complex DP featurally, as there are no lexical nouns with 1st or 2nd person features in Romanian (or in any other language that we know of).

In **Experiment 1** (N=62 Romanian speakers), a speeded forced choice continuation task [9], we sought to first establish whether **3rd person pronouns** can create number agreement attraction in Romanian by comparing them with two other types of distractors: **bare Ns** (the only form in which simple nouns can occur after prepositions in Romanian) and **full DP intervenors** (i.e., Det-Noun-Adj). Participants had to choose between a 3rd singular and a 3rd plural verbal form. **Materials:** There were 24 items with 6 conditions (see Table 1): MATCH (Match/ Mismatch) x INTERVENOR TYPE (Bare N/ Full DP/Pronoun). These were combined with 72 fillers. **Results** (see Table 2 & Fig 1). We ran a parsimonious mixed-effects logistic regression with accuracy as a dependent variable. In the (mis)match conditions, there were fewer errors with bare Ns and 3rd person pronouns than with full DP intervenors. This suggests that bare Ns and pronouns may not be ideal attractors: bare Ns are not subject-like, being typically used as non-referring Ns [10, 11, 12], and pronouns differ from full DPs through their lack of specified lexical context [13].

Having established that pronoun intervenors attract in number (to a certain extent), we further tested **person and number attraction** in **Experiment 2** (N=51) another speeded forced choice continuation task. **Materials:** There were 24 items with 4 conditions (see Table 3): MATCH (Match/ Mismatch) x PERSON (1/2 or 3). These were combined with 72 fillers. **Results** (see Table 4 & Fig 2). We ran a parsimonious mixed-effects logistic regression with accuracy as a dependent variable. Contrary to [3], we found that 1st and 2nd person pronouns behaved differently (i.e., led to significantly fewer errors) than 3rd person pronouns.

We conclude that (a) (3rd person) pronoun intervenors do allow number attraction, though less so than full DPs, (b) (1st and 2nd) pronoun intervenors create significantly less attraction than 3rd person pronouns; in the present experiment, we observed no reliable person attraction at all. Our results are easily explained by representational accounts of attraction, while cue-based theories would require further modifications to allow retrieval processes to distinguish between interference from person and number features. Our results dovetail with the widely observed asymmetry between 1st/2nd and 3rd person pronouns [14-18, a.o.] and the Feature Hierarchy Hypothesis [19], according to which Person is cognitively more significant than Number. In an agreement attraction context, it seems that the more salient a feature is, the more accurate people are.

Experiment 1 (Num attraction with 3rd Pron, Ns, DPs)

Table 1. Example items per conditions

Conditions	Example sentences
Match/ Mismatch x Bare Noun/Full DP/3 rd Person Pronoun Intervenor	Pisica/Pisicile de lângă fete/ fetele brunete/ ei adesea au /are Cat-the/ Cats-the near girl/ girls.the brunette/ they often have.3pl/have.3sg

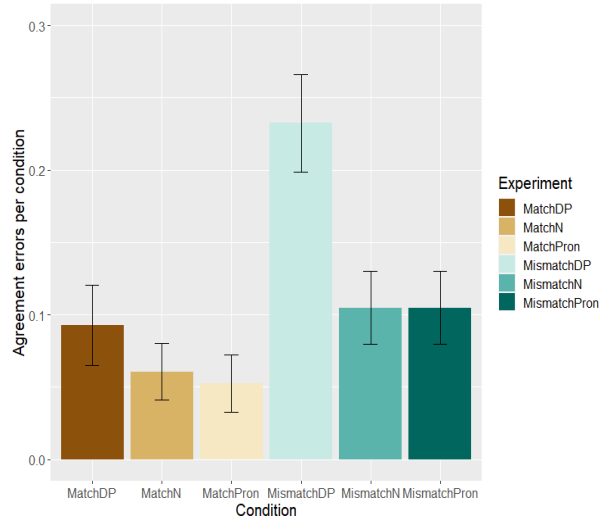


Figure 1. Agreement errors per condition (Experiment 1)

Table 2. Results of a generalized linear mixed effects model (Experiment 1)

Parameter	Estimate	Std. error	z	p
Intercept	-3.195	0.289	-11.056	<2e-16***
IntervenorA (N&Pron vs Full DP)	-0.323	0.106	-3.042	0.00235**
Matching	-1.227	0.349	-3.509	0.00045***
IntervenorB (Pron vs N)	-0.048	0.147	-0.327	0.744
IntervenorA:Matching	0.224	0.174	1.291	0.197
Matching:IntervenorB	-0.089	0.351	-0.254	0.799

Helmert coding schemes:

Intervenor A (N&Pron vs Full DP): N=1, Pron=1, Full DP=-2

Intervenor B (Pron vs N): Noun=1, Pron=-1, Full DP=0

Experiment 2 (Num & Person Attraction with Pron)

Table 3. Example items per conditions

Conditions	Example sentences
Number (Mis)match x 1 st /2 nd PL OR 3 rd PL Pron Interv	Pisica/Pisicile de lângă noi/voi/ei adesea avem/aveți/au/are Cat-the/Cats near we/you/they often have.1pl/2pl/3pl/3sg

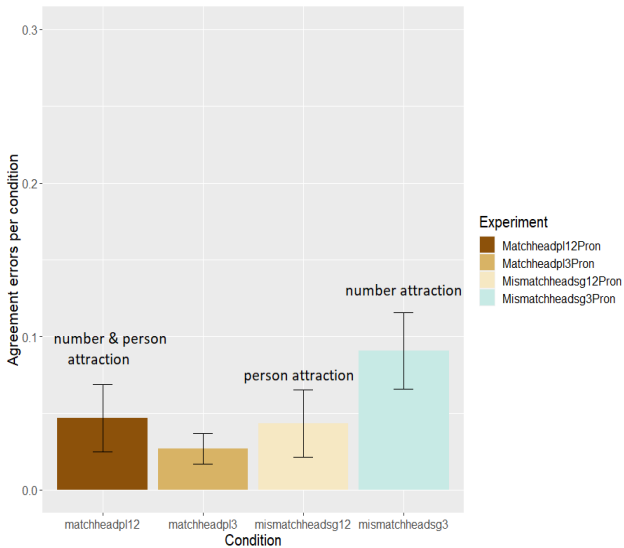


Figure 2. Agreement errors per condition (Experiment 2)

Table 4. Results of a generalized linear mixed effects model (Experiment 2)

Parameter	Estimate	Std. error	z	p
Intercept	-4.296	0.443	-9.710	< 2e-16 ***
Intervenor	-0.615	0.505	-1.217	0.224
Matching	-0.644	0.399	-1.613	0.107
Intervenor:Matching	1.664	0.701	2.375	0.0176

References: [1] Bock & Miller, 1991. *Cognitive Psychology* [2] Slioussar & Malko, 2016. *Frontiers in Psychology*. [3] Laurinavichyute & Vasishth, 2016. Agreement attraction in Person is symmetric. Poster CUNY. [4] Badeker & Kuminiak, 2007. *JML* [5] Slioussar, 2018. *JML* [6] Dillon et al., 2013. *JML*. [7] Wagers et al., 2009. *JML* [8] Eberhard et al., 2005. *Psychol. Rev.* [9] Staub, 2009. *JML* [10] Chierchia, 1998. *Natural Language Semantics*. [11] Dobrovie-Sorin, 2013. In *A Reference Grammar of Romanian*. [12] Tănase-Dogaru, 2014. *BWLP*. [13] Ritter, 2008. *NLLT*. [14] Silverstein, 1985. In *Features and Projections*. [15] Harley & Ritter, 2002. *Language* [16] Nevins, 2007. *NLLT*. [17] Mancini et al., 2011. *Brain Research* [18] Mancini et al., 2014. *Lingua*. [19] Carminati, 2005. *Lingua*.

Dynamics of referent demotion and promotion: Consequences for pronoun interpretation

Jina Song, Elsi Kaiser (University of Southern California)

Implicit causality (IC) research shows that some verbs bias subject-position pronouns to refer to preceding *subjects*, while other verbs bias reference to preceding *objects* (e.g. [1,2,3,5]). We use these IC verb effects, known to be associated with thematic roles, as a backdrop for new work testing how pronoun interpretation is guided by the referential dynamics of the transitions between clauses – i.e., the consequences of promoting vs. demoting referents to more or less salient positions. We consider grammatical *and* thematic roles, as both influence referent salience.

We test *referential structure effects in the pronoun-containing clause*: whether one or both referents from the preceding clause are re-mentioned. The **Referential Structure Hypothesis** states that a subject pronoun in clause 2 is more likely to refer to the clause1 subject when both clause 1 referents are re-mentioned in clause 2 (2-*pro*), compared to only one (1-*pro*, ex.1-2). This is based on the idea that **demoting** a higher-salience referent (clause1 sub) to a less-privileged position (clause2 obj), while **promoting** a lower-salience referent (clause1-obj) to a privileged position (clause2 sub), yields a less-coherent transition (Tbl2) (for related ideas, see [4]).

(1) **Exp 1 Exp-Stim/Stim-Exp verbs** (all-male name items (50%), all-female name items (50%))

a. Henry {surprised_{IC1} (SE) / respected_{IC2} (ES)} Kevin because he daxed him. [2-*pro*]

b. Henry {surprised_{IC1} (SE) / respected_{IC2} (ES)} Kevin because he daxed Tom. [1-*pro*]

(2) **Exp 2 Agent-Patient verbs** (all-male name items (50%), all-female name items (50%))

a. Henry {cheated_{IC1} (AP1) / saluted_{IC2} (AP2)} Kevin because he daxed him. [2-*pro*]

b. Henry {cheated_{IC1} (AP1) / saluted_{IC2} (AP2)} Kevin because he daxed Tom. [1-*pro*]

If we find referential structure effects, this would mean that models of pronoun interpretation need to incorporate more relational information about the *transitions between clauses*: specifically, not only the semantics of cross-clausal transitions [7], but also the referential properties of the transitions between clauses (Table 2). We report two studies testing the Referential Structure Hypothesis with IC1/IC2 verbs. We also test if **thematic roles** modulate referential structure effects, to better understand the relation between thematic roles and discourse salience.

Exp1 (n=40) tested Stimulus-Experiencer verbs whose IC biases change when the thematic role mapping changes: Stim_{subj}-Exp_{obj} verbs (e.g. *surprise*) elicit a subject bias (IC1); Exp_{subj}-Stim_{obj} verbs (e.g. *respect*) elicit an object bias (IC2) ([1,2,3,5]). Changes in IC bias are associated with a change in thematic roles. **Exp2** (n=60) tested Agent_{subj}-Patient_{obj} verbs. Some Ag-Pat verbs (e.g. *cheat*) elicit a subject bias (IC1); others (e.g. *salute*) elicit an object bias (IC2), ([1,3,5]). With this verb class, changes in IC bias do *not* involve any changes in thematic roles.

Method: Exp1-2 had 24 targets, 36 fillers. We manipulated (i) the referential structure of clause 2 (2-*pro*: *He...him*, 1-*pro*: *He...Tom*, ex.1-2) and (ii) the verb in clause 1 (IC1/IC2, Table 1). Nonce verbs in clause 2 minimized semantic variability. We used a picture task (Fig.1): People typed the names in the boxes such that the picture matches the event of the underlined part.

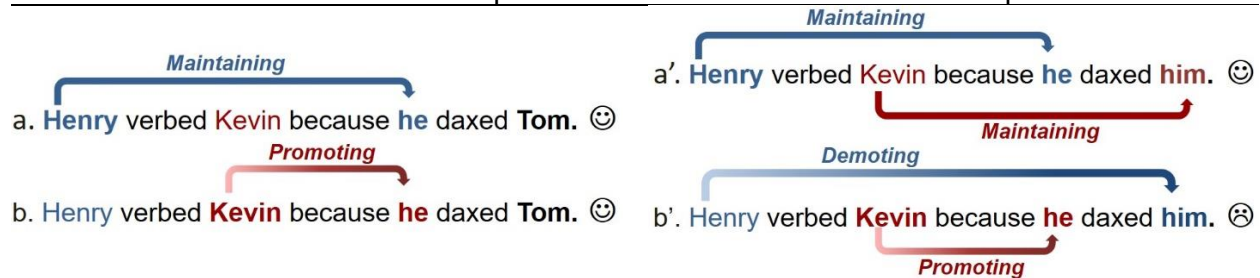
Results: **Exp1** (Stim-Exp, **Fig.2**) shows referential structure effects with both ES and SE verbs (more obj choices, less subj choices, in 1-*pro* than 2-*pro*, lmer, $p < .001$). SE conditions elicit fewer object choices than ES conditions (IC effect: $p < .001$). Strikingly, SE conditions show weaker effects of referential structure than ES (interaction, $p < .01$). This asymmetry may stem from Experiencers being inherently more topical than Stimuli (due to animacy, sentience, [8,9]): Demotion of Stimulus subjects (SE condition) may be less problematic than demotion of more salient Experiencer subjects (ES), yielding weaker referential structure effects with SE verbs.

Exp2 (Ag-Pat, **Fig.2**) replicates referential structure effects with both AP1 and AP2 verbs ($p < .001$), and IC effects ($p < .05$). Now, there is **no** interaction (p 's $> .3$): Referential structure effects are equal with AP1 and AP2 verbs. Between-experiment analyses yield a marginal 3-way interaction (exp x IC1/2 x ref.str.; $p = 0.057$), and effects of referential structure, IC1/2, exp, and interactions (ref.str. x IC1/2; IC1/2 x exp) (p 's $< .02$). **IN SUM:** Exp1-2 support the Referential Structure Hypothesis, showing that (i) its effects generalize across verb classes and that (ii) the thematic roles and their relative topicality also play a role by modulating discourse salience.

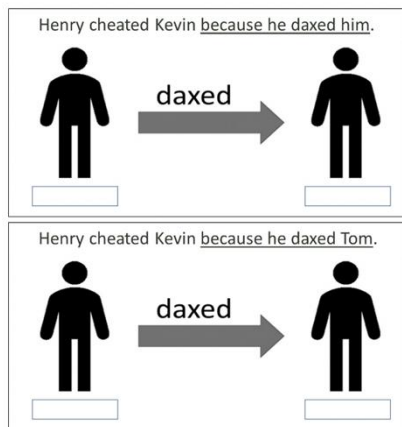
Table 1. IC bias of verb types used (All targets used 'because')

	Exp 1		Exp 2	
	S-biased Stim-Exp	O-biased Exp-Stim	S-biased Ag-Pat	O-biased Ag-Pat
IC bias [3],[5]	S bias: M=67.4%, sd=13.6	O bias: M=76.3%, sd=11.7	S bias: M=67.6%, sd=9.16	O bias: M=72.1%, sd=5.53

Table 2. Referential structure with 1 pronoun Referential structure with 2 pronouns

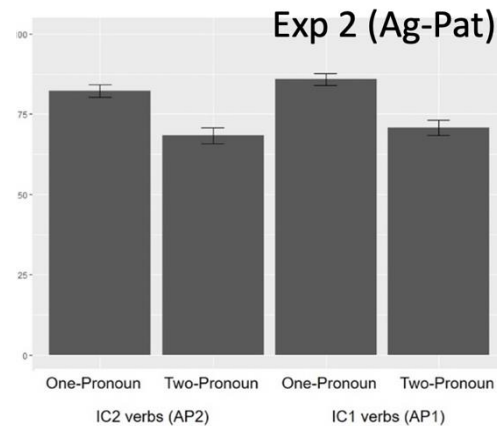
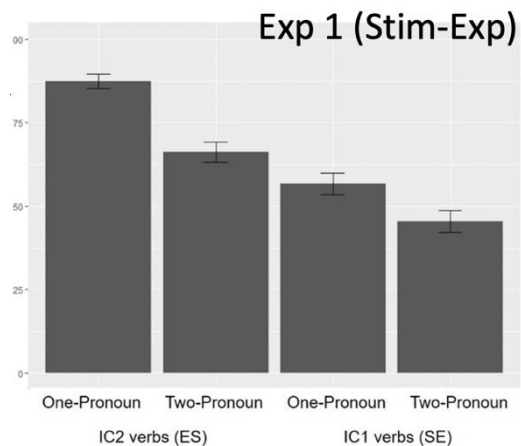


Both (a) and (b) yield coherent transitions. (b') yields a less coherent transition than (a').



<= Fig.1 Example items: 2-pro (top), 1-pro (bottom)

Fig.2 => Proportion of trials where subject-position pronoun refers to preceding object



Selected references [1] Au 1986. A verb is worth a thousand words. JML. [2] Caramazza et al. 1977. Comprehension of anaphoric pronouns. J of VLVB. [3] Ferstl et al. 2011. Implicit causality bias in English: A corpus of 300 verbs. BRM. [4] Grosz et al 1995. Centering: A framework for modeling the local coherence of discourse. CL [5] Hartshorne and Snedeker. 2013. Verb argument structure predicts implicit causality. LCP. [6] Kaiser. 2009. Investigating effects of structural and information-structural factors on pronoun resolution. [7] Kehler. 2002. Coherence, reference, and the theory of grammar. [8] Plank. 1979. Ergativity, syntactic typology, and universal grammar. [9] Verhoeven. 2014. Thematic prominence and animacy asymmetries.

Antecedent prominence and the Chinese reflexive *ziji*

Jun Lyu & Elsi Kaiser (University of Southern California)

Introduction Chinese reflexive *ziji* ('self') can be bound by a long-distance (LD) antecedent when the antecedent is an *internal* perspective center or an empathy locus (Pan'97; Huang & Liu'01; Charnavel et al.'17; a.o.), the person whom one stands in the shoes of. This property of *ziji* necessitates two predictions: (i) if the non-local antecedent is made more salient as a discourse topic, LD binding should be more acceptable; (ii) if an *external* perspective center (e.g. a 1st person referent representing the perspective of the comprehender) is introduced as the local antecedent, there should be more local binding as a first-person referent is presumably more salient on the discourse level than the 3rd NP (Kuno'87; Pan'97), a phenomenon called the "blocking effect". In this work, besides testing these two hypotheses, we also checked whether syntactic prominence contributes to greater blocking effects, thus helping us understand how top-down (discourse level) and bottom-up (syntactic level) cues guide antecedent retrieval. Furthermore, in all experiments in this study, we examined the influence of verb bias as another bottom-up cue, the interaction of which with discourse-level prominence is not quite clear.

Methods To assess whether *ziji* is sensitive to topic prominence, **Exp.1** (N=45) crossed factors *Context* (biased vs. neutral) and *Verb bias* (self- vs. other-directed). This is used to contrast with **Exp.2** (N=46) on *ta-ziji* ('s/he-self') which supposedly does not involve perspective-taking (Pan'98; Pan & Hu'02). This contrast should lead us to expect more non-local binding in biased contexts for *ziji* than *ta-ziji*. For an example, see (1a-b). **Exp.3** (N=50) and **Exp.4** (N=48) test whether, in the presence of a 1st person pronoun (blocker), the strength of the blocking effect is sensitive to the grammatical role of the blocker (i.e. whether the 1st person pronoun is in subject vs. object position), as subjects are more structurally prominent than objects. *Blocker type* (1st-person vs. 3rd person) and *Verb bias* was crossed in a factorial design. See (2a-b) for examples. Twenty target sentences with 20 fillers were presented to participants in forced choice judgment tasks.

Results See Fig.1-4 for the proportion of local antecedent choices. **Exp. 1** (*ziji*) reveals main effects of *Context* and *Verb bias* (*glmer* in R, $ps < .005$), but no interaction. The preference for the local antecedent was weaker in the presence of a *topical* non-local referent, suggesting *ziji* exhibits prominence sensitivity to discourse topicality. Additionally, self-directed verbs elicited overwhelmingly more local interpretations than other-directed verbs, a strong effect of verb bias. Surprisingly, **Exp.2** (*ta-ziji*) also revealed a significant effect of *Context* ($p < .05$), although cross-experimental analysis suggests that in biased contexts *ziji* showed higher probability of non-local binding than *ta-ziji* ($p < .05$). Additionally, there was a robust *Verb bias* effect ($p < .001$) in the predicted direction (contra Lu'11). **Exp. 3** (*blocker "I" in subj position*) showed main effects of *Verb bias*, *Blocker type*, and an interaction ($ps < .001$). Pairwise comparisons suggest that the 1st-person blocking effect only reached significance with other-directed verbs. This is to be expected, given that the blocker is only relevant when there is a non-local binding tendency in the first place (if *ziji* is interpreted as having a local antecedent, the blocker is redundant). When we compare **Exp. 4** (*blocker "me" in obj position*) to Exp. 3, it becomes clear that the blocking effect was weaker in Exp.4 with other-directed verbs compared to Exp.3: When verb semantics pushes people to look for the non-local antecedent, 'me' is less effective at blocking the non-local search than 'I'. Cross-experimental comparison showed that the subject blocking effect was stronger than the object blocking effect ($p < .01$). *This provides evidence for prominence sensitivity on the grammatical level (subject vs. object.)* (Exp4 shows no Blocker x Verb bias interaction, possibly due to differences in source/perceiver structure of the matrix verbs used in Exp.4. See Kaiser et al.'09).

Discussion In line with the idea that *ziji* prefers higher-prominence referents, we find that *ziji* is sensitive to prominence on (i) the *discourse level* (topicality and perspective), and on (ii) the *grammatical role level* (subj/obj blocker). The grammatical-role effects show that the blocking effect varies as a function of the blocker's structural prominence: when the 1st person blocker is a non-local/non-c-commanding object, blocking is weaker. Verb bias plays a crucial role with both *ziji* and *ta-ziji*, highlighting the impact of pragmatic/semantic information.

Stimuli for topicality effect

(1a) *Biased context (non-local antecedent = discourse topic)*

Ming is a good student in the class. [小明是班级里的优秀学生]

During class, he heard Prof. Wang just {published_{SELF}/graded_{OTHER}} (ta)ziji-GEN academic paper.

[课上, 他听说王教授刚刚{发表了/批改了}(他)自己的学术论文]

(1b) *Neutral context (no discourse topic)*

Today is the day for the literature class. [今天是上文学课的日子]

During class, Ming heard Prof. W. just {published_{SELF}/graded_{OTHER}} (ta)ziji-GEN academic paper.

[课上, 小明听说王教授刚刚{发表了/批改了}(他)自己的学术论文]

Stimuli for blocking effect

(2a) *First-person vs. third-person referent in subject position:*

Ming is a good student in the class. [小明是班级里的优秀学生]

During class, he heard {I/Prof. Wang} just {published/graded} ziji-GEN academic paper.

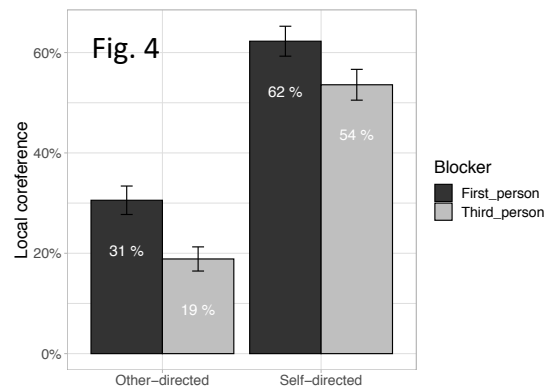
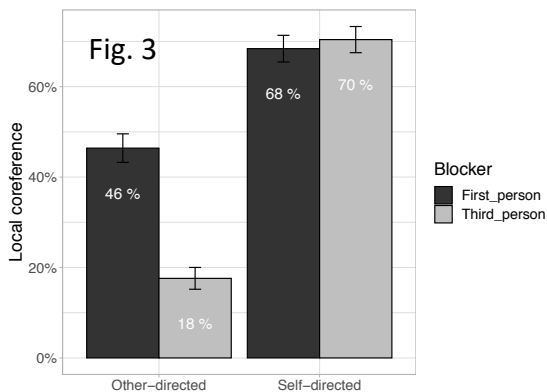
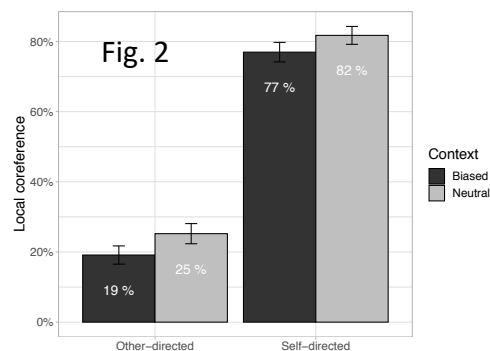
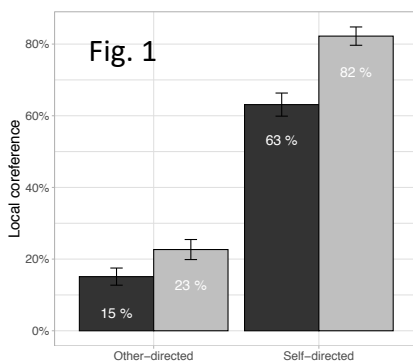
[课上, 他听说{我/王教授}刚刚{发表了/批改了}自己的学术论文]

(2b) *First-person vs. third-person referent in object position:*

Ming is a good student in the class. [小明是班级里的优秀学生]

During class, he told {me/others} Prof. W. just {published/graded} ziji-GEN academic paper.

[课上, 他告诉{我/别人}王教授刚刚{发表了/批改了}自己的学术论文]



Selected references

- Charnavel, I., Cole, P., Hermon, G., & Huang, C.-T. (2017). Long-distance anaphora: Syntax and discourse. In Everaert, M. & van Riemsdijk, H. C. (eds.), *The Wiley Blackwell Companion to Syntax*, 2321–2402. Hoboken, NJ: John Wiley & Sons, Inc.
- Huang, C.-T. J., & Liu, C.-S. L. (2001). Logophoricity, attitudes, and ziji at the interface. In P. Cole, G. Hermon, & Huang, C.-T. J. (eds.), *Long-distance reflexives*, 141–195. New York: Academic Press.
- Pan, H. (1997). *Constraints on Reflexivization in Mandarin Chinese*. New York: Garland Publishing, Inc.
- Pan, H. (1998). Closeness, prominence, and binding theory. *Natural Language & Linguistic Theory*, 16(4), 771–815.

Anaphora resolution in causal coherence relations in Mandarin Chinese

Jun Lyu & Elsi Kaiser (University of Southern California)

Introduction: In a sentence involving causal relations like “Jane angered Peter because...”, the ensuing pronoun typically refers to the *cause* of the event *Jane* (e.g., Hartshorne, 2014). However, the processing effort depends, among other factors, on the saliency of the intended antecedent. As a structurally salient constituent, the subject tends to be privileged for coreference, known as the *subject preference* (Frederiksen, 1981; Crawley et al., 1990). A less explored factor is *thematic role*. Previous research shows that thematic roles higher on the Thematic Hierarchy (Jackendoff, 1972/1987; Grimshaw, 1990) are more prominent, as a higher ranked role (e.g., agent vs. patient) often takes the subject position given syntactic flexibility, which makes one wonder if thematic-role-related prominence affects anaphora resolution. This study probes the influence of both subject preference and thematic role on real-time pronoun processing in Chinese.

Methods: In a self-paced reading (SPR) study (163 natives), we crossed *grammatical position* (subject/non-subject) of the cause and its *thematic role* (agent/patient) in the context sentence, followed by a target sentence (20 target items and 20 fillers). The pronoun in target sentences refers to the gender-unambiguous antecedent, the cause of the event (see (1) for an example). To promote patients to subject position, passive BEI-construction in Chinese was used. Note that the “agent” and “patient” in this study are defined in line with Dowty (1991) who outlined several properties of proto-agents/proto-patients but argued that agents and patients should be determined based on their similarities to proto-agents/proto-patients. Thus, the agents in this study are more agent-like relative to patients and patients are more patient-like relative to agents.

Predictions: The *subject preference account* predicts a main effect of *grammatical position* while the *thematic hierarchy account* predicts a main effect of *thematic role*.

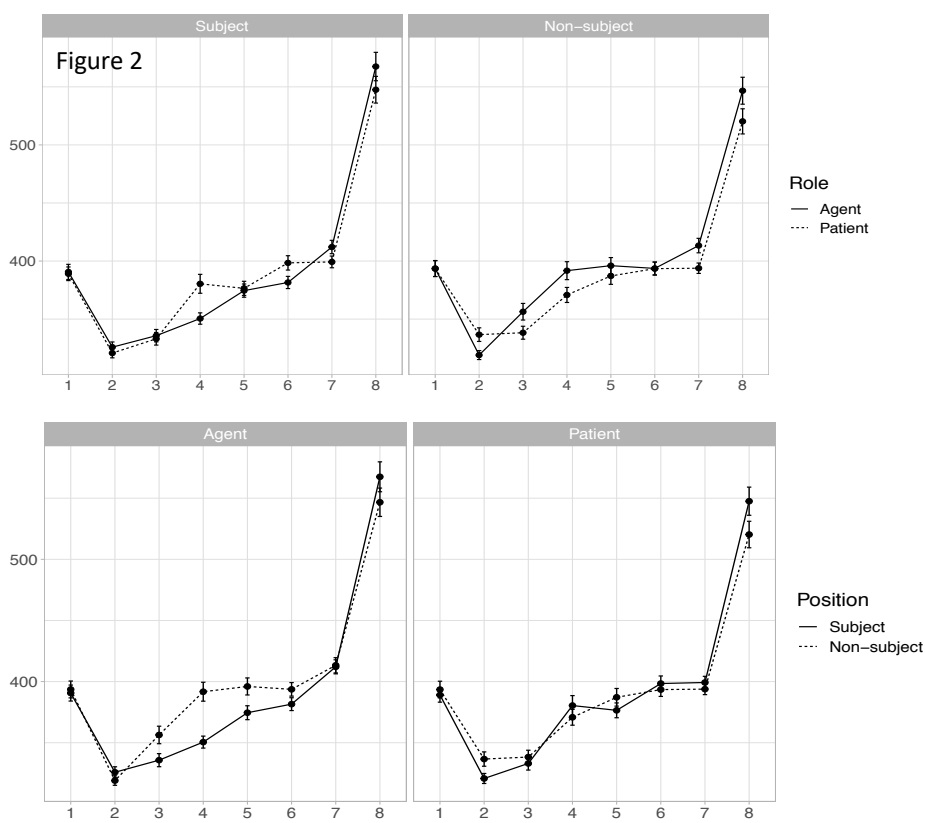
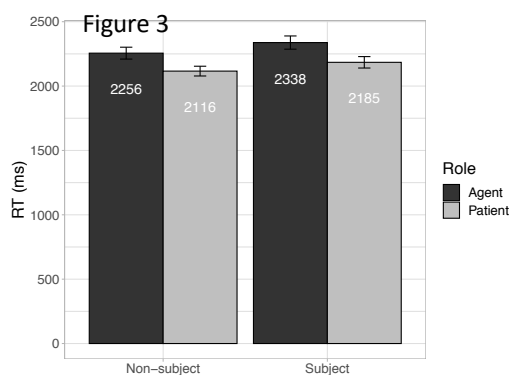
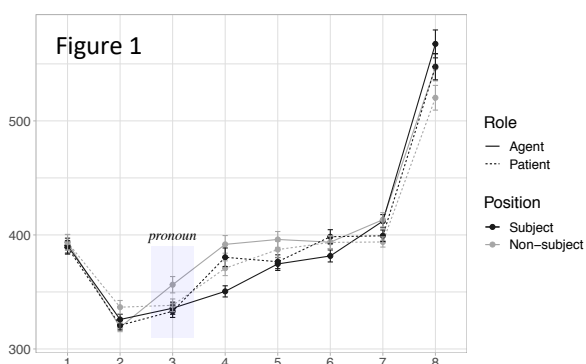
Results: **Figure 1** shows the RT patterns (stats: *lmer* in R) for all conditions in the target sentence. For better visualization that isolates the impact of grammatical positions and thematic roles, see **Figure 2**. RTs for comprehension question response are plotted in **Figure 3**. The critical pronoun region 3 showed that both main effects and the interaction were significant ($ps < .05$), because the non-subject/agent condition in (1b) elicited longest RTs compared to other conditions ($ps < .001$). Region 4 revealed a main effect of *grammatical position* ($p < .01$) and a significant interaction ($p < .001$), because while patient-hood lead to reading slowdowns in subject conditions ($p < .001$), agent-hood lead to reading slowdowns in non-subject positions ($p < .01$). Region 5 is characterized only by a main effect of *grammatical position* ($p < .005$), demonstrating a subject preference. Statistical model for region 6 was not significant ($ps > .05$). Interestingly, the final two regions showed that patient conditions lead to faster RTs ($ps < .05$). These results overall only support the *subject preference account* at best, but not the *thematic role account*. The divergent thematic role effect at Region 4 is highly intriguing, shown clearly by Figure 2 (top row). In fact, based on the effect size and the timing of the effect, the following processing ease hierarchy can be derived: Subject/Agent > Subject/Patient = Non-subject/Patient > Non-subject/Agent. Finally, comprehension question response latencies showed that patient conditions overall lead to faster responses ($p < .05$), similar to the SPR reading patterns at the last two regions.

Discussion: To account for the unexpected contrast in Figure 2 (top row), we propose a Mapping Principle, similar to Ferreira (1994): an agent in the sentence must be mapped to the subject position and a non-agent must be mapped to the non-subject position. Thus, when the pronoun refers to the cause of the event that violates the Mapping Principle, parsing difficulty occurs. Crucially, the Mapping Principle alone cannot fully account for the data as the bottom-right panel in Figure 2 suggests an absence of the Mapping Principle penalty, as the subject/patient condition did not lead to longer RTs compared to the mapped non-subject/patient condition. However, as shown by Table 1, when Subject Preference and Mapping Principle are acknowledged to play independent roles, the processing ease pattern we observed earlier can be explained. As to the response latencies, we hypothesize that the semantic representation of the event has a “agent-verb-patient” configuration, which helps explain why retrieval of patients is easier.

- (1) Example target stimuli (translated from Chinese; 20 sets in total; left: context; right: target sentence)
- | | | |
|--|---|--|
| a. Subject/Agent: Jane upset Peter. | } | This is/ because/ she / in/ the game/ not know/ team/ work. |
| b. Non-subject/Agent: Peter BEI Jane upset. | | |
| c. Subject/Patient: Peter BEI Jane blamed. | } | This is/ because/ he / in/ the game/ not know/ team/ work. |
| d. Non-subject/Patient: Jane blamed Peter . | | |

Table 1. Constraints active in pronoun resolution in a causal discourse.

Condition	Subject Preference	Mapping Principle	Penalty	Processing ease: a. Agent/Subject > c. Subject/Patient, d. Non-subject/Patient > b. Non-subject/Agent
a. Subject/Agent	-	-	0	
b. Non-subject/Agent	1	1	2	
c. Subject/Patient:	-	1	1	
d. Non-subject/Patient	1	-	1	



Selected references

- [1] Crawley, Rosalind A., Rosemary J. Stevenson & David Kleinman. 1990. The use of heuristic strategies in the interpretation of pronoun. *Journal of Psycholinguistic Research*.
- [2] Dowty, D. 1991. Thematic proto-roles and argument selection. *Language*.
- [3] Ferreira, F. 1994. Choice of passive voice is affected by verb type and animacy. *JML*.
- [4] Hartshorne, J. K. 2014. What is implicit causality? *LCN*.
- [5] Jackendoff, R. S. 1972. *Semantic interpretation in Generative Grammar*. MIT press.

Investigating perspective-sensitivity during the resolution of Korean anaphors

Sarah Hye-yeon Lee & Elsi Kaiser (University of Southern California)

Understanding perspective-sensitivity is central for theories of cognition and language processing. We test how interpretation of two perspective-sensitive elements—subjective adjectives (e.g. *tasty*, *fun*) and certain anaphors (e.g. *picture of herself/her*)—interacts in Korean.

Perspective-sensitivity of anaphors: Contrary to syntactic Binding Theory, reflexives and pronouns have been claimed to be perspective-sensitive (e.g.[7,11]). [4] found that in English picture-NPs (**PNPs**) like “*Nora told/heard from Amy about the picture of her/herself*”, reflexives’ interpretation is modulated by a preference for **sources-of-information** (subj of *told*, obj of *heard*, [7]), while pronoun interpretation is modulated by a preference for **perceivers-of-information** (subj/*heard*, obj/*told*, [10]). However, the crosslinguistic robustness of these effects is unknown.

Perspective-sensitivity of subjective adjectives: Adjectives expressing opinions (e.g. *funny*, *scary*) are inherently perspective-sensitive, interpreted relative to attitude-holders (e.g.[8])

Perspectival Uniformity: It has been claimed that perspective-taking is a monolithic process, in that *all* perspective-sensitive elements in the same linguistic domain must **uniformly refer to the same perspectival center** (e.g.[1]). Under this uniformity view, in *Nora told/heard from Amy about the funny picture of herself*, the perspective-sensitive *herself* refers to whoever finds the picture *funny* (referent of *herself* = attitude-holder of *funny*). However, experiments suggest that English does not fit this prediction ([3]). Thus, the status of Perspectival Uniformity is debated.

Korean allows us to test the crosslinguistic robustness of Perspectival Uniformity and the source/perceiver biases of pronouns and reflexives. Unlike English, Korean reflexives include (Table1): **cakicasin** (‘self’), commonly viewed as needing a local antecedent (cf.[6]) and **caki** (‘self’) which can be bound by a long-distance (LD) antecedent and is commonly viewed as requiring antecedents that are perspectival centers (e.g. [12,10]; cf.[2]). In addition to a richer reflexive paradigm, Korean personal pronouns **ku/kunye** (‘he/she’) have both pronominal and demonstrative properties ([10,5]), differing grammatically from English personal pronouns.

We used a forced-choice experiment to (i) identify which Korean forms show perspective-sensitivity along the source/perceiver dimension, to assess the broader validity of [4]’s claims that reflexives and pronouns exhibit complementary perspective sensitivity, and to (ii) test whether Korean perspective-sensitive anaphors and subjective adjectives exhibit Perspectival Uniformity.

Method (N=90, 36 targets, 42 fillers: People read sentences with PNPs modified by subjective adjectives (Table1), and answered two questions: *Who is shown on the photograph?* (anaphor resolution) and *Whose opinion is it that the photograph is [subjective adjective]?* (attitude-holder identification). We manipulated (i) the verb (*told/heard from*)—to manipulate *source*- and *perceiver* status of the subject and object—and (ii) whether the PNP contains *caki*, *cakicasin* or *ku/kunye*.

Results: Who-shown questions (anaphor resolution; Fig.1): Both reflexives (*cakicasin*, *caki*) show a *source preference* (more obj choices with *hear* than *tell*; glmer, $p's < .001$). But pronouns (*ku/kunye*) show no signs of perspective-sensitivity ($p > 0.1$). Interactions confirm only *caki* and *cakicasin* are sensitive to the verb manipulation, not *ku* (interactions, $p's < .001$). **Whose-opinion questions** (attitude-holder identification; Fig.2) reveal a strong preference to interpret the *source* (subj/*tell*, obj/*hear*) as the attitude-holder of the adjective, regardless of form or verb ($p's < 0.001$).

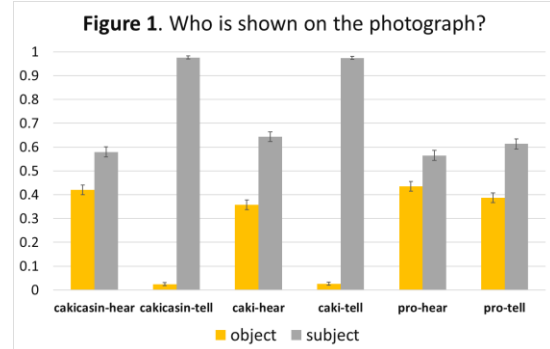
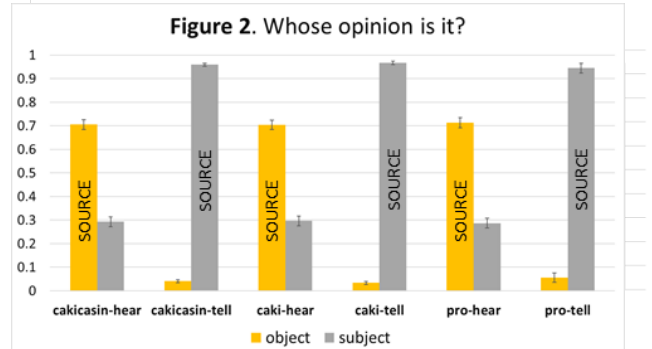
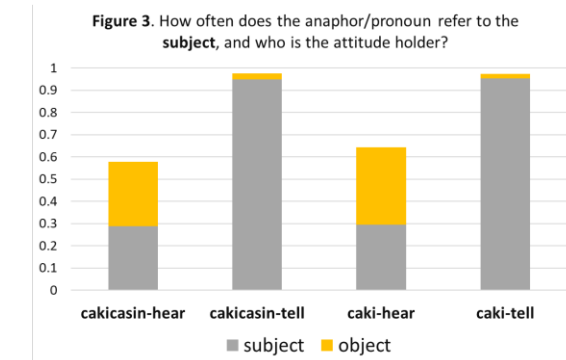
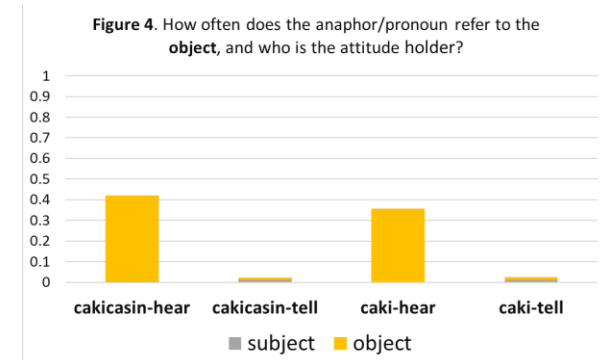
Do perspective-sensitive caki/cakicasin and subjective adjectives show Perspectival Uniformity? Not fully. Fig.3 shows that though antecedent choice and attitude-holder identification converge with *tell* (subject is antecedent \Rightarrow subject is opinion-holder), they clearly diverge with *hear* (subject is antecedent $\neq \Rightarrow$ subject is opinion holder). Similar patterns are found when the reflexive is interpreted as referring to the object (Fig.4; *tell* conditions).

Our results both replicate *and* identify limits of source/perceiver effects, suggesting that the source bias may be a core property of the entire class of reflexives crosslinguistically, while the perceiver bias may not extend to elements that are not purely anaphoric. Lack of Perspectival Uniformity challenges views that analyze perspective-taking as a monolithic process, and favors accounts acknowledging different sub-types (e.g. referential vs. evaluative perspective-taking).

Table 1. Example target stimuli (*pro*-drop/null *pro* not possible in this context)

Mina-ka	Senguni-{hantey/hanteyse}	sinmwun-ey	
Mina-NOM	Senguni-to/from	newspaper-DAT	
{caki/cakicasin/kunye}-uy	mwusewun	sacin-i	iss-ta-ko
{self _{LD} /self _{local} /she _{pronoun} }-GEN	scary	photograph-NOM	exist-DECL-COMP
{malhaycwu-ess-ta/tul-ess-ta}			
tell-PAST-DECL/hear-PAST-DECL			
'Mina {told/heard from} Sengun that there is a scary photograph of {herself _{LD} SELF/herself _{local} SELF/her _{pronoun} } on the newspaper.'			

Questions: Who is shown on the photograph? Mina Sengun
 Whose opinion is it that the photograph is scary? Mina Sengun Narrator
 (Instructions explained term 'narrator.' Results yielded 1 narrator-response, 0.003% of data.)

**Fig.1** Who is shown? (anaphor resolution)**Fig.2.** Whose opinion? (attitude holder)**Fig.3** Trials where people interpret reflexives as referring to the **subject**, as a function of whose opinion the adjective reflects (*pronoun* conditions omitted, not perspective-sensitive)**Fig.4** Trials where people interpret reflexives as referring to the **object**, as a function of whose opinion the adjective reflects (*pronoun* conditions omitted)

[1] Bylinina et al.'14. Landscape of perspective shifting. [2] Han/Storoshenko'12. Semantic binding of long-distance anaphor caki in Korean. [3] Kaiser'20. Investigating predicates of personal taste and perspectival anaphors. [4] Kaiser et al.'09. Structural and semantic constraints on the resolution of pronouns and reflexives. [5] Kim/Han'16. Inter-speaker variation in Korean pronouns. [6] Kim/Yoon'09. Long-distance bound local anaphors in Korean. [7] Kuno'87. Functional syntax. [8] Lasnik'05. Context Dependence, Disagreement, PPTs [9] Park'18. Attitudes de se and logophoricity. [10] Sohn'01. The Korean Language. [11] Tenny'03. Short distance pronouns in representational noun phrases. [12] Yoon'89. Long-distance anaphors in Korean and their crosslinguistic implications.

Interpretation of null pronouns in Mandarin Chinese does not follow a Bayesian model

Suet Ying Lam and Heeju Hwang (The University of Hong Kong)

lsy317@connect.hku.hk

INTRODUCTION There are at least three ways to model a speaker's interpretation of a pronoun. The Mirror Model (MM) argues that the interpretation bias of a pronoun toward a referent is proportional to the likelihood that a pronoun is used to refer to that referent (production bias). The Expectancy Hypothesis (EH, e.g., Arnold, 2001) suggests that the interpretation bias of a pronoun toward a referent is correlated with the likelihood that the referent is re-mentioned regardless of its referential form (next-mention bias). A Bayesian Model (BM, e.g., Kehler et al., 2008) proposes that pronoun interpretation is determined by both the production bias and the next-mention bias. Previous work suggests that BM best explains the interpretations of English pronouns (Rhode & Kehler, 2014), Chinese overt pronouns (Zhan et al., 2020) and German personal pronouns (Patterson et al., 2020). The current study tests the validity of the three models on Mandarin null pronouns. Zhan et al. (2020) assume that the interpretation of Mandarin null pronouns would follow BM just like overt pronouns, given that both are subject-biased. Yet study suggests that Mandarin null pronouns exhibit a much stronger bias toward the subject than overt pronouns (Zhang, 2018). This raises the possibility that the interpretation of null pronouns is less sensitive to the semantically-driven biases such as the next-mention bias and may not be best explained by the models that incorporate the next-mention bias as a predictor of pronoun interpretation, i.e., EH and BM.

EXPERIMENTS We conducted two story-continuation experiments. Exp. 1 aims to replicate previous findings on overt pronouns. Exp. 2 assesses the validity of the models on null pronouns. We included both subject (N1)- and object (N2)-biased verbs to investigate the effect of the next-mention bias, and both implicit causality (IC) and transfer-of-possession (TOP) verbs to examine if the best model generalizes across verb types. We controlled coherence relations by using 'because' for IC verbs and 'so' for TOP verbs to maximize our chance of detecting a potential effect of the next-mention bias. Each experiment contained two versions of prompts: free prompts (to measure the next-mention bias and the production bias) and pronoun prompts (to measure the interpretation bias). We indicated the presence of null pronoun with a verb 'want to/think' in Exp. 2. The below are example stimuli using (1) N1-/N2-biased IC verbs and (2) N1-/N2-biased TOP verbs.

- (1) 小玲吓到了/害怕嘉怡, 因为 (free)... /因为她 (overt)... /因为想 (null)...
- Xiaoling frightened/fears Jiayi, because... /because she... /because \emptyset wants to/think...
- (2) 立强从小刚那里收到了/向小刚寄了一个包裹, 所以 (free)... /所以他 (overt)... /所以想 (null)...
- Liqiang received/sent a package from/to Xiaogang, so... /so he... /so \emptyset wants to/think...

MODEL EVALUATION Following Zhan et al. (2020), we compared the predicted data against the observed data on an item-by-item basis, using R^2 (correlation between the predicted and the observed data), and MSE/ACE (prediction error compared to the observed data). Larger R^2 and smaller MSE/ACE imply better performance. Sometimes pronouns did not occur in an item at all, so we used additive smoothing to avoid zero-probability estimates (see Appendix B).

RESULTS For overt pronouns, the mixed effect logistic regression models showed that the interpretation bias was sensitive to both the next-mention bias and the production bias, consistent with BM. Fig. 1 also shows that BM works the best for overt pronouns, whereas EH underestimates the N1-bias and MM overestimates it. In terms of statistical metrics (see Table 1), although EH has a higher R^2 , BM has a much smaller prediction error. For null pronouns, however, the next-mention bias affected only TOP but not IC verbs. As can be seen in Fig. 2, the interpretation of null pronouns is strongly N1-biased compared to overt pronouns. Although BM outperforms EH and MM in statistical metrics, it systematically underestimates the N1-bias. Our results suggest that the existing models do not accurately capture the interpretation of null pronouns, and BM may only apply to overt pronouns across languages.

A. Quantitative models used:

- Bayesian: $P(\text{referent}|\text{pronoun}) = \frac{P(\text{pronoun}|\text{referent}) P(\text{referent})}{\sum_{\text{referent} \in \text{referents}} P(\text{pronoun}|\text{referent}) P(\text{referent})}$
- Mirror: $P(\text{referent}|\text{pronoun}) \leftarrow \frac{P(\text{pronoun}|\text{referent})}{\sum_{\text{referent} \in \text{referents}} P(\text{pronoun}|\text{referent})}$
- Expectancy: $P(\text{referent}|\text{pronoun}) \leftarrow P(\text{referent})$

B. Additive Smoothing:

$$\hat{P}(\text{NP}_j) = \frac{\text{Count}(\text{NP}_j) + 3}{\text{Count}(\text{NP}_1) + \text{Count}(\text{NP}_2) + 2 \times 3}$$

$$\hat{P}(\text{pronoun}|\text{NP}_j) = \frac{\text{Count}(\text{NP}_j \wedge \text{pronoun}) + 1}{\text{Count}(\text{NP}_j) + 3}$$

C. Item-by-item quantitative model evaluation collapsing over IC and TOP

Figure 1: Overt pronoun

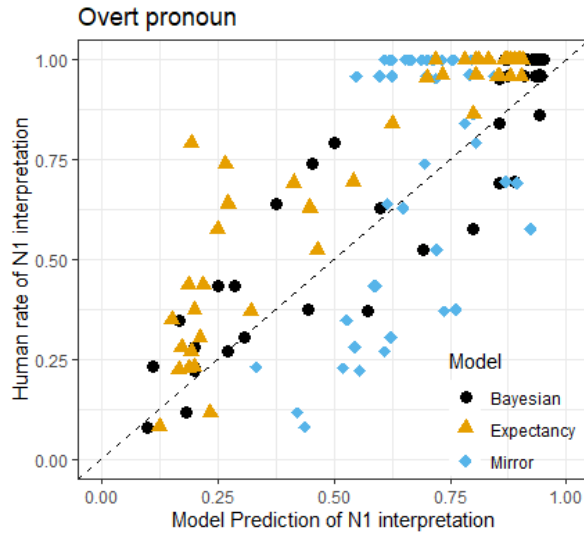


Figure 2: Null pronoun

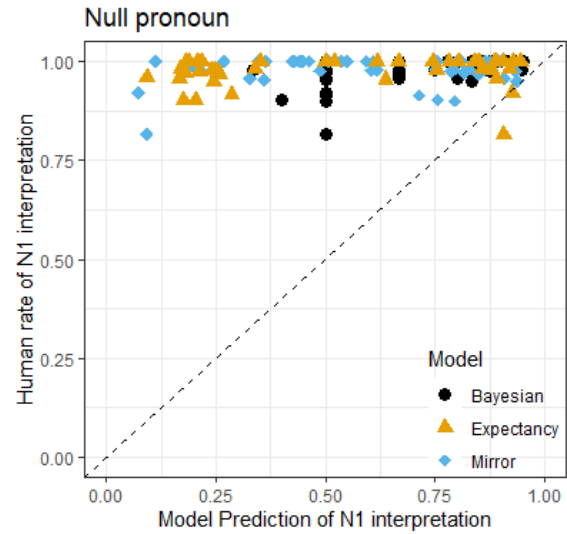


Table 1. Statistical metrics of model evaluation

IC Overt	BM	EM	MM	TOP Overt	BM	EM	MM
R ²	0.950***	0.952***	0.491***	R ²	0.585***	0.772***	0.080
MSE	0.009	0.016	0.085	MSE	0.019	0.060	0.068
ACE	0.804	0.888	0.521	ACE	0.253	0.575	0.309
IC Null	BM	EM	MM	TOP Null	BM	EM	MM
R ²	0.253*	0.004	0.045	R ²	0.488***	0.383**	0.041
MSE	0.114	0.297	0.262	MSE	0.067	0.197	0.137
ACE	0.445	0.858	0.820	ACE	0.277	0.606	0.452

***: $p < .001$; **: $p < .01$; *: $p < .05$;

References [1] Arnold (2001). The Effect of Thematic Roles on Pronoun Use and Frequency of Reference Continuation. *Discourse Processes*, 31(2), 137-162. [2] Kehler et al. (2008). Coherence and Coreference Revisited. *Journal of Semantics*, 25(1), 1-44. [3] Rohde & Kehler (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8), 912-927. [4] Patterson et al. (2020). A Bayesian approach to modelling German personal and demonstrative pronouns. Poster presented at the 33rd annual CUNY Human Sentence Processing Conference. [5] Zhan et al. (2020). Pronoun interpretation in Mandarin Chinese follows principles of Bayesian inference. *Plos One*, 15(8), e0237012. [6] Zhang (2018). Interpretational biases and processing of null and overt pronouns in Chinese. Unpublished doctoral dissertation, Konkuk University.

Source monitoring and false information endorsement in native and foreign language: an online study with Russian-English bilinguals

Aleksandra Dolgoarshinnaia, Beatriz Martín-Luengo (National Research University Higher School of Economics, Russia)

Language is tightly connected to the information processing and memory functioning. The effect of language on memory can be drastic as even slight lexical or grammar variations in statements can alter an individual's recollection of the event and incorporate false information (Loftus & Palmer, 1974). To this day, only a handful of studies investigated false information endorsement in people speaking more than language (Calvillo & Mills, 2018). Evidence from research on bilingualism suggests that bilinguals may have enhanced executive functioning, specifically, inhibitory control, when engaged in such cognitive processes as decision-making, attention, and memory processing (Bialystok, et al., 2004). At the same time, source monitoring, which is considered crucial for false information rejection, also heavily relies on inhibitory control (Ruffman, et al., 2001). Furthermore, bilinguals can rely more on reasonable and deliberate System-2 processing than heuristic System-1 (Caldwell-Harris, 2014). This evidence suggests that bilinguals can be more analytical when processing information in their second language and thus will endorse less false information when it is presented in their second language compared to the first.

To test this suggestion, we conducted a 2 x 2 x 3 within-subjects online-experiment with the language of misleading information (Russian, English), the type of item (true or false) and the source (English, Russian, or None) as our independent variables. We recruited 56 Russian-English unbalanced bilinguals (40 females, mean age = 24.1 SD = 4.66) who demonstrated high levels of English proficiency (mean score = 20 out of 25 points). Participants completed a classical misinformation paradigm in which they watched a recording of a crime, read a pair of English and Russian narratives describing the crime, and performed a yes/no recognition task and a source monitoring test (English narrative, Russian narrative, None).

Higher accuracy for the false control items ($M = .8$) than false misleading information ($M = .68$, $p < .0001$) confirmed endorsement of misleading information. However, interaction between item type and language was not observed. Testing differences between correct source attributions (Tab.1), correct attributions to the None source were higher ($M = .747$) compared to two other sources (Ru ($M = .275$); En ($M = .295$)). To facilitate the interpretation of the differences, we ran univariate ANOVAs for each of the sources. While the Russian and the None sources demonstrated similar patterns of higher correct attributions and no difference between incorrect ones, for the English source significantly more ($M = .1$, $p = .018$) incorrect attributions were made in favor of the Russian source ($M = .18$) and not the None source ($M = .08$).

The study examined the influence of the foreign language on the acceptance of false information and source monitoring. We observed the misinformation effect, however, our expectations that participants would accept more false information in their native tongue was not confirmed. It is possible, that the absence of the expected interaction might be attributed to high level of foreign language proficiency in our participants, indicating that with increasing levels of foreign language proficiency bilinguals' information processing becomes similar in both languages. At the same time the analysis of source monitoring revealed that for English source participants favored the Russian source rather than the None source when they made incorrect attributions. This suggests that our participants might have used more resources to process the information in English, which led to better recognition, although wrong attribution. Together the results in misinformation and source monitoring infer that people proficient in a second language might be susceptible to particular types of memory errors but not all of them.

Supplementary materials

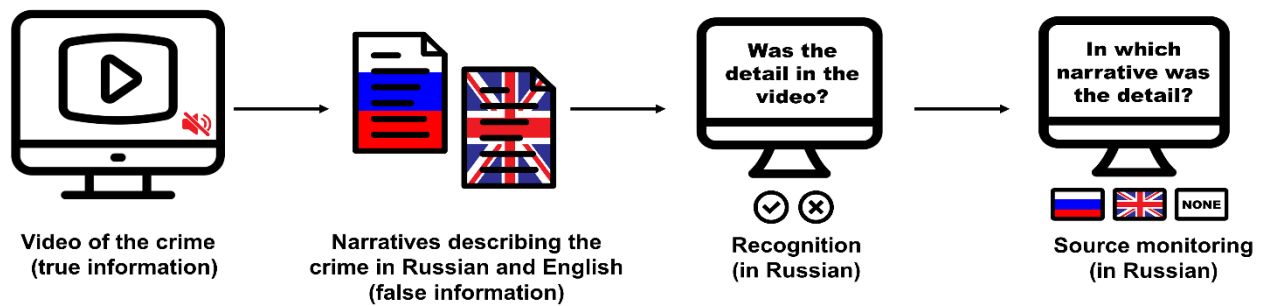


Figure 1. Experimental design. Stage 1: encoding of true information from a video. Stage 2: introduction of misinformation from 2 narratives (Russian and English). Stage 3: true/false recognition + source monitoring (Russian narrative, English narrative, none)

Table 1. Mean (standard deviation) of the proportions in the source monitoring for selected sources and actual sources.

Correct source	Source attribution		
	Russian narrative	English narrative	None (for true and false control items)
Russian narrative	.393 (.275)	.154 (.166)	.109 (.063)
English narrative	.182 (.149)	.482 (.295)	.081 (.061)
None (for true and false control items)	.131 (.114)	.123 (.102)	.747 (.178)

References

- Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology and aging, 19*, 290–303.
- Caldwell-Harris, C.L. (2014). Emotionality differences between a native and foreign language: Theoretical implications. *Frontiers in Psychology, 5*, 2-4.
- Calvillo, D.P., & Mills, N.V. (2018). Bilingual witnesses are more susceptible to misinformation in their less proficient language. *Current Psychology, 39*, 673–680.
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of auto-mobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589
- Ruffman, T., Rustin, C., Garnham, W., & Parkin, A. J. (2001). Source monitoring and false memories in children: Relation to certainty and executive functioning. *Journal of Experimental Child Psychology, 80*, 95–111.

ERP decoding shows bilinguals represent the language of a code-switch *after* lexical processing
Anthony Yacovone, Moshe Poliak, Harita Koya, & Jesse Snedeker (Harvard University)
anthony_yacovone@g.harvard.edu

Background. For decades, research using ERPs has revealed how and when comprehenders respond to unexpected linguistic material. For example, N400 effects often occur after hearing an unexpected word, e.g., I like my coffee with cream and **salt**.¹ N400 effects can also occur after hearing an unexpected switch into another language or *code-switching*, e.g., I like my coffee with cream and **azúcar** (sugar in Spanish).² In a recent study, Yacovone and colleagues (2019) tested whether or not these two N400 effects are functionally distinct. To do this, they used spoken English stories with target words that varied in language (English, Spanish), contextual fit (Strong-fit, Weak-fit), or both. They reasoned that, if there were two distinct N400 effects, the weak-fitting code-switches would result in an additive effect. Results indicated that all weak-fitting conditions (regardless of language) elicited the same N400 effect. The strong-fitting code-switches, however, only elicited N400 effects in their most predictable contexts (see **Figure 1**). After initial lexical processes, all Spanish words elicited a late positive complex (LPC) and all weak-fitting words elicited a sustained negativity. Given these findings, the authors concluded three things: 1) code-switches do not elicit a *unique* N400 effect; 2) listeners can predict a particular lexical item (not just semantic features) in highly predictable contexts; and 3) bilinguals only notice that a word is in another language after the N400 time window—thus, after lexical processing.¹

A tempting conclusion. This study demonstrates that the N400 is not sensitive to the language of the unexpected words per se. The only component that differentiated the language of the words was the LPC, which fully emerged *after* the N400. A tempting conclusion from these findings would be that bilinguals do not initially represent the language of words (or detect the language switch) in the earliest stages of lexical processing. However, there are two issues with this conclusion: First, traditional ERP methods cannot disentangle overlapping components; thus, it is possible that early ERP signals of the language switch were present but simply overwhelmed by the robust N400 effects. Second, ERPs cannot tell us anything about the type of information being represented in an individual's neural signals. The core issues are that 1) overlapping components cannot be disentangled, and 2) that the sensitivity of an ERP component does not reveal what information is or is not being processed at any given moment. In order to answer such questions, we would need to use neural decoding techniques.

The present study. We used *information-based* decoding^{5,6} to assess if (and when) information about a word's language is present in each bilingual's neural signal. To do this, we decoded each participants' data separately at each time point between -200 to 2000ms. First, we collapsed all conditions into groups of English and Spanish words. Then, we split a participant's data into a training set and a testing set. The training set was then fed to a support vector machine (SVM) classifier, which used 3-fold cross-validation to create a model of the data. This model was used to predict the language of the words in the testing set given the ERP data. We recorded the accuracy of the model's predictions at each time point (20ms intervals). After analyzing all of the participants' data, we averaged together the decoding accuracies. Finally, we tested the performance of this decoding procedure against chance using a cluster-mass permutation test. Decoding accuracy was significantly above chance at distinguishing between English and Spanish words from 740-1600ms ($t = 126.55$, $p < .001$; see **Figure 2**). This cluster coincides with the LPC effect, which occurred between 750-2000ms in the original study. These results show that information about a word's language is not represented in bilinguals' neural signals until *after* lexical processing. One potential limitation is that not *all* cognitive processes are captured by EEG, leaving the possibility that language representations are present but not observable in the EEG data. Our findings have many implications for bilingualism and language processing research: We show that a bilingual is not someone with two separate and competing languages living in their mind. Rather, bilinguals are individuals with a language system optimized to handle two coding systems, where a single lexical concept can be readily mapped onto two distinct forms.

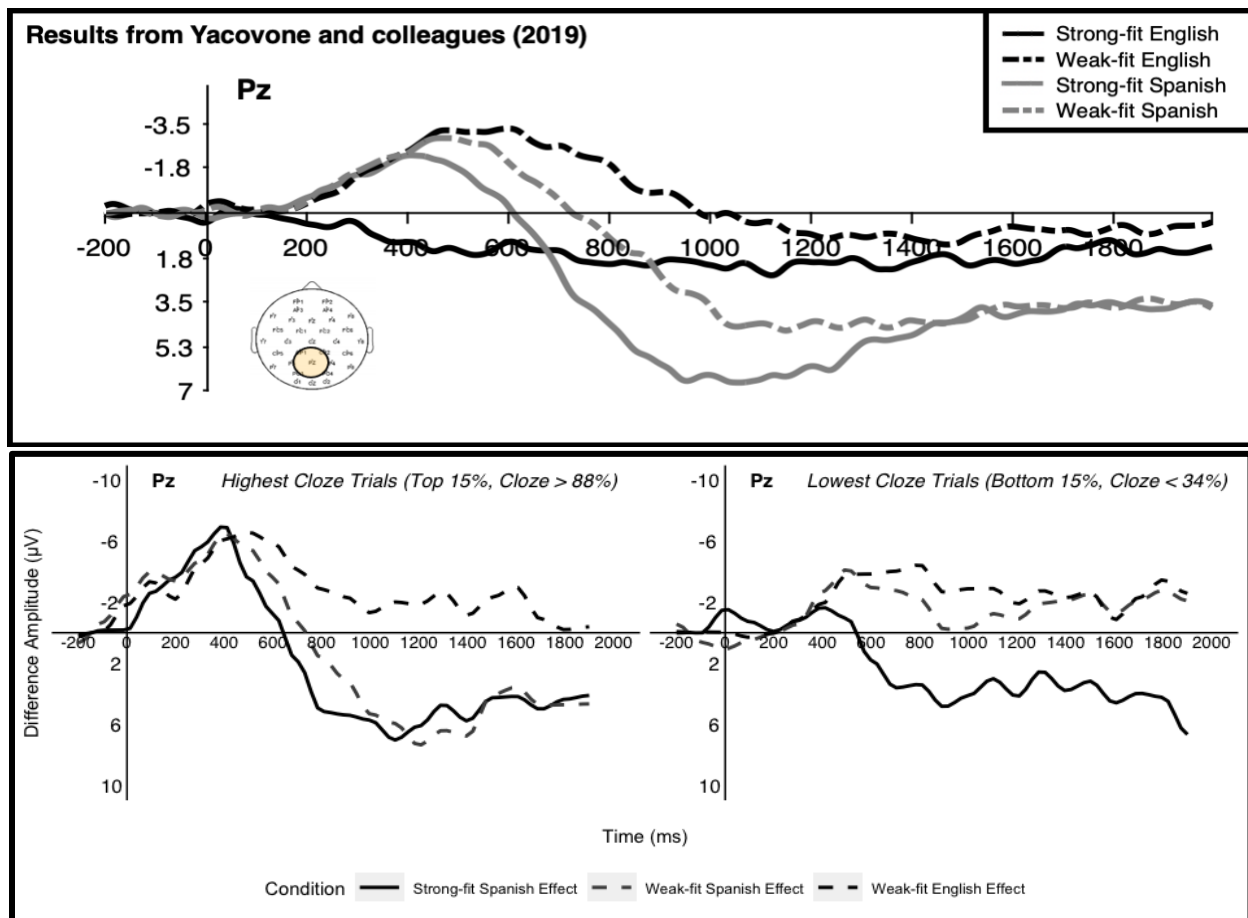


Figure 1: (Top panel) *Grand Averages from Yacovone and colleagues (2019)*—Spanish-English bilinguals heard sentences like “And the wig was so hot and heavy on my **head** / cranium / cabeza / cráneo.” All violations resulted in the same effect during the N400 time window (250-500ms) when collapsing across all trials. All Spanish conditions elicited LPCs (750-2000ms) and all weak-fitting words elicited sustained negativities (550-1300ms). (Bottom panel) When split by cloze probabilities, the N400 effect for the Strong-fitting Spanish condition was prominent in highest cloze trials and absent in the lowest cloze trials. Thus, bilinguals predicted a specific lexical item (the strong-fitting English word) in highly predictable sentences and only semantic features (or nothing) in lowest ones.

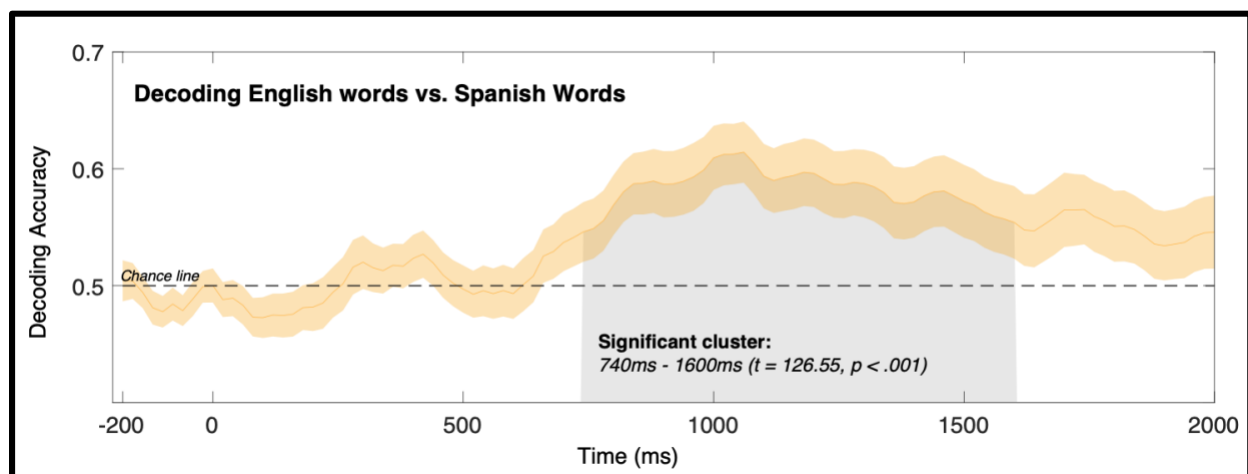


Figure 2: *Results from the information-based decoding technique.* This graph represents the grand average of the decoding accuracies from all 30 participants. Decoding accuracy is at chance for the first 740ms. The cluster in gray represents when decoding accuracy was significantly above chance as indicated by a cluster-mass permutation test with 1000 iterations. This timing of this cluster nicely corresponds to the emergence of the LPC effect, which was the only ERP signature that distinguished between English and Spanish stimuli in the original study.

References: ¹Kutas & Federmeier, 2011; ²Litcofsky & van Hell, 2017; ³Yacovone, Moya, & Snedeker, 2019; ⁴Connolly & Phillips, 1994; ⁵Luck, Bae, & Simmons, 2020; ⁶Bae & Luck, 2019.

The use of pronoun interpretation biases in unbalanced Spanish-English bilinguals: the role of language experience

Carla Contemori & Alma L. Armendariz Galaviz (University of Texas at El Paso)

Recent research on pronoun interpretation has shown that the strength of pronoun interpretation biases in English correlates with comprehenders' print exposure, demonstrating that language experience influences anaphora resolution in adults (e.g., Arnold, Strangmann, Hwang, Zerkle, & Nappa, 2018) and that individual variability among comprehenders may exist (e.g., Arnold, 2015). A question that remains open is how language experience affects anaphora resolution biases in languages other than English. In the present study, we focus on Spanish, a null-subject language where null pronouns typically refer to topic antecedents and overt pronouns refer to non-subject antecedents:

(1) Pedro_i saludó a Carlos_j cuando él_{i/pro_j} cruzaba la calle

Pedro greeted Carlos when he crossed the street

We look at a population of speakers that presents variability and optionality in pronoun interpretation biases, i.e., unbalanced bilinguals whose first language (L1, Mexican Spanish) is a minority language. Importantly, these speakers learn the minority language in the family and do not receive formal school education in the L1. In addition, their dominant language (L2, English) is a non-null subject language, which may interfere in the acquisition of L1 interpretation biases. We recruited seventy-four Spanish-English unbalanced bilinguals with different levels of proficiency in Spanish and we analyze individual factors that may determine variability in pronoun interpretation (i.e., language proficiency measured with a naming task, reading exposure measured with self-reported measures). Sixty-three monolinguals who speak the same regional variety of Mexican Spanish were recruited for the control group.

We used a sentence comprehension task where participants had to choose the referent of an ambiguous null/explicit pronoun in anaphoric or cataphoric position (Table 1). The comprehension question included a subject referent interpretation for the pronoun, an object referent interpretation and the external referent interpretation (i.e., someone else).

First, we compared the subject antecedent interpretations in bilingual and monolingual speakers, using mixed-effects logistic regression. The results showed that bilinguals chose the subject antecedent significantly more often than monolingual speakers for anaphoric and cataphoric pronouns ($p < .0001$), and for null and explicit pronouns ($p < .0001$). The strong subject preference for all pronoun types is a new result found in Spanish-English bilinguals, indicating high variability in pronoun preferences in this population, and cross-linguistic interference from English.

We analyzed the bilingual data separately, to investigate the effects of individual variables on pronoun interpretation in the L1, using mixed-effects logistic regression. The analysis of the bilingual data revealed a Reading Exposure*Pronoun Position interaction indicating that bilinguals who read more in Spanish chose fewer subject-antecedent interpretations for cataphoric pronouns ($p < .0001$), approaching the monolingual pattern of interpretation. A Reading Exposure*Anaphora Type interaction demonstrates that bilinguals who read more in Spanish chose fewer subject-antecedent interpretations for explicit pronouns ($p < .0001$). No significant effect of reading exposure emerged for null and anaphoric pronoun. (Figure 1).

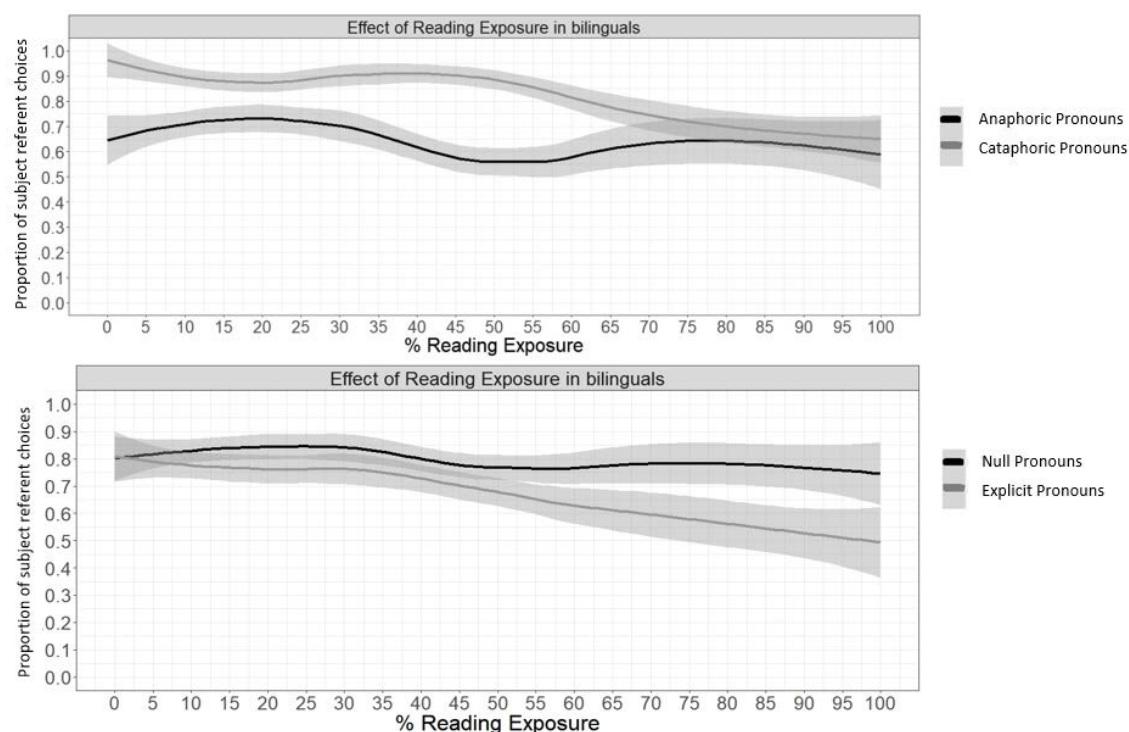
A main effect of Proficiency ($p < .04$) also indicated that bilinguals with higher Spanish proficiency chose overall fewer subject-antecedent interpretations than bilinguals with lower proficiency.

The results demonstrate an interplay between print exposure and proficiency on the acquisition of pronoun interpretation biases in L1 Spanish. Bilinguals who read more and have higher proficiency in Spanish show more monolingual-like pronoun interpretations, demonstrating that pronoun comprehension preferences are acquired by language experience (e.g., Arnold et al., 2018). The effect of reading exposure for pronouns that are more infrequent in the input (explicit pronouns, cataphora) demonstrates that reading exposure can provide discourse input that adds to the development of pronoun interpretation biases in bilinguals lacking L1 literacy.

Table 1. Average subject (=Pedro), object (=Carlos) and external (=someone else) referent pronoun interpretation in unbalanced bilinguals and monolinguals.

	Unbalanced Bilinguals			Spanish monolinguals		
	Subject	Object	External	Subject	Object	External
Anaphora Null Pronoun Pedro greeted Carlos when (he) crossed the street	0.73	0.26	0.01	0.62	0.37	0.02
Anaphora Explicit Pronoun Pedro greeted Carlos when he crossed the street	0.57	0.41	0.02	0.37	0.60	0.04
Cataphora Null Pronoun When (he) crossed the street, Pedro greeted Carlos	0.88	0.05	0.06	0.65	0.06	0.29
Cataphora Explicit Pronoun When he crossed the street, Pedro greeted Carlos	0.86	0.07	0.07	0.47	0.12	0.42

Figure 1. Average subject-antecedent interpretations for anaphoric/cataphoric pronouns (top panel) and null/explicit pronouns (bottom panel) based on average Spanish reading exposure percentage in unbalanced bilingual speakers.



Changing pronoun interpretations across-languages: discourse priming in Spanish-English bilingual speakers

Carla Contemori & Natalia Irene Minjarez Oppenheimer (University of Texas at El Paso)

Different languages have different referential expressions and interpretation biases. For example, in English, referents that are more accessible are usually expressed as pronouns in discourse and pronouns often refer to a subject/first-mentioned referent, which is often the most salient in the previous discourse (e.g., (1) John_i met Paul while he_i was in high school)).

In Spanish, native speakers show a preference for interpreting the null pronoun as referring to the subject antecedent (i.e., John in (1)), while explicit pronouns are more likely to refer to a non-subject antecedent (i.e., an explicit pronoun is interpreted towards the preceding object in (1) about 60% of time).

We know that comprehenders can adapt their pronoun resolution biases to the likelihood of occurrence of a specific type of pronouns in the input. For example, previous research has demonstrated that pronoun resolution biases are sensitive to immediate priming and adaptation in monolingual and bilingual individuals (e.g., Contemori, 2019; Fernandes et al., 2018). While pronoun interpretations can be primed in bilingual speakers using a single-language priming task (e.g., Contemori, 2019), it is not clear if pronoun interpretation biases can be primed cross-linguistically. Cross-linguistic priming effects have been shown at the phonological, semantic and syntactic level in bilinguals, demonstrating cross-language activation and shared abstract representations (e.g., Koosra & Muysken, 2017). However, existing research has not yet investigated the discourse level cross-linguistically using this methodology.

The goal of the present study is to understand if (i) bilingual speakers' statistics about likely referents are independent in the two languages or if (ii) probabilistic inference in tracking referents in one language (Spanish, the L1) can affect how referential expressions are resolved in the other language (English, the L2). In the present study, using a sentence comprehension experiment that implements the priming technique, unambiguous pronouns referring to the second mentioned-referent are presented in Spanish (2) with the aim of decreasing first noun phrase (NP1) interpretations in English (John in (1)) in potentially ambiguous sentences like (3).

(2) Spanish priming sentence: Ana invitó a Alvaro al cine porque él era un buen chavo.

Ana invited Alvaro to the movies because he was a good kid.

(3) Target English sentence/ambiguous pronoun: John met Paul while he was in high school
In a sentence comprehension task adapted from Contemori (2019), forty-five sequential Spanish-English bilinguals read English sentences containing an ambiguous pronoun ((3) and answered comprehension questions (in (3), Who was in high school?). Half of the sentences were preceded by a Spanish sentence that did not contain a pronoun ((4) baseline condition=Al final de el año escolar, Ryan compró un estéreo de Sheila/At the end of the school year, Ryan bought a stereo from Sheila. Who bought the stereo?). The other half of the ambiguous stimuli were preceded by a sentence with an unambiguous pronoun referring to the second-mentioned entity ((2) NP2 priming). The results of the comprehension questions did not show a significant effect of immediate priming (Table 1), demonstrating that bilinguals were as likely to interpret an ambiguous pronoun as referring to the second NP (e.g., Paul in (1)) after encountering a NP2 priming sentence (2) than a baseline sentence (4) ($p=.1$). In addition, no main effect of Order of the Items emerged ($p=.1$), indicating that participants were not adapting to the higher occurrence of Spanish NP2 interpretations when comprehending ambiguous English pronouns. The study shows that the English pronoun interpretation bias is not susceptible to priming from Spanish, suggesting that Spanish-English bilinguals keep separate statistics about probability of pronominal forms interpretations occurring in the two linguistic environments.

Current ongoing research is looking at cross-linguistic referential priming using a different pronominal form in Spanish (i.e., null pronouns) to prime English ambiguous pronoun interpretations to confirm the results of the present study.

Table 1. Proportion of NP1 choices (he=John) for the English sentences with ambiguous pronouns by priming type (SD in parenthesis)

	Spanish-English bilinguals
Baseline condition	0.7 (0.45)
(NP2) Priming condition	0.65 (0.47)
Total average NP1 choices	0.67 (0.46)

References

Contemori C. (2019). Changing comprehenders' pronoun interpretations: immediate and cumulative priming at the discourse level in L2 and native speakers of English. *Second Language Research*. DOI: <https://doi.org/10.1177/0267658319886644>.

Fernandes, E., Luegi, B., Correa Soares, de la Fuente, I. & Hemforth, B. (2018). Adaptation in pronoun resolution: Evidence from Brazilian and European Portuguese. *Journal of Experimental Psychology: Learning Memory and Cognition*. doi: 10.1037/xlm0000569.

Koosra, G., & Muysken, P. (2017). Cross-linguistic priming in bilinguals: Multidisciplinary perspectives on language processing, acquisition, and change. *Bilingualism: Language and Cognition*, 20(2), 215-218.

Similarity-Based Interference in Native and Non-Native Comprehension

Ian Cummings and Hiroki Fujita (University of Reading)

Similarity-based interference has played an important role in informing our understanding of the memory access mechanisms during sentence processing [6,8]. One example of similarity-based interference is observed in subject and object relative clauses (SRCs and ORCs), where the difficulty associated with processing ORCs is attenuated if the two noun phrases in the relative clause are dissimilar (e.g. a proper name and a noun), compared to when they are similar (e.g. two nouns) [3,4]. Such effects are believed to index difficulty in encoding information in memory that is similar along a particular dimension [9]. Although similarity-based interference has been widely studied in native (L1) comprehension, less is known about interference during non-native (L2) processing. L2 processing is generally more difficult than L1 processing, though the precise nature of this difference is debated [1,2,5]. If L2 learners are more susceptible to interference during processing than L1 speakers [2], L2 learners may show larger similarity-based interference effects during processing than L1 speakers.

We examined similarity-based interference in relative clauses in 80 L1 English speakers and 80 L2 English speakers from different L1 backgrounds (upper-intermediate to advanced English L2ers with mean English proficiency 46/60). Participants read sentences as in (1/2) while their eye-movements were monitored and completed an offline comprehension task as in (3/4). The offline task was conducted in a separate experimental session after the main experiment. Experimental items in both tasks manipulated clause type, ORC vs. SRC, and noun similarity, similar (two common nouns, e.g. 'the boy' and 'the girl') vs. dissimilar (one common noun and one proper name, e.g. 'the boy' and 'Rebecca'). At the relative clause, we expected longer reading times in ORCs when the two nouns were similar, as in (1a), than dissimilar, as in (1b) [3,4]. We also investigated processing at the matrix verb, as retrieval of a subject for this verb may be more difficult following ORCs, because the noun inside the relative clause ('the girl' / 'Rebecca') is itself also a subject [7]. For comprehension accuracy, we expected lower accuracy for ORCs with two similar nouns, as in (3a) than dissimilar nouns (3b), but no differences in SRCs (4a/b). If L2ers are more susceptible to interference than L1ers [2], they should show larger similarity-based effects during processing and in offline comprehension.

We pre-registered analyses of first-pass, regression path and total viewing times (<https://osf.io/awxju>). At the relative clause ("that the girl saw" / "that saw the girl"), we found significant interactions between clause type and noun similarity (p s < .033) in regression path times and total viewing times, with reading times being particularly long in ORCs when the two nouns were similar, as in (1a), than dissimilar, as in (1b) (Figure 1). At the matrix verb ('walked'), we did not find longer reading times following ORCs than SRCs in any measure. If anything, SRCs caused difficulty, especially in the dissimilar condition for L2ers, as evidenced by a significant group by clause type by noun similarity interaction in regression path times (p = .004). For comprehension accuracy, we observed a significant main effect of group (p < .001), with lower accuracy in the L1ers, and a significant clause type by noun similarity interaction (p = .005), with lower comprehension accuracy rates for similar than dissimilar ORCs, while the SRC conditions did not differ (Figure 2).

Our results at the relative clause replicate [3,4], indicating ORCs are easier to process when the two nouns are dissimilar, and extend these results to L2 learners. We did not find evidence of ORCs causing processing difficulty at the matrix verb, but noun similarity did influence comprehension accuracy for sentences containing ORCs. While we did find some L1/L2 differences, the pattern of results was not consistent with L2ers being more susceptible to interference than L1ers (cf [2]), and we did not find significant interactions with group at either the relative clause region, or in comprehension accuracy rates. Finding similarity-based interference in both groups suggests L1 and L2 comprehension utilise similar mechanisms when encoding and retrieving information from memory during sentence processing.

Eye-Tracking Experiment Items (n = 24)

(1a) Object Relative Clause, Similar

The boy that the girl saw yesterday afternoon, walked through the park.

(1b) Object Relative Clause, Dissimilar

The boy that Rebecca saw yesterday afternoon, walked through the park.

(2a) Subject Relative Clause, Similar

The boy that saw the girl yesterday afternoon, walked through the park.

(2b) Subject Relative Clause, Dissimilar

The boy that saw Rebecca yesterday afternoon, walked through the park.

Comprehension Task Experiment Items (n = 24)

(3a) Object Relative Clause, Similar

The passenger that the pilot saw before the flight, seemed to be nervous.

(3b) Object Relative Clause, Dissimilar

The passenger that Joseph saw before the flight, seemed to be nervous.

(4a) Subject Relative Clause, Similar

The passenger that saw the pilot before the flight, seemed to be nervous.

(4b) Subject Relative Clause, Dissimilar

The passenger that saw Joseph before the flight, seemed to be nervous.

Who seemed to be nervous?

(The passenger – The pilot / Joseph)

Figure 1. Reading times.

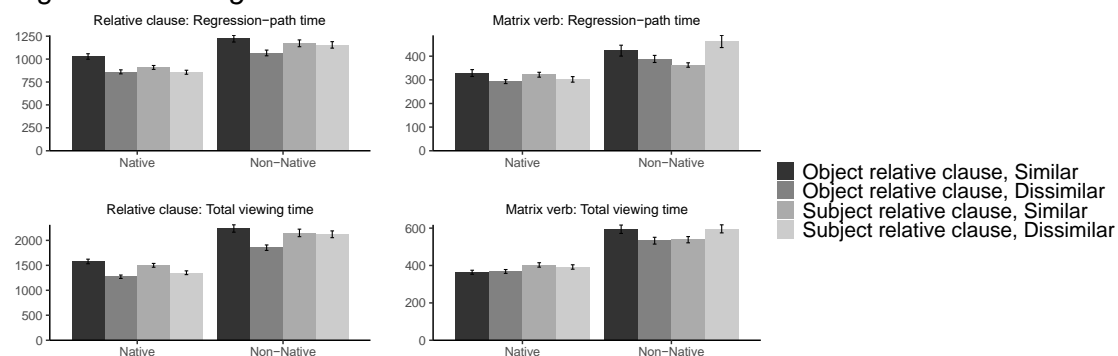
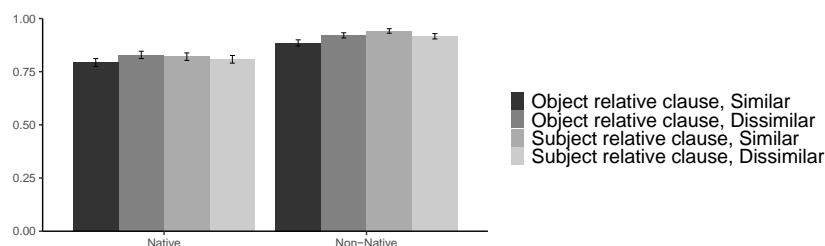


Figure 2. Comprehension accuracy.



References

- [1] Clahsen & Felser (2006). *TiCS*, 10, 564-570; [2] Cunnings (2017). *BLC*, 20, 659-678; [3] Gordon et al. (2001). *JEP: LMC*, 27, 1411-1423; [4] Gordon et al. (2006). *JEP: LMC*, 32, 1304-1321; [5] Hopp (2018). *SL*, 17, 5-27; [6] Jäger et al. (2017). *JML*, 94, 316-339; [7] Van Dyke (2007). *JEP: LMC*, 33, 407-430; [8] Vasishth et al. (2019). *TiCS*, 23, 968-982; [9] Villata et al. (2018). *FiP*, 9, 2.

How do structural predictions operate between languages for multilinguals? Evidence from cross-language structural priming in comprehension

Xuemei Chen & Robert J. Hartsuiker
Ghent University

Many cross-language studies showed structural priming effects: in particular, speakers tended to re-use the prime structure in a target sentence after processing the prime in a different language. This suggests that multilinguals have a syntactic representation that is shared across their languages or separate but interacting representations for each language. However, it is unclear whether multilinguals can rely on such language non-specific representations to predict structure in language *comprehension*.

To answer this question, we conducted two visual-world eye-tracking priming experiments with multilinguals (Cantonese-L1, Mandarin-L2, English-L3). Participants were instructed to read prime sentences in either Cantonese, Mandarin, or English; then they heard a target sentence in Mandarin while looking at the corresponding target picture. The sentences either had a double object (DO) structure (e.g., “Gushou di **You**chai yizhang **You**piao”, the drummer passed the mailman a stamp) or a prepositional object (PO) structure (e.g., “Gushou di **You**piao gei **You**chai”, the drummer passed a stamp to the mailman); Note that in the DO, the verb is followed by the recipient (“Youchai”, mailman), whereas in the PO, the verb is followed by the theme (“Youpiao”, stamp). The priming effect is expressed as the proportion of looks to the predicted referent (i.e., the recipient after a DO-prime, the theme after a PO-prime), for two critical time windows during target sentence processing: the verb and the first syllable of the first post-verbal noun (which was identical in theme and recipient). In Experiment 1 (N=72), we used six prime verbs (see **Table1**) that differed in their bias for DO and PO (verb bias) in each language and four relatively unbiased target verbs in Mandarin. There was within-language structural priming only (from Mandarin to Mandarin, see **Figure1A**). There was no interaction between verb bias and prime structure. In Experiment 2 (N=72), we held the verb in prime and target constant (i.e., the verb was identical between prime and target within Mandarin, shared meaning, orthography and partly phonology in Cantonese and Mandarin, and shared meaning in English and Mandarin). Now there was not only within-language priming but also between-language priming, albeit only from Cantonese to Mandarin (see **Figure1B**).

These results indicated that the structure prediction system between languages in comprehension: 1) is independent, so that prediction errors within a specific language do not generalize to another language; 2) is interactive, so that cognate languages (e.g., Cantonese and Mandarin) show cross-linguistic priming whereas non-cognate languages (e.g., English and Mandarin¹) do not; 3) is at least partly lexically-based, so that cross-linguistic structural priming only occurred with cognate verbs.

Table 1
Structure bias of prime verb in Experiment 1

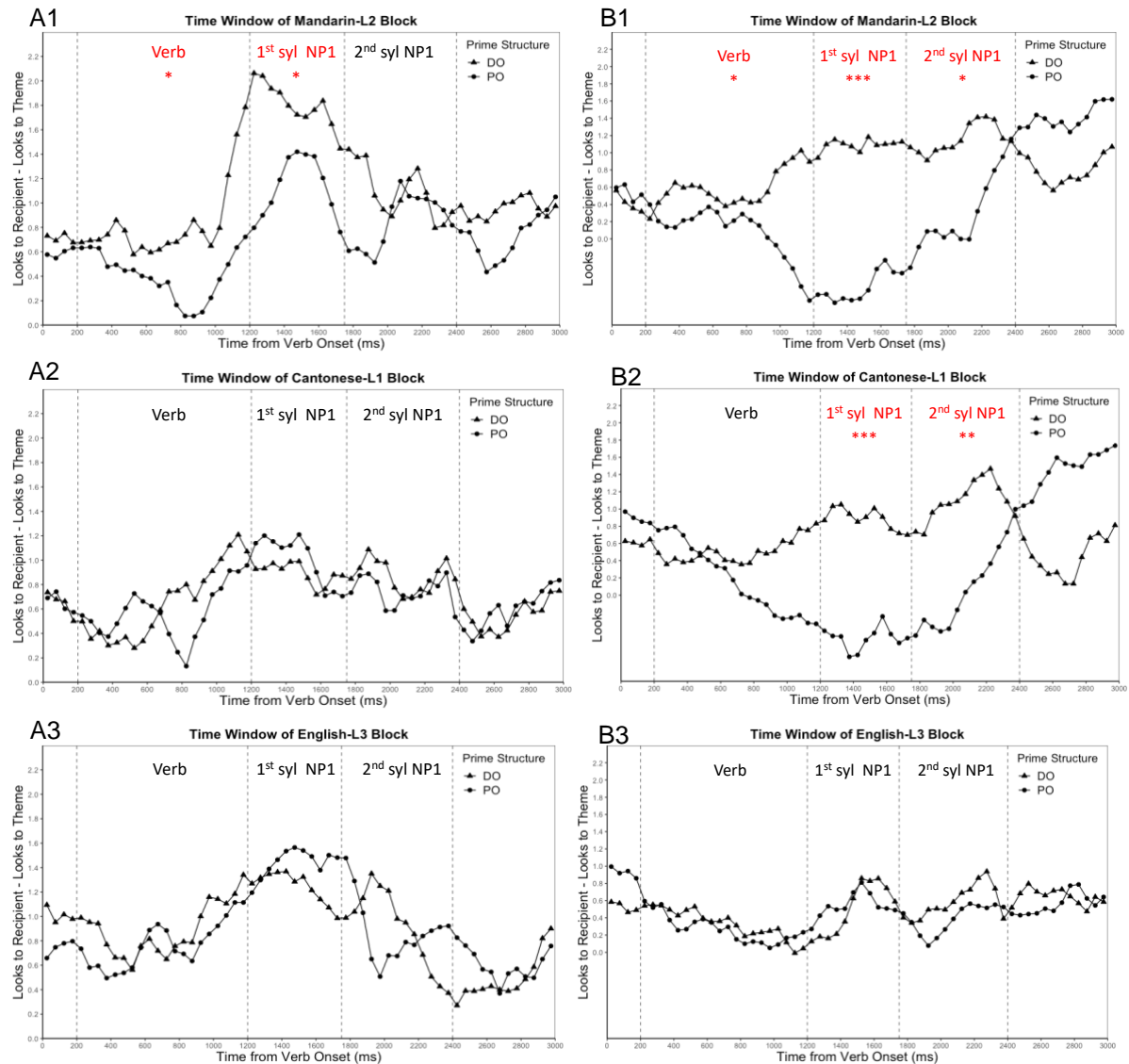
Verb (E)	English Corpus (G&S, 2004)	English Norming	Verb (C/M)	Cantonese Norming	Mandarin Norming
grant	0.69	0.69	赏	-0.65	0.99
award	0.69	0.47	赐	-0.61	1.21
send	-0.56	-2.20	发	-2.71	-1.83
threw	NA	-3.69	丢	-3.09	-2.43
leave	-1.10	-2.08	留	-3.40	-1.54
bring	-2.34	-1.64	带	-3.04	-1.91

¹ Age of acquisition (AOA) of Mandarin was earlier than English; and self-reported proficiency of Mandarin was higher than English.

Note. Structure bias was calculated as the log-odds for the DO responses following the verb divided by the PO responses (i.e., $\log[(\#DO+1)/(\#PO+1)]$, Bernolet & Hartsuiker, 2010; Jaeger & Snider, 2008). Therefore, values larger than 0 indicate a DO-biased verb and values below 0 indicate a PO-biased verb. We chose 11 dative verbs which have the same structure preference in both English Corpus (Gries & Stefanowitsch, 2004) and Mandarin norming data (N=367, Chen et al., 2020). Then we performed a norming study of verb bias in Cantonese (40 native Cantonese speakers) and in English (51 high-proficient Mandarin-English bilinguals). We selected 6 verbs with similar structure bias in Mandarin, Cantonese and English (i.e., 11 dative verbs showed an overall preference of PO, so we selected two less PO-biased verbs; negative values correspond to PO-bias).

Figure 1

Difference in proportion of looks to recipient and theme for each time bin (50ms) from onset of target verb in three language blocks of Experiment 1 and 2



Note. The time window of verb is from 200ms to 1200ms and the time window of the first syllable of the first noun phrase is from 1200ms to 1750ms. The unambiguous time window of the second syllable of first noun phrase is from 1750ms to 3600ms. Six plots indicate the difference in the proportions of looks to recipient (predicting DO structure) and to theme (predicting PO structure) after prime sentences with different structure (DO vs. PO) in Experiment 1 when the prime and target have different verbs (A1, A2, A3 on the left) and in Experiment 2 when prime and target shared the translation-equivalent verbs (B1, B2, B3 on the right). The first two plots (A1, B1) suggest the priming effect for within-language block of Mandarin. The following four plots suggest the priming effect for between-languages blocks of Cantonese-to-Mandarin (A2, B2) and English-to-Mandarin (A3, B3). The red label of time window indicates significant priming effect (* $p < .05$; ** $p < .01$; *** $p < .001$).

What to expect when you are expecting an antecedent: processing cataphora in Dutch

Anna Giskes (NTNU) and Dave Kush (University of Toronto; NTNU)

Background: During incremental processing, the parser cannot fully interpret cataphors like *he* in (1) until their antecedent is encountered. Past research has argued that the parser expects the antecedent in the next available syntactic position, often the main subject [e.g., 1-4]. Evidence comes from Gender- and Number Mismatch Effects (G/N-MME): in manipulations like (2), readers slow down at a gender-mismatching subject NP compared to a matching NP.

(1) When he_i had sown the field, the farmer $_i$ baked pancakes for the children.

(2) When **he/she** had sown the field, the boy...

Previous research is uninformative about how far in advance the parser predictively commits to an antecedent in a specific position. MMEs in (2) are compatible with (i) a parser that predictively builds the antecedent in subject position *before* receiving bottom-up input of the subject. However, MMEs may also reflect a parser that (ii) waits until *after* encountering a subject NP to posit coreference (but before gender features are processed bottom-up) [3]. Previous studies do not allow us to tease these two options apart, because MMEs occur at/after the subject NP. We constructed a test of the two hypotheses in Dutch, a V2 language with subject-verb number agreement. We reasoned that if the parser predictively commits to and builds an antecedent in main clause subject position, this should trigger a prediction of matching number agreement on the main verb. Because Dutch is V2, the finite verb will precede the main subject in sentences with fronted adjunct phrases like (1). We therefore looked for N-MMEs between a cataphor and the main verb as evidence for advance prediction of the subject.

Self-paced reading experiments: (exp 1: $n=80$; exp 2: $n=160$) We manipulated number-match between the main clause subject and a cataphor in a fronted adjunct clause (Table 1: main clause verb *bakte* underlined). In a control we replaced the finite subordinate clause with a participial clause without a cataphor. The participial clause was ambiguous regarding the number of its implicit 'PRO' subject, thus providing a baseline without an expectation for number. In experiment 2, we added 10 separate items manipulating the gender of the main clause subject (underlined in Table 1), in order to replicate the classic G-MME.

Results See Fig. 1-3. In both SPR experiments, the number manipulation did not yield a significant mismatch effect at the verb or in the spillover regions (maximal LMEMs on log-transformed RTs; for all models, $t < 1.5$). The largest trend towards a N-MME (19 ms) was observed in experiment 2 at the critical main verb ($t = 1.36$). In contrast, we observed a large G-MME in the spillover region for the gender manipulation (85 ms GGME, $t = 5.98$).

Conclusion: The absence of significant NMMEs at the V2 verb suggest that cataphors do not trigger an 'early' prediction of a matching NP in main subject position. The strong G-MME in the same studies suggest that participants still had strong expectations for an antecedent in main subject position. These results are consistent with a parser that does not make a predictive syntactic commitment to locate an expected antecedent in subject position. They are also in line with a parser that does predict the subject to some extent, but does not execute all knock-on consequences following from that prediction. Furthermore, is possible that the (degree of) prediction varies for number- and gender features, in line with relatively small and late N-MMEs in previous research [2].

The results suggest that at least some active parsing strategies triggered by long-distance dependencies do not reliably entail predictive building of syntactic structure.

References: [1] Kazanina, N. et al. (2007). *JML* 56.3 (2007): 384-409. [2] Van Gompel, R.P. & Livsedge, S.P. (2003). *JEP: Learning, Memory, and Cognition*, 29(1), 128. [3] Drummer, J.D. & Felser, C. (2018). *JML*, 101, 97-113. [4] Kush & Dillon, B. (2020). OSF Preprints, 14 Aug 2020.

Table 1: SPR item set (24 items) for SPR (exp. 1&2). Critical regions underlined. Items were counterbalanced for main verb number (1&2) and subject gender (2). Items for exp. 2 had one additional spillover region following the main clause verb (*the extremely | friendly | farmer...*)¹

Number-match/ mismatch	Nadat <u>hij/zij</u> de akker had/hadden ingezaaid, <u>bakte</u> de vriendelijke boer pannenkoeken voor de kinderen. After <u>he/they</u> the field had.SG/PL sown, <u>baked.SG</u> the friendly farmer pancakes for the children. <i>After he/they had sown the field, the friendly farmer baked pancakes for the children.</i>
PRO	Na de field te hebben ingezaaid, <u>bakte</u> de vriendelijke boer... After the field have.INF sown, <u>baked.SG</u> the friendly farmer... <i>After having sown the field, the friendly farmer baked...</i>
Gender-match/ mismatch (only exp. 2)	Nadat <u>hij/zij</u> de vliegtickets had gekocht, schreef <u>Diana</u> meteen de datum van haar/Jans aankomst op. After <u>he/she</u> the airline tickets had.SG bought, wrote.SG <u>Diana</u> immediately the date of her/Jan's arrival up.

Figure 1

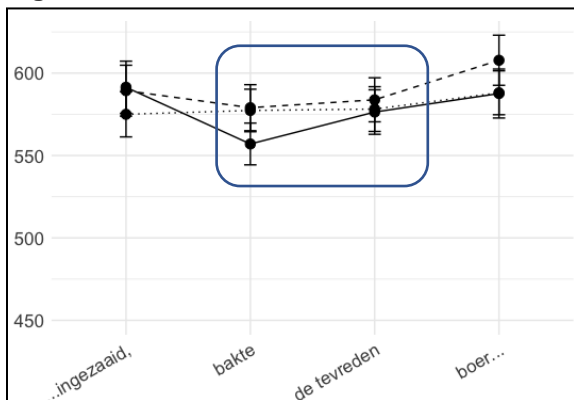


Figure 2

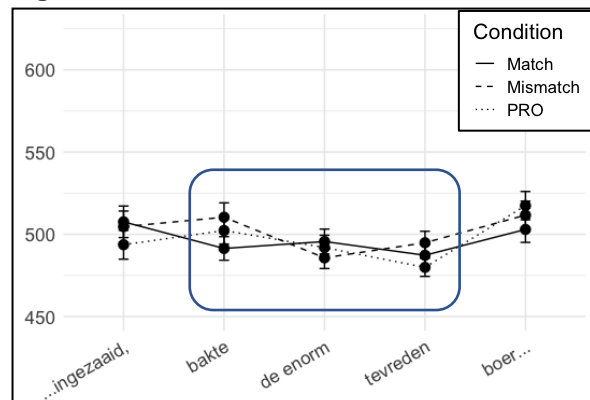


Figure 3

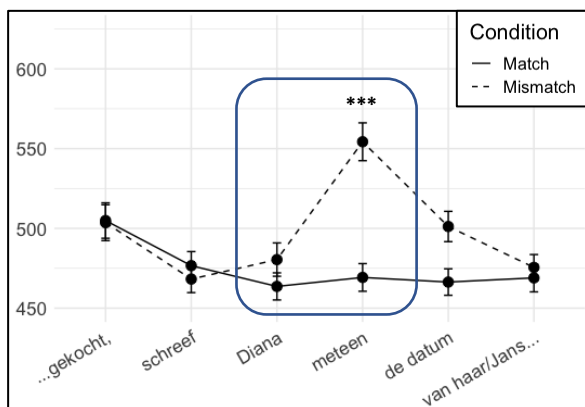


Figure 1: Average RTs + se for experiment 1. Analyzed regions (main verb *bakte* + spillover region) in the boxed area.

Figure 2: Average RTs + se for the number manipulation of experiment 2. Analyzed regions (main verb *bakte* + 2 spillover regions) in the boxed area.

Figure 3: Average RTs + se for the gender manipulation of experiment 2. Analyzed regions (main subject *Diana* + 1 spillover region) in the boxed area.

¹ The Dutch pronoun *zij* is ambiguous between sing-fem, and pl (both masc. and fem.) The number on the auxiliary in the embedded clause (*had/hadden*) disambiguates the pronoun.

The COMP-trace effect and sentence planning: Evidence from L2

Boyoung Kim (KAIST) and Grant Goodall (UC San Diego)

McDaniel et al. (2015) propose that the COMP-trace phenomenon, illustrated in (1), stems ultimately from properties of sentence planning. Their account has three basic components:

- The clause is the default major planning unit and filler-gap dependencies across clauses, as in (1), are inherently difficult.
- Reduced clauses (e.g. without *that*) can more easily be part of the matrix clause planning unit, so filler-gap dependencies into them are less difficult.
- Complex material is dispreferred at the beginning of a planning unit (Principle of End Weight). Gaps are highly complex, so strongly dispreferred in this position.

(1c) is bad because the filler-gap dependency extends into a *that*-clause, a separate planning unit by default, and because the gap is at the beginning of that planning unit; (1d) is better because the gap is not at the beginning. In (1a-b), the embedded clause (without *that*) can be part of the matrix clause planning unit, so in neither case is the gap at the beginning of that unit.

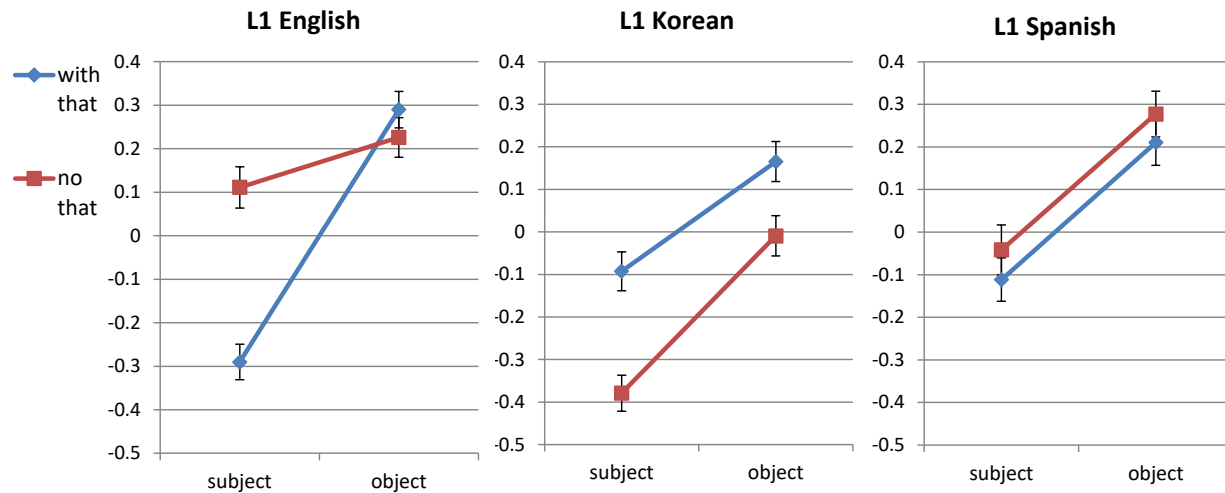
Under this account, speakers with a low capacity for sentence planning will be unable to do cross-clausal planning even with reduced clauses, so embedded subject gaps will always be at the beginning of a planning unit and will thus be severely degraded. Do such speakers exist? L2 speakers are likely candidates. Relatively little is known about their sentence planning in particular (Konopka et al. (2018)), but their sentence production in general is widely thought to be less efficient than that of L1 speakers (Runnqvist et al. (2011)). If this affects their sentence planning as well, as seems plausible, then the McDaniel et al. analysis predicts that they will find both (1a) and (1c) severely degraded in acceptability relative to (1b) and (1d), respectively.

The experiments: We test the prediction by means of three formal sentence acceptability experiments. Experimental stimuli were identical in all three. They consisted of long-distance *wh*-questions in a 2x2 design, crossing the factors THAT (+/-) and GAPSITE (subject vs. object), as in the samples in (1). Subjects saw 5 tokens of each condition, along with 81 filler items, resulting in 101 total stimuli (fully counterbalanced and pseudo-randomized) and rated them on a scale from 1 (“very bad”) to 9 (“very good”). The participants were 72 L1 English speakers (Exp. 1) and two groups of L2 English speakers: 72 L1 Korean speakers (Exp. 2) and 49 L1 Spanish speakers (Exp. 3). All L2 participants immigrated to the U.S. between ages 6 and 15 and had lived in the U.S. for at least 7 years.

The results, transformed to z-scores, are presented below. In the L1 English group, a linear mixed effects model reveals a significant interaction between THAT and GAPSITE ($p \leq 0.001$; lmerTest), indicating a classic COMP-trace effect. In the L1 Korean and L1 Spanish groups, there is no such interaction (Korean: $p = 0.219$; Spanish: $p = 0.946$). Instead, there is a significant main effect for GAPSITE (Korean: $p \leq 0.001$; Spanish: $p \leq 0.001$; cf. L1 English: $p = 0.067$), resulting from subject gaps being uniformly worse than object gaps, regardless of the presence or absence of *that*. In the two L2 groups, then, there is a “subject effect” rather than a COMP-trace effect.

Discussion: These results are exactly as predicted, under the plausible assumption that L2 speakers are less able to plan filler-gap dependencies into an embedded clause, regardless of the form of that clause. Why are these results of interest? First, they document a very clear and initially mysterious contrast between L1 and L2 speakers in the extent to which they allow gaps in embedded clauses. Second, they provide tentative evidence for an analysis of the COMP-trace phenomenon in which the effect derives from basic properties of sentence planning, as we have seen. Many questions remain, but the fact that L2 speakers show the “subject effect” that the McDaniel et al. analysis would seem to predict for speakers with a reduced capacity for cross-clausal planning is intriguing and worthy of further exploration.

- (1)a. Who do you think [__ saw Mary] ?
 b. Who do you think [Mary saw __] ?
 c. *Who do you think [that [__ saw Mary] ?
 d. Who do you think [that [Mary saw __] ?



References

- Cowart, W. and McDaniel, D. (to appear). The *that*-trace effect. In G. Goodall (ed.), *The Cambridge Handbook of Experimental Syntax*.
- Konopka, A., A. Meyer, T. A. Forest (2018). Planning to speak in L1 and L2. *Cognitive Psychology*, 102:72-104.
- McDaniel, D., McKee, C., Cowart, W., & Garrett, M. F. (2015). The role of the language production system in shaping grammars. *Language*, 91(2), 415-441.
- Runnqvist, E., Strijkers, K., Sadat, J., & Costa, A. (2011). On the temporal and functional origin of L2 disadvantages in speech production: A critical review. *Frontiers in Psychology*, 2, 379.

Prominence guides incremental interpretation: Lessons from obviation in Ojibwe

Christopher Hammerly (University of Minnesota), Adrian Staub, & Brian Dillon (UMass Amherst)

Existing work has shown that *animate* nouns are more likely to be predictively encoded as agents compared to *inanimate* nouns under incremental ambiguity [1,2,3]. The present study investigates how a previously unexplored type of “prominence” information, *obviation*, affects argument structure processing. Obviation organizes *animate third persons* according to their discourse prominence: The noun that refers to the entity “in the spotlight” is designated PROX(IMATE), while all others are marked OBV(IATIVE). Like animacy, obviation can be described through the *Person-Animacy Hierarchy* (1; PAH). The question explored here is whether the PAH is *generally* employed such that **higher ranked nouns are more likely to receive predictive agent interpretations**. Using a visual world paradigm that allows interpretations to be incrementally probed, we ask if the PAH is recruited in Border Lakes Ojibwe, an Algonquian language of Ontario, to process argument structure. We show that **PROXIMATE arguments are predictively interpreted as agents** in an analogous fashion to what has been claimed for animate nouns.

The critical stimuli (2) are RCs crossed by two factors: HEAD obviation (PROX/OBV) and VOICE (DIR/INV). To interpret the sentences, the combination of obviation and voice must be used. DIRECT (-aa) indicates PROX acting on OBV, and INVERSE (-igo) the reverse. 32 experimental sentences were interspersed with 16 fillers. Sentences were recorded by a speaker of Ojibwe and played auditorily. The sentences include a critical period of ambiguity where the obviation of the head noun has been encoded, but the disambiguating voice information has not yet been encountered. The question is **whether listeners make assumptions about the thematic role of the head noun** during this period. 16 speakers of Ojibwe participated in a visual world task schematized in (3). Participants first saw a fixation cross, followed by three visual stimuli. Two of the images were role-reversals, where the head noun was either the agent or patient. A third distractor image depicted the same action but excluded the head noun. After familiarization, a sentence played. Participants then selected the image associated with their final interpretation via a touch screen. During the trial, a webcam recorded gaze direction, which was used to observe which image participants looked at as the sentence unfolded to determine incremental interpretation.

The ROI is the period of ambiguity. Look proportions towards each image collapsed across levels of VOICE (which has not been encountered) are in (4). The analysis consisted of a series of cluster-based permutation tests [7]. The main comparison was between looks towards agent versus patient images. There was an effect of HEAD ($p = .005$), with contrasts showing a cluster of significance ($p = .013$) such that increased looks towards the agent image occurred following proximate heads, but no differences following obviative heads. The findings support the hypothesis that **PROXIMATE nouns are incrementally interpreted as agents under ambiguity**.

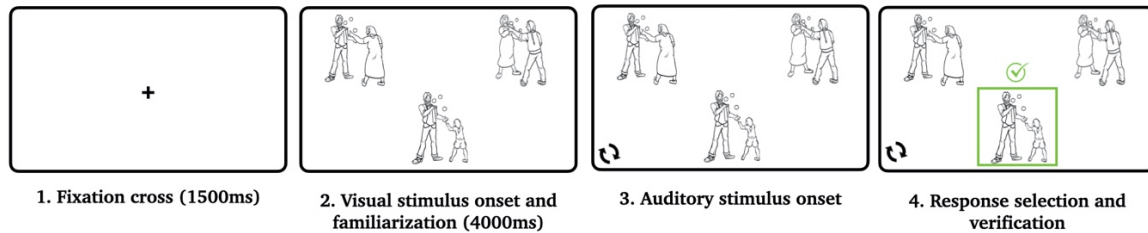
A logistic mixed effects model on picture selection accuracy (5) revealed a main effect of HEAD ($p < .001$) such that proximate is more accurate than obviative, and an interaction between HEAD and VOICE ($p < .001$) such that inverse was associated with increased accuracy with obviative heads, and decreased accuracy with proximate. The main effect of obviation is consistent with a passive-like analysis of the inverse (e.g. [4]), where proximate patients are promoted to subject position. This leads to increased accuracy via the “Subject Gap Advantage” [e.g. 5], as proximate nouns always occupy the syntactic subject position. The interaction between HEAD and VOICE is interpreted as an *agent-first preference*: Assign the agent role before non-agentive roles [e.g. 6]. When voice is congruent with the head being the agent, accuracy is high as reanalysis is not necessary. This also suggests an analysis of the *lack of looking preference with obviatives*: There is a conflict between a patient encoding based the PAH, and an agent encoding based on the agent-first preference—these preferences cancel out. This differs with proximates, where both the PAH and agent-first preference point towards agent encodings. The findings support a model where prominence effects are unified under the PAH, providing an explanation for why the same types of effects appear with different types of prominence information (i.e. animacy, obviation) and across a typologically diverse set of languages (e.g. Indo-European, Algonquian).

(1) 1/2 (PARTICIPANTS) > 3 (PROXIMATE) > 3' (OBLIVATIVE) > 0 (INANIMATE)

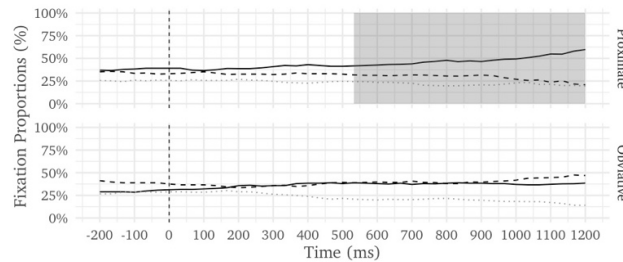
(2) a. ... **gichi-aya'aa** gaa-baapi' **-aa/-igo** -d inini -wan
 ...**elder.PROX** REL-laugh **-DIR/-INV** -3 man -OBV
 '...the elder (PROX) who is {laughing at the man/being laughed at by the man}'

b. ... **gichi-aya'aa -n** gaa-baapi' **-aa/-igo** -d inini
 ...**elder** **-OBV** REL-laugh **-DIR/-INV** -3 man.PROX
 '...the elder (OBV) who the man {is laughing at/is being laughed at by}'

(3) *Outline of task. Images were randomly generated in the left, right, or bottom of the screen. Initial responses could be changed, with final responses registered by pressing the check mark. Sentences could be repeated by pressing the icon in the lower left corner.*



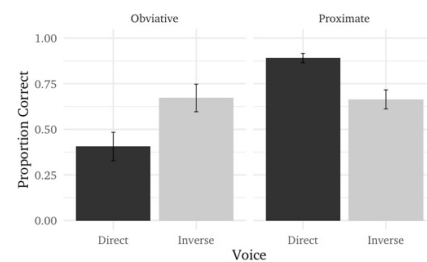
(4) *Critical ROI looking results*



Main Effect of Head	Cluster (ms)	CMS (z)	p-value
Agent v. Patient	433–1200	55.55	*0.005
Distractor v. Agent	—	—	—
Distractor v. Patient	0–1200	29.02	0.078

Contrast	Head	Cluster (ms)	CMS (z)	p-value
Agent v. Patient	Proximate	533–1200	48.54	*0.013
	Obvative	—	—	—
Distractor v. Agent	Proximate	0–1200	−112.39	*0.001
	Obvative	367–1200	−74.17	*0.009
Distractor v. Patient	Proximate	0–133	−7.60	0.185
		267–933	−38.52	*0.010
	Obvative	0–100	−5.96	0.221
		267–1200	−96.87	*< 0.001

(5) *Picture selection results*



Effect	z	p-value
HEAD	3.39	*< 0.001
VOICE	0.60	0.548
HEAD:VOICE	3.67	*< 0.001

[1] Gennari & McDonald (2008) Semantic indeterminacy in object relative clauses. [2] Wagers & Pendleton (2016) Structuring expectation: Licensing animacy in relative clause comprehension. [3] Wagers et al. (2018) Grammatical licensing and relative clause parsing in a flexible word-order language. [4] Bruening (2005). The Algonquian inverse is syntactic: Binding in Passamaquoddy. [5] Kwon et al. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in Korean. [6] Bornkessel-Schlesewsky & Schlewsky (2009). The role of prominence information in the real-time comprehension of transitive constructions: a cross-linguistic approach. [7] Barr et al. (2014) Using a voice to put a name to a face.

Interference and Filler-Gap Dependencies in Native and Non-Native Comprehension

Hiroki Fujita and Ian Cunnings (University of Reading)

The resolution of filler-gap dependencies, as in (1), where the displaced filler ('the beer') must be interpreted as the complement of 'drank' for successful comprehension, has been widely examined in native (L1) and non-native (L2) sentence processing. Both L1 and L2 speakers 'actively' fill gaps at the first available position during processing and use syntactic constraints to guide when a dependency can be formed [5,8,9]. This study extends on this research by examining whether L1 and L2 readers are susceptible to interference during dependency resolution. Cue-based parsing predicts that dependency resolution utilises a cue-based retrieval mechanism that is susceptible to interference (for review, see [7,10]). [4] reported interference in filler-gap dependencies in L1 readers, but whether L2 readers also exhibit such interference is not yet known. Finding increased difficulty in dependency resolution for L2ers would be compatible with the Shallow Structure Hypothesis of L2 processing [1,2]. The claim that L2ers are more susceptible to interference [3] as a result of how they weight structural and semantic retrieval cues, would also predict L1/L2 differences in the resolution of filler-gap dependencies.

80 L1 English speakers and 80 L2 English speakers from different L1 backgrounds (upper-intermediate to advanced English L2ers with mean English proficiency 46/60) read sentences like (1/2) as their eye-movements were monitored. Sentences manipulated the plausibility of both the retrieval target ('the beer'/'the cake') and a linearly closer distractor ('the wine'/'the food'). In a separate experimental session after the main experiment, participants also completed an offline comprehension task as in (3/4), which manipulated the plausibility of a distractor ('the cake'/'the milk') in sentences either with a filler-gap dependency (3) or without (4). For eye-tracking, we expected longer reading times at 'drank' in implausible (2) than plausible (1) sentences. Interference was expected, such that implausible sentences should have shorter reading times when the distractor is plausible, as in (2a), than implausible, as in (2b) [4]. If L2ers are more susceptible to interference, they should show a larger difference between plausible and implausible distractor conditions. For the offline task, we expected interference in filler-gap dependency conditions only, with lower accuracy in (3a) than (3b), but no differences between (4a/b).

In a pre-registered analysis (<https://osf.io/5up4f>), we analysed first-pass, regression-path and total viewing times at the critical verb ('drank') and spillover region ('during the party'). Reading times were significantly longer for implausible than plausible sentences in regression-path and total viewing times ($p < .001$). We observed a significant plausibility by distractor interaction in regression path times ($p = .003$), where reading times for implausible sentences were significantly shorter ($p < .001$, estimated difference 45ms [19ms, 72ms]) when the distractor was plausible (see Figure 1). Although this effect was most clearly visible at the spillover region, the relevant interaction was not significant ($p = .054$). We did not find evidence of significantly more interference in L2ers in any measure. In the comprehension data (see Figure 2), we observed significant main effects of group ($p = .002$), with higher accuracy in the L2ers, and distractor ($p < .001$), with lower accuracy when the distractor was plausible. Additional (non pre-registered) analyses also indicated that individual differences in L2 proficiency, lexical processing ability (see [6]) or L1 background (wh-movement vs wh-in-situ L1) did not significantly influence the interpretation of our L2 results.

The eye-tracking results replicate and extend [4], indicating retrieval interference during L1 and L2 processing of filler-gap dependencies. In the offline task, we did not find the expected interference pattern in dependency conditions only, and interpret these results as suggesting interference in dependency and no dependency conditions during the post-trial comprehension question phase. Although we did not find evidence of increased interference in L2 as compared to L1 processing (cf. [3]), our results suggest both L1 and L2 readers utilise a cue-based memory retrieval mechanism that combines structural and semantic cues during sentence processing.

Eye-Tracking Experiment Items (n = 24)

(1a) *Plausible Target, Plausible Distractor*

Mary saw the beer that the man with the wine very happily drank during the party.

(1b) *Plausible Target, Implausible Distractor*

Mary saw the beer that the man with the food very happily drank during the party.

(2a) *Implausible Target, Plausible Distractor*

Mary saw the cake that the man with the wine very happily drank during the party.

(2b) *Implausible Target, Implausible Distractor*

Mary saw the cake that the man with the food very happily drank during the party.

Comprehension Task Experiment Items (n = 24)

(3a) *Filler-Gap Dependency, Plausible Distractor*

Kevin saw the sandwich that the boy by the cake quickly ate during lunch.

(3b) *Filler-Gap Dependency, Implausible Distractor*

Kevin saw the sandwich that the boy by the milk quickly ate during lunch.

(4a) *No Dependency, Plausible Distractor*

Kevin saw the boy by the cake who quickly ate the sandwich during lunch.

(4b) *No Dependency, Implausible Distractor*

Kevin saw the boy by the milk who quickly ate the sandwich during lunch.

What did the boy eat during lunch? (The sandwich / The cake)

Figure 1. *Reading times.*

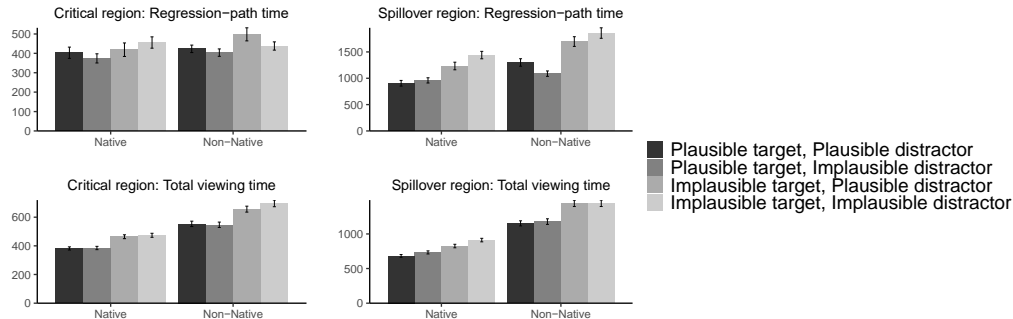
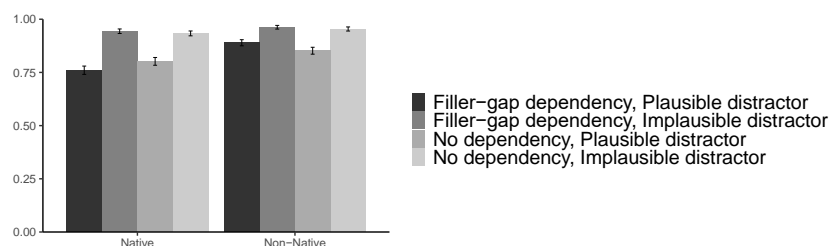


Figure 2. *Comprehension accuracy.*



References

- [1] Clahsen & Felser (2006). *TiCS*, 10, 564-570; [2] Clahsen & Felser (2018). *SSLA*, 40, 693-706; [3] Cunnings (2017). *BLC*, 20, 659-678; [4] Cunnings & Sturt (2018). *JML*, 102, 16-27; [5] Felser et al. (2012). *SSLA*, 34, 67-98; [6] Hopp (2018). *SL*, 17, 5-27; [7] Jäger et al. (2017). *JML*, 94, 316-339; [8] Omaki & Schulz (2011). *SSLA*, 33, 563-588; [9] Williams et al. (2001). *AP*, 22, 509-540; [10] Vasishth et al. (2019). *TiCS*, 23, 968-982.

Inference in the processing of complement control: an eye-tracking study on lexically determined long-distance dependencies

Iria de-Dios-Flores^{1,2}, Juan Carlos Acuña-Fariña¹, Simona Mancini² and Manuel Carreiras^{2,3,4}

¹*Universidade de Santiago de Compostela*, ²*Basque Center on Cognition, Brain and Language*,

³*Ikerbasque, Basque Foundation for Science*, ⁴*Euskal Herriko Unibersitatea*

Background: This investigation focuses on the resolution of lexically-driven anaphoric dependencies in Spanish complement control constructions. This dependency, illustrated in Table 1, involves an interpretative relation between the null subject of the non-finite clause (PRO) and its antecedent: the subject or the object of the matrix clause, depending on certain lexico-semantic properties of the matrix clause verbs (e.g. *promise* = subject control, *order* = object control). Previous eye-tracking studies have contended that whereas control information is immediately accessed and used to retrieve an antecedent, distance effects also influence antecedent selection processes at the point of dependency formation (Betancort et al. 2006; Kwon and Sturt 2016). In these works, object control dependencies were found to be processed faster at the infinitive region, which was interpreted as evidence for a recency effect (or locality advantage). Furthermore, other studies have respectively shown that adjunct control dependencies and subject nominal (rather than verbal) control dependencies exhibit interference effects by irrelevant but feature matching antecedents (Parker et al., 2015; Sturt & Kwon 2015). Here we replicate previous works by examining whether object control dependencies are facilitated over subject control ones at the point of retrieval (the infinitive verb) due to a locality advantage. Furthermore, by fully crossing the type of control verb and the gender of the NPs in the matrix clause we are able to investigate whether the integration of the embedded adjective is subject to facilitatory and/or inhibitory interference effects in both subject and object control dependencies.

Method (n=48): The effects of the experimental factors –CONTROL, GRAMMATICALITY and DISTRACTOR– on the different eye-tracking measures are analyzed in five regions using LMEM: the NP2, infinitive verb, the adverb the adjective, and PP following the adjective. The materials consisted of 96 item sets like the one in Table 1.

Results: No differences between subject and object control dependencies were found at the infinitive verb. Significant interactions between the three experimental factors were found in first-pass times of the adjective region (Figure 1) and the PP region (Figure 2). An interaction between GRAMMATICALITY and DISTRACTOR was found in go-past times at the PP (Figure 3).

Discussion: First, in contrast with the results from previous works, in this study we found no evidence for a facilitation effect for object control dependencies. Instead, the two types of sentences were read similarly at the NP2, the infinitive and the adverb region. This discrepancy with previous works is possibly due to a confound identified in the materials by Betancort et al. (2006) and differences between control nominals (used in Kwon and Sturt 2016) and control verbs. Second, the significant interactions indicate that control-irrelevant antecedents are temporarily considered during the adjective's integration. The effect found in first-pass times of the adjective region (Figure 1) is suggestive of inhibitory interference processes in subject control sentences. The effect found in first-pass times of the PP region (Figure 2) is consistent with facilitatory interference processes in subject control sentences. Effects for facilitatory interference processes for both types of dependencies are only found in the go-past times of the PP region (Figure 3). Furthermore, the lack of grammaticality effects independently of the type of distractor (match/mismatch) appears to indicate that there is a tradeoff between grammatical sensitivity and facilitatory interference. These findings show that verbal control dependencies are also affected by interference effects and, what is more interesting, these effects emerge for both types of control structures. Nonetheless, the fact that interference effects appear earlier and more pervasively in subject control sentences seems to indicate the proximity of the NPs with respect to the adjective plays a role in the adjective's integration.

Table 1: Experimental materials*

Subject control		
Gram.	D. Match	María _i prometió a Cristina _j PRO _i ser mucho más ordenada con los apuntes del instituto.
	D. Mismatch	María _i prometió a Francisco _j PRO _i ser mucho más ordenada con los apuntes del instituto.
Ungr.	D. Match	Antonio _i prometió a Cristina _j PRO _i ser mucho más ordenada con los apuntes del instituto.
	D. Mismatch	Antonio _i prometió a Francisco _j PRO _i ser mucho más ordenada con los apuntes del instituto.
Object control		
Gram.	D. Match	María _j ordenó a Cristina _j PRO _j ser mucho más ordenada con los apuntes del instituto.
	D. Mismatch	Antonio ordenó a Cristina _j PRO _j ser mucho más ordenada con los apuntes del instituto.
Ungr.	D. Match	María _j ordenó a Francisco _j PRO _j ser mucho más ordenada con los apuntes del instituto.
	D. Mismatch	Antonio _j ordenó a Francisco _j PRO _j ser mucho más ordenada con los apuntes del instituto.
<i>NP1 promised/ordered NP2 PRO to be much more organized with the notes from high school.</i>		

*Note that *María* and *Cristina* are feminine names and *Antonio* and *Francisco* are masculine names. In this example, the sentences become ungrammatical when the feminine adjective *ordenada* (*organized*) does not agree in gender with the appropriate antecedent of the null subject (PRO). The regions of interest are underlined in the English translation at the bottom of the table.

Figures: The y-axis represents the transformed RTs for the different eye-tracking measures. The power transformation was determined using the Box-Cox procedure. Asterisks indicate significant post-hoc contrasts after applying Hochberg's correction.

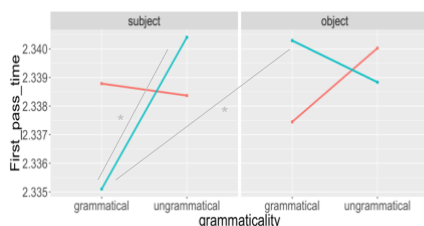


Figure 1: CONTROL X GRAMMATICALITY X DISTRACTOR interaction in the first-pass times at the adjective.

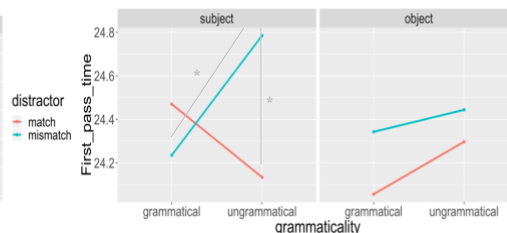


Figure 2: CONTROL X GRAMMATICALITY X DISTRACTOR interaction in first-pass times at the PP.

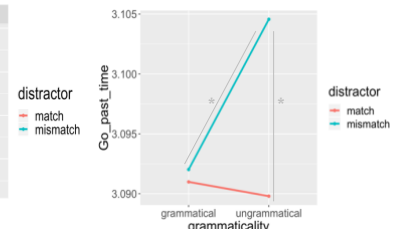


Figure 3: GRAMMATICALITY X DISTRACTOR interaction in the go-past times of the PP.

References:

- Betancort, Moises, Manuel Carreiras & Carlos Acuña-Fariña.** 2006. Processing controlled PROs in Spanish. *Cognition* 100(2). 217–282.
- Kwon, Nayoung & Patrick Sturt.** 2016. Processing Control Information in a Nominal Control Construction: An Eye-Tracking Study. *Journal of Psycholinguistic Research* 45(4). 779–793.
- Parker, Dan, Sol Lago & Colin Phillips.** 2015. Interference in the processing of adjunct control. *Frontiers in Psychology* 6.
- Sturt, Patrick & Nayoung Kwon.** 2015. The processing of raising and nominal control: an eye-tracking study. *Frontiers in Psychology* 6.

Processing embedded clauses in Korean: silent element or a dependency formation?

Nayoun Kim (Sungkyunkwan U.), Keir Moulton (U Toronto), and Daphna Heller (U. Toronto)

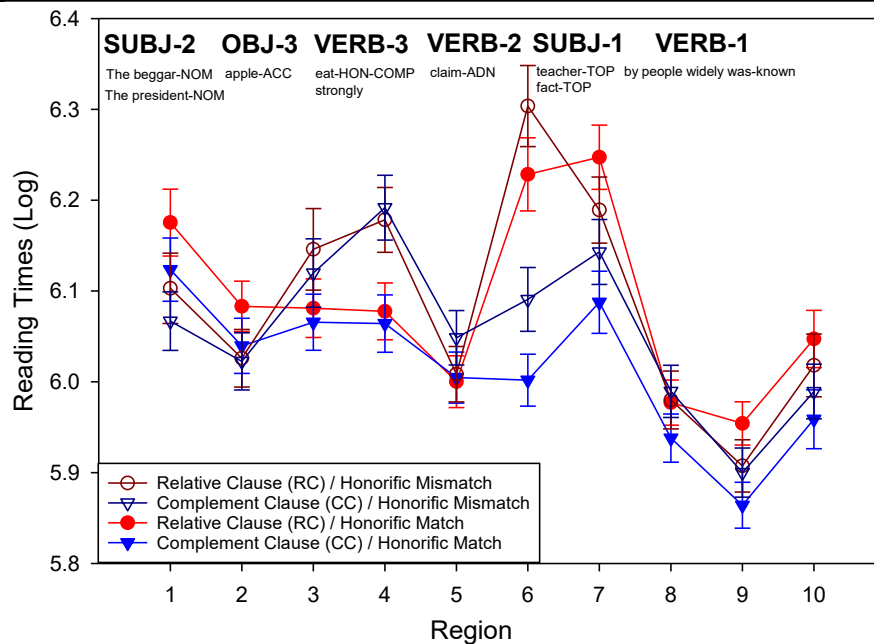
Long-distance dependencies such as Relative Clauses (RCs) are difficult to process, as they involve linking a silent element and an overt phrase (e.g., *the fact that __ worried the president*). Indeed, readers are known to prefer Complement Clauses (CCs) which do not involve a long-distance dependency (e.g., *the fact that the deficit worried the president*) [1,2,3]. The difficulty of long-distance dependencies could arise from (i) the need to *identify* a silent element, or from (ii) the need to create a non-local link between the silent element and the head noun (or both).

We disentangle the two by examining RC/CC ambiguities in Korean, a head-final language where RCs and CCs *precede* the noun. We exploited the fact that Korean has 'null' pronouns and created a temporary ambiguity between **RCs** and **CCs** [Regions 1-5]: the head noun [Region 6] disambiguates as a RC in (A,C) (the *teacher* can eat an apple) and as a CC in (B,D) (a *fact* cannot eat an apple). If the difficulty of long-distance dependencies is due to the silent element, our manipulation should eliminate the asymmetry between RCs and CCs. Second, we exploited the fact that Korean uses an honorific marker on the verb when the subject is honorable [Region 3: eat-HON-COMP]. Our RC/CC had two embeddings (=the teacher/fact that the **beggar/president** claimed __ ate an apple); we manipulated the honorable status of the embedded subject (beggar/president). Because of the word order [Regions 1-3], the Mismatch (A-B) could cue readers early on into the (correct) possibility of there being another discourse referent, whereas the Match could lead readers to (wrongly) assume that the president is eating the apple (C-D).

SELF-PACED READING RESULTS (n=56). (i) **What happens when readers encounter the honorific-marked verb (Region 3)?** A main effect of Honorific ($\beta=-0.11$, $SE=0.03$, $t=-3.90$), with Mismatch conditions being read significantly slower than the Match conditions, an effect that continued into the spillover region ($\beta=-0.11$, $SE=0.02$, $t=-5.03$). This indicates that the mismatch between the subject (beggar) and the honorific-marked verb led to processing difficulty. In the Match conditions, readers probably (wrongly) assumed that the president was the one eating the apple. The Mismatch cases may have simply been parsed as an error, but it is also possible that it led readers to consider the more complex parse of a double embedding. (ii) **What happens when readers encounter the disambiguating head noun (Region 6)?** Here we still observe a main effect of Honorific ($\beta=-0.08$, $SE=0.03$, $t=-2.71$), with Mismatch sentences read slower, but, importantly, there is also a main effect of Clause Type ($\beta=0.22$, $SE=0.05$, $t=4.64$), with RCs being read significantly slower than CCs. This is our central finding: because, for both clause type, encountering the head noun reveals the need to identify and interpret a silent element, and so the difference in reading times can be traced to the difference between RCs and CCs, namely the cost of creating a link between the head noun and the silent element. The main effect of Clause Type continued in the spillover region ($\beta=0.10$, $SE=0.03$, $t=4.05$), but, interestingly, here it was accompanied by a Clause Type X Honorific interaction ($\beta=0.11$, $SE=0.05$, $t=2.24$). At this point, the difference between the RC and the CC in the Mismatch cases was no longer significant ($\beta=0.05$, $SE=0.03$, $t=1.23$), suggesting the non-local link in the RC was formed easily when an additional (silent) discourse referent was predicted earlier. In contrast, the RC-Match sentences were still read significantly slower than the CC-Match sentences ($\beta=-0.16$, $SE=0.04$, $t=4.45$), reflecting a continued cost of linking the silent element to the head noun (teacher) after it was linked to another discourse referent (president), a reanalysis that is not needed in the CC case.

These results are inline with previous findings that Relative Clauses are harder to process than Complement Clauses (cf. [1,2,3]). We extend prior results by showing that this difference holds in a head final language, where the silent element appears before the head noun [4-11]. Most importantly, our findings disentangle difficulties of long-distance dependencies by isolating the difficulty of forming a non-local link (in the RC) from the difficulty of identifying silent elements (present in both the RC and CC conditions). These findings suggest that over and above the costs of managing a silent element, linking that element to form a long-distance dependency with the head noun is costly (cf. [1]).

Region: 1 [[SUBJ-2	2 [OBJ-3	3-4 VERB-3]	5 VERB-2]	6 SUBJ-1]	7	8-10 VERB-1
(A) Relative Clause (RC) / Honorific Mismatch						
거지가	사과를	드셨다고 강하게	주장한	선생님은	사람들에	의해 널리 알려졌다.
The beggar-NOM	apple-ACC	eat-HON-COMP strongly	claim-ADN	teacher-TOP	people	by was widely known
The teacher who the beggar claimed ate an apple was widely known by people.						
(B) Complement Clause (CC) / Honorific Mismatch						
거지가	사과를	드셨다고 강하게	주장한	사실은	사람들에	의해 널리 알려졌다.
The beggar-NOM	apple-ACC	eat-HON-COMP strongly	claim-ADN	fact-TOP	people	by was widely known
The fact that the beggar claimed an honorable person ate an apple was widely known by people.						
(C) Relative Clause (RC) / Honorific Match						
회장님이	사과를	드셨다고 강하게	주장한	선생님은	사람들에	의해 널리 알려졌다.
The president-NOM	apple-ACC	eat-HON-COMP strongly	claim-ADN	teacher-TOP	people	by was widely known
The teacher who the president claimed ate an apple was widely known by people.						
(D) Complement Clause (CC) / Honorific Match						
회장님이	사과를	드셨다고 강하게	주장한	사실은	사람들에	의해 널리 알려졌다.
The president-NOM	apple-ACC	eat-HON-COMP strongly	claim-ADN	fact-TOP	people	by was widely known
The fact that the president claimed an honorable person ate an apple was widely known by people.						



References

- [1] Gibson (1998). *Cognition*. [2] Staub, Foppolo, Donati, & Cecchetto (2018). *JML*. [3] Konrad, Burattin, Cecchetto, Foppolo, Staub, & Donati (in press). *Syntax*. [4] Aoshima, Phillips, & Weinberg (2004). *JML*. [5] Mazuka (1991). *JPR*. [6] Yamashita (1995). *JPR*. [7] Kwon (2008). UCSD dissertation. [8] Yun, Chen, Hunter, Whitman, Hale (2015). *JEAL*. [9] van Gompel, & Liversedge (2003). *JEP*. [10] Hirose (2009). *The Handbook of East Asian Psycholinguistics*. [11] Jäger, Chen, Li, Lin, & Vasishth (2015). *JML*.

Does negation influence the choice of sentence continuations? Evidence from a four-choice cloze task

Elena Albu*, Carolin Dudschig*, Tessa Warren**, Barbara Kaup*

(*University of Tübingen, **University of Pittsburgh)

Although there has been considerable investigation of lexical expectations in affirmative sentences (see Kuperberg & Jaeger, 2016), little is known about how negative sentence fragments are completed. In four experiments, we used a four-choice cloze task to investigate how negation might interact with world and linguistic knowledge to influence the choice of continuations. We structured the word choices to shed light on three possibilities: do participants prefer negation to be used strictly logically, or is their preference influenced by the plausibility of the event described? If the second, are they more likely to make a choice that denies a plausible positive event or that describes a plausible negative event?

Participants saw sentence fragments (*The child will (not) eat the ...*) and clicked on one of four alternatives: a plausible word (*yoghurt*), a weak world knowledge violating word (*shellfish*), a severe world knowledge violating word (*branch*) or a semantic violation inducing word (*minivan*). In the affirmative condition, the plausible word should be the overwhelming choice. In the negative condition, if participants prefer negation to be used logically, the four choices should be equally likely. If they prefer negation to be used as the denial of a plausible positive event, they should favor the plausible word (*The child won't eat the yoghurt*). If they prefer it to be used as a description of a plausible negative event, they should favor the weak world knowledge violating word (*The child won't eat the shellfish*). We also included a 3-level manipulation (Trio, They, and LexAss) of the association between the agent, the verb, and the patient, which drove the plausible words to have either high or low conditional probability. This manipulation appears in Figure 1, but it had no effects, so we will not discuss it further.

In Experiment 1 (N=60 in German in lab), there was a clear difference in the frequency of the four words ($\chi^2(3) = 4744.4$, $p < .01$; see Figure 1 for all data); the plausible word was chosen overwhelmingly. However, there was no effect of sentence polarity ($\chi^2(3) = 4.74$, $p = .19$). This suggests that participants preferred negation to deny a plausible positive situation, but the lack of a polarity effect raises the concern that participants may not have integrated negation into the sentence meaning. In Experiment 2 (N=60, English online), we added 48 fillers that could only be answered correctly if negation was considered (*Which animals don't live in dens? sharks/foxes/rabbits/skunks*). A polarity effect emerged ($\chi^2(3) = 44.17$, $p < .001$), suggesting participants processed the negation. We also replicated Experiment 1's word choice finding ($\chi^2(3) = 4594.5$, $p < .001$); The plausible word was chosen most frequently in both affirmative and negative conditions. In Experiment 3 (N=64, English online), we added hedges to the experimental sentences (*Of course/obviously/certainly/definitely the child will (not) eat the yoghurt/shellfish/branch/minivan*) to render the violations more expected in the negative conditions. Consistent with this, polarity now influenced the frequency of each of the four word choices (all $ps < .001$). The plausible word was still most likely in the negative conditions, but the distribution was flatter. A polarity effect was also apparent ($\chi^2(3) = 472.33$, $p < .001$). Experiment 4 (N=66, English online) investigated whether the observed plausibility effects can be generalized to other aspectual forms (*The child has (not) eaten the yoghurt/ shellfish/ branch/minivan*). The pattern was similar to Experiment 2, with a polarity effect ($\chi^2(3) = 22.45$, $p < .001$) and a strong preference for the plausible word in both affirmative and negative conditions ($\chi^2(3) = 5128.7$, $p < .001$).

This body of findings suggests that upcoming continuations are chosen based on plausibility in both affirmative and negative sentences, with negation inspiring a robust preference that a plausible situation will be denied. Experiment 4 shows that this preference is not modulated by the internal representation of events, but Experiment 3 confirms that it can be

modulated by changes to the expected informativity of the sentence. Overall, these results are in line with a pragmatic account of negation which supports the idea that negation favors a context of plausible denial (Wason 1965).

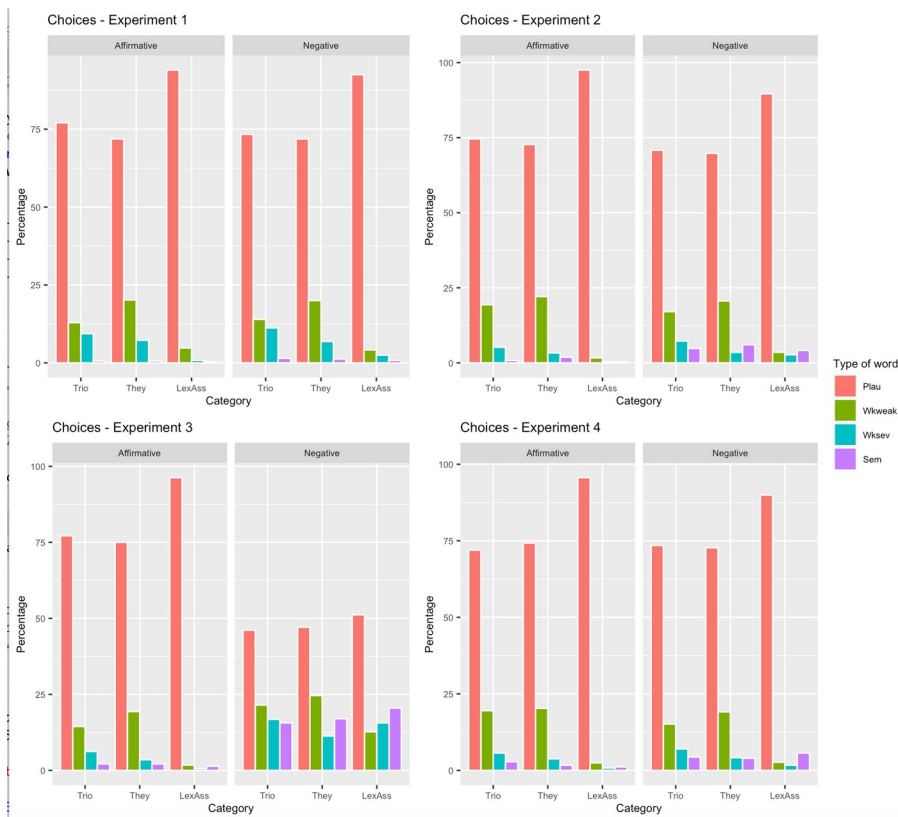


Figure 1: Choices of chosen type of word depending on fragment polarity (aff/neg) and category (trio, they, lexical association)

Contrary to expectations: Is negation more difficult than affirmation?

Elena Albu, Oksana Tsaregorodtseva, Barbara Kaup (University of Tübingen)

Research question. In comparison with affirmation, the processing of negation is said to be more difficult when presented out of context (for an overview, see Kaup & Dudschig, 2020). When embedded in a supportive context, i.e. narrative stories where the proposition denied is either explicitly stated or strongly inferred (Lüdtke & Kaup, 2006) or the relevant attribute dimension is highlighted (Glenberg et al., 1999), the difficulty associated with negation is reduced or completely eliminated. Based on the premise that the processing of negation requires a context of plausible denial (Wason, 1965), we investigated whether negation is facilitated in a minimal context provided by discourse connectives which deny contextual expectations (in the following: “denial contexts”).

Experiment 1. We compared the response times (RT) of negative and affirmative sentences (*[Contrary to expectations/ Surprisingly/ Unexpectedly/ Unpredictably], John has/hasn't eaten the soup*) in a sensibility-judgement task (see Table 1). We expected an interaction between the factors *Context* and *Polarity* with (a.) significantly longer RTs for negative sentences in comparison with affirmation in the non-denial contexts and (b.) similar RTs for affirmative and negative sentences in the denial contexts.

Results. We analyzed the data of 79 participants (32 females; $M_{age} = 38.13$, $SD_{age} = 11.32$) by means of repeated measures ANOVA with the factors *Polarity* (affirmative/negative) and *Context* (non-denial/denial). There was a main effect of Polarity ($F(1,78) = 22.14$, $p < .001$), with shorter RTs in the affirmative condition, and a main effect of Context ($F(1,78) = 145.1$, $p < .001$), with shorter RTs in the non-denial contexts. The interaction was not significant ($F < 1$), invalidating our second prediction. However, the sentences in the two contexts differed in length, an aspect which may have confounded the findings.

Experiment 2 addressed the length confound and investigated the effect of context in denial and non-denial contexts. Expressions reporting people's beliefs with the same number of syllables were added to non-denial contexts (*Everybody is convinced that/ Everyone thinks that/ We believe that/ Based on what we know, John has/hasn't eaten the soup*). The design and predictions were identical to those in Exp. 1.

Results. The data of 62 participants were analyzed (26 females; $M_{age} = 39.96$, $SD_{age} = 11.13$). As in Exp. 1, the ANOVA revealed a main effect of Polarity ($F(1,61) = 21.02$, $p < .001$) and a main effect of Context ($F(1,61) = 21.41$, $p < .001$). This time, however, there were longer RTs in the non-denial contexts, possibly reflecting the complexity of the grammatical structures employed. Similarly to Exp. 1, there was no polarity-by-context interaction ($F < 1$).

Experiment 3. To rule out that the previous results were an artefact of the task, as the RTs in the sensibility-judgement task included the time required for response decision and preparation, a self-paced reading paradigm was employed, where the participants read the sentences fragment by fragment (*Contrary to expectations, // John has/hasn't eaten the soup*). In the attempt to avoid the assumed complexity disparity of the expressions used, connectives with similar complexity were added to the non-denial context (*By all accounts/ Reportedly/ Apparently/ Supposedly, // John has/hasn't eaten the soup*). The predictions were identical to those in Exp. 1.

Results. The analysis of the data (59 participants, 22 females; $M_{age} = 39.76$, $SD_{age} = 13.11$) revealed the same patterns: a main effect of Polarity ($F(1,58) = 56.31$, $p < .001$), and a main effect of Context ($F(1,58) = 14.27$, $p < .001$), but no significant interaction ($F < 1$).

Conclusions. To sum up, this study aimed at investigating whether negation is facilitated when presented in denial contexts provided by discourse connectives. Both affirmative and negative sentences were designed similarly around the mismatch between the polarities of contextual expectations and sentence meaning. The discourse connectives were meant to provide the context of interpretation by activating and accommodating the expectations as part of the hearers' common ground. The findings in all three experiments showed that the relevant interaction was not significant, indicating that polarity and context do not influence each other. In other words, the denial context provided by discourse connectives does not facilitate the processing of negation. In contrast to previous work, our behavioral study suggests that the contextual licensing of negation is not enough to reduce the processing difficulty associated with negation. By comparison, factors like relevance and informativeness which are triggered in longer narrative stories may be responsible for the facilitation of negation.

Table 1: Conditions Experiment 1 - 3

Exp.	Context	Affirmative	Negative
Exp. 1	non-denial	<i>John has eaten the soup.</i>	<i>John hasn't eaten the soup.</i>
	denial	<i>Contrary to expectations, John has eaten the soup.</i>	<i>Contrary to expectations, John hasn't eaten the soup.</i>
Exp. 2	non-denial	<i>Everybody is convinced that John has eaten the soup.</i>	<i>Everybody is convinced that John hasn't eaten the soup.</i>
	denial	<i>Contrary to expectations, John has eaten the soup.</i>	<i>Contrary to expectations, John hasn't eaten the soup.</i>
Exp. 3	non-denial	<i>By all accounts, John has eaten the soup.</i>	<i>By all accounts, John hasn't eaten the soup.</i>
	denial	<i>Contrary to expectations, John has eaten the soup.</i>	<i>Contrary to expectations, John hasn't eaten the soup.</i>

Table 2: Means per condition (standard errors in parentheses) in the four conditions of Experiment 1 - 3

Context	Experiment 1		Experiment 2		Experiment 3	
	Affirmative	Negative	Affirmative	Negative	Affirmative	Negative
non-denial	1683(65)	1827(60)	2119(95)	2270(94)	1462(60)	1621(72)
denial	2099(76)	2208(80)	2009(90)	2117(89)	1709(69)	1709(76)

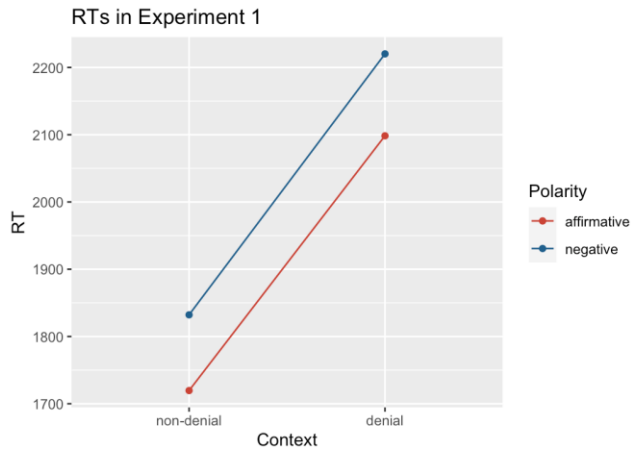


Figure 1. RTs in Experiment 1

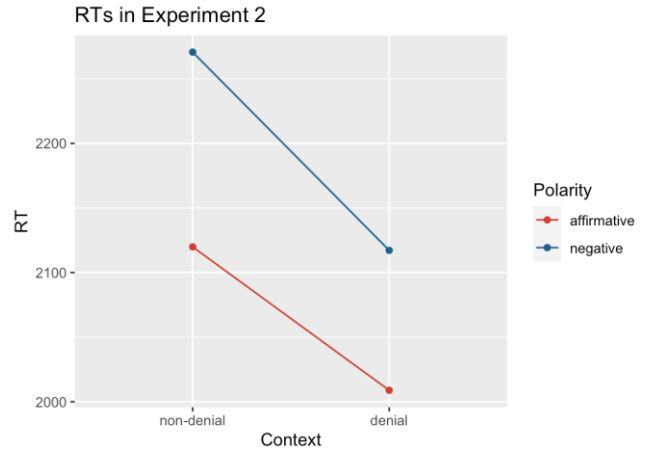


Figure 2. RTs in Experiment 2

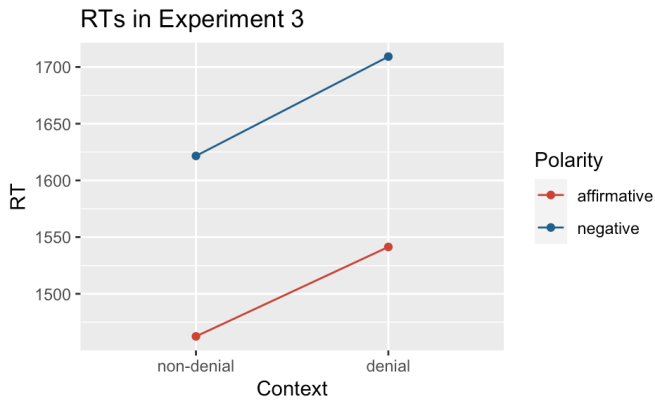


Figure 3. RTs in Experiment 3

Negation cancels discourse-level processing differences: Evidence from reading times in concession and result relations

Ludivine Crible (University of Edinburgh)

Negative sentences are difficult to process since they require an extra mental step (e.g. Wason, 1959; Kaup, Lüdtke, & Zwaan, 2006), although the prior context can reduce this difficulty in some conditions (e.g. Tian, Ferguson, & Breheny, 2016). The present study takes a novel approach to negation by switching the focus from “what makes negation easier to process” to “what is made easier thanks to negation.” Negation can act as a predictive cue at sentence level (Staab, 2007), but its role beyond the sentence remains to be uncovered. The study reports four self-paced reading experiments that investigated the effect of negative vs. positive polarity on the processing of two discourse relations, namely result and concession. In result relations, the link between the two clauses is logical and expected (e.g. *My sister is an excellent cook so she made a delicious cake for dessert*), while in concession, the second clause is unexpected (e.g. *My sister is an excellent cook but she made a disgusting cake for dessert*). Both relations thus involve a causal inference, except that, in concession, the inference is denied. This denial of expectation leads to a higher processing cost for concession compared to other relations (e.g. Townsend, 1983). By making this denial explicit, negative polarity is expected to be preferred in concession than in result, as reflected in corpus data. In processing, however, the affinity between concession and negation has so far only been demonstrated in materials where negation occurs in the second clause of the relation (e.g. Lyu, Tu, & Lin, 2019). The present study instead manipulated the polarity of the first clause, following the hypothesis that negation facilitates the processing of an upcoming concession and reduces the baseline difference between concession and result.

To test this hypothesis, 40 experimental items were created where the overt verb polarity of the first clause was manipulated (e.g. *knew* vs. *didn't know*). In addition, the type of discourse relation (result vs. concession) was controlled by changing one disambiguating word from the second clause (cf. Table 1, in bold). All relations were connected by *and* in order to avoid implausible conditions, and were preceded by a neutral sentence setting up the context. In addition, 60 filler items were created following the same structure (30 nonsensical, 30 neutral), half of those expressing negative polarity. Using the self-paced reading paradigm, 80 English-speaking participants were recruited on Prolific.co and performed a sense rating task after each trial. The data was analyzed with linear mixed effect models. The results support the central hypothesis that negation cancels the processing difference between result and concession, with a significant interaction between relation and polarity ($\beta = -15.999$, $SE = 6.024$, $t = -2.656$, $p < .01$), as shown in Figure 1. This facilitation is reflected in the offline ratings, which show a preference for negation in concession and for affirmation in result ($\beta = 0.12154$, $SE = 0.02626$, $t = 4.628$, $p < .001$).

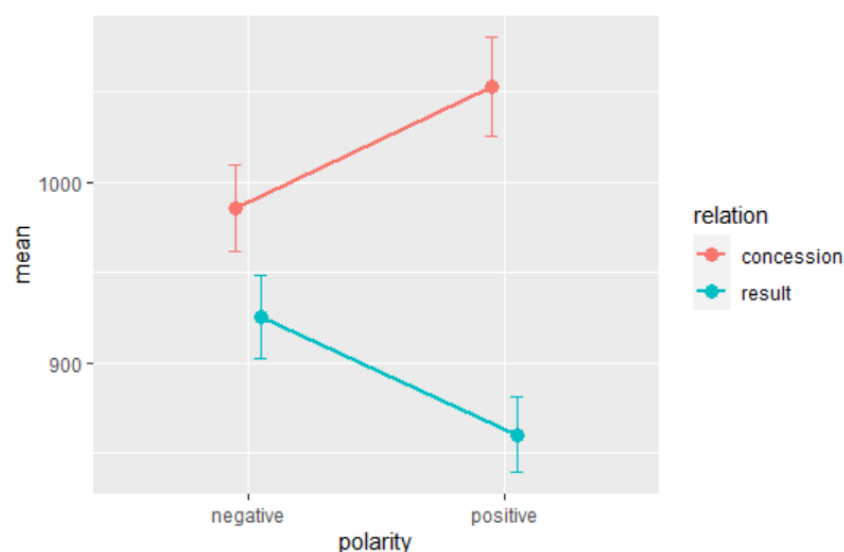
Experiments 2 and 3 further investigated the time-course of the effect of negation on concession by adding a 1,500ms delay (as in Kaup et al., 2006) and by adding a second spill-over region (as in Lyu et al., 2019), respectively. These manipulations did not remove the interaction between relation and polarity and confirmed in particular that positive concession is significantly more difficult to process than positive result, while the slow-down effect of negation in result disappeared over time. Finally, Experiment 4 replicated the findings by replacing *and* with *but* and so in order to address a potential ceiling effect in concession. Despite these more explicit connectives, concession remained more difficult than result overall, and the same interaction was once more observed on the critical region, thus confirming the robust facilitation effect of negation on concession. We can therefore conclude that the interaction between polarity and discourse relations is mutual and bi-directional: not only do some relations facilitate the processing of negation, but initial (i.e. first-clause) negation itself modulates the processing of an upcoming relation and acts as a concessive facilitator.

Table 1. Example materials (context sentence: “The students had an upcoming exam.”)

positive-result	They all knew their coursework well // and they // were confident // about their performance.
negative-result	They didn’t know their coursework // and they // were anxious // about their performance.
positive-concession	They all knew their coursework well // and they // were anxious // about their performance.
negative-concession	They didn’t know their coursework // and they // were confident // about their performance.

Double forward slashes “//” represent the segmented regions. In Experiment 2, the delay was added before the connective region. In Experiment 3, the second spill-over region contained neutral commentaries.

Figure 1. Mean reading times by condition on the spill-over region (Experiment 1)



References

- Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *J of Prag*, 38, 1033–1050.
- Lyu, S., Tu, J.-Y., & Lin, C.-J. (2019). Processing plausibility in concessive and causal relations : Evidence from self-paced reading and eye-tracking. *Disc Proc*, 57(4), 320–342.
- Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle. An event-related potential study on the pragmatics of negation. *Psych Sci*, 19(12), 1213–1218.
- Staab, J. (2007). Negation in context: Electrophysiological and behavioral investigations of negation effects in discourse processing. Doctoral dissertation, UC San Diego.
- Townsend, D. J. (1983). Thematic processing in sentences and texts. *Cognition*, 13(2), 223–261.
- Wason, P. (1959). The processing of positive and negative information. *Quart J of Exp Psych*, 11, 92–107.

Verifying negative sentences - How context influences which strategy is used

Shenshen Wang (University College London), Chao Sun (Leibniz Centre for General Linguistics), Richard Breheny (University College London)

When given a sentence to verify against a state of affairs (soa), the natural strategy would be to use the semantics of the sentence to infer what kinds of states of affairs make the sentence true and to check that the target soa is among those (1-step strategy). However, in the case of sentential negation, its truth-functional semantics offers another route – which is to first verify the prejacent of negation and then reverse the response (2-step strategy). Several studies (e.g. [1-2]) show that participants adopt both strategies in verification tasks, which results in different patterns in response time between participants. The psychological processes underpinning the use of negation has been debated. The use of 2-step strategies has been argued to provide support for Composite models [1-3]. These say that the process of representing an interpretation for a negative sentence is composed of parts which reflect what we see at the level of linguistic structure - negation and its argument. By contrast, [4,5] says that incremental and probabilistic language processes have two simultaneous aims: to compute the sentence content and the intended Source of Relevance (SoR - often described in terms of QUDs). Language processes thus exploit information in the linguistic stimulus, in addition to any contextual information, to infer *both* sentence content *and* SoR. In the case of processing negative sentences, when presented in the absence of other information, sentential negation is a strong cue to a specific class of SoRs, in which the prejacent is a live possibility which the speaker intends to exclude (Default context). However, the presence of other cues (e.g. information structure or a preceding question) can override this. This account finds support in probe-response and visual world paradigms [4,5]. Here we extend this account to sentence-picture verification: In Default contexts, attention can be drawn to the prejacent and this may interfere with a 1-step verification strategy, resulting in the adoption of the 2-step strategy. Typically, the 2-step strategy leads to an interaction between polarity and truth value ($TA < FA$, $FN < TN$), whereas 1-step strategy leads to only main effects ($TA < FA$, $TN < FN$) – see [1-2] among many other references.

Experiment: We manipulated contexts using two types of question. See Table 1. A positive polar question spells out the Default context. Wh-questions with Congruent positive or negative predicates cue a SoR which would not interfere with a 1-step strategy. We predict a greater use of 2-step strategy in Default context than Congruent. Participants ($N=64$) evaluated positive or negative statements in the presence of an image. The statements take the form of an elliptical answer to either a positive polar question (Default Context) or a congruent wh-question (Congruent Context). Shown in Figure 1, the statement and image are constant wrt polarity and truth value, but their elliptical form varies to conform with question context.

Results: We constructed a linear mixed-effects model predicting reaction time (RT) from polarity (affirmative or negation), truth value (true or false), and context (default or congruent). All main effects were highly significant and there was a significant three-way interaction (all $ps < .001$). See Figure 2. The default context showed an interaction between polarity and TV, suggesting a greater effect of negation on True than on False trials ($TA < FA$, $p < .001$; $FN < TN$, $p = .06$). The congruent context however showed only main effects (all $ps < .001$). To examine whether participants adopted different strategies, we divided participants into two distinct groups based on their response patterns in the default context using K-means clustering, and then fitted a mixed-effects model predicting RT from polarity and TV for each group in each context. See Figure 3. Group 1 ($N=28$) in the default context showed an interaction between polarity and TV ($TA < FA$, $FN < TN$, all $ps < .001$), whereas Group 2 ($N=29$) in the same context showed only main effects (all $ps < .01$). By contrast, both groups showed only main effects in the congruent context (all $ps < .001$).

Discussion: Our results provide further evidence that it is context which is responsible for the use of 2-step strategy and cast doubt on composite models for negative sentence comprehension. Particularly as the same group (Group 1) switch strategy depending on context.

Default context

Condition	Polar question	Elliptical answer	Display
TA	Is the apple peeled?	It is.	
FN	Is the apple peeled?	It isn't.	
FA	Is the apple peeled?	It is.	
TN	Is the apple peeled?	It isn't.	

Congruent context

Condition	Congruent question	Elliptical answer	Display
TA	Which one is peeled?	The apple.	
FN	Which one isn't peeled?	The apple.	
FA	Which one is peeled?	The apple.	
TN	Which one isn't peeled?	The apple.	

Table 1 Example items. 2(Polarity) * 2(Truth value) * 2(Context) within-participants design.

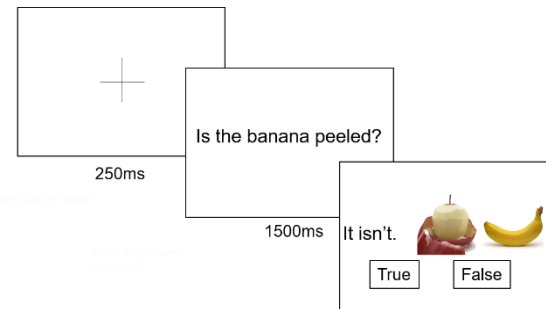


Figure 1 Procedure (True-Negative-Default trial). Context questions appear for 1500ms prior to target screen. In the target screen, the elided statement appears on the left and image on the right.

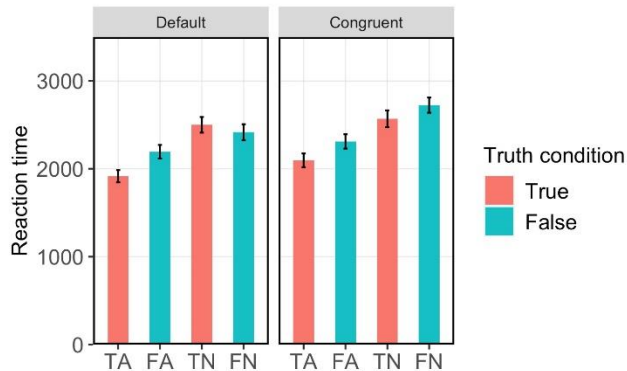


Figure 2 Mean RT for each polarity, truth value, and context. Error bars represent standard errors of the mean.

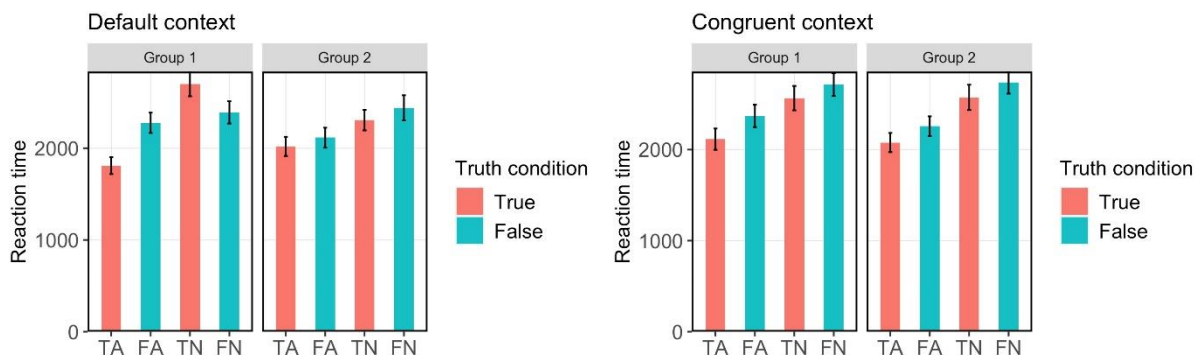


Figure 3 Mean RT for each polarity, truth value, and group in two different contexts. Error bars represent standard errors of the mean.

References: [1] Clark & Chase (1972). *Cognitive Psychology*, 3(3), 472–517; [2] Mathews, Hunt, & Macleod (1980). *J. Verbal Learning & Verbal Behavior*, 19, 531-548. [3] Kaup, Yaxley, Madden, Zwaan, & Lüdtke (2007). *QJEP*, 60, 976–990. [4] Tian, Breheny, & Ferguson, (2010). *QJEP*, 63(12), 2305–2312. [5] Tian, Ferguson, & Breheny, (2016). *LCN*. 31: 683-698.

Processing polar questions in contexts with varying epistemic biases in English

Vinicius Macuch Silva (Osnabrück University), E Jamieson (University of Southampton)

Background: In English, there are two possible ways to form a polar question with negation (NPQ): the negation marker can be “low” (LNPQ) (1) or “high” (HNPQ) (2).

1) Is there **not** a vegetarian restaurant around here?

2) **Isn't** there a vegetarian restaurant around here?

While (1) questions the negative proposition, (2) seems to be more complex. Ladd (1981) claims there is an “ambiguity” whereby (2) *can* question the negative proposition, but it can also be used to indicate the questioner has a prior belief that the *positive* proposition is true (see also e.g. Gärtner & Gyuris, 2017; Krifka, 2015; Romero & Han, 2004; Sudo, 2013).

However, Domaneschi et al. (2017) found that in production English natives have preferences as to which question to produce, depending on their epistemic state and the evidential context surrounding the discourse (Table 1). Their results suggest no ambiguity: LNPQs question the negative proposition, and HNPQs express a belief. In this study, we report results from a self-paced reading experiment investigating whether Ladd’s hypothesized “ambiguity” holds in processing.

Design and procedure: We carried out a word-by-word self-paced reading experiment, with 120 self-reported English natives recruited through Prolific. Participants read LNPQs and HNPQs against short background discourses designed to target the effects of prior belief given negative evidence (e.g. 3-4). All vignettes were normed by an independent sample of participants for the presence or absence of a prior belief.

3) *Prior belief:* Someone told you I won the marathon at the weekend. However, I start telling you I am disappointed with my performance. You say:

HNPQ: Hold | on | a | minute. | Didn't | you | win | the | marathon?

LNPQ: Hold | on | a | minute. | Did | you | not | win | the | marathon?

4) *No belief:* We are talking about baths. I say I haven't had one in 3 years. You say:

LNPQ: I | love | a | bath. | Have | you | not | got | one | at | home?

HNPQ: I | love | a | bath. | Haven't | you | got | one | at | home?

We hypothesize that HNPQs will be facilitated in contexts where there is a prior belief about the proposition, whereas LNPQs will be facilitated in contexts where there is no prior belief about the proposition, following Domaneschi et al. (2017).

Results: As per pre-registered protocol, we compare the reading times (RTs) at each region up to the main verb (Figure 1), which serves as the spillover for the negation marker in the LNPQs. We model our RT data using Bayesian hierarchical regression models, regressing the log-transformed RTs as a function of the belief and negation type for each region of interest (Table 2).

In HNPQs, we find no evidence for an effect at the critical region nor at its immediate spillover. However, we do find strong evidence for an effect at the VERB region, such that HNPQs are read faster against contexts with a prior belief compared to contexts without a prior belief. This is in line with our original hypothesis. In the case of LNPQs, we find no evidence for an effect at the regions up to the VERB, which contradicts our hypothesis. However, the descriptive results at the region immediately following the VERB suggest that LNPQs are read more slowly against contexts with a prior belief compared to contexts without a prior belief. While we did not have predictions about regions following the verb, this result suggests difficulty in integrating the question form with information from the verb when there is a prior belief in the discourse context.

Discussion: Our results show that, at least in the case of HNPQs, comprehenders process NPQs differently depending on whether or not the prior discourse context sets them up with a belief about the truth of a proposition. This partially supports Domaneschi et al.’s (2017) results and challenges the idea of Ladd’s (1981) ambiguity in HNPQs. We discuss these findings against the results from a replication where we revised our items to re-assess the case of LNPQs.

Table 1: Production preferences from Domaneschi et al. (2017). Shaded cells are not investigated in this study.

			belief: \emptyset		belief: p
			evidence: \emptyset	PosQ	HighNegQ
			evidence: $\neg p$	LowNegQ	HighNegQ
Negation	Region	Term	Posterior mean	95% CrI	$P(\beta < 0)^a$
High	Critical	Intercept (no belief)	5.73	[5.66; 5.80]	
	Critical	Prior belief	0.01	[-0.03; 0.06]	.29
	Critical +1	Intercept (no belief)	5.70	[5.65; 5.75]	
	Critical +1	Prior belief	-0.01	[-0.05; 0.03]	.62
	VERB	Intercept (no belief)	5.70	[5.60; 5.80]	
	VERB	Prior belief	-0.18	[-0.31; -0.05]	> .99
Low	Critical	Prior belief	0.00	[-0.07; 0.06]	.59
	Critical +1	Prior belief	0.03	[-0.02; 0.08]	.99
	Critical +2	Prior belief	0.19	[0.00; 0.37]	.75
	VERB	Prior belief	0.04	[-0.20; 0.28]	.40

Table 2: Model coefficients for Bayesian regressions.

^aIn the case of the low negation the hypothesis tested was $P(\beta > 0)$, i.e., Prior belief > No belief, as per the hypothesis indicated in the text.

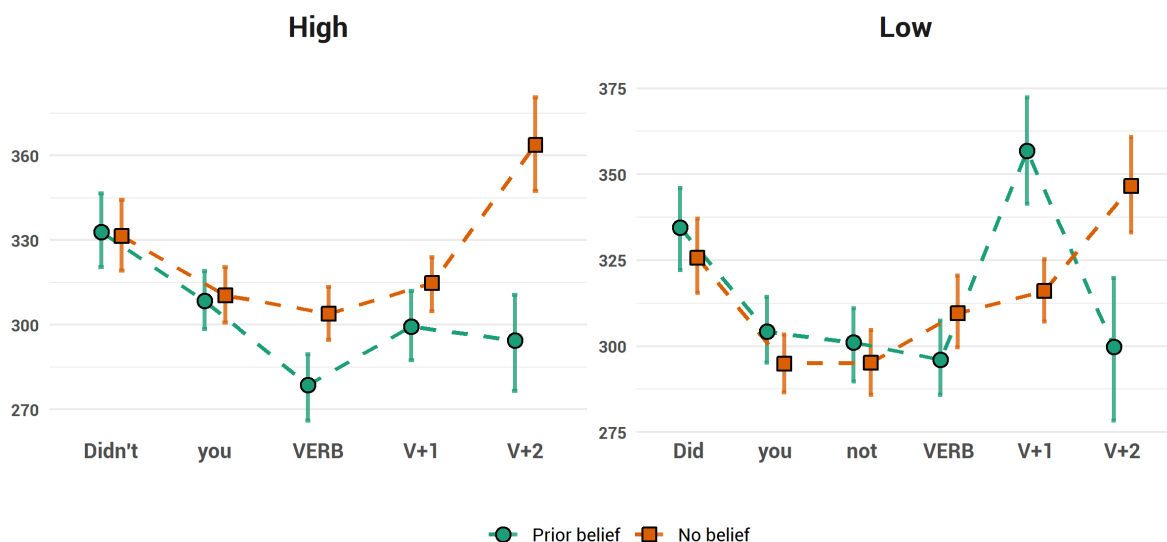


Figure 1: Reading times at the different sentence regions as a function of the negation type (*low* vs. *high*) and belief (*prior belief* vs. *no belief*).

REFERENCES: Domaneschi, F., Romero, M. & Braun, B. 2017. Bias in polar questions: Evidence from English and German production experiments • Gärtner, H. & Gyuris, B. 2017. On delimiting the space of bias profiles for polar interrogatives. • Krifka, M. 2015. Bias in commitment space semantics: Declarative questions, negated questions and question tags. • Ladd, R. 1981. A first look at the semantics and pragmatics of negative questions. • Romero, M. & Han, C. 2004. On negative yes/no questions. • Sudo, Y. 2013. Biased polar questions in English and Japanese.

Uniformity and variability in the understanding of expletive negation across languages

Yanwei Jin (University at Buffalo) & Jean-Pierre Koenig (University at Buffalo)

Expletive negation (EN) is a construction where a negator in the complement clause of certain lexical items (EN-triggers; ‘fear’ in (1)) does not change the polarity of the complement proposition. Jin & Koenig (2019, 2020) found it to occur widely in languages of the world and that the same set of predicates or operators trigger EN. They propose a language production model where EN arises from the production of a lexically entailed negative proposition $\neg p$ rather than the intended proposition p . They suggest that what starts out as an interference between the argument proposition that is part of the message (p) and its entailed dual ($\neg p$) can become entrenched or even grammaticized depending on the trigger or language.

In this paper, we report the results of a French and a Mandarin experiment that test the hypotheses that (i) the same semantic interference effects occur in comprehension across languages but that (ii) the propensity of speakers to interpret a negator expletively can vary from language to language. We model our experiments after Jin & Koenig (2020) who designed a semantic interference effect experiment that tested whether English speakers include in their representations the expletive use of negators. Participants in their experiment read short paragraphs and judged the logical consistency of continuations given the paragraphs. Continuations fell into 4 conditions: (3a) non-EN-trigger + logically inconsistent negation, (3b) EN-trigger + logically inconsistent negation, (3c) non-EN-trigger + logically consistent negation, and (3d) EN-trigger + logically consistent negation. If participants interpret the negator in the complement clause of EN-triggers (‘prevent’ in (3b, 3d)) *expletively*, determining whether the continuation is consistent should be more difficult than for non-EN-trigger continuations, as the logical and expletive interpretations support conflicting answers. They found EN-trigger continuations elicited less logically accurate answers and longer response times. Importantly, they found that there was a high correlation ($r=.66$, $p<.01$) between EN interpretation across triggers in a corpus and logical inaccuracy for EN-trigger continuations, which suggests that English speakers keep track of the likelihood that a negator is interpreted expletively after each trigger.

Our French and Mandarin experiments mirror Jin & Koenig’s English experiment’s logic and stimuli. We chose French and Mandarin because EN is, according to grammars, more entrenched in both languages and, additionally, French has one negator *ne* which is a grammaticalized marker of EN and one which is not (*ne*) ... *pas*. First, we predicted and found that our French and Mandarin participants, like Jin & Koenig’s English participants, made more logical errors ($p<.01$ for both French and Mandarin participants) and took longer to respond ($p<.01$ for French and $p=.07$ for Mandarin participants) for continuations that contained EN-triggers than for continuations that did not. We also found a high correlation ($r=.75$, $p<.01$) between logical inaccuracy for EN continuations in the Mandarin experiment and frequency of EN interpretations in a Mandarin corpus study (French corpus study pending), confirming that the more EN interpretation for a particular EN-trigger a speaker has encountered, the more likely she is to interpret expletively a new occurrence of a negator in the scope of that EN-trigger. Second, we predicted and found an interaction between language and trigger condition. A logistic regression showed that French and Mandarin speakers made more logical errors ($p<.01$) than English speakers after EN-triggers (22.5% EN interpretation for English, 54.2% for French, and 58.5% for Mandarin speakers), but not after non-EN-triggers, suggesting that speakers of both languages were more likely to interpret negators expletively after EN-triggers than English speakers. Third, we predicted and found an effect of negator form for French. French speakers made more logical errors ($p<.01$) when the negator in the argument proposition of an EN-trigger was *ne* (82.04%) than either English speakers (29.83%) or Mandarin speakers (71%) for corresponding triggers and conditions, but less logical errors ($p<.01$) than Mandarin speakers (64.75%) and roughly the same number of logical errors ($p=.11$) as English speakers (24.12%) when the negator was the standard negation (*ne*)...*pas* (29.05%). Overall, the results of our experiments suggest that

although there is uniformity across languages in the availability and triggers of EN interpretation of negators, entrenchment can vary across, languages, triggers, and negator form.

(1) A French example of EN marked with grammaticalized negator *ne*

J'ai peur qu'il ne pleuve demain.
I.have fear that.it NEG rain.SBJV tomorrow
'I fear that it will rain tomorrow.'

(2) A French example of EN marked with low-entrenchment negator (*ne*)...*pas*

Vous avez oublié de ne pas nommer Jacques Stephen Alexis,
You have forgotten INF NEG nominate PN
un grand des grands savants.
one great of.the great savants

'You have forgotten to nominate Jacques Stephen Alexis, one of the greatest savants.' (Jin & Koenig 2019: 173; such examples sound like errors to native speakers)

(3) A stimulus set with four different conditions in Jin & Koenig's (2020) English experiment

(a) Non-EN-trigger + logically inconsistent negation

My husband and I were high school classmates and we graduated ten years ago. Several days ago, we both got an invitation for our 10-year high-school reunion. I think it'll be fun to get together for the first time after so many years. But my husband said he won't go because he didn't like most people in his class. I want him to go with me. **I'll persuade him to not go there.**

(b) EN-trigger + logically inconsistent negation

Every time when my husband comes back from his annual high-school reunion, he is unhappy. I know this is because he thinks he has accomplished the least among his classmates. Now this year's reunion is approaching and he said he would go. **I'll prevent him from not going there.**

(c) Non-EN-trigger + logically consistent negation

Every time when my husband comes back from his annual high-school reunion, he is unhappy. I know this is because he thinks he has accomplished the least among his classmates. Now this year's reunion is approaching and he said he would go. **I'll persuade him to not go there.**

(d) EN-trigger + logically consistent negation

My husband and I were high school classmates and we graduated ten years ago. Several days ago, we both got an invitation for our 10-year high-school reunion. I think it'll be fun to get together for the first time after so many years. But my husband said he won't go because he didn't like most people in his class. I want him to go with me. **I'll prevent him from not going there.**

(4) **Table 1.** Mean accuracy and response time in the Mandarin experiment

Trigger Condition	Logical consistency	Mean accuracy of judgments	Mean RT
non-EN-triggers	Logically inconsistent	89.80%	4907.79
EN-triggers	Logically inconsistent	35.88%	5773.35
non-EN-triggers	Logically consistent	90.80%	5418.28
EN-triggers	Logically consistent	47.18%	6170.93

Table 2. Mean accuracy and response time in the French experiment

Trigger Condition	Mean accuracy of judgments	Mean RT
EN-triggers that take <i>ne</i> as EN	17.96%	5162.84
Non-EN-triggers used as controls	90.51%	4127.28
EN-triggers that take (<i>ne</i>)... <i>pas</i> as EN	70.95%	7123.64
Non-EN-triggers used as controls	90.80%	3760.88

TITLE: Testing the influence of the listener's perspective in the epistemic step.

Blanche Gonzales de Linares, Napoleon Katsos (University of Cambridge)

INTRODUCTION: In the traditional Gricean theory of quantity implicature derivation, the consideration of the speaker's epistemic state is a necessary step before a full implicature can be derived (Sauerland, 2004). A psycholinguistic model based on the Gricean theory would therefore predict that if a speaker is not considered sufficiently knowledgeable by a listener, no implicature will be derived. Empirical evidence matching this prediction has been found (Bergen & Grodner, 2012; Breheny, Ferguson & Katsos, 2013). However, a factor that was not actively manipulated in these studies is the listener's perspective, and the question of whether a better pragmatic match being visible only to the listener would lead to implicatures being derived erroneously. The present study attempts to explore this gap in the literature, by creating a situation where the listener has to avoid choosing a referent that is hidden from the speaker but that matches the most informative interpretation of the speaker's instructions. In this case, the listener must not only consider what the speaker does and does not see, like in the existing literature, but must also inhibit the better pragmatic match in their perspective.

METHOD: The experiment was a computer version of the director task, a paradigm commonly used to study perspective taking (e.g. Keysar, Barr, Balin & Brauner, 2000). In the experiment, the instruction was given in text form over the image and was presented as spoken by an unseen person. The displays featured a 2x2 grid in which there were cards with either one or two types of item. An example of the displays seen by the participants can be found in Figure 1. Participants were trained to know that the card in the grey box was hidden from the speaker. In the critical condition (Figure 1, Display A), the instruction required one kind of item (e.g. "pick the card with apples") and the display featured two cards featuring that item: one featured that item alone, and the other featured that item with another item. The card with only apples is a better pragmatic match for the utterance, as a more informative sentence to describe the mixed card would have been "Pick the card with apples and oranges". However, the card with only apples is hidden from the speaker. Therefore, the Gricean theory predicts that if the participant correctly does the epistemic step, the implicature will be blocked and they will choose the card in common ground. This condition was directly compared with Display B (Figure 1), in which the card hidden from the speaker was the same as the card with apples that was in common ground. Indeed, the only difference between these conditions is that the critical condition there is a pragmatic preference for the card in hidden ground, so if in Display B participants always choose the common ground card but in Display A they sometimes choose the hidden card, it shows that an implicature has been derived despite the speaker's insufficient knowledge.

RESULTS AND DISCUSSION: Results showed that in the condition where the hidden card is a better pragmatic match for the utterance (Figure 1, Display A), the accuracy rates were significantly lower than in cases where the card in hidden and privileged ground were identical (Figure 1, Display B) (76.9% vs. 91.95%). This brings preliminary evidence to the prediction that when the listener's perspective contains a potential referent that is a better pragmatic match, implicatures can be derived even if that referent is not visible to the speaker. This fits with a constraint-based view of implicature derivation, in which the speaker's perspective is one of many factors in the probability of an implicature being derived (Degen & Tanenhaus, 2019), rather than a fixed step which can block an implicature. A future avenue of research would be to integrate the listener's perspective as a factor into existing models of implicature such as the Rational Speech Act model (Goodman & Stuhlmüller, 2013), for example by drawing inspiration from a Bayesian model of perspective taking which calculates probabilities of a referent being chosen based on the simultaneous integration of both the listener's perspective and the common ground (Heller, Parisien & Stevenson, 2016).

REFERENCES:

- Bergen, L., & Grodner, D. J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5), 1450.
- Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, 126(3), 423-440.
- Degen, J., & Tanenhaus, M. K. (2019). Constraint-based pragmatic processing. *The Oxford handbook of experimental semantics and pragmatics*, 21-38.
- Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, 149, 104-120.
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1), 32-38.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1), 173-184.
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and philosophy*, 27(3), 367-391.

FIGURES:

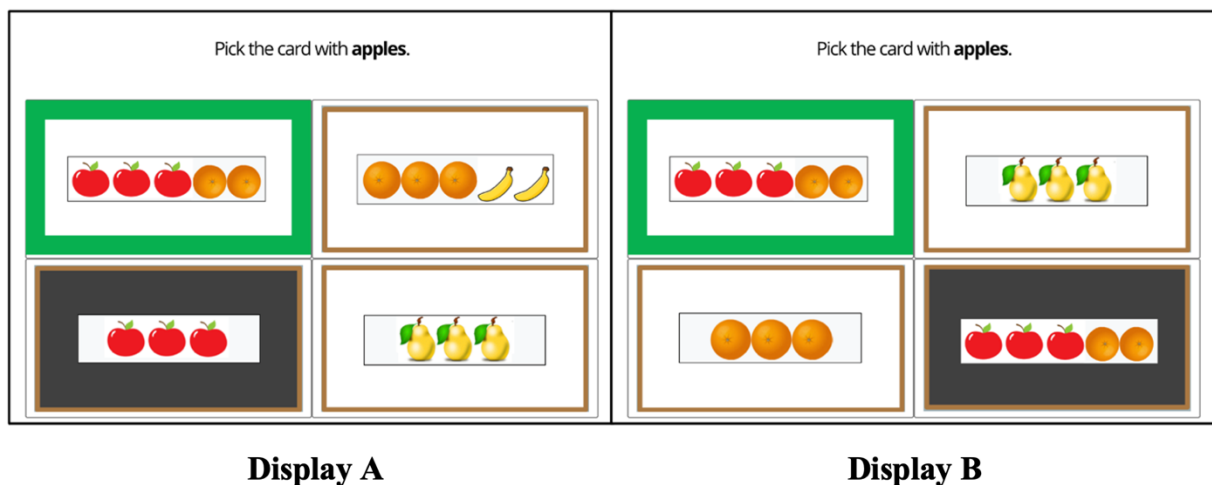


Figure 1: Comparison of the critical displays for the utterance “Pick the card with apples”. Cards highlighted in green are the target cards for accuracy measures.

The costs and benefits of different metaphoric structures: evidence from pupillometry

Camilo R. Ronderos^{1,2}, Ernesto Guerra³, Pia Knoeferle^{1,4,5}

¹Humboldt-Universität zu Berlin, ²University of Oslo, ³Universidad de Chile, ⁴Berlin School of Mind and Brain,

⁵Einstein Center for Neurosciences Berlin

c.r.ronderos@ifikk.uio.no

The difference between understanding metaphoric and literal expressions has long been at the center of metaphor research. Noveck et al. (2001), for example, argued that understanding a metaphor carries additional costs and benefits (quantified as longer reading times and higher comprehension accuracy, respectively) than a literal equivalent, resulting in greater processing load. However, it has not been investigated whether a metaphor's processing load is stable or whether it varies as a function of the sequence of its elements.

This question is critical for theory development. Some accounts, for example, posit that metaphors are processed asymmetrically (e.g., Chiappe et al., 2003; Glucksberg, 2001; Wilson & Sperber, 2008): In a nominal metaphor (e.g., 'my cat is a princess'), the topic ('my cat') must appear prior to the vehicle ('a princess'), and the order may not be reversed. However, German verb-object metaphors (see 1a and 1b), can reverse the topic-vehicle sequence and still be felicitous. Does this mean that verb-object metaphors are not asymmetrically processed? One possibility is that topic-vehicle metaphors are understood better than vehicle-topic metaphors, resulting in an increased cost-benefit balance (following Noveck et al. 2001) which should translate to a greater processing load for topic-vehicle compared to vehicle-topic metaphors.

The present work thus investigates the impact of the sequence of the elements on a metaphor's processing load via pupil dilation. This has been associated with increased processing load of linguistic stimuli (e.g., Engelhardt et al. 2010; Just & Carpenter, 1993), but has not been previously used to study figurative language comprehension.

We re-analyzed the data of a previous study on the processing of German verb-object metaphors such as (1). 32 participants read 4 sentences that either biased towards a literal or a metaphoric interpretation of the target utterance (literal and metaphoric conditions, see 1). They then heard the utterance (1a or b) while looking at four pictures, two of which represented the literal and the metaphoric interpretation of the sentence (a princess and a cat). We reasoned that if metaphors carry more processing load compared to literals regardless of element sequence, both utterances (1a&b) should cause more pupil dilation in the metaphoric compared to the literal condition. If, however, topic-vehicle metaphors (1a) are understood better than vehicle-topic metaphors (1b), there might a lesser processing load for (1b) relative to literal controls.

We computed pupil dilation using the R package PupilPre (Kyröläinen et al., 2019) for preprocessing the data, time-locked to the onset of the main verb ('füttert', Figure 1) and the onset of the direct object ('Prinzessin', Figure 2) (i.e. the regions where the metaphor is understood in the metaphoric-late condition and the metaphoric-early condition, respectively). We followed Engelhardt et al. (2010) for data pre-processing. We fitted mixed-effects linear models (with treatment contrast coding) to the verb and vehicle regions, with verb position (early vs. late), context (literal vs. metaphoric) and their interaction as fixed effects and pupil dilation (measured in abstract units) as dependent measure.

In the early verb metaphoric condition (i.e., topic-vehicle order), participants showed more pupil dilation when hearing the vehicle compared to the early-verb literal condition ($t=2.7$, $p<0.05$). A significant interaction was also found ($t=14.3$, $p<0.001$), suggesting that this difference was unique to the metaphoric conditions. No significant differences were found in the verb region. We interpret this as suggesting that topic-vehicle metaphors are associated with a higher processing load compared to literals, but this does not hold for vehicle-topic metaphors. To confirm these preliminary findings, a follow-up replication experiment is underway. Overall, we see this finding as complementing Noveck et al. (2001) and as being in line with asymmetric accounts of metaphor comprehension.

(1) Example of critical item. The verb is considered the topic since it is the only element that refers to the nominal topic ('the cat'). In German, the verb 'füttern' (feed) has a strong selectional preference for taking an animal as its accusative object. All items had verbs with strong selectional preferences biasing towards the nominal topic of the metaphor.

(1a, early-verb conditions) Sebastian **füttert** VERBAL TOPIC eine **Prinzessin** VEHICLE und wird unablässig der Adligen/der Katze beistehen.

(1b, late-verb conditions) Sebastian wird eine **Prinzessin** VEHICLE **füttern** VERBAL TOPIC und wird unablässig der Adligen/der Katze beistehen.

'Sebastian feeds/will feed a princess and will continuously support the noble woman/the cat.'

Example of linguistic context

(Literal context) Sebastian liebt eine berühmte Adlige. Er hat sie in einem Schloss kennengelernt und seitdem sind sie unzertrennlich. Die Adlige ist schwach und abhängig, und kann sehr hilfsbedürftig sein. Deswegen tut Sebastian alles für sie, wenn sie etwas braucht. Er wird sich immer um sie kümmern wollen. **(English translation:** 'Sebastian loves a famous noble woman. He met her in a castle and they have been inseparable since. The noble woman is weak and dependent and can be very needy. That's why Sebastian would do anything for her when she's hungry. He will always want to take care of her.')

(Metaphoric context) Sebastian liebt eine wunderschöne Katze. Er hat sie in einem Tierheim adoptiert und seitdem sind sie unzertrennlich. Die Katze ist verwöhnt und launisch, und kann sehr wählerisch sein. Deswegen würde Sebastian alles für sie tun, wenn sie etwas braucht. Er wird sich immer um sie kümmern wollen. **(English translation:** 'Sebastian loves a beautiful cat. He adopted her in a shelter and they have since been inseparable. The cat is spoiled and moody and can be very fussy. That's why Sebastian would do anything for her when she's hungry. He will always want to take care of her.')

Figure 1

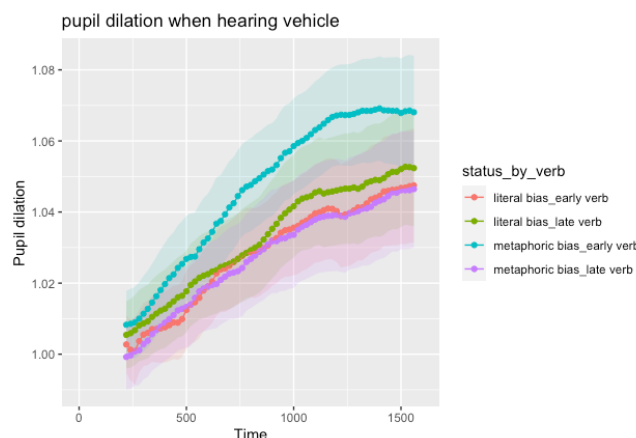
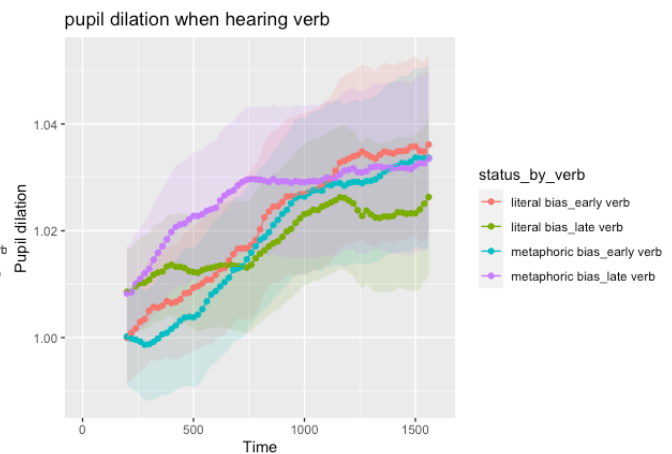


Figure 2



References

- Chiappe, D., Kennedy, J. M., & Smykowski, T. (2003). Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol*, 18(2), 85-105.
- Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly journal of experimental psychology*, 63(4), 639-645.
- Glucksberg, S. (2001). *Understanding figurative language: From metaphor to idioms* (No. 36). Oxford University Press on Demand.
- Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(2), 310.
- Kyröläinen, A.-J., Porretta, V., van Rij, J., & Järviö, J. (2019). PupilPre: Tools for preprocessing pupil size data [Version 0.6.1, updated 2019-12-18]. URL: <https://CRAN.R-project.org/package=PupilPre>
- Noveck, I. A., Bianco, M., & Castry, A. (2001). The costs and benefits of metaphor. *Metaphor and Symbol*, 16(1-2), 109-121.
- Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. *The Cambridge handbook of metaphor and thought*, 84, 105.

Ageing and communication in face-threatening contexts

Madeleine Long (U Oslo), Sarah MacPherson (U Edinburgh), Paula Rubio-Fernandez (U Oslo)
(madeleine.long@ifikk.uio.no)

Most research on face-saving focuses on whether listeners adjust their interpretation according to the degree to which their face is threatened [1-4]. For example, in a recent study, participants judged the probability of 'possibly' developing cancer (vs insomnia) as more likely because they assumed their doctor would use the term 'possibly' as a face-management device to diagnose what they perceived to be the more severe diagnosis [1]. While numerous studies show this sensitivity in comprehension, much less is known about the type of adjustments made during production [5-6] and how that might vary by age. To address these questions, we recruited adults over the lifespan to test how they relayed bad news to others. In keeping with the audience design literature, which shows younger adults use more partner-specific language than older adults [7-10], we predicted that younger adults would be more sensitive to a listener's perspective and thus save face to a greater extent. However, we also predicted that while speakers should consider both the recipient and event severity when giving bad news, adjusting the message to the recipient would be more important (and hence prevalent) than adjusting for event severity.

EXP 1 presented participants (N=100, ages 18-72) from Prolific (an online crowdsourcing platform) with 20 severe scenarios in which the recipient of the news varied. Participants were asked to convey the news in an open text box, then through multiple choice options (see Table 1). The inclusion of open text alongside multiple choice (the most commonly used method in this line of work [5-6]) allowed us to conduct nuanced analyses by coding for Indirectness, Uncertainty, and Emotion (Table 2). Our LMER model of Indirectness with Recipient (Face-threat, Non-face-threat) and Age as FE and max RE structure revealed a main effect of Recipient ($p=.033$), whereby indirectness increased when the listener's face was threatened. This mirrors results from the multiple-choice model where lower probabilities were selected in the face-threatening context ($p<.001$). Our model of Uncertainty also revealed a main effect of Recipient ($p=.003$), with greater uncertainty expressed in the face-threatening context. Supporting our prediction, there was a Recipient x Age interaction ($p=.036$), whereby younger adults expressed greater uncertainty when the recipient's face was threatened, while older adults did not (Fig. 1). Finally, our model of Emotion revealed a main effect of Recipient ($p<.001$), with less emotion conveyed when the recipient's face was threatened (perhaps to mitigate the discomfort of the situation). Similar to the Uncertainty model, a Recipient x Age interaction ($p=.048$) revealed that younger adults modulated their use of emotion based on the recipient's face, unlike older adults (Fig. 1).

EXP 2 presented a new set of Prolific participants (N=100, ages 19-70) with 20 face-threatening scenarios in which the severity of the outcomes varied. Participants again conveyed the news through both text responses and multiple choice and the same coding was used from Exp 1. Here our LMER model of Indirectness with Severity (Severe, Less Severe) and Age as FE and max RE structure revealed no main effects or interactions (all p 's $>.05$). These results are in contrast to the multiple choice, where a main effect of Severity ($p=.001$) revealed that participants selected lower probability statements for the severe outcomes. In the Uncertainty model, there was also a main effect of Severity ($p=.047$), with greater uncertainty conveyed for the severe outcomes. Finally, the Emotion model revealed a main effect of Severity ($p=.027$), whereby more emotional language was used for the severe outcomes (perhaps as a way to convey sympathy).

Our study is the first to demonstrate age-related differences in how speakers relay news in face-threatening contexts. Confirming our main hypothesis, younger adults were more likely to adjust their speech along a number of dimensions (from indirectness to emotion) based on who the recipient was, likely due to enhanced audience design [7-10] or a difference in conversational goals [11]. We also found more speech modifications for Recipient than Severity. The absence of an effect of Severity on Indirectness suggests that estimates of severity may be perceived as less important than face-threat. Alternatively, adjusting for Recipient may be computationally easier than for Severity. Future work should further investigate these questions across the adult lifespan.

Table 1. Example trial from Exps 1 and 2

Exp 1 (Recipient manipulation)	
Scenario	Imagine that the company you work for has not been doing well financially. After a meeting with your boss, you are anxious that your co-worker will be made redundant. Later that day your co-worker (face-threat)/someone from a different department (non-face-threat) asks how the meeting went.
Open text	You tell your co-worker/the person from the other department: _____
Multiple choice	Out of the following options what would you tell your co-worker/the other person? <ul style="list-style-type: none"> ○ (1) It is highly unlikely you/my co-worker will be made redundant. ○ (2) It is somewhat unlikely you/my co-worker will be made redundant. ○ (3) It is possible you/my co-worker will be made redundant. ○ (4) There's a good chance you/my co-worker will be made redundant. ○ (5) It is almost certain you/my co-worker will be made redundant.

Exp 2 (Severity manipulation)	
Scenario	Imagine that the company you work for has not been doing well financially. After a meeting with your boss, you are anxious that your co-worker will be made redundant (severe)/receive a salary decrease (less severe) . Later that day your co-worker asks how the meeting went.
Open text	You tell your co-worker: _____
Multiple choice	Out of the following options what would you tell your co-worker? <ul style="list-style-type: none"> ○ (1) It is highly unlikely you will be made redundant/receive a salary decrease. ○ (2) It is somewhat unlikely you will be made redundant/receive a salary decrease. ○ (3) It is possible you will be made redundant/receive a salary decrease. ○ (4) There's a good chance you will be made redundant/receive a salary decrease. ○ (5) It is almost certain you will be made redundant/receive a salary decrease.

Table 2. Coding of variables

Coding	
Indirectness	1= Relay the bad news and give the reason for the bad news, 2= Relay the bad news only, 3= Give the reason in a way that requires an inference, 4= Don't give the bad news or lie
Uncertainty	1= Convey uncertainty (e.g. might, could, possible), 0= Don't convey uncertainty
Emotion	1= Convey emotions (e.g. I'm worried, afraid, etc.), 0= Don't convey emotions

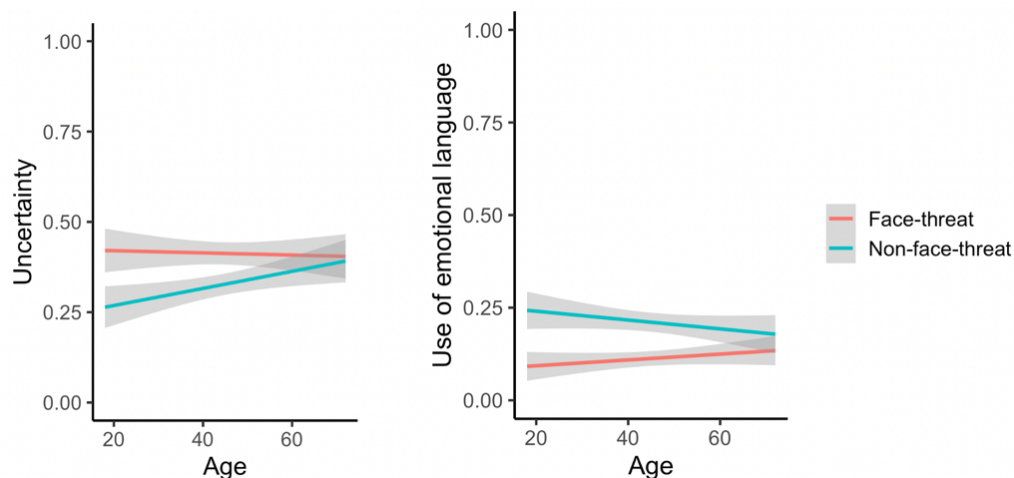


Figure 1. Recipient x Age interactions for Uncertainty (left) and Emotion (right) from Exp 1.

References: [1] Bonnefon & Villejoubert, 2006. *Psych Sci*. [2] Juanchich & Butler, 2012. *Organ Behav Hum Decis Process*. [3] Bonnefon et al., 2009. *Cognition*. [4] Feeney & Bonnefon, 2013. *J Lang So. Psychol*. [5] Juanchich & Sirota, 2013. *Q J Exp Psychol*. [6] Holtgraves & Perdue, 2016. *Cognition*. [7] Horton & Spieler, 2007. *Psychol Aging*. [8] Healey & Grossman, 2016. *Exp Aging Res*. [9] Long et al., 2018. *Cognition*. [10] Schubotz et al., 2019. *Lang Cognit Process*. [11] James et al., 1998. Production and perception of verbosity in younger and older adults. *Psychol. Aging*.

The social benefits of being a non-native speaker

Martin Ho Kwan Ip^{1,2}, Anna Papafragou^{1,2}

¹Integrated Language Sciences and Technology (ILST), University of Pennsylvania

²Department of Linguistics, University of Pennsylvania

Speaking in a foreign accent has often been thought to carry several disadvantages. Compared to native speech, accented utterances are less intelligible¹ and may make the non-native speaker appear more unpleasant². Foreign-accented speakers are more likely to face workplace discrimination³ and are less likely to be considered reliable or 'morally upright'⁴. Even infants are less likely to learn from, and be friends with, social partners who speak in a foreign accent^{5,6}.

Here we take the position that non-native speech sometimes carries a social *advantage*. We examined how listeners process underinformativeness, the pragmatic phenomenon of saying less than is conversationally required. Speakers are underinformative either because they are unable or unwilling to say more⁷. A recent study found that readers were more likely to seek information from an underinformative character after they read that she had a heavy foreign accent compared to a character with a native accent, presumably because underinformativeness is linked to inability in the non-native character⁸. Here, we probe the social evaluation of foreign-accented vs. native speakers more directly, using spoken stimuli to test if listeners form different impressions of underinformative native and non-native speakers.

EXPT1. Monolingual English speakers ($N = 576$, age range: 19-84 years) from MTurk viewed an illustrated story. The story took place in a mansion that had been robbed and vandalized and showed a woman calling the owner to tell her about the robbery. Her utterances were recorded by the same bilingual speaker, who produced three different speaker versions: native-accented (NS), non-native accented without grammatical errors (NNS), and non-native accented with grammatical errors (NNS with errors). We manipulated informativeness at the end of the story, where the young woman saw crates of apples and pineapples in an otherwise empty kitchen and said (referring to the robbers): "*They left some apples and pineapples*" (informative) or "*They left some apples*" (underinformative). This critical utterance was identical across all conditions. Both Speaker and Informativeness were between-speakers factors. Participants saw a single story and had to rate the woman (1-7 scale) on various personal attributes (i.e., honesty, likability, competence, likelihood of becoming their friend, and a good witness for the police). An ANOVA for each attribute with Informativeness and Speaker as factors revealed only an interaction of the two factors in honesty ratings, $F = 7.63$, $p = .001$ (Fig.1); the NS and NNS - but not the NNS with errors - were judged to be less honest when being underinformative compared to informative. A final question confirmed that people explained underinformativeness differently across Speaker types (Table 1).

EXPT2. We replicated Exp.1 with a new set of participants ($N = 576$, age range: 14-83 years) but replaced pineapples with money (a more desirable object). The interaction of Informativeness and Speaker for honesty remained, $F = 9.10$, $p < .001$ (Fig.2); the NNS and the NNS with errors showed smaller decreases in honesty ratings compared to the NS (cf. Table 1: unwillingness/deception was less likely to be invoked as the reason for omitting the money for the two NNSs). Additionally, participants indicated that they were less likely to be friends with the woman in underinformative contexts, but such a dip in likelihood was smaller when she was a non-native speaker, $F = 9.34$, $p = .015$ (Fig.3). Underinformativeness also led to lower competence, $F = 19.15$, $p < .001$, likeability, $F = 86.25$, $p < .001$, and witness potential ratings, $F = 120.34$, $p < .001$, but these did not vary by Speaker. Speaker type also affected likability, $F = 5.24$, $p = .006$, with the NNS with errors being better liked than both the NNS and the NS. At the end of both our experiments, listeners rated the woman's English to be better in the NS case than in the NNS case, which in turn was better than the NNS with errors, $F = 178.85$, $p < .001$.

Our findings show that listeners are less suspicious of underinformative speakers with heavy foreign accents, even in contexts where not saying what is required can be detrimental to or misleading for the listener. Contrary to previous studies, we also show no consistent global bias against non-native speakers. Thus the fact that non-native speakers have imperfect control of the linguistic signal can affect pragmatic interpretation and lead to unexpected social benefits.

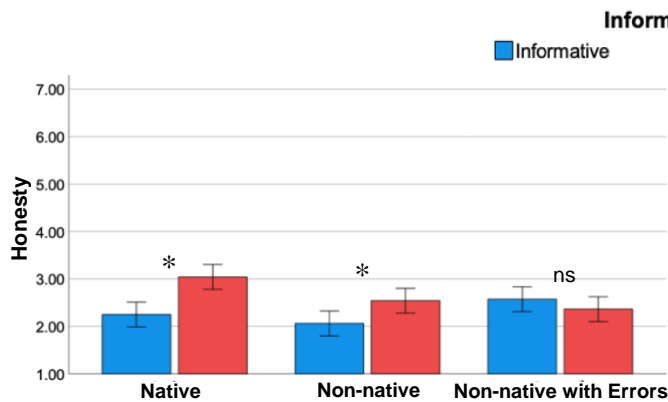


Figure 1. Honesty ratings in Experiment 1
(1=Extremely honest; 7=Extremely dishonest)

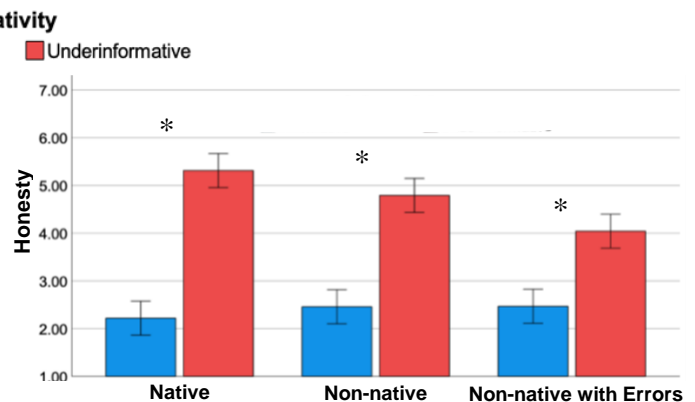


Figure 2. Honesty ratings in Experiment 2
(1=Extremely honest; 7=Extremely dishonest)

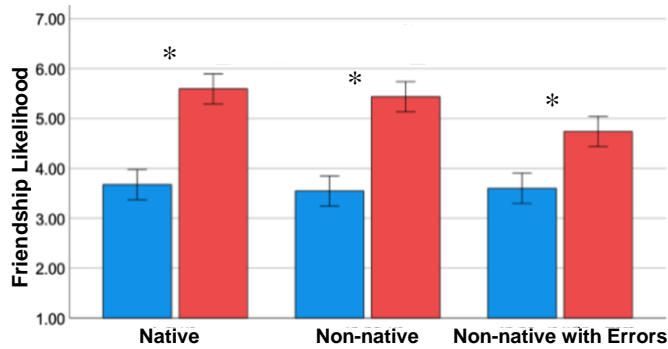


Figure 3. Friendship likelihood ratings in Experiment 2
(1=Extremely likely; 7=Extremely unlikely)

Table 1. Percentage of responses invoking unwillingness and incompetence as explanations for the omission of the second object (pineapples/money) in Experiments 1 and 2.

EXPT1	Unwilling	Unable	Other
NS	14.89%	21.27%	63.84%
NNS	7.95%	39.77%	52.25%
NNS Errors	4.00%	57.00%	39.00%
EXPT2	Unwilling	Unable	Other
NS	83.33%	2.08%	14.59%
NNS	73.96%	7.29%	12.79%
NNS Errors	53.13%	12.50%	34.37%

References

1. Munro, M.J., & Derwing, T.M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73-97.
2. Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? *Journal of Experimental Social Psychology*, 46, 1093-1096.
3. Kaylin, R., & Rayko, D.S. (1978). Discrimination in evaluative judgments against foreign-accented job candidates. *Psychological Reports*, 43, 1203-1209.
4. Tsurutani, C. (2012). Evaluation of speakers with foreign-accented speech in Japan. *Journal of Multilingual and Multicultural Development*, 33, 589-603.
5. Begus, K. et al. (2016). Infants' preferences for native speakers are associated with an expectation of information. *PNAS*, 113, 12397-12402.
6. Kinzler et al. (2007). The native language of social cognition. *PNAS*, 104, 12577-12580.
7. Grice, H.P. (1975). Logic and conversation. In P. Cole, J.J. Morgan (Eds.), *Syntax and semantics*, 3.
8. Fairchild, S. et al. (2020). Pragmatics and social meaning. *Cognition*, 200, 104171.

Viewing the Metaphor Interference Effect in context

Shaokang Jin & Richard Breheny (University College London)

In literal truth-value judgement tasks, participants take longer to judge metaphors as *literally false* than scrambled counterparts (control sentences). This is known as the *metaphor interference effect* (MIE) [1]. Two models of metaphor derivation provide competing accounts of the MIE. The attributive categorization model argues that, rather as in a Stroop task, the MIE results from automatic metaphorical meanings, whose truth-value conflicts with the literal [1,2]. The structure-mapping model proposes that the interference is caused by an initial alignment to find the basis of an analogy that underpins the figurative meaning (not interference from figurative meaning itself) [3,4]. Here we assume an automatic attribution of figurative meaning but propose that an important factor contributing to delay in task response is uncertainty over which figurative meaning a sentence has, due to lack of context in typical MIE-task stimuli. (1.a-b) illustrates how metaphors are typically ambiguous without context. It is well-established that unresolved ambiguity can tap resources [5,6] and this could delay selection of the literal. Thus, we predict that a constraining context will eliminate or decrease the delay. [1,2] predicts, if anything, context will increase delay due to greater salience of figurative meaning. [3,4]'s initial process is context independent [7] and so does not predict difference with context. Results of Exp.1 confirm our prediction but we still find an MIE with context. Exp 2 explores the timecourse of participants deriving metaphorical meaning(s) and shows that, with context, figurative meanings are available at the same time as verification RTs in Exp.1. We conclude that the MIE can result from uncertainty over figurative meaning computation, or stroop-like interference where context strongly constrains.

Experiment 1: We follow the general design of [1] except we add a Context condition. 24 metaphors and their scrambled counterparts plus context sentences were employed in a 2*2 between groups design. Participants (N=48) were instructed to judge the literal truth-value of target sentences in either a no-context or a context condition (see Table 1). The context sentence was formulated so that target sentence was an elaboration and thus it strongly constrained figurative meaning. Literal fillers counterbalance response biases. *Results:* We found main effects of form and context and an MIE in both conditions, although RTs for metaphors in Context are sig. lower than no-Context (see Fig. 1).

Experiment 2: 48 participants made comprehensibility decisions (*comprehensible* or *incomprehensible*) to the same set of target sentences in either the no-context or the context condition; occasionally, they were asked to paraphrase the target sentence they had read. Unsurprisingly, decisions took longer with no context. Analysis of the data from the four groups across Exps. 1&2 showed RTs for comprehensibility decisions were later than verification RTs only in no-Context condition (replicating [3]) – see Fig. 2.

Discussion: The effect of strong constraining context on the MIE is surprising for different theoretical accounts of the effect ([1,3]). Our results support the marriage of an attributive model with current models of language processing under uncertainty.

1. He is a cactus.
 - a. Mary's boyfriend is an awkward character and often says unkind things. He is a cactus.
 - b. Mary's boyfriend loves spending the day in the desert, in the hot sun. He is a cactus.

CONDITIONS	SAMPLE ITEMS	
	CONTEXT	TARGET
CONTEXT	The man who lives next door is a grubby, shifty person.	Metaphor: Some men are cockroaches.
	The man who lives next door is a grubby, shifty person.	Scrambled counterpart: Some men are duvets.
NO-CONTEXT	/	Metaphor: Some men are cockroaches.
	/	Scrambled counterpart: Some men are duvets.

Table 1: Metaphor or scrambled counterpart followed a context sentence or not. Scrambled forms were constructed by paring the topic ('some men') with the vehicle of another of the 24 metaphors to yield a sentence that is low in sensicality.

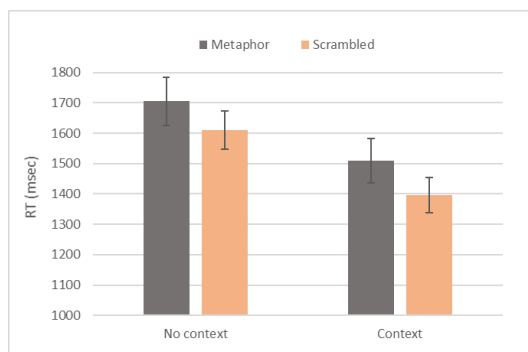


Figure 1

Mean RT (and standard errors of the means) to make *literally-false* decisions to metaphors and their scrambled counterparts in the two conditions.

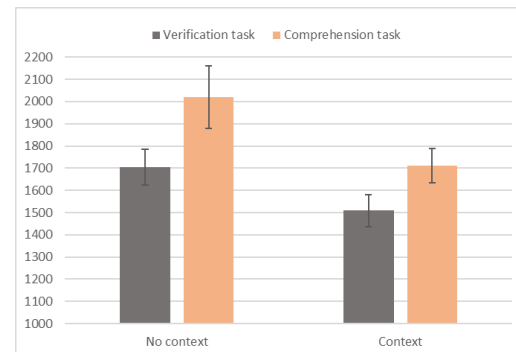


Figure 2

Mean RT (and standard errors of the means) for the sentence-verification task (metaphors only) and the metaphor comprehension task in the two conditions.

References: [1] Glucksberg, Gildea & Bookin (1982). *Journal of Verbal Learning and Verbal Behavior* 21, 85-98. [2] Gildea & Glucksberg (1983). *Journal of Verbal Learning and Verbal Behavior* 22, 577-590. [3] Wolff & Gentner (2000). *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26, 529-541. [4] Wolff & Gentner (2011). *Cognitive Science* 35, 1456-1448. [5] Duffy, Morris & Rayner (1988). *Journal of Memory and Language* 27, 429-446. [6] Griffiths, Steyvers & Tenenbaum (2007). *Psychological Review* 114, 211-244. [7] Gentner (1989). In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199-241).

How many response options in a TVJT? It depends

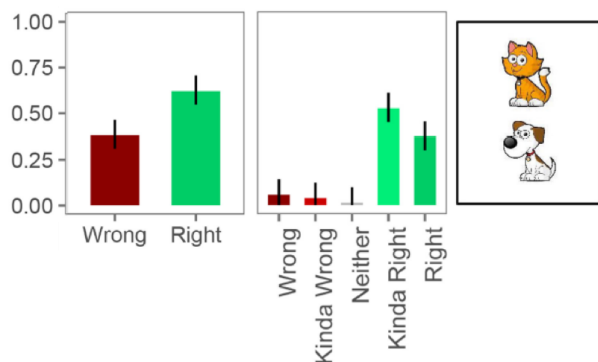
Yuhan Zhang, Giuseppe Ricciardi & Kathryn Davidson (*Harvard University*)

Investigations of compositional semantics and pragmatics rely on quantitative data collection, and there is increasing awareness that basic task features affect participant behavior in ways that crucially bear on theoretical conclusions (e.g. Katsos & Bishop 2011; Sprouse & Almeida 2017; Jasbi et al. 2019; Davidson, 2020; Marty et al. 2020; Waldon & Degen 2020). Here, we focus on the effect of the number of response options in sentence evaluation tasks with a context (i.e., TVJTs/“truth value judgment tasks”) by comparing adult participant behavior with two and five options across five different semantic phenomena: three data points come from Jasbi et al. 2019 (scalar implicature of ‘or’, *ad hoc* scalar implicature, conjunction) and two from novel experiments (*de re* and *de dicto* definite DPs). We argue that when it comes to deciding how many response options to offer in a TVJT, the most insightful practice is to manipulate their number.

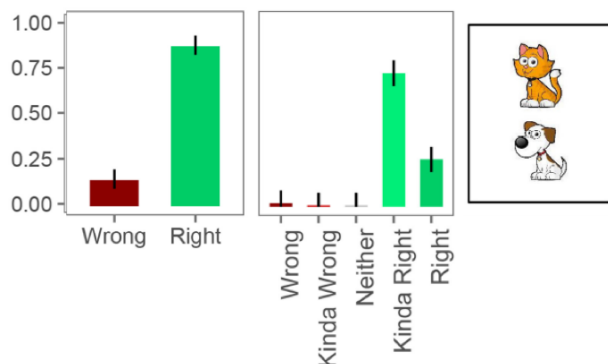
A: lexical scalar implicature. In Jasbi et al. 2019, participants evaluated the guesses of a blindfolded character about the content of a card. When the guess was “There is a cat or a dog” and the card contained a cat and a dog (Fig 1), 38% of the participants selected “wrong” in the binary condition, but in the quinary almost everyone chose either “kinda right” or “right”. This suggests that choosing binary “wrong” was a resistance to choosing “right”; in the quinary, this judgment was realized instead as “kinda right”. **B: *ad hoc* scalar implicature.** When the character guessed “There is a cat” when the card had a cat and a dog (Fig 2), most participants selected “right” in the binary condition, while in the quinary, instead, the majority selected “kinda right”. The binary pattern (falsely) suggests that many English speakers are not sensitive to the pragmatic oddity of the sentence; in contrast, in the quinary condition participants are clearly aware that it is not entirely “right”. **C: conjunction.** Here, the character guessed “There is a cat and a dog” when only a cat was pictured (Fig 3). Participants overwhelmingly judged this “wrong” in the binary condition, as expected based on the semantics of the boolean connective; however, in the quinary condition many participants chose “kinda wrong” and “kinda right”, perhaps induced by the presence of intermediate options to analyze the conjunctive statement as a sequence of independent statements, one being true (“there is a cat”) and the other false (“there is a dog”). In this case, intermediate options create a task demand not fitting the immediate theoretical goal. **D: *de re* definite DP.** In the novel task, participants were asked to evaluate a belief statement where the subject of the embedded clause was a definite DP interpretable as *de re* or *de dicto* depending on the context (Table 1). As in Jasbi et al. 2019, we manipulated between-subjects the number of response options (binary vs. quinary). In the binary condition of *de re* trials (Fig 4), one can observe a bimodal pattern, similar to that of case A. However unlike case A, this pattern persists in the quinary condition, suggesting that in the *de re* case what is underlying participants behavior is inherent disagreement about the truth/falsity of the sentence, rather than sensitivity to the pragmatic oddity of the sentence as in case A. **E: *de dicto* definite DP.** In the *de dicto* trials (Fig 5), most participants accepted the sentence in the binary condition, superficially similar to the binary condition of case B; however in this quinary condition, participants did not shift to the intermediate option, which we take to be due to *de dicto* sentences being judged as both true and pragmatically felicitous.

The overall picture is that more intermediate response options can reveal multiple underlying patterns, which may be due to TVJTs relying sometimes on primarily pragmatic judgments (see cases A & B) and other times on semantic ones (see cases D & E). Based on this, one might be inclined to give up binary TVJTs altogether. However, as shown by case C, the presence of more response options can, depending on the properties of the phenomenon under investigation and the specific experimental setup, induce additional unintended inferences. Therefore, we conclude that, especially for understudied phenomena in semantics and pragmatics, it may be most informative to design TVJTs by manipulating the number of response options and draw conclusions based on a comparison across conditions.

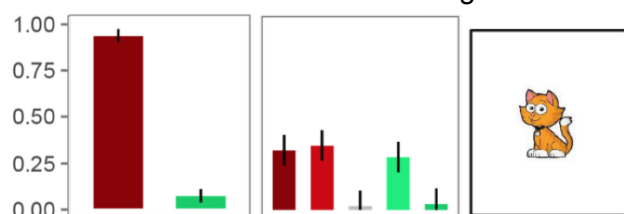
(A) Fig. 1 Lexical scalar implicature
“There is a cat or a dog.”



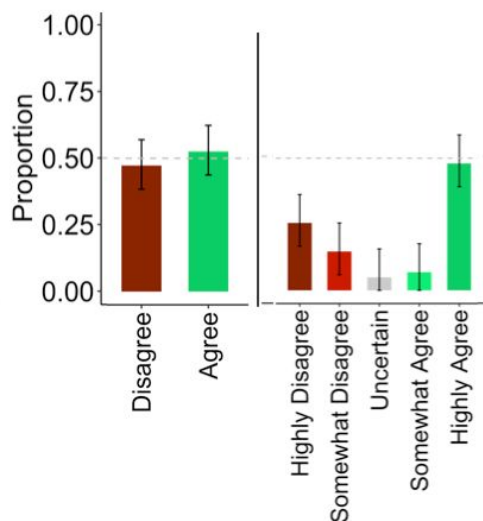
(B) Fig. 2 *Ad hoc* scalar implicature
“There is a cat.”



(C) Fig. 3 Conjunctive statement
“There is a cat and a dog.”



(D) Fig. 4 *De re* reading of Definite DPs
“...believe...Elizabeth’s poem...”



(E) Fig. 5 *De dicto* reading of Definite DPs
“...believe...Nicole’s poem...”

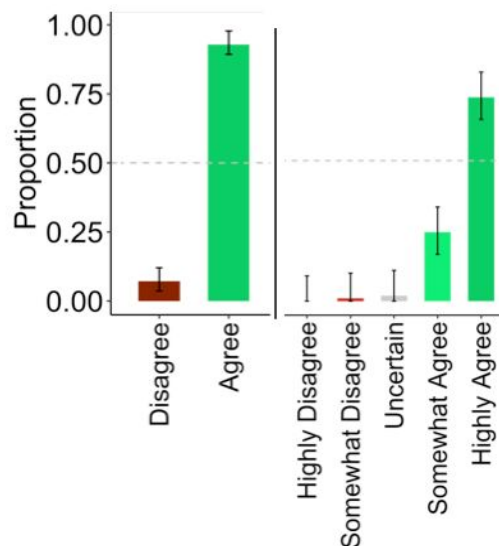


Table 1 An example context of experiments testing *de re/de dicto* readings of definite DPs

Context: Julie is one of several judges of an ongoing poetry competition. The best poem that she’s read so far is an extremely intriguing poem about the ocean. She believes that this poem will win the competition. Julie remembers being told that Nicole, one of the best-known contemporary poets, submitted a poem about the ocean to the competition. Therefore, Julie concludes that the first prize will be going to Nicole. However, this poem was actually written by Elizabeth, a younger and lesser-known poet. It is just a coincidence that the two poets wrote on the same topic.

Please indicate whether/to what extent you agree or disagree with the following statement.

S_{Target}: Julie believes that [Elizabeth’s poem] is going to win the competition. (*de re*)

S_{Target}: Julie believes that [Nicole’s poem] is going to win the competition. (*de dicto*)

Are there segmental and tonal effects on syntactic encoding? Evidence from structural priming in Mandarin

Chi Zhang (Ghent University), Sarah Bernolet (University of Antwerp), Robert J. Hartsuiker (Ghent University)

Numerous studies have established that speakers tend to form utterances by reusing previously experienced sentence structures (i.e., structural priming, Bock, 1986). Repetition of lexical items enhances such structural priming (i.e., lexical boost, Pickering & Branigan, 1998). This facilitation effect occurs not only when there is a full overlap of verbs, but also when one level of the lexical representation (semantic or phonological representation) overlaps between the prime and the target (e.g., Santesteban, Pickering, & McLean, 2010). It is unclear however which levels of representation drive the phonological boost, as the critical items in alphabetic languages like Dutch and English (e.g., *bat*[animal] – *bat*[sports]) overlap in orthography as well as phonology. Further, studies in these languages did not tease apart effects of segmental overlap and overlap in metrical structure. Here, we used Mandarin to scrutinize phonological effects on structural priming. This logographic language allowed us to test whether the phonological boost is independent of orthographic overlap, and whether it is driven by overlap of segments, tone, or both.

In five structural priming experiments (three lab-based, two web-based experiments), native Mandarin speakers described transitive pictures after receiving SVO or SOV “ba” prime sentences (see Table 1). In Experiment 1 ($n = 40$), prime and target verbs had lexical overlap (e.g., 脱[tuo1, to take off]-脱[tuo1], 1a-b), semantic overlap (e.g., 卸[xie4, to remove]-脱[tuo1], 2a-b), phonological overlap (e.g., 拖[tuo1, to mop]-脱[tuo1], 3a-b), or no overlap (e.g., 洗[xi3, to wash]-脱[tuo1], 4a-b) while similarities at other levels were carefully avoided. The phonological overlap condition consisted of verb pairs that overlapped in their full phonological representation or only overlapped in syllable. There was an overall structural priming (16.4%) and a lexical boost (14.8%, see Fig.1A), but semantic or phonological overlap did not boost priming (non-significant modulation of priming for semantic overlap = 2.5%; phonological overlap = -3.9%).

The next two experiments tested the full phonological boost and segmental boost effects in a lab-based experiment (Experiment 2a; $n = 72$) and a large-scale online replication (Experiment 2b; $n = 216$). Verbs in prime and target had full phonological overlap (segmental+tonal, e.g., 拖[tuo1]-脱[tuo1], 1a-b), syllabic overlap only (e.g., 驮[tuo2, to carry]-脱[tuo1], 5a-b), or no overlap. Both experiments showed significant overall structural priming (18.6% and 34.0%) and a boost from full phonological overlap (full phonological boost = 4.9% and 8.0%, see Fig.1B). The effect of syllabic overlap (3.3%) was not significant in Experiment 2a. However, more decisive evidence for a syllabic boost effect (6.5%) was found in the well-powered online replication.

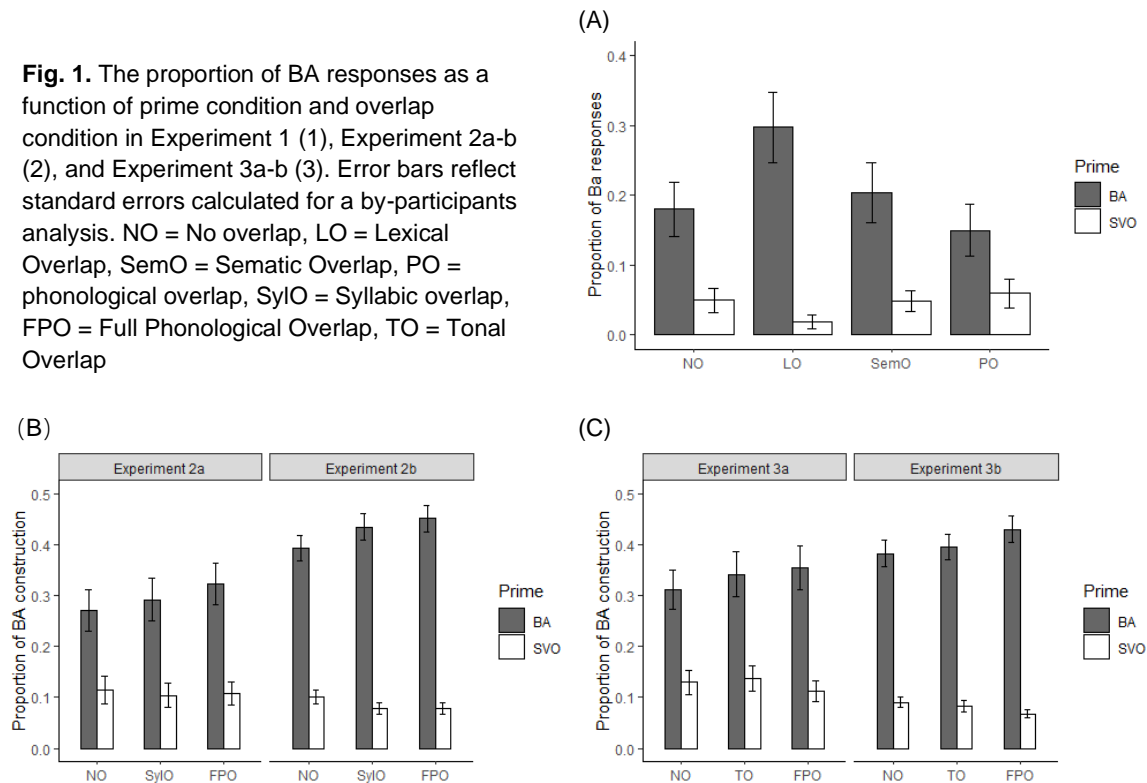
Experiments 3a-b ($n = 72$; $n = 216$) replicated Experiments 2a-b but replaced the syllabic-only condition with a tonal-only condition (称[cheng1, to weigh]-脱[tuo1], 6a-b). The facilitation effect of full phonological overlap was replicated in both experiments (full phonological boost effect = 5.9% and 7.1%, see Fig.1C), demonstrating once more that there is a phonological boost on priming even in the absence of orthographic overlap. However, no evidence of a tonal overlap effect was observed (tonal boost effect = 2.2% and 2.1%).

Together, these results indicate that processing at the phonological level feeds back to syntactic encoding in sentence production, which further supports an interactive view of language production. Phonological feedback effects on syntactic choice seem to be restricted to feedback from the syllabic level. We speculate that this is because feedback from the metrical level (tone) is less specific than syllabic feedback; an activated representation of tone would feed back to thousands of word forms sharing that tone, whereas a syllable would feed back to only a few.

Table 1: Exemplar prime sentences in each condition. The corresponding to a target picture that depicts a secretary taking off a jacket.

Exemplar SVO prime sentence		Exemplar SOV "ba" prime sentence	
(1a)	<i>Mama tuo1-LE yurongfu.</i> Mum take-off-LE the down jacket. [Mum took off the down jacket.]	(1b)	<i>Mama BA yurongfu tuo1-LE.</i> Mum BA the down jacket take-off-LE. [Mum took off the down jacket.]
(2a)	<i>Shibing xie4-LE toukui.</i> The soldier remove-LE the helmet. [The soldier removed the helmet.]	(2b)	<i>Shibing BA toukui xie4-LE.</i> The soldier BA the helmet remove-LE. [The soldier removed the helmet.]
(3a)	<i>Qingjiegong tuo1-LE yangtai.</i> The cleaner mop-LE the balcony. [The cleaner mopped the balcony.]	(3b)	<i>Qingjiegong BA yangtai tuo1-LE.</i> The cleaner BA the balcony mop-LE. [The cleaner mopped the balcony.]
(4a)	<i>Siji xi3-LE che</i> The driver wash-LE the car [The driver washed the car.]	(4b)	<i>Siji BA che xi3-LE.</i> The driver BA the car wash-LE. [The driver washed the car.]
(5a)	<i>Xiaoniao tuo2-LE shuiguo.</i> The bird carry-LE the fruit. [The bird carried the fruit.]	(5b)	<i>Xiaoniao BA shuiguo tuo2-LE.</i> The bird BA the fruit carry-LE. [The bird carried the fruit.]
(6a)	<i>Hushi cheng1-LE ying'er.</i> The nurse weigh-LE the baby. [The nurse weighed the baby.]	(6b)	<i>Hushi BA ying'er cheng1-LE.</i> The nurse BA the baby weight-LE. [The nurse weighted the baby.]

Fig. 1. The proportion of BA responses as a function of prime condition and overlap condition in Experiment 1 (1), Experiment 2a-b (2), and Experiment 3a-b (3). Error bars reflect standard errors calculated for a by-participants analysis. NO = No overlap, LO = Lexical Overlap, SemO = Semantic Overlap, PO = phonological overlap, SyO = Syllabic overlap, FPO = Full Phonological Overlap, TO = Tonal Overlap



References

- Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3), 355-387. [https://doi.org/10.1016/0010-0285\(86\)90004-6](https://doi.org/10.1016/0010-0285(86)90004-6)
- Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39(4), 633-651. <https://doi.org/10.1006/jmla.1998.2592>
- Santesteban, M., Pickering, M. J., & McLean, J. F. (2010). Lexical and phonological effects on syntactic processing: Evidence from syntactic priming. *Journal of Memory and Language*, 63(3), 347-366. <https://doi.org/10.1016/j.jml.2010.07.001>

The dynamic prominence status of thematic roles in simulated Mandarin conversations

Fang Yang, Martin Pickering, Holly Branigan (University of Edinburgh)

In discourse the prominence status of an entity changes across time. Language systems employ syntactic and information-structural operations to reflect such dynamic status (see [1] for a review). For example, Mandarin constructions (2) – (6) all assign prominence to the Patient (Beckham) but with different magnitude. Specifically, BA-construction (6) encodes the Patient before the verb rendering it conceptually more prominent than a neutral Patient in a canonical SVO structure (1) but still less prominent than the sentence-initial Agent (Obama in (6)), whereas topicalisation (TOP), left-dislocation (LDT), focalisation (FOC) or passive encodes the Patient in the sentence-initial position ranking it more prominent even than the Agent (2-5). However, we still know little about how speakers in conversation accommodate discourse constraints when generating messages that reflect the dynamic prominence status of thematic roles in an event [2]. Do they maintain the prominence status of one particular thematic role across different messages? Do they take into consideration their interlocutors' information-seeking goals?

We investigated this with Mandarin speakers in three experiments (N=48, 64 & 39) using a confederate-scripted priming paradigm in which participants and a confederate took turns to describe pictures to each other and used a keyboard to indicate whether their pictures matched or mismatched their interlocutor's descriptions (mismatched pictures had one difference in either Patient or Agent). The confederate always gave descriptions first using SVO, TOP, LDT or an intransitive baseline in Expts 1&3, or using SVO, TOP, FOC or an intransitive baseline in Expt 2. Participants then described a different picture depicting the same action with different Agent (animate) and Patient (inanimate). Additionally, in Expt 3 interlocutors asked each other a scripted question before the other gave descriptions and the Patient in the target picture was always topicalised in a question (e.g. *the table, who knocked-over?*). Across all experiments, participants showed a tendency to maintain the prominence status of the Patient when generating different messages: they were more likely to produce patient-prominent responses after a TOP ($p < .001$ in Expts 1&2; $p < .01$ in Expt3) or FOC ($p < .001$ in Expt 2) than an SVO prime. Interestingly, LDT led to more patient-prominent responses than SVO did ($p < .01$) but less than TOP did ($p < .05$) in Expt1, however, both differences disappeared in Expt 3 ($p = .52, .28$). Given that LDT shares prominence representation with TOP and (at least partially) syntactic representation with SVO, and that the topic-setting question interfered with primes in Expt 3, these results cannot be explained by purely syntactic priming but better explained by a priming effect of prominence independent of syntax.

Moreover, even while maintaining prominence status, participants used constructions that were not used by their interlocutor. In Expts 1&2, they tended to use BA-construction (98% of patient-prominent responses in Expt 1; 86% in Expt 2) to elevate the prominence status of the Patient to a higher gradient but not as high as the animate Agent, suggesting that while maintaining prominence of the Patient, speakers adjust its magnitude to accommodate discourse constraints (e.g. animacy hierarchy). In contrast, in Expt 3 where participants' descriptions constituted an answer to their interlocutor's topic-setting questions, when producing patient-prominent responses they tended to use an ellipsis (45%), passive (20%) or TOP construction (25%) to rank the Patient more prominent even than the Agent despite the constraints of animacy hierarchy (significant effect of experiment in a combined analysis of Expts 1&3: $p_{MCMC} < .01$). This suggests that speaker's knowledge of their addressee's communicative goals influences their encoding of entity prominence in message planning in a top-down fashion that outweighs animacy.

Taken together, our studies show that speakers maintain the prominence status of a thematic role across different messages and in doing so they accommodate pragmatic constraints in dialogue.

Table 1. Prominence status of the Patient in different constructions in Mandarin

Example						Construction	Prominence status of the Patient
(1)	<i>Aobama</i> Obama	<i>ti-dao</i> kick-fall	<i>le</i> aspect-marker(ASP)	<i>Beikehanmu.</i> Beckham		SVO	Neutral
(2)	<i>Beikehanmu,</i> Beckham	<i>Aobama</i> Obama	<i>ti-dao</i> kick-fall	<i>le.</i> ASP-LE		TOP	Topicalised
(3)	<i>Beikehanmu,</i> Beckham	<i>Aobama</i> Obama	<i>ti-dao</i> kick-fall	<i>le</i> ASP-LE	<i>ta.</i> him	LDT	Left-dislocated
(4)	<i>Shi</i> Focus-marker	<i>Beikehanmu</i> Beckham	<i>bei</i> BEI	<i>Aobama</i> Obama	<i>ti-dao</i> kick-fall	<i>le.</i> ASP-LE	FOC Focalised
(5)	<i>Zhuozi</i> table	<i>bei</i> BEI	<i>Chenglong</i> Jackie Chan	<i>ti-dao</i> kick-fall	<i>le.</i> ASP-LE	BEI (Passive)	BEI-subject
(6)	Chenglong Jackie Chan	<i>ba</i> BA	<i>zhuozi</i> table	<i>ti-dao</i> kick-fall	<i>le.</i> ASP-LE	BA	BA-object
(7)	Chenglong Jackie Chan	<i>ti-dao</i> kick-fall	<i>de.</i> ASP-DE			Ellipsis	Null-pronominalised

References

- [1] Von Heusinger, K., & Schumacher, P. B. (2019). Discourse prominence: Definition and application. *Journal of Pragmatics*, 154, 117-127. <https://doi.org/10.1016/j.pragma.2019.07.025>
- [2] Ünal, E., Ji, Y., & Papafragou, A. (2019). From Event Representation to Linguistic Meaning. *Topics in Cognitive Science*. <https://doi.org/10.1111/tops.12475>

Morphological boost in structural priming: Evidence from Czech

Maroš Filip^{1,2}, Filip Smolík^{2,1}

¹ Faculty of Arts, Charles University, Prague

² Institute of Psychology, Czech Academy of Sciences, Prague

Today's research shows that structural priming effects are often supported by non-structural aspects of language. E. g. Ziegler and colleagues (2019) suggested that solely abstract structure is not sufficient to elicit syntactic priming and that other factors are usually needed, e.g. effects of animacy, semantic structure, information structure, shared phonology or others.

These findings suggest that morphology could also play some role in syntactic priming. However, to our knowledge, this issue has been addressed only in two studies, which yield contradictory results. Santesteban and colleagues (2015) did not find the evidence that case endings in ergative Basque language contribute to structural priming. On the other hand, Chung and Lee (2017) successfully primed the use of case markers in Korean. In Czech the same case can be encoded with different endings in different nouns. We can therefore address the question whether the repetition of the same morpheme used for marking a grammatical function can enhance the priming more than the using different morpheme coding the same case.

We executed two experiments masked as memory tests, in which participants read prime sentences and described following target pictures for later recall. We modified two independent variables in primes – type of the sentence and noun case endings. Sentence type had three levels – double-object construction with dative-accusative order or with accusative-dative order, or a neutral intransitive prime sentence. Case endings for nouns in accusative and dative had two conditions – in same-suffix condition the prime nouns had suffixes identical to suffixes used in the target nouns. In the different-suffix condition, the dative and accusative nouns in the target were inflected using different suffixes. In Experiment I (N=59), primes with four different dative markers were used.

Prime:

Same suffix: Kráva olizuje ovečc-e hlav-u/Cow licks sheep-DAT head-ACC –

Different suffix: Průvodce popisuje návštěvník-ovi ulic-i/Guide describes visitor-DAT street-ACC

Target: Klaun nabízí baletc-e žvýkačk-u/Clown offers ballerina-DAT chewing gum-ACC

Generalized linear mixed-effect model with random intercepts for subjects and items revealed that in different-suffix sentences, no significant effect of priming against the neutral condition, neither acc/dat ($p=0.239$) nor for dat/acc sentences ($p=0.527$) was found (Table 1). In same-suffix sentences we found significant effect of acc/dat primes compared to dat/acc ($p=0.005$, Table 2). The results suggest that effect like lexical boost exist also on morphological level, and that the ordering of case forms is easier to prime if these forms share case-marking suffixes.

In Experiment II (N=60), only two dative markers in primes were used (-ovi and -e). The results show similar pattern as the first experiment but are not significant. For different-suffix sentences, we again did not observe any significant effect against the neutral condition for acc/dat ($p=0.284$) nor for dat/acc sentences ($p=0.454$) (Table 3). For same-suffix sentences we found marginal effect of acc/dat structures compared to dat/acc ($p=0.068$) (Table 4). When tested for interactions between prime word order and marker agreement, it was marginally significant in Experiment 1 but not significant in Experiment 2.

Together, the experiments do not confirm that repeating the same case markers enhances priming, but they are suggestive of this possibility. However, the effect appears to be weak and perhaps limited to some markers.

Table 1

Model estimates in different ending condition in Experiment 1

Parameter	Est.	SE	P-value
Intercept (neut.)	-0.298	0.308	0.333
DA	0.154	0.243	0.527
AD	-0.288	0.245	0.239

Table 2

Model estimates in same ending condition in Experiment 1

Parameter	Est.	SE	P-value
Intercept (DA)	0.683	0.432	0.114
AD	-1.395	0.495	0.005 **

Table 3

Model estimates in different ending condition in Experiment 2

Parameter	Est.	SE	P-value
Intercept (neut.)	-0.040	0.304	0.896
DA	0.223	0.297	0.454
AD	-0.325	0.303	0.284

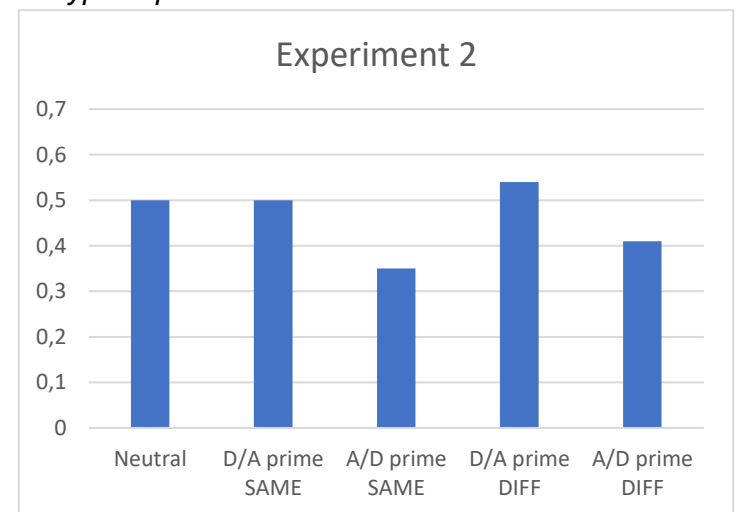
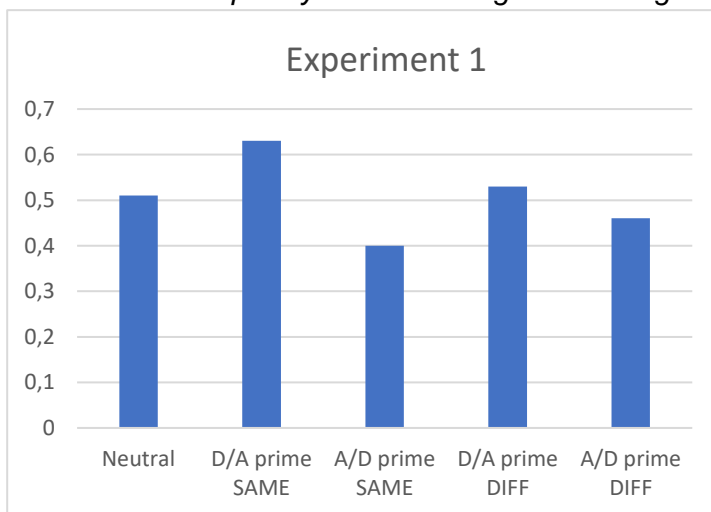
Table 4

Model estimates in same ending condition in Experiment 2

Parameter	Est.	SE	P-value
Intercept (DA)	-0.350	0.594	0.555
AD	-1.182	0.648	0.068

Graph 1

Relative frequency of dat/acc targets following different type of primes



(D/A = prime where dative precedes accusative, A/D = prime where accusative precedes dative
 SAME = sentences with same suffixes for accusative and dative nouns, DIFF = sentences with different suffixes for accusative and dative nouns)

References

- Chung, E. S., & Lee, E. K. (2017). Morpho-syntactic processing of Korean case-marking and case drop. *Linguistic Research*, 34(2), 191-204.
- Santesteban, M., Pickering, M. J., Laka, I., & Branigan, H. P. (2015). Effects of case-marking and head position on language production? Evidence from an ergative OV language. *Language, Cognition and Neuroscience*, 30(9), 1175.
- Ziegler, J., Bencini, G., Goldberg, A., & Snedeker, J. (2019). How abstract is syntax? Evidence from structural priming. *Cognition*, 193, 104045.

Syntactic Rule Frequency as a measure of Syntactic Complexity:

Insights from Primary Progressive Aphasia

Neguine Rezaii (Harvard Medical School), Rachel Ryskin (University of California), Kyle Mahowald (University of California), Bradford Dickerson (Harvard Medical School), Edward Gibson (Massachusetts Institute of Technology)

We investigate syntactic processing in the language of patients with primary progressive aphasia (PPA), a neurodegenerative clinical syndrome where language is the predominant initial impairment. Depending on the primary region of brain atrophy, PPA can have different psycholinguistic presentations. The nonfluent variant of PPA (nfvPPA) is characterized by simple and impoverished syntactic structures and/or effortful speech. In contrast, the other two variants of PPA are described based on lexico-semantic deficits: Individuals with the logopenic variant of PPA (lvPPA) exhibit difficulty with sentence repetition and lexical retrieval. In the semantic variant of PPA (svPPA), difficulties in object naming and word comprehension are the hallmark of the disorder (Gorno-Tempini 2011). In this work, we aim to use the frequency of syntactic rules as a measure of syntactic complexity based on the psycholinguistic literature suggesting that language comprehension is sensitive to the probability distribution of words and syntactic rules. The consequence of this finding is that the complexity of any utterance corresponds to the probability of the utterance in context. Thus one production complexity metric is one based on the frequency of combinatory syntactic rules: compared to control participants, nfvPPA patients might have relatively weaker access to lower frequency syntactic rules.

Methods. Clinical and language assessments and MRI scans were used to characterize 79 patients with PPA and its subtypes (29 nfvPPA, 26 lvPPA, and 24 svPPA). We also included 51 age matched healthy controls. Participants were asked to describe a drawing of a family having a picnic from the Western Aphasia Battery–R (Kertesz, 2007) using as many full sentences as they could. The recorded responses were transcribed by a researcher blind to the subtypes. Disfluencies were removed from the analyses. These language samples were then parsed using the Stanford Probabilistic Context-Free Grammar (PCFG) parser (Klein and Manning, 2003). We examined binary syntactic rules using the output from the dependency grammar parse. For this metric, we take each dependency in the dependency structure as a separate rule (e.g., amod-NOUN) (Figure 1).

Results. Figure 2 shows the 20 most common binary rules for binary dependency grammar in nfvPPA and healthy controls. Fitting a maximal mixed effect model with random effects for subject and sentence that predicts log syntactic rule frequency with patient subtype and sentence length as predictors, we found a main effect of *patient subtype* while controlling for sentence length, with higher binary syntactic rules likely to occur in nfvPPA than other subtypes ($\beta=0.18$, $SE=0.06$, $t(16234)=2.78$, $p<0.01$). To better control for the effect of sentence length, we sampled sentences from each of the four groups so that all groups would have similar sentence length distributions. We continued to find a main effect of *patient subtype*, with nfvPPA patients producing higher frequency syntactic rules ($\beta=0.29$, $SE=0.06$, $t(16234)=4.40$, $p<0.001$).

Conclusions. Language production in nfvPPA is characterized by use of high frequency syntactic rules, when compared with control, svPPA, and lvPPA language production. Our results suggest a syntactic rule-specific locus of impairment, in line with proposals of a syntax-specific component of language production (Garrett, 1980; Bock, 1995), perhaps localized to a particular brain area (Fedorenko, Williams & Ferreira, 2018).

Figure 1. Constructing dependency rules. Sentences were parsed using the Stanford Lexicalized Parser Package (v3.9.2). Dependency heads, underlined, were identified by the parser. The combination of a head and its dependent was considered as a binary dependency rule.

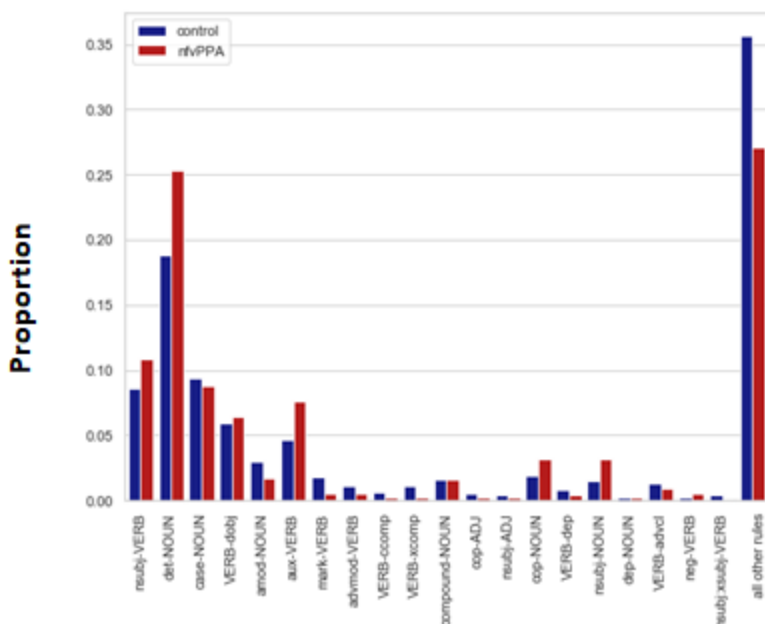
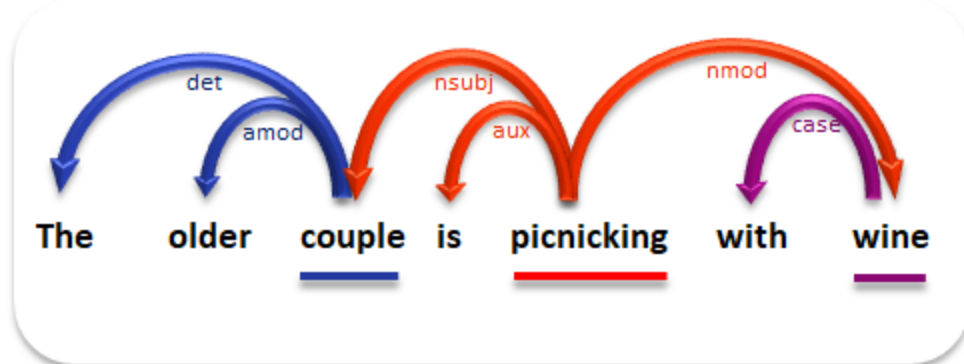


Figure 2. The proportion of 20 most common binary dependency grammar rules in nfVPPA and healthy controls. The last two bars show the proportion of all other binary dependency grammar rules combined.

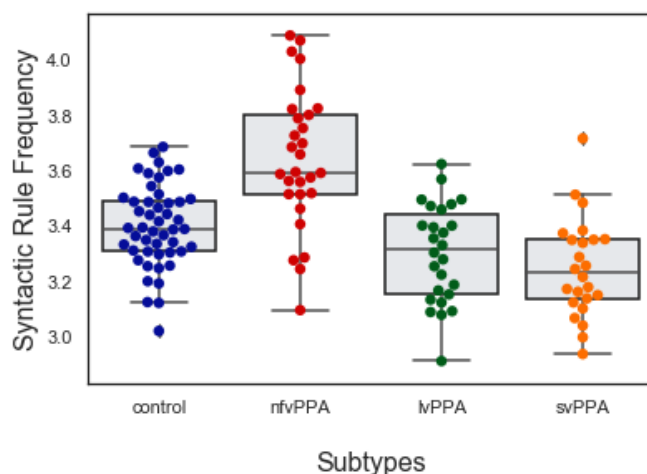


Figure 3. The box plots of syntactic rule frequency for the four groups

Reference

- Fedorenko, E., Williams, Z.M. & Ferreira, V.S. (2018). Remaining puzzles about morpheme production in the posterior temporal lobe. *Neuroscience*, 392, 160-163.
- Kertesz A. (2007). *Western Aphasia Battery (Revised)*. San Antonio: PsychCorp;
- Gorno-Tempini et al. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, 76(11), 1006-1014
- Klein, D. & Manning, C. D. (2003) Accurate unlexicalized parsing. *ACL*. 423–430
- Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production*, Vol. 1, (pp. 177-220). London: Academic Press.
- Bock, J. K. (1995). Sentence production: From mind to mouth. In J. L. Miller, & P. D. Eimas (Eds.), *Handbook of perception and cognition*.

Does deciding what to say involve deciding how to say it?

Ruth Corps^{1,2}, Holly Abercrombie², Ellie Demengeli², Alix Dobbie², Luke Raben², & Martin Pickering²

¹ Psychology of Language Department, Max Planck Institute for Psycholinguistics,
Ruth.Corps@mpi.nl

² Department of Psychology, University of Edinburgh

Answering a question involves conceptualisation (i.e., message preparation), formulation (i.e., linguistic encoding), and articulation [1], and selecting an answer is an aspect of conceptualisation. The answer then has to be formulated – the words have to be retrieved from the lexicon, assigned to a grammatical structure, and converted into phonological representations. But how are these processes related?

One possibility is that speakers select a single answer before formulating it, thus they make a final decision about the message without converting that message into words (selection-before-formulation). If this is the case, then selection should be unaffected by linguistic properties of unselected, but plausible, answers because they are not formulated. Alternatively, speakers could select a single answer only after they have formulated different potential answers (selection-after-formulation). Thus, speakers consider how they will produce their answer before settling on what they will produce. If this is the case, selection should be affected by linguistic properties of unselected answers because multiple answers are formulated.

We tested between these two possibilities in two question-answering experiments that exploited the fact that to-be-expressed answers vary in their linguistic complexity (e.g., *Harry Potter and the Philosopher's Stone*, *Dracula*). We manipulated the ease of selecting an answer by manipulating whether the questions were constraining (with most participants providing a particular answer concept) or unconstraining (with participants providing different answer concepts; see Table 1, all stimuli were pre-tested). We also manipulated the length of answers, so that they were short and linguistically simple, or long and linguistically complex.

If answer selection occurs before formulation, then there should be an interaction between question constraint and answer length. In particular, we expect stronger effects of answer length when the question is unconstraining compared to constraining because speakers will tend to activate and formulate a larger set of linguistically complex items. In contrast, if answer selection occurs after formulation, then participants should be equally slower to answer unconstraining than constraining questions regardless of whether the set of potential answers is long or short, because they will formulate only one answer.

In Experiment 1, native English monolinguals (N=40) answered more quickly when questions were constraining ($M=647$ ms) rather than unconstraining ($M=1279$ ms; $t=-7.26$), suggesting they found it easier to answer when they did not have to extensively search through a number of potential answers during retrieval. Consistent with previous research [2], participants also answered more quickly when answers were shorter rather than longer ($t=3.73$). Importantly, there was no interaction between these two predictors ($t=-0.09$; Bayes Factor=0.88), suggesting that speakers were not affected by the complexity of unselected, but plausible, answers.

We replicated these findings in Experiment 2, in which we increased the cognitive load of the task by recruiting L2 English speakers (N=41), who are likely to have more difficulty accessing concepts and preparing answers, particularly if they consider the complexity of unselected concepts. Participants answered more quickly when questions were constraining ($M=1177$ ms) rather than unconstraining ($M=1816$ ms; $t=-5.49$) and when answers were shorter rather than longer ($t=-3.20$). Consistent with Experiment 1, there was no interaction between question constraint and answer length ($t=-1.02$, Bayes factor=0.78).

We conclude that speakers decide what to say early, during pre-linguistic planning, so that only a single answer is processed further during formulation. These findings suggest that speakers can decide what they want to produce without considering the complexity of how they are going to produce it.

References

- [1] Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
[2] Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30, 210-233.

Table 1. Example stimuli for the four conditions in both experiments.

Question Constraint	Answer Length	Question
Constraining	Short	What is the capital of France?
	Long	How did The Titanic sink?
Unconstraining	Short	What is your favourite city?
	Long	What is your favourite book?

Early preparation during question-answering: Speakers prepare content but not form

Ruth Corps^{1,2}, Laura Lindsay, & Martin Pickering²

¹ Psychology of Language Department, Max Planck Institute for Psycholinguistics, Ruth.Corps@mpi.nl, ² Department of Psychology, University of Edinburgh

Conversation is a puzzle: Formulating an utterance takes at least 600 ms [1], but interlocutors' turns are so finely coordinated that there is often little gap between their contributions [2]. Most theories agree that interlocutors achieve such timing by predicting what the current speaker is likely to say, so that they can prepare a response early while still comprehending (the early-planning hypothesis; [3]). But do speakers prepare as much of their response as they can?

One possibility (an *early-form* account; [4]) is that speakers complete all stages of formulation early, and so they prepare both the content and the form of their turn while still comprehending. Preparing in this way removes the timing burden of response preparation from language production: Speakers know what they will say and how they will say it before articulating. But dual-tasking production and comprehension is cognitively demanding [5] and preparation may interfere with concurrent comprehension [6]. As a result, speakers may minimise these cognitive demands by preparing the content of their turn early, but the form late (a *late-form* account).

We tested between these hypotheses in two experiments using a verbal question-answering task using questions with high answer agreement (as determined by pretest). In both experiments, the critical information necessary for response preparation was available either early, so that participants could prepare their answer before question end, or late, so that they could not (see Table 1; [7]). To determine whether participants who prepared their answer early did so all the way up to form, we manipulated the length of to-be-prepared answers, so that they were either short (single word) or longer (multi-word) answers. We analysed answer times using linear-mixed effects models, with maximal random structure.

In Experiment 1, participants (N=42) answered more quickly when the critical information necessary for preparation occurred early ($M=388$ ms) rather than late ($M=824$ ms; $t=-4.85$), suggesting they prepared the content of their answer early. Participants also answered more quickly when their answer was short ($M=578$ ms) rather than long ($M=631$; $t=-1.93$), and there was some evidence that this effect depended on when participants prepared the content of their answer ($t=2.11$): They were affected by answer length when they prepared late ($t=-2.83$), but not when they prepared early ($t=-0.54$).

Experiment 1 provides some evidence that participants prepared the form of their answers early, supporting an early-form account and suggesting participants completed all stages of formulation. However, the effect of answer length was small and the effect was only marginally significant. This weak effect could have occurred because the difference in the average word length of answers in the short-answer and long-answer conditions was also quite small ($M^{\text{difference}}=1.26$). In Experiment 2, we therefore increased the word length of answers in the long-answer condition (from M of 2.27 words in Experiment 1 to 3.64 words in Experiment 2).

In Experiment 2, participants (N=92) again answered more quickly when the critical information necessary for preparation occurred early ($M=252$ ms) rather than late ($M=852$ ms; $t=-8.68$). Participants also answered more quickly when answers were short ($M=405$ ms) rather than long ($M=698$; $t=-2.79$). Unlike Experiment 1, however, there was no interaction ($t=0.07$): The difference between the two answer conditions was 270 ms for early questions and 297 ms for late questions. The Bayes Factor for this interaction was 0.49, providing no evidence for the alternative hypothesis.

Together, our findings are consistent with a late-form account and suggest that participants prepared the content of their answers early, but prepared the length late. These results provide insight into how speakers manage the cognitive demands of overlapping production and comprehension. In particular, speakers adopt a strategy that enables partial, but not complete, preparation, so that they can still allocate resources to comprehension.

References

- [1] Indefrey, P. & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92, 101-144.
- [2] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., . . . & Levinson, S. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106, 10587-10592.
- [3] Bögels, S. & Levinson, S. C. (2017). The brain behind the response: Insights into turn-taking in conversation from neuroimaging. *Research on Language and Social Interaction*, 50, 71-89.
- [4] Levinson, S. C. & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, <https://dx.doi.org/10.3389/fpsyg.2015.0073>.
- [5] Fairs, A., Bögels, S., & Meyer, A. S. (2018). Dual-tasking with simple linguistic tasks: Evidence for serial processing. *Acta psychologica*, 191, 131-148.
- [6] Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia*, 109, 295-310.
- [7] Bögels, S., Magyari, L., & Levinson, S. C. (2018). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, 5, <https://dx.doi.org/10.1038/srep12881>.

Table 1. Example stimuli for both Experiments 1 and 2. The critical information for preparation for the short conditions is *Barks*, while the critical information for the long conditions is *Harry Potter*

Answer Length	Critical Information	Question	Mean RT E1 (ms)	Mean RT E2 (ms)
Short	Early	Which animal barks and is also a common household pet?	427	109
	Late	Which animal is a common household pet and also barks?	711	701
Long	Early	Which platform, that appears in Harry Potter, can be found at Kings Cross Station?	330	379
	Late	Which platform can be found at Kings Cross Station and appears in Harry Potter?	933	998

Source of processing costs of indirect anaphors – self-paced reading and ERP data

Magdalena Repp, Petra B. Schumacher (Universität zu Köln)

Indirect anaphors (*Lisa went to a wedding in Italy. The bride was beautiful.*) encompass two different dimensions of newness: they represent new information and they introduce a new discourse referent into the mental model. Previous event-related potential (ERP) studies show an enhanced Late Positivity effect for indirect anaphors relative to (coreferential) direct anaphors (*A bride_i bought a wedding gown. The bride_i was very happy.*), which has been associated with the processing of newness (Burkhardt 2006). An open question remains whether the increased processing costs of indirect anaphors arise from the integration of a *new informational aspect* or through the integration of a *new discourse referent*. This question was addressed in two experiments via a comparison of indirect anaphors and so-called specification anaphors (*Marie_i bought a wedding gown. The bride_i was very happy.*). Specification anaphors resemble indirect anaphors in as far as they convey new information (about an already given referent, i.e. *the bride=Marie*) and they resemble direct anaphors by indicating a coreference relation. By contrast, indirect anaphors require the introduction of a new discourse referent.

First a self-paced reading (SRP) experiment was conducted where the reading times (RTs) of direct and indirect anaphors were compared to the RTs of specification anaphors (see Table1). The results indicate that the RTs of specification anaphors pattern with indirect anaphors in the critical region (see Fig.1), suggesting that the increased processing costs of indirect anaphors arise from the integration of new information. However, the RTs in the spill over regions show longer RTs for specification anaphors. We suggest that this indicates that specification anaphors are initially analyzed as new discourse referents and are subsequently recognized as being coreferential with an already given entity, when discourse unfolds. This reanalysis exerts costs. This leads to the conclusion that the increased processing costs of indirect anaphors observed in previous investigations of direct and indirect anaphors arise from the integration of a *new discourse referent*.

To follow up on this, an ERP study was carried out to contrast the three different types of anaphors and shed more light on the functional contribution to newness of the Late Positivity observed in previous research. The material for the ERP study was adapted to preclude that the specification anaphor was interpreted as a new referent: names of famous personalities were used as antecedents and commonly known information about them as the specification anaphor (e.g., *Joanne K. Rowling and the author* in Table1). The ERPs revealed a three-way modulation in the N400-window (300-500ms: indirect anaphor > specification anaphor > direct anaphor), reflecting different degrees of predictability, and a more pronounced Late Positivity over left-anterior electrode sites (600-800ms) for indirect anaphors relative to the other two anaphors (see Fig.2). This confirms the conclusions from the SPR study: The increased Late Positivity of indirect anaphors is associated with the establishment of a new discourse referent, lending support to the view that the positivity signals *newness of the discourse referent* rather than of information per se.

References

Burkhardt, Petra. 2006. Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language* 98(2). 159–168.

Table 1: Experimental Design & Sample Stimuli

	Informational aspect	Discourse Referent	Example
Indirect anaphor	new	new	<i>Theo read an article about waste disposal. I heard that the author wrote very well about that topic.</i>
Specification anaphor	new	given	<p>SPR Item: <i>Lisa_i worked the whole night through. I heard that the author_i is going to publish a new book soon.</i></p> <p>ERP Item: <i>Joanne K. Rowling_i worked the whole night through. I heard that the author_i is going to publish a new book soon.</i></p>
Direct anaphor	given	given	<i>An author_i worked the whole night through. I heard that the author_i is going to publish a new book soon.</i>

Figure 1: Reaction times of the SPR Experiment. Standardized logarithmic reaction times on the y-axis and each region of the target sentences on the x-axis. Region 5 (“die Professorin” – the female professor) is the region of interest.

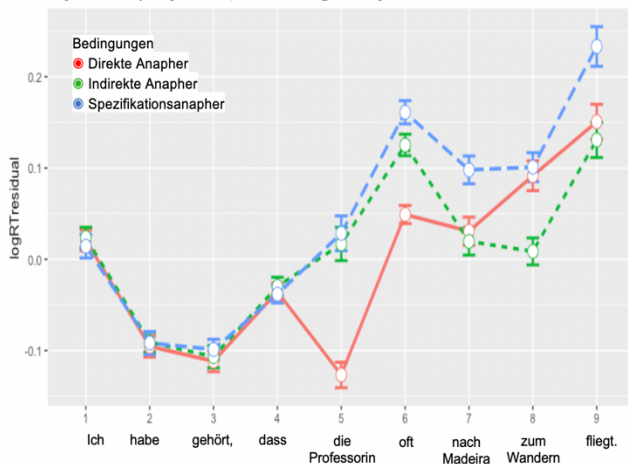
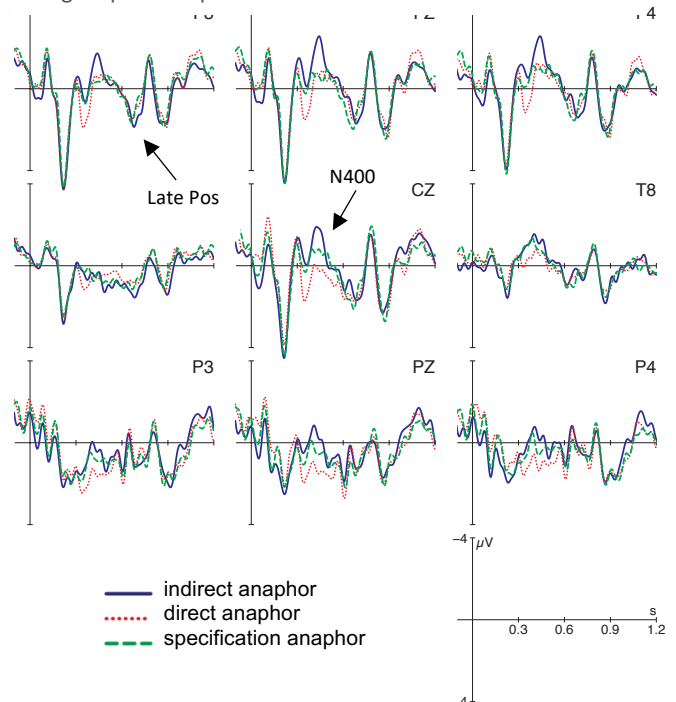


Figure 2: Grand-average-ERPs recorded to the onset of the critical anaphor (onset at the vertical line). Window presentation spans from 200 ms before until 1200 ms after onset of the anaphor. The voltage scale ranges from -4 to 4 μ V and negative voltage is plotted upward.



What reaction times can reveal behind acceptability judgments

Eunkyung Yi (Ewha Womans University) and Sang-Hee Park (Duksung Women's University)

Acceptability judgment experiment is one of the most common methods used to investigate one's syntactic knowledge. Based on speakers' judgments on a sentence (e.g., on a gradient scale), syntacticians decide which syntactic rule is relatively more (or less) operative in the grammar of a particular language. While what counts as important in syntactic studies is the *product* of judgments, or judgment scores, we focus in this study on cognitive measures such as reaction times in the *process* of judgments. Specifically, we investigated what reaction times can implicate in acceptability judgment where judgment scores ultimately do not make much difference.

We conducted an auditory acceptability judgment experiment with 2x2 conditions, i.e., two syntactic variants of the Korean ditransitive construction (1) and two semantic verb types (2). Participants were asked to judge acceptability of each stimulus on a seven-point Likert scale and their reaction time (i.e., end of an auditory stimulus ~ judgment selection) was recorded in milliseconds.

- | | |
|------------------------|---|
| (1) Syntactic variants | a. Canonical (<i>John-NOM Mary-DAT book-ACC gave</i>)
b. Double-Acc (<i>John-NOM Mary-ACC book-ACC gave</i>) |
| (2) Verb types | a. Caused-possession verbs (CP, e.g., <i>give</i>)
b. Caused-motion verbs (CM, e.g., <i>send</i>) |

Previous research showed Korean speakers tend to judge the Double-Acc structure (1b) to be highly unacceptable as opposed to the Canonical one (1a). Theorists endorse both as grammatical, though. In this context, Lee (2018) reported a small verb type effect in a written judgment experiment. Namely, CP verbs slightly improve the Double-Acc structure. In addition, the Canonical structure is perceptually even better with CP verbs than with CM verbs, since the dative case marker is more often used to mark a recipient than a goal (Yun & Hong, 2014). In this context, we expect subjects to produce a gradient acceptability across conditions as indicated in (3) and more specifically, based on Nagata (1990) and McElree (1993), we expect them to be faster in judging the best and worst combinations at either end than judging the less obvious ones in the middle. Namely, we predict that CM verbs make judgment on the Canonical structure relatively slower while making judgment on the Double-Acc structure faster, which is the opposite for CP verbs.

- (3) Canonical+CP > Canonical+CM >> Double-Acc+CP > Double-Acc+CM

We analyzed the data using mixed-effects regression models with structure, verb type and their interaction as predictors. In the first model, where judgment scores set as outcome, we found the main effect of structure ($b=-4.73$, $p<.001$) but found no effects of verb type and the interaction. In the second, where reaction time was the outcome, we found no main effects but found a marginally significant interaction between structure and verb type ($b=1244.85$, $p=.086$). An examination of the interaction showed, as predicted, CP and CM verbs made judgments relatively slower on the Double-Acc and on the Canonical structure, respectively (Figure 1). We further examined whether highly (un)acceptable and less-so judgments are correlated with reaction times and found a significant correlation within the Canonical structure ($r=-0.27$, $p<.01$) as well as within the Double-Acc structure ($r=.35$, $p<.001$), i.e., faster for the highly (un)acceptable, confirming Nagata's (1990) results (Figure 2). This study shows, the small verb type effects on judgment scores observed in the written mode may disappear in the auditory judgment experiment, but the effects can survive in subjects' reaction times. The present study suggests that reaction time can be a meaningful remnant of such small effects left behind the process of acceptability judgment.

Example stimuli in Korean

- a. apeci-ka atul-**eykey** wuncen-**ul** kaluchy-ess-ta (Canonical)
 father-NOM son-DAT driving-ACC teach-PAST-DECL
- b. apeci-ka atul-**ul** wuncen-**ul** kaluchy-ess-ta (Double-Accusative)
 father-NOM son-ACC driving-ACC teach-PAST-DECL
 'A father taught his son how to drive.'

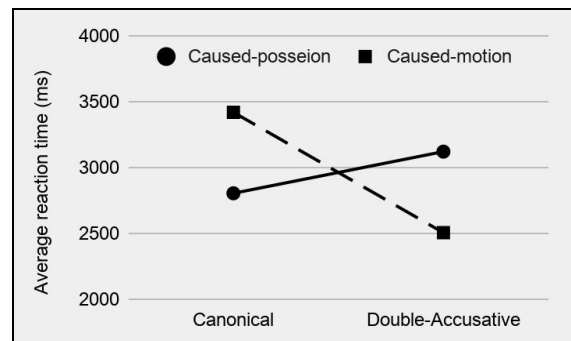


Figure 1. An illustration of the interaction between structures and verb-types

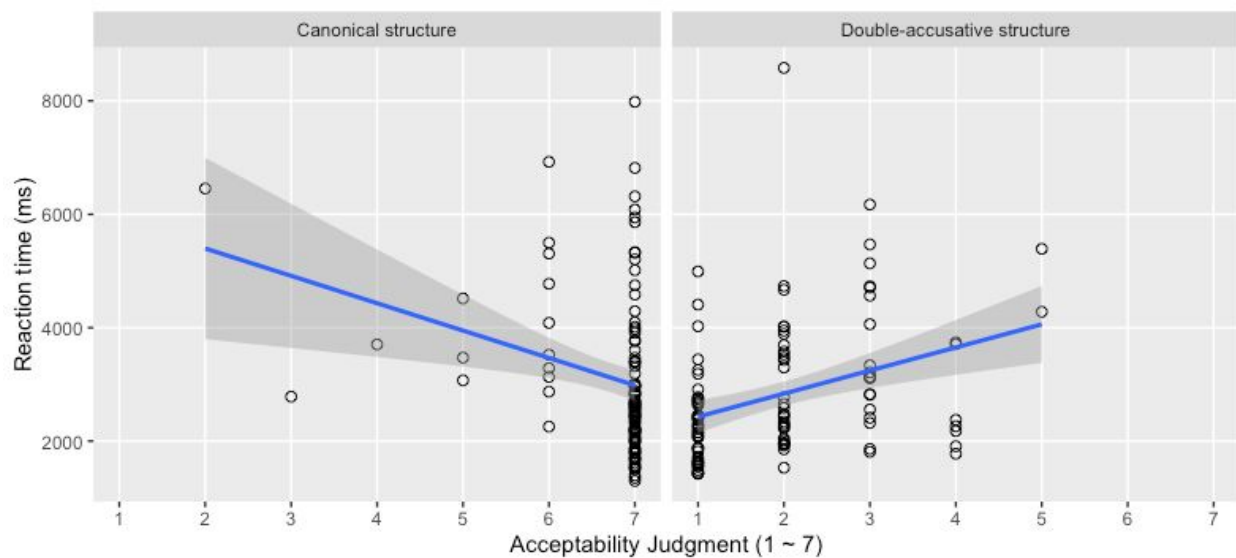


Figure 2. Relationship between acceptability ratings and reaction times

References

- Keller, F. (2000). Gradience in grammar. Doctoral dissertation. University of Edinburgh.
- Lee, Hanjung (2018). *Linguistic Research* 35(3), 449-482.
- McElree, B. (1993). *Journal of Memory and Language* 32, 536-571.
- Nagata, H. (1990). *Perceptual and motor skills* 70, 987-994.
- Schütze, C. T. (2016). *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.
- Yun & Hong, (2014). The effect of role predictability and word predictability on sentence comprehension. *Journal of Cognitive Science* 15, 349-390.

Limits on failure to notice word transpositions during sentence reading

Kuan-Jung Huang & Adrian Staub (University of Massachusetts Amherst)

Mirault et al. (2018) found that readers sometimes judge a sentence with transposed words to be grammatical (e.g., *The white **was** cat big*). They attributed these errors to noisy positional information resulting from parallel word processing. Their account predicts a higher error rate when the second transposed word is easier to recognize than the first, because this increases the probability that the second word will be identified before the first (Snell et al., 2018). Here we tested this prediction by manipulating the frequency of each of two transposed words; to avoid confounds with part of speech (as in Huang and Staub, 2020), both words were open-class.

Frequency was factorially manipulated for the first and second transposed word in sentences (Table 1); mean Zipf frequency was around 5 for high-frequency words and 3 for low-frequency words, based on the SUBTLEX Corpus (Brysbaert & New, 2009); the frequency distributions were non-overlapping. We used two sentence frames for each combination of levels of word frequency, one with a noun preceding a verb in the transposed order (e.g., *His sister **stuff** drew*) and one with a noun preceding an adjective (*A really **fellow** scary*), rendering 8 sub-conditions. Each subject read 7 transposed sentences and 7 un-transposed grammatical sentences in each of the 8 conditions. We also added a reference condition with an additional 7 transposed and 7 grammatical items, in which the transposition involved a pronoun (e.g., *It might **him** cure of the deadly disease*); this transposition was among the most frequently missed in our previous experiments. Finally, we also included twice as many grammatical fillers as critical items. Self-reported native English speakers participated on MTurk (N=69). For the critical items, the question to be answered after reading the sentence was an error-detection question, while for the fillers it was a comprehension question. Subjects could not predict which type of question would be asked until each sentence was removed from the screen (Fig. 1). Trials with RT to questions > 15s were discarded (0.4%).

Averaging across the critical conditions, subjects failed to detect transpositions only 9.1% of the time, while rejecting the corresponding grammatical versions 9.3% of the time. Thus, subjects failed to detect the transposition numerically less often than they rejected the grammatical counterparts; there was no transposition effect. In sharp contrast, in the reference condition, where one of the transposed words was a pronoun, they failed to detect transpositions 32% of the time, while rejecting the grammatical version 5.5% of the time (Fig. 2). Despite the apparent lack of a transposition effect in the target items, we assessed the prediction that word frequency should modulate the rate of failure to notice transpositions. We ran GLME models (Bates et al., 2015) testing effects of frequency, frame type, and their interaction (Table 2) on the probability of noticing the transposition. There was a main effect of frame type, with Frame 1 being more illusory, and a marginal effect of frequency, in the direction of less frequent failure to notice the error when the first transposed word was low frequency and the second word was high frequency (i.e., the low-high condition). This is the opposite direction from the prediction of the parallel processing account. This pattern was similar in a post-hoc analysis restricted to items that were highly acceptable in their grammatical version.

To explore the source of the difference in detectability of the transposition between the critical and reference conditions, we correlated (item-wise) error rate with word length, and with bigram frequency of the transposed words in their canonical order. While both factors explained between-condition variation, the latter also explained within-condition variation; the items in the reference condition that tended to elicit the highest rate of failure to notice the transposition were those that had high bigram frequency in the grammatical word order (Fig. 3).

In sum, we barely found a transposition effect with open-class words, while replicating a large effect in an additional reference condition in which one word was closed-class. Any effect of word frequency was in an unpredicted direction, with failure to notice errors being less common when the second word was higher frequency. Post-hoc analyses suggest that bigram frequency of the two critical words, in their grammatical order, may account for much of the item-level variability in the failure to notice transpositions. This finding aligns with a rational inference account emphasizing the role of the reader's prior for the underlying grammatical string (Gibson et al., 2013).

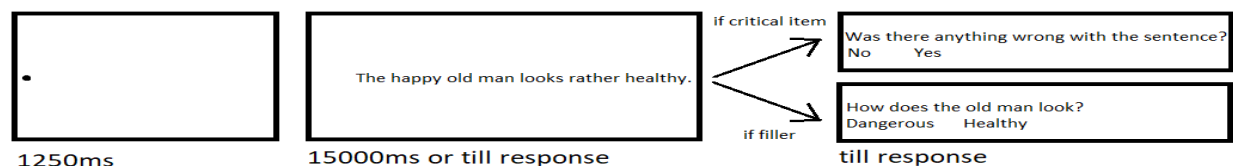


Figure 1. Procedure for each trial.

Freq condition	Frame 1	Frame 2
High-High	His sister stuff drew that was not recognizable.	A really fellow scary came into the room.
High-Low	The cells water absorb through their tiny pores.	A painfully sound eerie came from the woods.
Low-High	My nephew cider stores in the wine cabinet.	The particularly jester short will please the king.
Low-Low	The factories alloy refine using very high heat.	An especially hobbit rugged went into the cave.

Table 1. Critical items (grammatical not shown). All sentences were 8 words, and transposition always occurred between words 3 and 4. The transposed words were always shorter than 6 letters and with the length difference between them no greater than 1 letter.

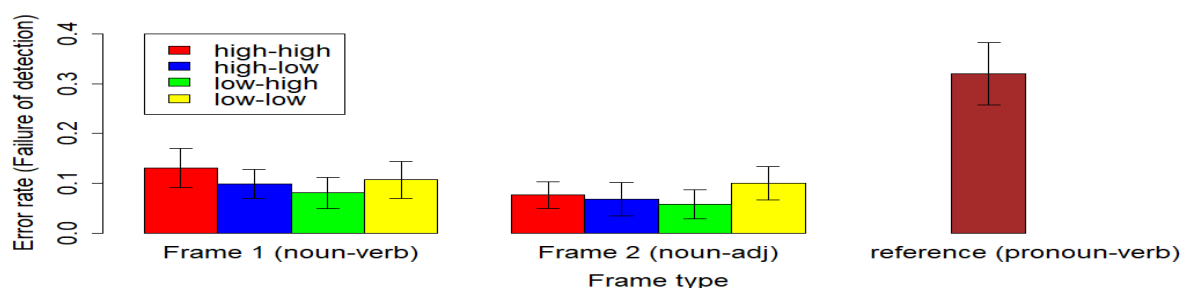


Figure 2. Error rate for transposed sentences, by condition (error bar = by-subject 95% CIs).

	Estimate	SE	Z	P
Intercept	3.32	0.253	13.13	<2e-16
Frame type (sum-coded)	0.45	0.186	2.39	.01
HH – LH (treatment-coded)	-0.52	0.268	-1.94	.05
HL – LH (treatment-coded)	-0.25	0.272	-0.93	.35
LL – LH (treatment-coded)	-0.50	0.27	-1.85	.06

Table 2. Estimation of effects of frame type and frequency on accuracy.

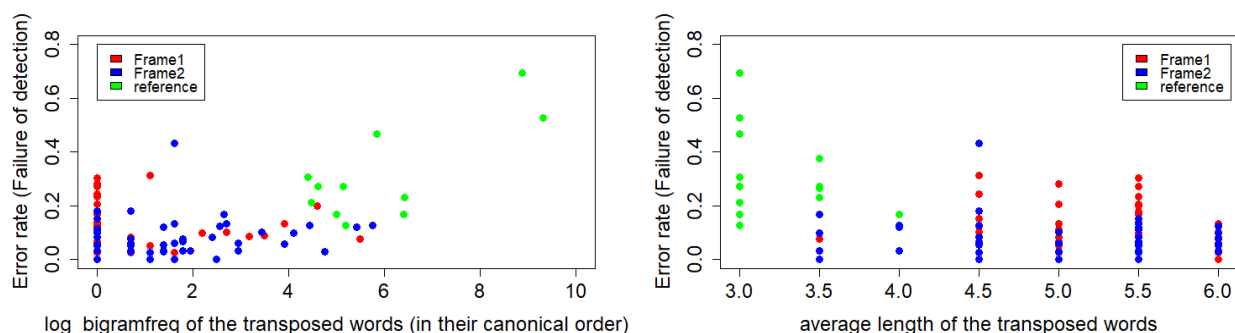


Figure 3. Scatterplots of error rate against bigram frequency (left) and word length (right).

References:

- [1] Mirault, J., Snell, J., & Grainger, J. (2018). *PsychScience*. [2] Huang, K-J. & Staub, A. (2020). Poster at CUNY. [3] Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). *PsychReview*. [4] Brysbaert, M. & New, B. (2009). *Behavioral Research Methods*. [5] Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *arXiv*. [6] Gibson, E., Bergen, L., & Piantadosi, S. (2013). *PNAS*.

Generalizing speaker-specific 'stylistic' preferences

Nitzan Trainin and Einat Shetreet (Tel Aviv University)

Speakers can recognize inter-speaker variability in various pragmatic phenomena (e.g., uncertainty expressions [1], or under-specification of adjectives [2]) and to adapt to the speakers' different preferences of language use. In these cases, the motivation seems clear: such distinction facilitates the derivation of meaning from the utterances of a specific speaker. In this study, we asked whether speaker-specific adaptation can occur when the language use of different individuals does not entail different meanings, but instead is based on differences in pragmatic-stylistic preferences (see [5] for an account of syntactic-stylistic adaptation). We utilized the weak adjective ordering preferences in Hebrew, where two orders for three-adjective phrases are preferred to the same extent [3]. Thus, choosing to use one over the other does not convey a meaning modification.

Methods: Native Hebrew speakers ($N=60$) took part in a learning paradigm consisted of an exposure phase, where one speaker used a certain order and the other a different order, and an explicit test phase that tested whether the participants learned these speaker-specific preferences. The exposure phase included 3 between-subject conditions, differing in the adjective orders which were used (the two most common and natural for this combination of adjective classes: Noun-Size-Color-Pattern and Noun-Color-Size-Pattern, and the most deviant one: Noun-Pattern-Size-Color, based on [3]). In each group, participants were visually presented with 96 images of shapes which had 3 distinctive visual features: size, color and pattern (Figure 1), and had to judge whether they matched an auditory description which used varying adjective orders (based on the condition/speaker). In half of the cases, the descriptions matched the image, and in the other half, they did not. The auditory descriptions were recorded by a male (Yoav, a common Israeli male name) and a female (Naama, a common Israeli female name) to ease their discrimination. The characters always used the same adjective order in their 48 descriptions (counter-balanced across participants and conditions). 12 pseudo-randomized lists of 4 interleaved speaker blocks were used, counterbalanced for the first speaker's identity and for the first used adjective order. In the test phase, participants had to decide which speaker could have uttered written three-adjective phrases, similar to those in the exposure phase (Figure 2). Half of the descriptions included the adjective order consistent with the male speaker and half included the order consistent with the female speaker, presented in a randomized sequence.

Results: The conditions in which one common order was presented with the most deviant order yielded substantially more successful distinction than the condition where the two common orders were used (mean accuracy: Noun-Size-Color-Pattern/Noun-Pattern-Size-Color = 74.38%; Noun-Color-Size-Pattern/Noun-Pattern-Size-Color = 68.75%; and Noun-Size-Color-Pattern/Noun-Color-Size-Pattern = 48.13%) (Figure 3). A logistic regression model revealed that both conditions in which one of the orders was the deviant one yielded more accurate identifications of the speaker than the condition in which both speakers produced a common order ($ps < 0.03$). Post-hoc pairwise comparisons revealed that there was no significant difference between these two conditions ($p = 0.82$).

Discussion: When both speakers produced the most common and natural adjective orders in Hebrew (when using color, size and pattern adjectives), almost all the listeners were somewhat unaware of the different speakers' preferences, and could not attribute, on the test phase, a certain order to a certain speaker. However, when one of the speakers produced the most deviant order in Hebrew, most listeners correctly assigned each speaker with their preferred order. This suggests that listeners can detect speaker-specific language use, when such use deviates from common or natural use, at least when speakers are easily distinguishable from one another in their non-linguistic characteristics (male vs. female). It remains an open question whether successful adaptation is mediated by increased attention to the deviance (e.g., through surprisal-driven learning [4]), or whether the deviant order conveys not only a stylistic preference, but also a subtle change in meaning (e.g., changing the focus).

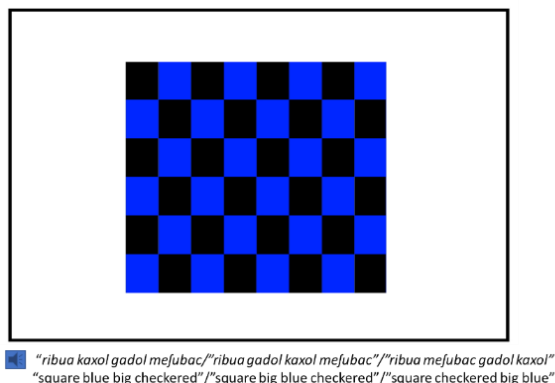


Figure 1. An example of a stimulus in the training phase. Each of the orders was produced either by Speaker A or by Speaker B. In half of the trials in the training phase the description and the image mismatched (one of the features was inappropriate for the image). Participants were required to press F if the description matched the image or K if it did not.

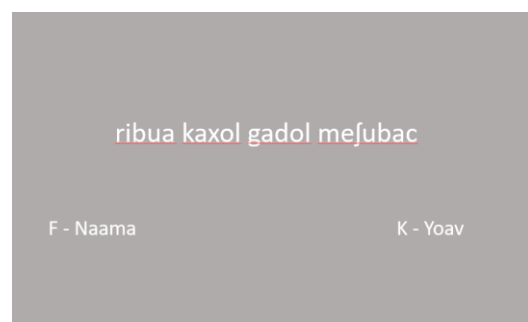
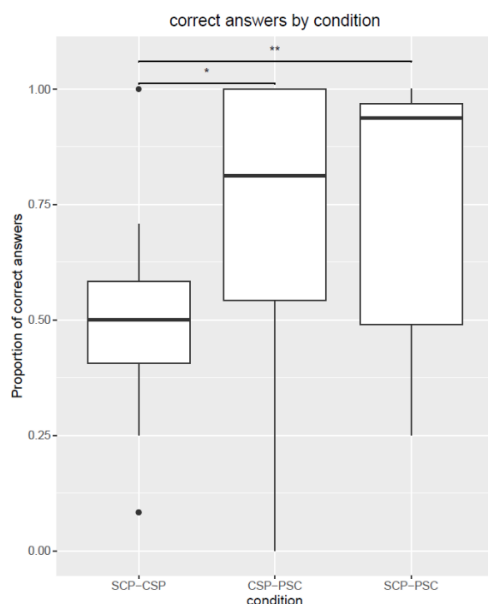


Figure 2. An example for a trial in the test phase. Participants were instructed to choose who of the speakers could have uttered the written descriptions. Originally, descriptions were presented in Hebrew with Hebrew letters.



References

- [1] Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, 203, 104285.
- [2] Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in psychology*, 6, 2035.
- [3] Trainin, N., & Shetreet, E. (2020). It's a dotted blue big star: on adjective ordering in a post-nominal language. *Language, Cognition and Neuroscience*, 1-22.
- [4] Lai, W., Rácz, P., and Roberts, G. (2020) Experience with a linguistic variant affects the acquisition of its sociolinguistic meaning: An alien-language-learning experiment. *Cognitive Science* 44(4): e12832.
- [5] Ostrand, R., & Ferreira, V. S. (2019). Repeat after us: Syntactic alignment is not partner-specific. *Journal of memory and language*, 108, 104037.

Figure 3. Correct answers in the trial phase, by condition. SCP = Noun-Size-Color-Pattern; CSP = Noun-Color-Size-Pattern; PSC = Noun-Pattern-Color-Size. SCP and CSP are the most common and natural adjective orders and PSC is the most uncommon and unnatural adjective order in Hebrew.

Variability in the agreement attraction effect

Sanghee Kim & Ming Xiang (University of Chicago)

Agreement attraction is an effect that has been extensively reported across different structures and languages [1-3]. This effect predicts that, for sentences like (1a), the ungrammatical verb (*were*) will be read faster and be judged more acceptable if there is an intervening plural noun (*sharpshooters*) between the singular subject and the verb, compared to when the intervening noun is singular. Some recent findings, however, found cases of null attraction effects and questioned the robustness of the effect [4]. Through a series of four experiments, the current study also found a null result for the standard number agreement attraction effect. The null results call for more work in understanding the variability and reliability of the agreement attraction effect.

Design and procedure. All four experiments are self-paced reading experiments conducted on Ibex Farm [5], with native English speakers recruited from Prolific.co. After reading each sentence, participants gave an acceptability judgment on a 1-7 scale (with 7 being the most acceptable). **Experiments 1-3** have a subject relative clause (SRC) structure, in the form of *NP1 who VERB NP2 (adverb) was/were ...*, with the subject NP1 always in the singular form. In a 2x2 design, we varied Grammaticality (*was* vs. *were*) and Distractor Noun (singular vs. plural NP2). The stimuli for **Experiment 1** (subj n=58; item n=48) are adapted from [3], and an example is given in (1a). **Experiment 2** (subj n=59; item n=48) removed the adverb before the critical verb to reduce the distance between the verb and the distractor NP2 (1b). **Experiment 3** (subj n=61; item n=48) further removed the adjective modifiers on both NP1 and NP2, so that there are fewer encoding features to distinguish the two nouns (1c). The modifications on Experiment 2 and 3 are designed to maximize the chances of the agreement attraction effect. **Experiment 4** (subj n=81; item n=48) has the same 2x2 design as Experiment 3, but we looked at the object relative clause structure (ORC), in the form of *NP1 who NP2 VERB*, which placed the distractor noun NP2 in the embedded subject position within the relative clause (1d).

Results. For the **acceptability judgment tasks**, we performed a linear mixed-effects model on the ratings, with Grammaticality, Distractor Noun, and their interaction as fixed effects, and maximum by-participant and by-item random intercepts and random slopes that led to model convergence. Both Grammaticality and Distractor NP factors are sum coded. We found a standard number agreement attraction effect in all four experiments (Fig. 1), which appeared as an interaction between Grammaticality and Distractor Noun, such that ungrammatical conditions were rated higher when the distractor was a plural noun ($|t| > 2$ for all experiments). However, we found no evidence of attraction in **the online reading measures**. For the reading time (RT) analysis, we performed linear mixed-effects models on the log-transformed RTs in the critical region (*was/were*) and the spill-over region. The models included the same fixed effects and random effects as the ones used in the judgment tasks. As shown in Fig. 2, across all four experiments, the only consistent effect is the grammaticality effect on the critical verb region, with ungrammatical conditions read slower than the grammatical conditions ($|t| > 2$ for all experiments). There were no other consistent effects, and there was no evidence in any experiment for the standard number agreement attraction effect, which would have been demonstrated by faster RTs on the ungrammatical condition with a plural distractor noun than the ungrammatical condition with a singular distractor noun.

In summary, we did not find the standard number agreement attraction effect in online RTs in SRC and ORC constructions. This is in line with the recent results in [4] but is inconsistent with many previous studies that showed the effect (e.g., see a review in [6]). This difference in results across studies on agreement attraction might be due to differences in the constructions tested or the tasks used in these tasks (e.g., [4, 6]). The variability of the number agreement attraction effect across different studies calls for more nuanced investigations of the processing mechanisms involved in this phenomenon. We are currently running a follow-up study to better understand the process involved.

References. [1] Pearlmutter et al. 1999. JML. [2] Wagers et al. 2009. JML. [3] Dillon et al. 2013. JML. [4] Parker & An. 2018. Fpsyg. [5] Drummond. 2013. <http://spellout.net/ibexfarm> [6] Hammerly et al. 2019. Cog Psy.

(1) A sample of experimental sentences. *Slash (/) indicates the SPR regions

- Experiment 1:** The cruel hunter / who accompanied / the accurate sharpshooter(s) / surely / {was/were} / capable / of finding / deer / in / the forest.
- Experiment 2:** The cruel hunter / who accompanied / the accurate sharpshooter(s) / {was/were} / capable / of finding / deer / in / the forest.
- Experiment 3:** The hunter / who accompanied / the sharpshooter(s) / {was/were} / capable / of finding / deer / in / the forest.
- Experiment 4:** The sharpshooter / who / the hunter(s) / accompanied / {was/were} / capable / of finding / deer / in / the forest.

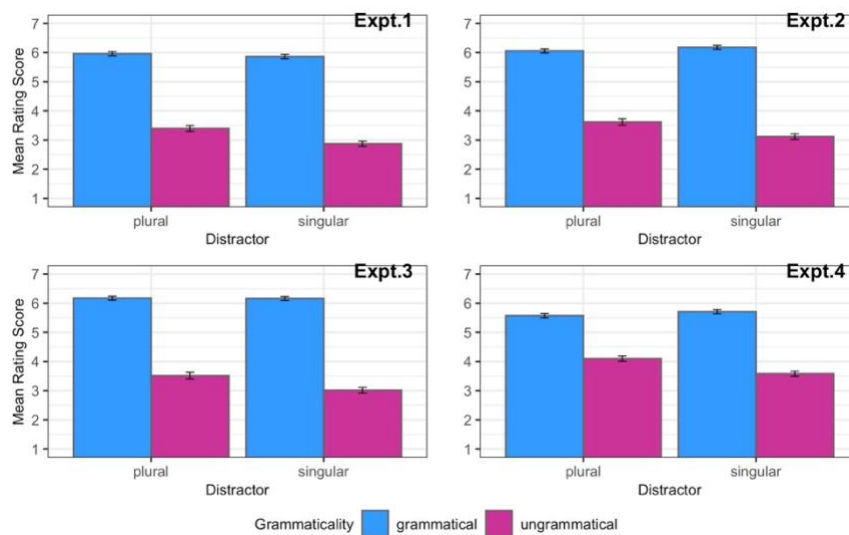


Figure 1. Mean acceptability rating scores (Experiment 1-4).

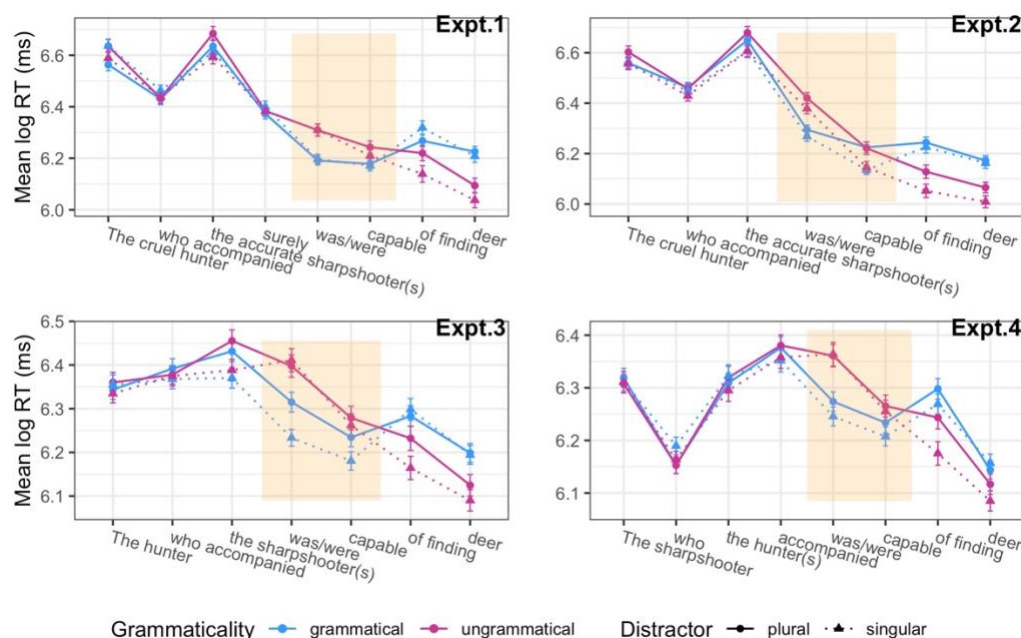


Figure 2. Mean log reading times (Experiment 1-4).

Bayesian surprise predicts incremental processing of grammatical functions

Thomas Hörberg (Department of Linguistics, Stockholm University), T. Florian Jaeger (Department of Brain and Cognitive Sciences, University of Rochester)

A central part of sentence understanding involves assigning grammatical functions (GFs) to NPs, thereby determining how participants are related to events. Cross-linguistically, GF assignment is signal by a variety of cues and their interactions, such as morpho-syntactic (e.g., word order, case), referential/semantic (e.g., animacy) and verb class (e.g., volitionality). Using data from Swedish transitive sentences (with SVO and OVS order), we test whether cues to GFs affect processing directly (as hypothesized by e.g. Bornkessel & Schlesewsky, 2006) or mediated through expectations based on complex statistical patterns created by those cues (e.g., Kempe & MacWhinney 1999; MacDonald, 2013). We first develop a Bayesian model of incremental GF-assignment, and fit it to a database of written Swedish. We use this model to derive estimates of the change in *syntactic* expectations at each sentence region (cf. Jurafsky, 1996). These predictions are then tested against reading times from a self-paced reading experiment, and compared to estimates of word-level expectations (i.e., word surprisal). We extend previous work by a) explicitly assessing the changes in expectations about GF-assignment, and b) expanding cross-linguistic coverage of computational theories of sentence understanding.

The Bayesian model of incremental GF-assignment (Figure 1) is trained on 16,552 transitive sentences from the Svensk Trädbank corpus (Nivre & Megyesi, 2007), consisting of Swedish texts from various genres. Sentences were annotated for word order (SVO vs. OVS), GF information (e.g., animacy, case/pronominality), and verb semantic properties (e.g., volitionality, sentience). Based on the distribution of these properties, estimates of the probability for SVO vs. OVS GF-assignment at each sentence region (NP1, verb, NP2) were calculated. These estimates are then used to predict incremental processing costs related to the change in the expectation for a GF-assignment at these regions. This is done in terms of *Bayesian surprise*—the relative entropy over the two possible GF assignments before and after seeing the constituent at hand (cf. Kuperberg & Jaeger 2016). Bayesian surprise over syntactic trees has been claimed to underlie the correlation between word surprisal and both processing times (Smith & Levy 2013) and neural responses (e.g., the N400, Frank et al. 2015).

In the self-paced reading experiment, 45 Swedish participants read 64 transitive sentences (with fillers) that varied in word order (SVO vs. OVS), NP1 animacy (animate vs. inanimate) and verb class (volitional vs. experiencer). Length-corrected by-region reading times (RTs) on NP1, verb, and NP2 were predicted by incremental Bayesian surprise (as shown by Bayesian LMMs with full random effect structures; Figure 2). Bayesian surprise also qualitatively captures interactions between morphosyntactic, animacy, and verb class cues. E.g., both RTs and Bayesian surprise in the NP2 region of locally ambiguous OVS sentences are mitigated when NP1 animacy and its interaction with the verb class bias towards OVS word order.

Comparisons of model's predictive accuracy (leave-one-out information criterion) found that Bayesian surprise explains *with a single degree of freedom* a substantial part of the variance in RTs that is explained by the many cues to GFs, suggesting that Bayesian surprise provides a plausible and parsimonious link function for the cognitive computations performed during sentence understanding. In a final step, we ask how much of the predictive power of Bayesian surprise over the GF-assignment can be explained by traditional word surprisal estimated by a neural network model (GPT2).

Summary. These findings indicate that incremental GF assignment draws on statistical regularities in the language input, as predicted by expectation-based accounts (MacDonald, 2013). Bayesian surprise—a measure of the prediction error experienced when new evidence is integrated into gradient expectations—provides a computationally plausible and empirically validated linking hypothesis.

References

- Bornkessel, I., and Schlesewsky, M. 2006. The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, 113, 787–821.
- Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11.
- Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20 (2), 137–194.
- Kempe, V., and McWhinney, B. 1999. Processing of Morphological and Semantic Cues in Russian and German. *Language and Cognitive Processes*, 14, 129–171.
- Kuperberg, G. R., & Jaeger, T. F. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59.
- MacDonald, M. C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology*, 4 (226), 1–16.
- Nivre, J., & Megyesi, B. 2007. Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection. *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories*, 97–102.
- Smith, N. J., & Levy, R. 2013) The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.

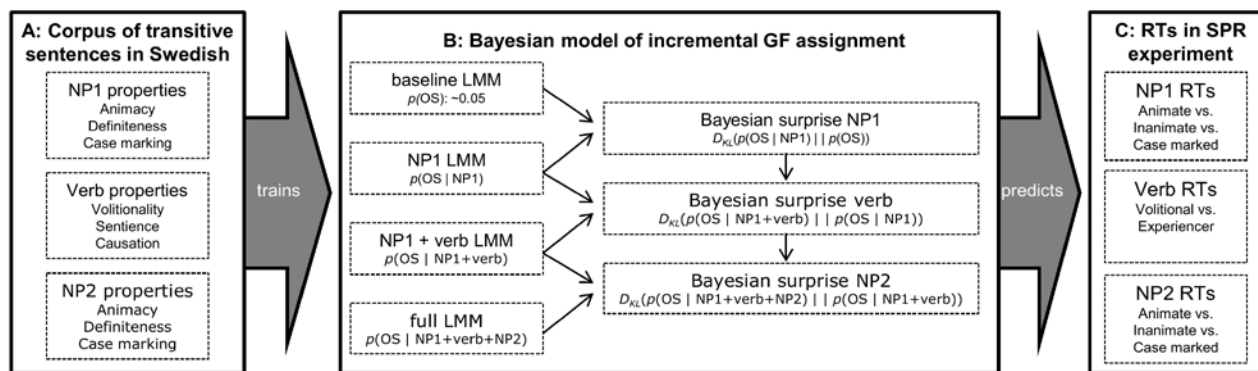


Figure 1. Illustration of the Bayesian model of incremental GF-assignment.

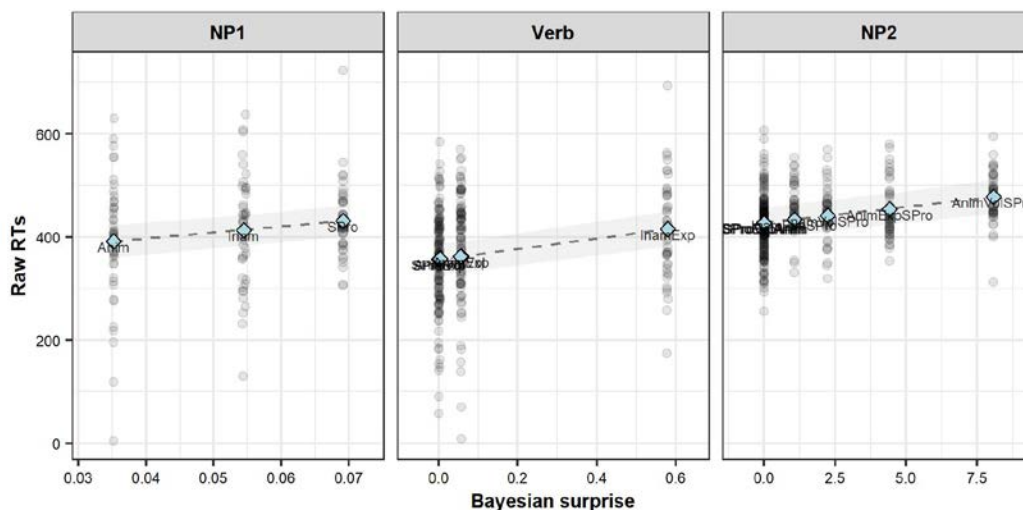


Figure 2. The relationship between Bayesian surprise as predicted by the model and raw reading times.

34th Annual CUNY Conference on Human Sentence Processing

Friday Evening March 5, 2021

Hour	Session	Time	Title	Authors
Hour 1	1	18:30	Experience and conceptual overlap modulate cross-language priming effects	Adel Chaouch-Orozco, Jorge González Alonso and Jason Rothman
Hour 1	1	18:30	Pseudorelatives and L1-Attrition	Alexander Cairncross, Margreet Vogelzang and Ianthi Tsimpli
Hour 1	1	18:30	Age effects in L2 processing of passive sentences	Candice Glenday, José Ferrari-Neto and Elisangela Nogueira Teixeira
Hour 1	1	18:30	★ A divergence between judgments and response times in L2 agreement attraction	Eun-Kyoung Rosa Lee and Colin Phillips
Hour 1	1	18:30	Meaning, but not grammatical features, is cross-linguistically mask-primed in sentential contexts	Jeonghwa Cho and Jonathan Brennan
Hour 1	1	18:30	★ Monolingual and bilingual processing at the syntax-discourse interface: Evidence from the English dative alternation	Joshua Weirick and Elaine Francis
Hour 1	2	18:30	★ Learning the generative principles of a linguistic system from limited examples	Lei Yuan, Violet Xiang, David Crandall and Linda Smith
Hour 1	2	18:30	If Memory Doesn't Serve: Timecourse of Syntactic Forgetting in Ellipsis and Recognition	Caroline Andrews
Hour 1	2	18:30	The Structure of Antecedent Influences Processing of Ellipsis	Hyosik Kim, Ming Xiang and Masaya Yoshida
Hour 1	2	18:30	I(nterpolated) Maze: High-sensitivity measurement of ungrammatical input processing	Pranali Vani, Ethan Wilcox and Roger Levy
Hour 1	2	18:30	Task influences on lexical underspecification: Insights from the Maze and SPR	John Duff, Adrian Brasoveanu and Amanda Rysling
Hour 1	2	18:30	The interaction of semantic information and parsing biases: An A-maze investigation	Xinwen Zhang and Jeffrey Witzel
Hour 1	3	18:30	Is the relationship between word probability and processing difficulty linear or logarithmic?	James Michaelov, Megan Bardolph, Seana Coulson and Benjamin Bergen

Hour	Session	Time	Title	Authors
Hour 1	3	18:30	Bridging language acquisition and processing via the integrated systems hypothesis: Evidence from self-paced reading of newly-learned words within sentential contexts	Laura Morett, Sarah Hughes Berheim and Jack Shelley-Tremblay
Hour 1	3	18:30	Word recognition in sentence processing is predicted by domain-general processing speed	Naomi Sellers, Shannon McKnight, Phillip Gilley and Albert Kim
Hour 1	3	18:30	Perceptual connectivity influences toddlers' attention to known objects and subsequent label processing	Ryan Peters, Justin Kueser and Arielle Borovsky
Hour 1	3	18:30	Virtual-World eye-tracking: The efficacy of replicating word processing effects remotely	Zoe Ovans, Jared Novick and Yi Ting Huang
Hour 1	4	18:30	★ Complex syntax and conversational turn-taking during toddler-adult picture book reading	Anastasia Stoops, Jane Hwang, Mengqian Wu and Jessica L Montag
Hour 1	4	18:30	★ Preferences for communicative efficiency in miniature languages are independent of learners' L1s	Lucy Hall Hartley and Masha Fedzechkina
Hour 1	4	18:30	★ The role of L1 and L2 frequency in cross-linguistic structural priming: An artificial language learning study	Merel Muylle, Sarah Bernolet and Robert Hartsuiker
Hour 1	4	18:30	★ Probability matching vs. regularization in contact-induced syntactic change	Ming Xiang, Christine Gu, Yixue Quan, Weijie Xu and Suiping Wang
Hour 1	4	18:30	★ When animacy trumps word order in sentence comprehension: The case of late first-language acquisition	Qi Cheng, Sheila Price and Rachel Mayberry
Hour 1	4	18:30	★ Distributed Morphology feature geometries crosslinguistically: Acquiring the copula	Shiloh Drake
Hour 1	5	18:30	★ Effects of word order on L1 and L2 semantic prediction	Carrie Jackson, Holger Hopp and Theres Grüter
Hour 1	5	18:30	Model-based estimates of predictability reveal brain's robust sensitivity to variation in semantic fit even among unexpected words	Jakub Szewczyk and Kara Federmeier

Hour	Session	Time	Title	Authors
Hour 1	5	18:30	★ Online cloze evidence for rapid use of lexical and grammatical cues	Masato Nakamura and Colin Phillips
Hour 1	5	18:30	★ Children with hearing loss use semantic and syntactic cues for prediction in sentence comprehension	Rebecca Holt, Benjamin Davies, Laurence Bruggeman and Katherine Demuth
Hour 1	5	18:30	Does reading unexpected words lead to engagement of cognitive control?	Suzanne Jongman, Yaqi Xu and Kara Federmeier
Hour 1	5	18:30	Prediction accuracy facilitates processing of visual word form	Yang Agnes Gao, Tamara Swaab and Matthew Traxler
Hour 1	6	18:30	★ Putting the pieces together: Two-year-olds hearing an unfamiliar accent recognize known words and learn new words, but do not use known words to learn new words	Alexander LaTourrette, Cynthia Blanco and Sandra Waxman
Hour 1	6	18:30	★ Inside the wug-test: phonological well-formedness and processing costs	Canaan Breiss
Hour 1	6	18:30	Talking, like, a Valley Girl? Online Processing of Sociolinguistic Cues	Daisy Leigh, Judith Degen and Robert Podesva
Hour 1	6	18:30	Structural Priming in the Comprehension of Non-Native Speech	Douglas Getty and Scott Fraundorf
Hour 1	6	18:30	★ Pre-schoolers process word onsets and codas similarly: A time-course analysis	Rosanne Abrahamse, Nan Xu Rattanasone, Katherine Demuth and Titia Benders
Hour 1	6	18:30	★ Having to predict a (native or non-native) partner's utterance increases adaptation in L2	Theres Grüter, Alice Zhu and Carrie N. Jackson
Hour 1	7	18:30	Does bilingual inhibitory control operate over structural representations?	Andrea Seanez, Alejandra Fanith and Iva Ivanova
Hour 1	7	18:30	Bilingual language control in connected speech	Kyle Wolff and Iva Ivanova
Hour 1	7	18:30	Gap-filler dependencies are sensitive to islands: The case of Japanese relative clauses	Maho Takahashi and Grant Goodall
Hour 1	7	18:30	★ Verb Metaphoric Extension during Sentence Processing	Daniel King and Dedre Gentner

Experience and conceptual overlap modulate cross-language priming effects

Adel Chaouch-Orozco (University of Reading), Jorge González Alonso (The Arctic University of Norway) & Jason Rothman (The Arctic University of Norway, Universidad Nebrija)

Studies examining translation priming between non-cognate words with lexical decision tasks (LDT) report a priming asymmetry (larger L1 prime-L2 target priming compared to L2-L1). This potentially reflects (qualitative and/or quantitative) differences in representation and processing between L1 and L2 words. Several models of bilingual lexical processing have sought to explain this finding, with the Revised Hierarchical Model (Kroll et al., 2010) and Multilink/BIA+ (Dijkstra et al., 2019) being the most prominent. The former explains the asymmetry through differential access to conceptual information by L1 vs. L2 words. Under Multilink, it is explained by slower L2 word processing, reflecting lower (subjective) word frequencies in the L2. Both models assume holistic, largely overlapping conceptual representations between translation equivalents.

The present study explores the role of L2 use and word frequency on cross-language priming. We implement a nuanced two-way operationalization of L2 use by (a) manipulating immersion in a novel design with three groups differing in both quality and quantity of L2 exposure, and (b) employing the Language and Social Background Questionnaire' (Anderson et al., 2018) score as continuous variables. Secondly, we employ stimuli with a large frequency range to thoroughly investigate this factor. Potential effects of conceptual overlap between cross-language related words would have consequences for both the RHM and Multilink (e.g., van Hell & de Groot, 1998). For this reason, our stimuli set contains concrete and abstract translation equivalents, as well as cross-language semantic associates (obtained from a norming study).

Three hundred late sequential (L1) Spanish-(L2) English bilinguals are being tested in three groups: an L2 immersed group (UK), a non-immersed group (Spain), and a group (Norway) where the participants' relative L2 use is higher than in the Spain Group but lower than in the UK Group. All participants are similarly (highly) proficient in the L2, allowing factoring out a potential confounding effect. 300 translation equivalent pairs and 110 semantic associative pairs were used. This large sample size and stimuli N reflects our effort to draw robust conclusions supported by large statistical power (Brysbaert, 2020). We employ an unmasked translation priming LDT (see Figure 1 and Table 1 for procedure and stimuli). Effects of L2 use, word frequency, and concreteness are investigated through linear mixed effects models (Baayen, 2008).

Preliminary results with a subset of 72 participants from the immersed group (UK) show significantly faster responses with related primes in all conditions (see Table 2). The priming asymmetry is replicated: priming effects are larger in the L1-L2 direction. Also, L1-L2 priming is significantly larger for concrete pairs than for abstract ones, suggesting that a higher degree of conceptual overlap (which is typically true of concrete vs. abstract words) might produce a larger stimulation of the L2 targets. Moreover, more L2 use leads to significantly slower responses in all conditions (a potential effect of competition), except from L1-L2 responses with unrelated primes (Figure 2). This suggests that participants with increased L2 use can cope more efficiently with the misleading information from the L1 unrelated primes. Finally, more frequent related primes yield larger priming effects in all conditions, indicating these primes are more efficient in facilitating target processing.

These partial and preliminary results highlight the importance of L2 exposure/use and prime frequency in the study of translation priming. Also, they suggest that future experimental research should further explore the degree of conceptual overlap between cross-language related words, which could imply a step forward in our current understanding of lexico-semantic effects in bilingual visual word recognition. We are cautious in interpreting these results, however, as they represent a fraction of the expected data. Collection (Norway group) and analysis (of non-immersed groups and semantic associative priming) are still ongoing, but we expect being able to report definitive results at CUNY 2021. Whatever the outcome, these will surely have broad implications for the role of speaker- and stimulus-level variables in bilingual lexical processing.

Table 1. Sample stimuli used in each translation direction.

L1-L2			
Related prime	Unrelated prime	Word target	Nonce target
<i>lápiz</i> ('pencil')	<i>bosque</i> ('forest')	PENCIL	SMOUNT
L2-L1			
Related prime	Unrelated prime	Word target	Nonce target
onion	clown	CEBOLLA ('ONION')	TUNGO

Table 2. Response times, standard deviations (in parentheses), and priming effects, in milliseconds, for all conditions.

	Concrete pairs		Priming	Abstract pairs		Priming
	Related	Unrelated		Related	Unrelated	
	RT	RT		RT	RT	
L1 to L2	641 (220)	731 (275)	90*	687 (267)	763 (324)	76*
L2 to L1	647 (257)	704 (259)	57*	657 (235)	712 (258)	55*

Figure 1. Presentation procedure.

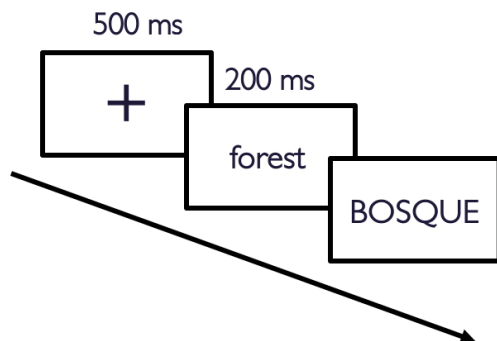
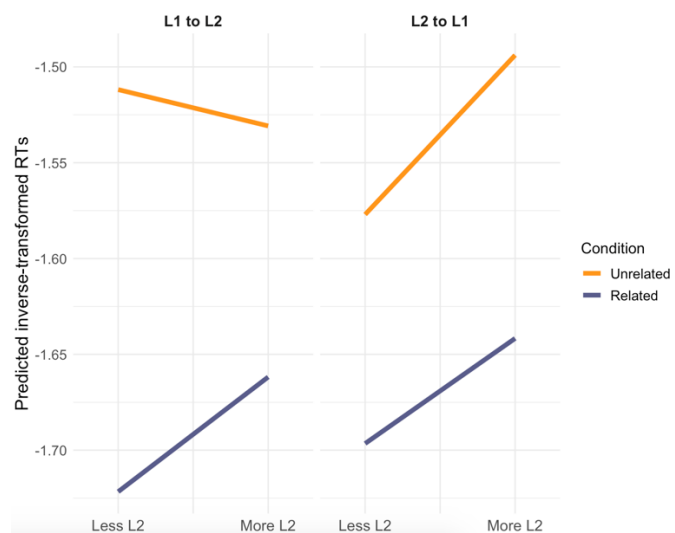


Figure 2. Effect of L2 use on the predicted inverse-transformed RTs in all conditions.

Note: smaller inverse RTs indicate slower responses



References

- Anderson, J., Mak, L., Keyvani Chahi, A., & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior research methods*, 50(1), 250–263. <https://doi.org/10.3758/s13428-017-0867-9>
- Baayen, R. H. (2008). *Analyzing linguistics data: a practical introduction to statistics using R*. Cambridge: CUP.
- Brysbaert, M. (2020). Power considerations in bilingualism research: Time to step up our game. *Biling: Lang & Cog*, 1-6. Ahead of print. DOI:10.1017/S1366728920000437.
- Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: a computational model for bilingual word recognition and word translation. *Biling: Lang & Cog*, 22(4), 657-679.
- Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Biling: Lang & Cog*, 13, 373–38.
- Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Biling: Lang & Cog*, 1(3), 193-211.

Pseudorelatives and L1-Attrition

Alex Cairncross, Margreet Vogelzang, and Ianthi Tsimpli
The University of Cambridge

Given ambiguous strings as in (1), speakers of languages like Spanish (or Italian) resolve this ambiguity by preferentially attaching the non-matrix clause to the first DP ('high attachment') while speakers of languages like English preferentially attach it to the second DP ('low attachment'; Cuetos & Mitchell, 1988).

(1) Pedro se enamoró de la hija₁ del psicólogo₂ que estudió en California.

'Peter fell in love with the daughter₁ of the psychologist₂ who studied in California.'

Following Grillo and Costa (2014), the difference in biases owes to a structural difference. Namely, Spanish and Italian admit pseudorelatives (PR) but English does not. PRs, while string identical to relative clauses, are a type of small clause and force attachment to the first DP (Grillo & Costa, 2014). When PRs are locally blocked, languages like Italian display a low attachment bias although PRs act as the online parsing default (Grillo & Costa, 2014; Pozniak et al., 2019). Under L2-immersion, these biases have been observed to change i.e. they attrite. Dussias (2003) explored items like (1) with L1-Spanish speakers in the United States (average years of residency = 7.5) via a sentence interpretation task. Results indicated that while monolingual Spanish speakers overwhelmingly selected the first DP (74%), the experimental group selected the first DP at a significantly lower rate (28%). As their experiment did not divide their items by PR availability, it is unclear whether the results indicate an across-the-board effect or a change only in PRs.

To explore this, an online interpretation task was conducted in Italian. Sentences were presented written alone and followed by a *who*-question with the 2 possible DP responses. **Critical items** like (2) consisted of 24 sentence pairs from Grillo and Costa (2014) in which PR availability is manipulated by the matrix predicate (PR-Condition: perceptive; non-PR-Condition: stative).

(2) Gianni (ha visto / vive con) il figlio del medico che correva.

'Gianni (saw / lives with) the son of the doctor who was running.'

Participants consisted of a control group (Italians in Italy, N = 25) and an experimental group (N = 32). The experimental group had lived in an English-speaking country for a minimum of 2 year (average > 4.45 years) and were proficient in their L2 English (average self rating = 8.69/10).

Global attachment preferences are presented in Table 1. Responses were coded as \pm high attachment and entered in a mixed effect logistic regression as the dependent variable with *condition* and *group* as predictors. The model also included random effects of *item* and *participant* and is reported in Table 2. **Results** indicated a main effect of condition ($\beta = 3.30$; $z = 12.51$; $p < 0.01$) with high attachment being significantly more frequent in PR taking items. They did not indicate an effect of group nor an interaction of group by condition. As such, these results do not replicate the findings in Dussias (2003) in Italian and the role of PRs in attrition remains unclear.

As the average L2-immersion of our participants was less than in Dussias (2003) and 25/32 participants in the experimental group reported recently having visited Italy prior to testing, the absence of a group effect in the present study may be due to attrition having a later onset or a re-exposure effect (cf. Chamorro et al., 2015). In response, a new experimental group is being collected (currently N = 23) who have been immersed in their L2-English for a minimum of 6 years (cf. Tsimpli et al., 2004, currently average = 15.92 years) and have not visited Italy in the 3 months prior to testing. Their global attachment rates are presented in Table 3. This new experimental group is noticeably older than the original control group (new experimental group average: 44.30 years; original control group average: 31.04 years) and the difference is significant under a Welch's t-test ($p < 0.01$). As such, a new control group is also being collected prior to statistical modelling.

Tables

Group	PR	RC-Only
Control	77.67%	25.67%
Experimental	70.05%	17.72%

Table 1: Experiment A High Attachment Rates

	Estimate	Std. Error	z-value	p-value
Intercept	1.53	0.36	-4.29	<0.01
Condition	3.30	0.26	12.51	<0.01
Group	-0.54	0.45	-1.21	0.23
Condition:Group	-0.16	0.33	-0.48	0.63

Table 2: Experiment A Regression Output

Group	PR	RC-Only
New Experimental	63.33%	17.73%

Table 3: Experiment B High Attachment Rates

References

- Chamorro, G., Sorace, A., & Sturt, P. (2015). What is the source of L1 attrition? the effect of recent L1 re-exposure on spanish speakers under L1 attrition. *Bilingualism: Language and Cognition*, 19(3), 520–532. <https://doi.org/10.1017/s1366728915000152>
- Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in spanish. *Cognition*, 30(1), 73–105. [https://doi.org/10.1016/0010-0277\(88\)90004-2](https://doi.org/10.1016/0010-0277(88)90004-2)
- Dussias, P. E. (2003). Syntactic ambiguity resolution in L2 learners. *Studies in Second Language Acquisition*, 25(4), 529–557. <https://doi.org/10.1017/s0272263103000238>
- Grillo, N., & Costa, J. (2014). A novel argument for the universality of parsing principles. *Cognition*, 133(1), 156–187. <https://doi.org/10.1016/j.cognition.2014.05.019>
- Pozniak, C., Hemforth, B., Haendler, Y., Santi, A., & Grillo, N. (2019). Seeing events vs. entities: The processing advantage of pseudo relatives over relative clauses. *Journal of Memory and Language*, 107, 128–151. <https://doi.org/10.1016/j.jml.2019.04.001>
- Tsimpli, I., Sorace, A., Heycock, C., & Filiaci, F. (2004). First language attrition and syntactic subjects: A study of greek and italian near-native speakers of english. *International Journal of Bilingualism*, 8(3), 257–277. <https://doi.org/10.1177/13670069040080030601>

Age effects in L2 processing of passive sentences

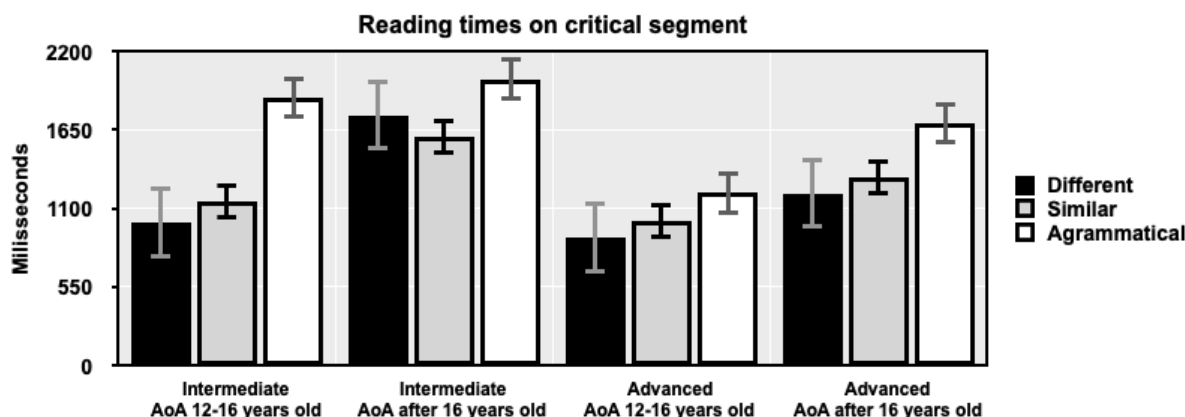
Candice H. Glenday (Universidade Estadual Vale do Acaraú), José Ferrari-Neto (Universidade Federal da Paraíba), Elisângela N. Teixeira (Universidade Federal do Ceará)

Introduction: There has been a long-standing debate concerning the sensitive period for first and second language acquisition. Recent studies have reported that the age factor is a significant marker for L2 language processing (Hartshorne, et al., 2018; Bonfieni, et al., 2019; Oh, et al., 2019), especially regarding syntactic processing. The claim is that early learners would process L2 and inhibit L1 more efficiently. However, during the initial stage of the L2 acquisition, late learners rely on their L1 knowledge to organize the syntactic structures of L2 (VanPatten, 2015). According to Hartsuiker et al. (2004), learners recognize L2 syntactic structures that are similar to their L1 structure, but when the L2 structure differs from their L1, they resort to their L1 to process the structure, which results in syntactic transfer. Considering that English allows the double-object structure while Portuguese does not, that Portuguese allows the prepositional phrase at the beginning of the sentence, and that no studies were found on the passive structure for both languages, we are interested in investigating whether there would be a difference between early and late learners regarding the L1 influence on L2 processing on these types of constructions. Assuming that age is significant for L2 syntactic processing, the aim of this study was to investigate to what extent the mother tongue (BP) of intermediate and advanced learners would influence the processing of English passive sentences with three-argument verbs.

Materials & Methods: The participants were native Brazilian Portuguese (BP) adult English learners who were divided into the following groups: intermediate and advanced learners matched for age of acquisition (12-16 years and after 16). A self-paced (moving window) reading task was run with 56 English adult learners (mean age: 23.64) that read, among 24 fillers, 12 experimental English passive sentences with three-argument verbs in three conditions: (i) **different** syntactic structure from L1-BP (E.g., Finally, *the girl* [critical segment] was given the book in the library), (ii) **similar** structure to L1 (E.g., Finally, *the book* [critical segment] was given to the girl in the library), and (iii) **agrammatical** syntactic structure (E.g., Finally, *to the girl* [critical segment] was given the book in the library), which is an acceptable and current structure in BP. The VST (Vocabulary Size Test) was used to determine the proficiency of the groups.

Results & Discussion: A multivariate ANOVA (SPSS) examined age of acquisition and English proficiency as covariates, the reading time as dependent variable, and the age of acquisition (AoA) and English proficiency as independent variables. The multivariate result was significant for English proficiency (Intermediate group: Pillai's trace = 0.142, $F(6.328)=4.169$, $p<0.05$; Advanced group: Pillai's trace = 0.142, $F(6.766)=9.737$, $p<0.001$). Further analysis using the Tukey test revealed a significant difference between the pairs **agrammatical-different** and **agrammatical-similar** for the intermediate group ($p=.002$ and $p=.001$) and the advanced group ($p<.0001$ for both). The reading times for **different** and **similar** conditions were faster for those who acquired English between the ages of 12-16 in both groups. For the **agrammatical** condition, the reading times were higher both groups, which suggests that they recognized the sentence as agrammatical in English, despite resembling their L1 structure. This suggests that there might have not been L1 transfer when processing this type of construction.

Conclusions: As expected, the results show a progressive increase in reading times proportional to the age of acquisition, the earlier the L2 was acquired the faster the critical segments were processed. Our findings revealed that learners in both groups increased the reading time in the agrammatical condition suggesting that they did not resort to their L1 to process the structure. The reading times of the similar, despite being structurally identical in both languages, and the different conditions were higher in both groups when acquisition occurred after the age of 16. It may be concluded that the mother tongue of these learners probably did not influence the processing of passive sentences with three-argument verbs.



Graph 1. Mean reading time of the critical segment in the three experimental conditions between the two groups and AoA.

References

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177, 263–277. <https://doi.org/10.1016/j.cognition.2018.04.007>

Bonfieni, M., Branigan, H. P., Pickering, M. J., & Sorace, A. (2019). Language experience modulates bilingual language control: The effect of proficiency, age of acquisition, and exposure on language switching. *Acta psychologica*, 193, 160–170. <https://doi.org/10.1016/j.actpsy.2018.11.004>

Oh TM, Graham S, Ng P, Yeh IB, Chan BPL and Edwards AM (2019) Age and Proficiency in the Bilingual Brain Revisited: Activation Patterns Across Different L2-Learner Types. *Front. Commun.* 4:39. doi: 10.3389/fcomm.2019.00039

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, 15(6), 409–414. <https://doi.org/10.1111/j.0956-7976.2004.00693.x>

VanPatten, B. (2015). Input Processing in Adult Second Language Acquisition. In B. VanPatten, & J. Williams (Eds.), *Theories in Second Language Acquisition: an introduction* (pp. 113-134). New York, NY: Routledge.

A divergence between judgments and response times in L2 agreement attraction

Eun-Kyoung Rosa Lee, Colin Phillips (University of Maryland at College Park)

This study addresses a puzzle in the second language processing literature about the use of L2 features that are absent in the L1, and in so doing, it uncovers evidence for a “hidden” L2 agreement attraction effect. Native speakers of languages with number agreement have shown to be susceptible to attraction effects in their L2 [1, 2]. However, there have been conflicting findings about L2 learners whose L1 lacks number agreement, based on studies that examined different languages, structures, and methods [3, 4]. A previous study [5] resolved the conflict by showing that Korean learners of English were prone to attraction with relative clause (RC) modifiers but not prepositional phrase (PP) modifiers, based on end-of-sentence judgments. In the present study we use a modified paradigm with speeded mid-sentence judgments that allow us to measure judgment errors as well as RTs in correctly judged sentences. The judgments replicated the structural contrast in L2 agreement attraction (attraction with RCs, not PPs), but the RTs in correctly judged sentences revealed attraction for both structures. We consider the implications of this hidden attraction effect for accounts of interference in L1 and L2 processing.

A group of advanced Korean learners of English ($N = 36$), with a control group of native English speakers ($N = 36$), participated in a speeded forced-choice task, where participants read English preambles in RSVP and judged whether the following target word was a good continuation or not, as quickly as possible. Critical trials included manipulations of grammaticality, attractor, and modifier type (Table 1), and fillers with different types of errors were included as distractors. The acceptance rates and the RTs for correctly judged trials were analyzed using mixed-effects logit models. The results showed increased acceptance rates for sentences with attractors in RCs but not PPs, only in the L2 group (Table 2, Figure 1), replicating the pattern found in [5] and the conflicting results in earlier studies [3, 4]. The RTs, however, did not show this contrast. There was an overall increase in RTs when an attractor was present, indicating an attraction effect, which did not interact with modifier structure in either group (Table 3, Figure 2).

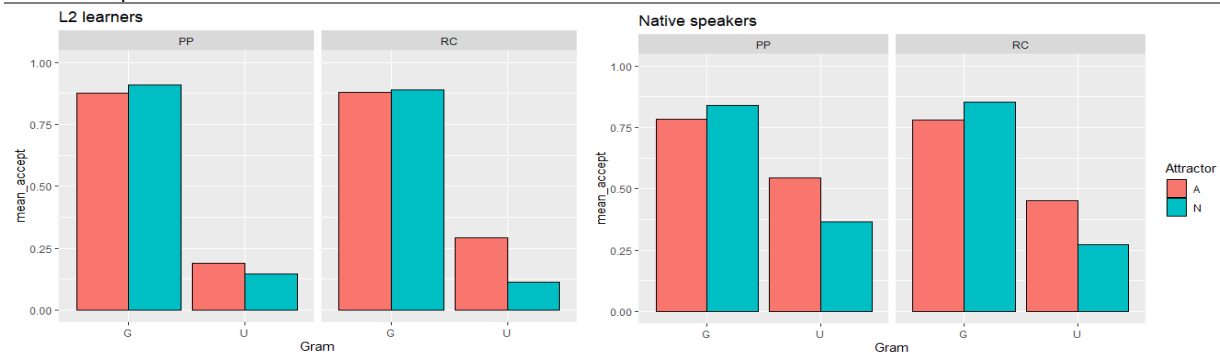
The unique structural contrast in L2 attraction found in the learners’ judgments challenges accounts that predict a general effect of no attraction [3] or similar [4] or greater [6] size of attraction compared to native speakers. Clear L2 attraction from RC modifiers in both the judgments and RTs suggests that speakers of a language that lacks number agreement can still readily use the number cue to compute L2 agreement. Using the number cue sometimes makes the learners incorrectly retrieve the attractor instead of the subject, leading to an attraction effect like native speakers. The case with PP modifiers presents an interesting puzzle: judgments did not show attraction while RTs did. Even though it may appear from the judgments that the attractors played no role in computing agreement, the increase in RTs in sentences with attractors is evidence that the attractors did interfere, even in cases where the learners made correct judgments and when their judgments did not show an attraction effect. We present two possible interpretations of this judgment-RT asymmetry. One is that the judgments and RTs are both products of the same process probed at different time points: the RTs reflect initial competition between the subject and attractor while the judgments reflect subsequent correct retrieval of the subject. RTs increase when there is competition between the [+subject] cue-matching subject and [+plural] cue-matching attractor, causing a delay in retrieval. However, this competition is not strong enough to pass the threshold for producing an incorrect judgment, possibly because the number cue, which is specific to the L2, is not a strong enough competitor for the subject cue that is shared between L1 and L2. Another possibility is that the RTs reflect an equally strong competition between the subject and attractor in the L1 and L2 groups, but there are additional mechanisms associated with the learners’ judgments, such as a self-monitoring system that the learners use to filter out errors and avoid attraction in their judgments. While the interpretation of the judgment-RT asymmetry is uncertain given that most previous works have relied on either one, the comparison between these measures can be particularly informative for cases where judgments show immunity to attraction effects.

Table 1. Experiment conditions and example stimuli

Type	Grammaticality	Attractor	Condition	Preamble	Target word
PP	Grammatical	No attractor	PGN	The artist with the tall sculpture	is
		Attractor	PGA	The artist with the tall sculptures	is
	Ungrammatical	No attractor	PUN	The artist with the tall sculpture	are
		Attractor	PUA	The artist with the tall sculptures	are
RC	Grammatical	No attractor	RGN	The artist who made the sculpture	is
		Attractor	RGA	The artist who made the sculptures	is
	Ungrammatical	No attractor	RUN	The artist who made the sculpture	are
		Attractor	RUA	The artist who made the sculptures	are

Table 2. L2 learners' and native speakers' mean acceptance rates (%)

Group	PGN	PGA	PUN	PUA	RGN	RGA	RUN	RUA
L2 learners	90.86	87.70	14.44	18.92	88.83	88.11	11.11	29.26
Native speakers	80.45	75.38	38.50	54.21	81.67	77.72	28.57	43.98



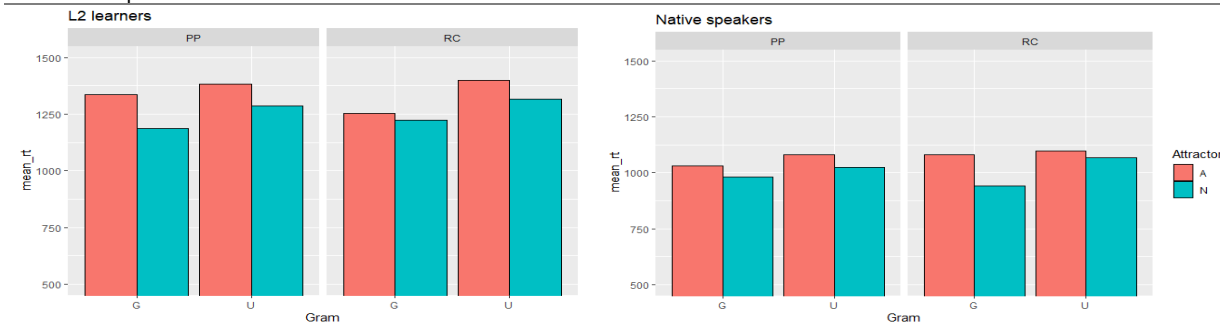
PP: Gram. x Attr.: $p > .05$ RC: Gram. x Attr.: $p < .001$

Gram. x Attr.: $p < .001$ (no interaction with Type)

Figure 1. L2 learners' (left) and native speakers' (right) mean acceptance rates

Table 3. L2 learners' and native speakers' mean response times (ms)

Group	PGN	PGA	PUN	PUA	RGN	RGA	RUN	RUA
L2 learners	1186	1336	1288	1382	1224	1252	1318	1400
Native speakers	978	1027	1018	1097	953	1058	1075	1078



Attr.: $p < .001$ (no interaction with Type)

Attr.: $p < .001$ (no interaction with Type)

Figure 2. L2 learners' (left) and native speakers' (right) mean response times (ms)

References

- [1] Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, 76, 195-215.
- [2] Lago, S. & Felser, C. (2017). Agreement attraction in native and nonnative speakers of German. *Applied Psycholinguistics*, 39(3), 619-647.
- [3] Schlueter, Z., Momma, S., & Lau, E. (2019). No grammatical illusions with L2-specific memory retrieval cues in agreement processing. Manuscript submitted for publication.
- [4] Lim, J. H., & Christianson, K. (2015). Second language sensitivity to agreement errors: Evidence from eye movements during comprehension and translation. *Applied Psycholinguistics*, 36(6), 1283-1315.
- [5] Lee, E. K. (2020). Agreement attraction in nonnative language processing: The effect of sentence complexity. Poster presented at *The 33rd Annual CUNY Human Sentence Processing Conference*.
- [6] Cunnings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4), 659-678.

Meaning, but not grammatical features, is cross-linguistically mask-primed in sentential contexts

Jeonghwa Cho, Jonathan Brennan (University of Michigan)

Background: One of the topics much debated in bilingual studies is how multiple languages are represented in bi/multilinguals' minds. Much of the key evidence comes from cross-linguistic priming effects on lexical and syntactic levels, although a consensus view has yet to emerge. At the lexical level, Chen and Ng (1989) and others report faster reading times in a lexical decision task when a prime in the L1 has the same meaning as the L2 target word. Cross-linguistic priming effects were observed for syntactic structure (Hartsuiker et al., 2004; Shin and Christianson, 2009; Kantola and van Gompel, 2011), which showed that an exposure to a certain construction in one language significantly increases the probability of using the same structure in another language. However, non-masked results do not rule out effects of conscious perception and controlled processes. Hence, the current study aims to broaden the understanding of cross-linguistic priming effects by investigating the priming effect of L1 case feature on L2 using a masked priming paradigm. In specific, Korean words that are marked for case are used as prime words for English accusative words presented in a sentential context.

Methods: 60 Korean-English bilinguals (19 males, age: 28.58 (18-46)) living in the United States (49 participants) or United Kingdom (11 participants) performed an online masked priming experiment. Participants self-reported their first language as Korean, and had lived in the United States or United Kingdom for at least one year at the time of participating in the study. A total of 104 experimental sentence sets and 104 filler sentences were used. All sentences were presented in English, comprised of a person's name, a past transitive verb, and a target word in accusative case (e.g. *Mary bought bread*). The target word was preceded by a prime word presented in Korean that either matched or mismatched in meaning and case (accusative vs genitive), resulting in four conditions (Table 1). The experimental sentences and prime words were distributed across four lists in a Latin-square design. Target words in filler sentences were chosen not to match the meaning of the rest of the sentence (e.g. *Gianna drank clocks*), and were preceded by random Korean prime words with either accusative or genitive case markers. In each trial of the experiment, a subject and a verb were presented in the center of the monitor for 30 frames at 60 Hz (approx. 480 ms), respectively. Then a forward mask was presented for 30 frames (approx. 480 ms), followed by a prime word for two frames (approx. 34 ms) and a target (Figure 1). After seeing the target word, participants judged whether the sentence they just read is semantically correct or not.

Results: Mean raw reaction times (RTs) for each condition are presented in Figure 2. For statistical analysis, log-transformed RTs were analyzed with generalized linear regression model with lexical identity, case feature identity and their interaction, and target word length as fixed effects. We found main effects of lexical identity ($t = -6.3, p < .001$) and target word length ($t = 3.0, p < .001$), such that primes that were translations (with any kind of case marker) had faster RTs than non-translations and longer words were read more slowly. On the other hand, the main effect of case feature identity ($t = 1.1, p = .27$) and the interaction of lexical identity and case identity ($t = -1.0, p = .31$) did not reach significance.

Discussion: The results show that lexical items show cross-linguistic masked priming for semantic repetition, especially when primes are in L1. This replicates previous findings (e.g. Chen and Ng, 1989; de Groot and Nas, 1991; Jiang, 1999; Hoshino et al., 2010). However, grammatical case features do not demonstrate cross-linguistic masked priming effects. We interpret this as indicating that grammatical features between languages may not form an integrated representation as lexical items do, at least as regards case features.

Table 1. Examples of experimental sentences

Condition		Prime	Target
(a) LexId, CaselD	Mary bought	빵을 <i>ppang-ul</i> bread-ACC	bread-ACC
(b) LexId, CaseDiff		빵의 <i>ppang-uy</i> bread-GEN	
(c) LexDiff, CaselD		빳을 <i>bis-ul</i> comb-ACC	
(d) LexDiff, CaseDiff		빳의 <i>bis-uy</i> comb-GEN	

LexId: lexical identity, LexDiff: lexical difference, CaselD: case feature identity, CaseDiff: case feature difference

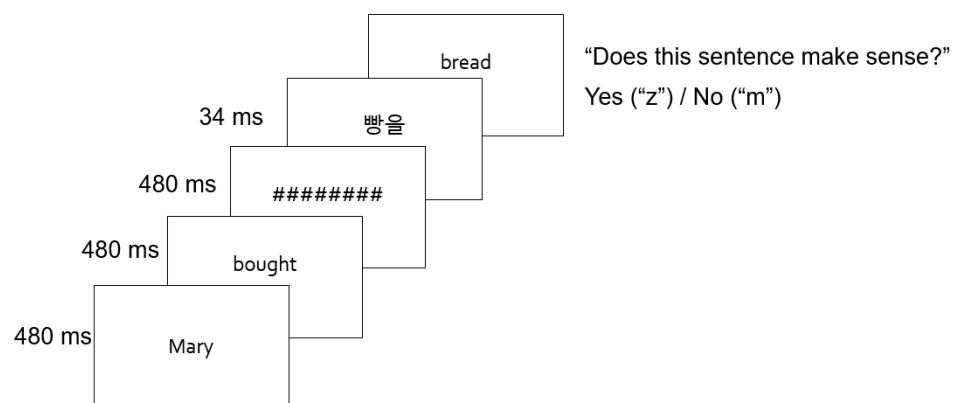


Figure 1. Experiment procedure

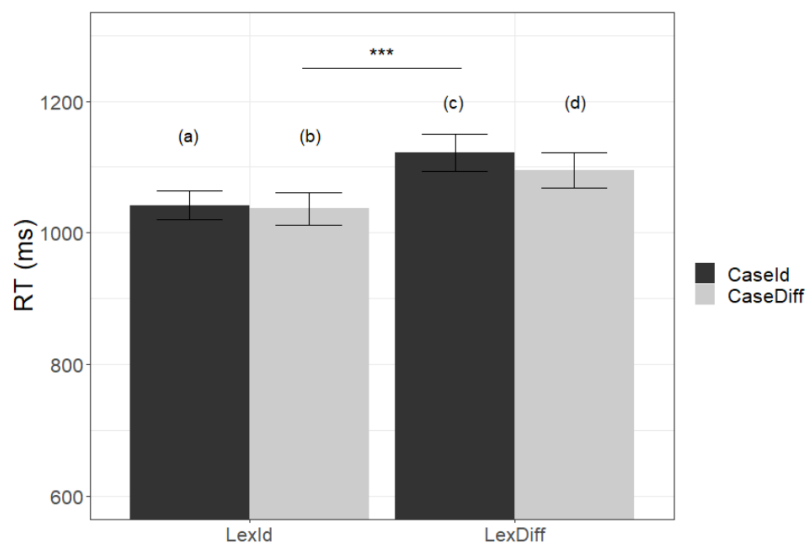


Figure 2. Mean RTs for lexically identical vs different conditions and case identical vs different conditions. Error bars indicate standard errors.

Monolingual and bilingual processing at the syntax-discourse interface: Evidence from the English dative alternation

Joshua D. Weirick & Elaine J. Francis (Purdue University)

In English, dative sentences where an agent causes the physical or metaphorical transfer of a theme to a recipient are expressed using different structural options, as in (1-3). A speaker's choice of structure is governed by a set of interacting, probabilistic constraints, one of which is the accessibility/givenness of the post-verbal arguments (Arnold et al 2000; Thompson 1990). Self-paced reading and judgment tasks have shown that DO and HNPS sentences but not PO sentences are more difficult to process when the referent of the first post-verbal argument is new/less accessible (Clifton & Frazier, 2004; Brown, Savova & Gibson, 2012). Similar studies of L2 English have shown variation based on English proficiency and task: In a forced preference task, Park (2011, 2014) found that advanced L2 English/L1 Korean speakers were sensitive to information order, but preferred PO more often than L1 speakers. Marefat (2004) found that acceptability ratings of intermediate and advanced L2 English/L1 Farsi speakers matched those of L1 speakers, while ratings of low proficiency speakers were always higher for PO. These authors point to L1 transfer due to lack of a DO equivalent in Korean/Farsi, but DO is also less frequent in the input, and speakers from other L1 backgrounds were not tested. Previous L2 studies also did not include HNPS structure or reading time measures and did not consider speakers' exposure to and use of English. The current study fills these gaps by examining the relative influence of sentence type, information order, L1 background, English proficiency, English exposure, and English use on the processing of English dative sentences.

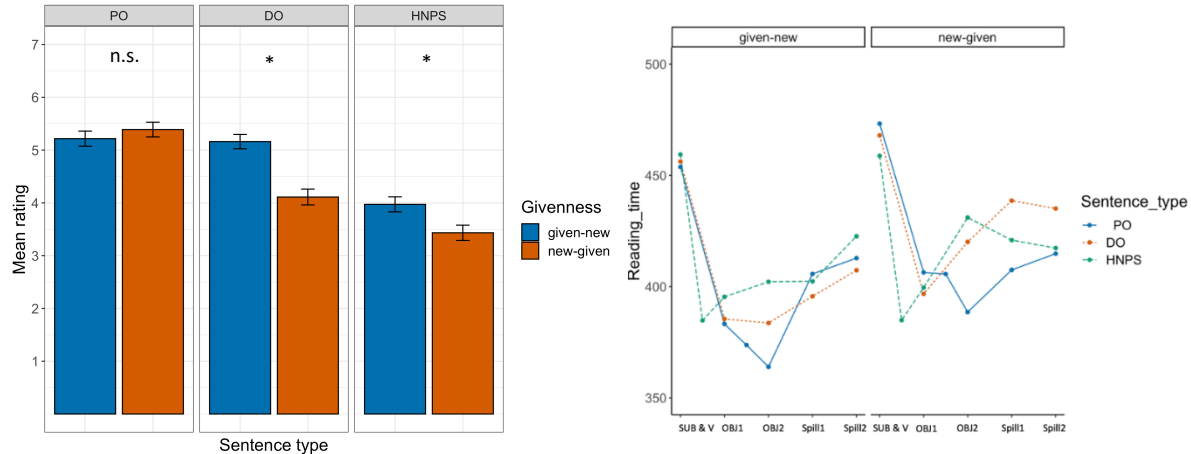
Hypotheses: (H1) for monolingual English speakers, DO and HNPS sentences will be significantly less acceptable/dispreferred/harder to process in the new-given order compared to the given-new order, while PO sentences will not differ (Brown et al., 2012; Clifton & Frazier, 2004). (H2) for bilingual L2 English speakers, sensitivity to information order will be more target-like as English proficiency, exposure, and use increase (Park 2011, 2014). (H3) Spanish but not German lacks a close DO equivalent. DO will be less acceptable/preferred less often by Spanish-English speakers than German-English speakers (Park 2011, 2014).

Experiment: 60 monolingual English speakers living in the United States, 60 bilingual English/German speakers living in Germany, and 60 bilingual English/Spanish speakers living in Mexico were recruited online via Prolific. Participants completed three experimental tasks (self-paced reading, scalar acceptability rating, forced preference, counterbalanced between participants), a language background questionnaire, and the LexTALE lexical decision task as a measure of proficiency (Lemhöfer & Broersma, 2012) in a single session of about one hour. Stimuli for the experimental tasks were adapted (with permission) from Brown et al. (2012). Each item contained a context sentence followed by a dative test sentence (PO, DO, or HNPS) with full NP post-verbal arguments in given-new or new-given order. Givenness was marked by a definite article and prior mention of the referent in the context sentence. Preliminary analyses given here are based on linear (self-paced reading), ordinal logistic (acceptability) and binomial logistic (forced preference) mixed-effects models for each participant group.

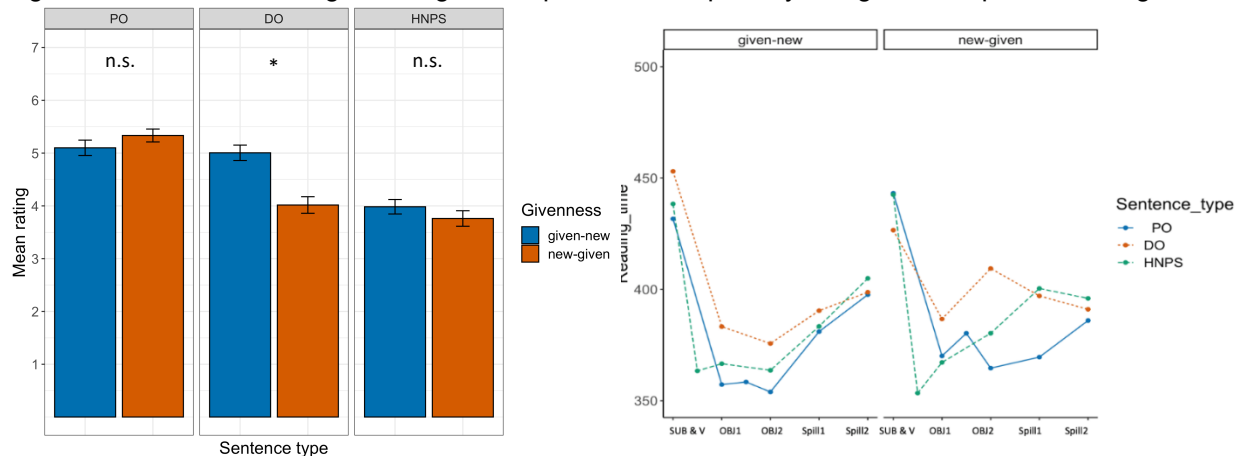
Results and discussion: For monolingual speakers, acceptability and forced preference results were as predicted in H1 (Figure 1; forced-preference omitted). In the self-paced reading task, RTs at the critical region (OBJ2) were slower for all three sentence types in the new-given condition (Figure 2). This was expected for DO and HNPS, but not for PO. Both bilingual groups were sensitive to information order in DO sentences in a similar manner to the monolingual group but showed some minor differences with respect to HNPS: German-English speakers showed no differences due to information order (Figures 3 & 4); Spanish-English speakers showed the expected dispreference for new-given order, but unlike the other groups rated HNPS sentences as no less acceptable than DO sentences (Figure 5). Statistical analyses to assess H2 and H3 are ongoing and will be reported in the presentation.

- (1) The student sent the botanist a photograph. (double object; DO)
- (2) The student sent a photograph to the botanist. (prepositional object; PO)
- (3) The student sent to the botanist a photograph. (heavy NP shift; HNPS)

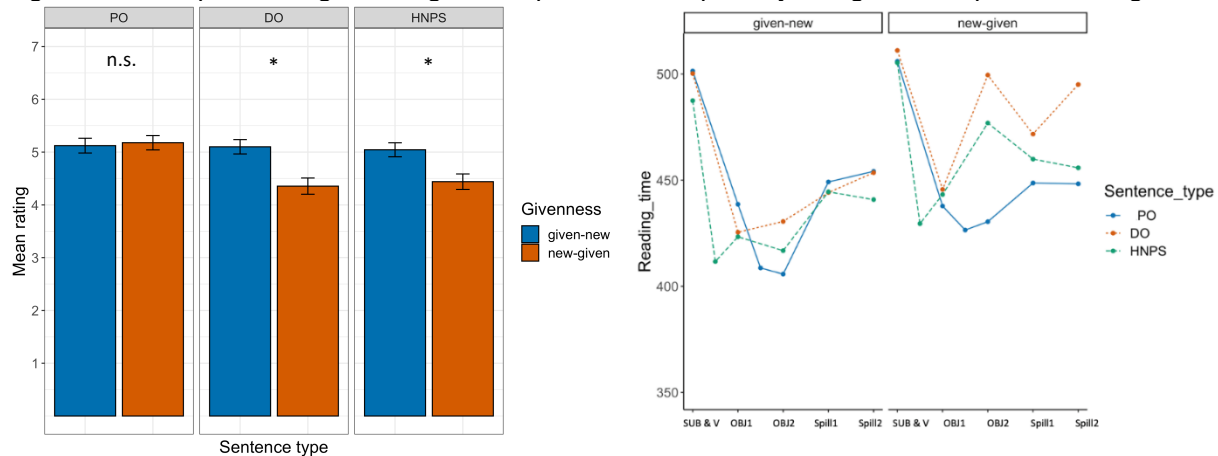
Figures 1 & 2: Monolingual English speakers' acceptability rating and self-paced reading results



Figures 3 & 4: German-English bilingual L2 speakers' acceptability rating and self-paced reading results



Figures 5 & 6: Spanish-English bilingual L2 speakers' acceptability rating and self-paced reading results



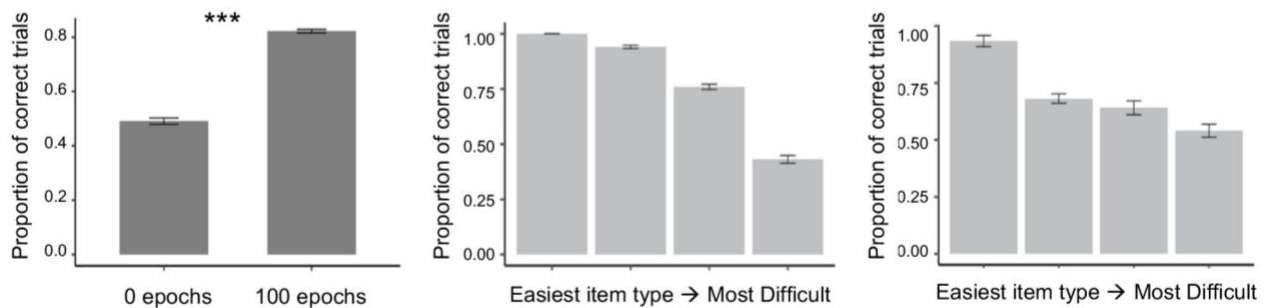
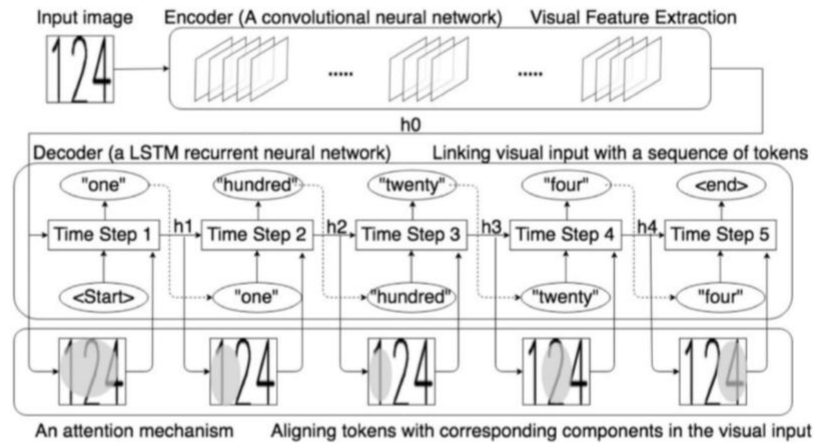
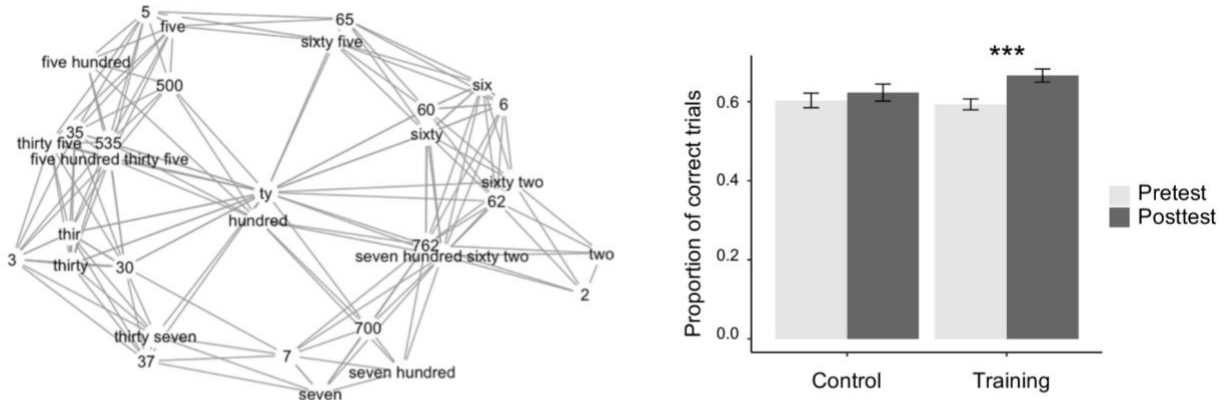
Learning the generative principles of a linguistic system from limited examples
Lei Yuan, Violet Xiang, David Crandall, & Linda Smith

One of the motivating questions of this year's special session is: How are children exposed to a small set of linguistic input but are able to master a complete linguistic system that allows for infinite generalization? Sharing the organizers' vision, we view this question as an important step to understand mature language processing and to arrive at important theoretical unifications about human cognition in general. Proposals to the above question typically either: a) emphasize rapid generalization based on prior principles that are language-specific and part of human core knowledge system, or b) highlight the generative principles discovered through more domain-general mechanisms, including associative learning mechanism. But even in its most advanced form of deep learning models, associative mechanism has been criticized as data-hungry and limited in generalization, thus falling short of accounting for human language learning. In these debates, little attention has been paid to the structure of input data—on which proposed learning mechanisms must operate. Thus, the goal of the current research is to better understand how associative mechanism interfaces with the statistical structure of input data to produce **far generalization** of a **linguistic system** based on **limited input data**.

We address the above question by focusing on the language system that underlies the Arabic multi-digit number symbols and their spoken names. As a relatively recent human-invention, this system is less likely to have innately-involved structures, with much evidence suggesting that formal education is necessary to acquire this system and that many school-aged children struggle to do so. But it is also a system with many overlapping features that exhibit a small-world like structure as shown in Fig. 1 (Left), in which multiple redundant, degenerate, imperfect but inter-predictive features offer multiple pathways to the to-be-learned generalizable principles. We ask: How are place value terms (e.g., "hundred") combined in conjunction with single-digit number names (e.g., "three") to create an in-theory infinite set of possible expressions such as "three hundred gazillions and five"? We hypothesize that a suite of these inter-correlated imperfect predictive components can allow an associative learning mechanism to generalizations that accord with generative principles and can do so despite limited training data.

In Study 1 and 2, 148 preschool children were randomly assigned to either a training condition in which they were given experience with alignable pairs of written multi-digit numbers and their corresponding spoken names in casual learning activities such as storybook reading (e.g., "Johnny wants to save money to buy a new toy car. How much does it cost? Look! It costs forty-two dollars [the experimenter pointed to the written "4" and "2" in sequence]."), or a control condition in which they saw the same material but were shown letters (e.g., "CAR") rather than numbers. Across three days of 15-minutes daily exposure to a small set of 36 unique numbers, only children in the training condition (but not the control condition) showed significant accuracy improvement from pre- to post-test in recognizing a novel, never-before-seen multi-digit number (e.g., Which is one hundred twenty-five? 125 vs. 251), as shown in Fig. 1 (Right). Study 3 used a deep learning model—as shown in Fig. 2—to provide evidence that the co-predictive properties between number names and their written forms, albeit imperfect and local predictors, are sufficient for an associative learner to make far and systematic generalizations without explicitly representing any rules or principles. As shown in Fig. 3, after training using the same small set of input data as children in Study 1 & 2, the models not only demonstrated significant learning (based on four complementary measures), but also showed the same error pattern as children in responding to items that vary in difficulty levels.

This result—that associative mechanism can lead to far generalization when operated on limited input data with statistical structure that is conducive to learning and in fact pervasive in many real-life domains—is important for understanding how limited data *with a particular statistical structure* gives rise to far generalizations. Implications for learning natural language will be considered.



If Memory Doesn't Serve: Timecourse of Syntactic Forgetting in Ellipsis and Recognition

Caroline Andrews (University of Zürich)

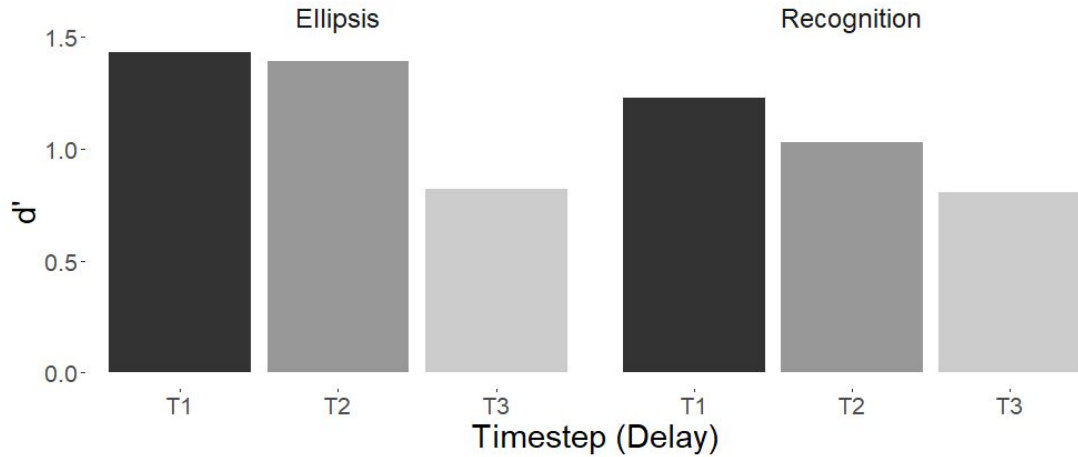
The past two decades have seen an increase in the prominence of memory-based theories of syntactic processing[1,2,3]. These theories use domain general memory architectures, which in turn were built on data from recognition/recall memory methods. Yet recognition/recall memory studies consistently find evidence that syntax is forgotten faster (<30s) than other information types, e.g., semantic and lexical information[4]. This is a puzzle for memory-reliant theories, because if syntax rapidly becomes unavailable in short-term memory, it should rapidly become unavailable to guide syntactic processing. Here we test whether the timecourse of forgetting for syntax in a sentence recognition task is similar to forgetting in Verb Phrase Ellipsis (VPE), a grammatical dependency that requires comprehenders to access the syntax of a prior clause.

Memory would be particularly advantageous when processing VPE, (1), because VPE can span across independent clauses but also requires strict matching of the syntactic structure of the antecedent and elided material[5]. Prior work on memory for VPE indicates that, despite the potential advantage, syntax is quickly forgotten[6], but these studies did not compare to recognition/recall to see if there was a *relative* advantage for syntax in ellipsis resolution. If there were, then recognition might not be a reliable indicator of access to syntactic memory.

This study directly compared memory decay in ellipsis resolution and recognition over time. Either the active or passive version of (1) was presented in RSVP. On a following screen, participants (N=54) were asked to either verify whether a new sentence matched what they had seen (Recognition task) or if the active or passive version of (2) was a possible continuation (Ellipsis task)(Items=144). In both tasks, if the voice matched the correct response was 'yes', and if voice mismatched 'no' was correct. In between (1) and the target task, participants completed 0, 2, or 5 math problems (TIME 1,2,3 respectively) so that we could track forgetting over time. If recognition tasks underestimate available memory for syntax, performance in the ellipsis task will be less impacted by memory decay over time (forgetting) than in recognition. **Results:** There was a reliable effect of TIME in Bayesian logistic models (dependent variable: accuracy), indicating that the design was able to measure forgetting as it occurred (95% CredibleInterval: -0.52 – -0.32). There was also an overall yes-bias, indicated by MIS/MATCH and MIS/MATCHxTASK having very low probability densities around zero (95%CI: 1.85 – 2.72 & -3.02 – -1.27 respectively). But, TASK was marginal (95%CI: -0.86 – 0.01) and the TIMExTASK posterior probability was centered near 0 (95%CI: -0.10 – 0.28). Fig.1 illustrates this with the decrease over time in d' (sensitivity or the ability to discriminate between the correct and incorrect items). The final sensitivity is comparable, demonstrating how quickly syntactic information is lost. However, a model of only T1 had a more reliable posterior for TASK differences (95%CI:-0.71 – -0.05), indicating that at T1 the Ellipsis task had better *initial* sensitivity to syntactic differences. This initial difference leads to a difference in the syntactic forgetting profiles of the two tasks.

The results suggest that syntax does play a privileged role in grammatical processing, but, in line with previous studies, found that the privileged role does not persist into memory. Domain general recognition memory may reasonably model rapid forgetting of syntactic information, but is not necessarily an accurate model of memory access in syntactic processing overall.

- (1) a. *Active*: The politician criticized the journalist over the presentation of the new bill.
 b. *Passive*: The journalist was criticized by the politician over the presentation of the new bill.
- (2) *Ellipsis Task Continuations*:
 a. *Active*: The T.V. pundit did too.
 b. *Passive*: The T.V. pundit was too.



Bayesian Logistic Model: T1-T3

	\hat{R}	N_{Eff}	Mean	SD	2.5%	97.5%
Timestep	1.00	20000	-0.42	0.05	-0.52	-0.32
Match/Mismatch	1.00	15867	2.28	0.22	1.85	2.72
Task	1.00	15326	-0.42	0.22	-0.86	0.01
Timestep x Match/Mismatch	1.00	20000	-0.26	0.10	-0.45	-0.08
Timestep x Task	1.00	20000	0.09	0.10	-0.10	0.28
Match/Mismatch x Task	1.00	14970	-2.14	0.45	-3.02	-1.27
Timestep x Match/Mismatch x Task	1.00	15872	0.33	0.19	-0.05	0.71

Bayesian Logistic Model: T1 Only

	\hat{R}	N_{Eff}	Mean	SD	2.5%	97.5%
Match/Mismatch	1.00	15309	2.28	0.17	1.95	2.62
Task	1.00	16104	-0.37	0.17	-0.71	-0.05
Match/Mismatch x Task	1.00	15044	-1.83	0.34	-2.50	-1.18

Tables 1 & 2: Bayesian logistic analysis posterior estimates for all three timesteps (Table1) and Timestep 1 only (Table 2). Dependent measure accuracy.

[1] Lewis & Vasishth, 2005. *CogSci* [2] Van Dyke & McElree, 2006. *JML* [3] Wagers, Lau, Phillips, 2009. *JML* [4] Potter & Lombardi, 1990. *JML* [5] Johnson, 2001. *Blackwell Syntax*. [6] Garnham & Oakhill, 1987. *QJEP*.

The Structure of Antecedent Influences Processing of Ellipsis

Hyosik Kim (Northwestern University), Ming Xiang (The University of Chicago) and Masaya Yoshida (Northwestern University)

[Introduction] One of the long-standing questions in the study of the processing of ellipsis constructions is whether processing of an ellipsis site is influenced by the structure of the antecedent of the ellipsis site. Some previous studies have shown that the structure of the antecedent does not influence the processing of the ellipsis site and suggested that structures may not be built in the ellipsis site [1,2,8,9]. On the other hand, other studies have suggested that the structure of the antecedent may influence the processing of the ellipsis site [7,10,11]. In the present study, we investigate whether structural properties of antecedent clauses influence the processing of the ellipsis site. The result of an eye-tracking while reading experiment shows that the structural complexity of the antecedent and the processing complexity of the ellipsis site correlate, i.e., when the antecedent involves more complex structures, the processing of the ellipsis site is slower. We argue this result suggests that the parser is accessing the structure of the antecedent when the ellipsis site is processed.

[Experiment] An eye-tracking while reading experiment ($n=77$) was conducted in which, the structure of the antecedent (*Antecedent*: NP vs. CP) x Structure of the second clause (*2nd Clause*: Ellipsis vs. Pronoun) were manipulated in a 2x2 factorial design (a sample set of stimuli is summarized in the table 1). Previous studies on the processing of wh-dependencies have shown that when the wh-phrase moves over a complex NP as in (1a), the processing of a wh-gap dependency is more difficult compared to when the wh-phrase moves out of a subordinate clause (CP) as in (1b) [3,6]. [3,6] argued that the different structure created different processing complexity effects.

- (1) a. ... **who** [_{NP} the consultant's denial about that the new proposal] had pleased GAP.
b. ... **who** [_{Clause} the consultant denied that the new proposal had pleased GAP].

Taking advantage of this paradigm, we can potentially test whether the structure of the antecedent of the ellipsis site influences the processing of the ellipsis site. If the parser accesses the structure of the antecedent during the processing of the ellipsis site, then when the antecedent involves more complex structure, the processing of the ellipsis site should be more difficult. On the other hand, if the parser does not access the structure of the antecedent, then the complexity of the antecedent should not create the difficulty of the processing of the ellipsis site. Pronoun conditions were included to serve as baseline since studies have shown that the parser does not access the structural information of the antecedent of the pronoun when the pronoun is processed [4,7]. A linear mixed effects model revealed that at the wh/pronoun region, a main effect of *Antecedent* in the Total Time Duration measure was found, such that the NP conditions were read significantly slower than the CP conditions ($\beta = 0.10$, $SE=0.03$, $t=2.71$, $p<0.01$) and an interaction between *Antecedent* x *2nd Clause* ($\beta = -0.12$, $SE=0.05$, $t=-2.26$, $p<0.05$) was observed (see Figure 2). Further subset analysis found a significant difference within the *Antecedent* conditions whereby the NP conditions were read significantly slower than the CP conditions ($p<0.05$), but there was no difference within the Pronoun conditions.

[Conclusion] Taken together, this study shows that readers were sensitive to the syntactic structure of antecedents when processing ellipsis sites. One potential objection to this conclusion is that the sentences tested in this experiment are overly long and thus, readers would have given up processing these sentences. However, we observe the difference in processing between the Ellipsis conditions and the Pronoun conditions, in which the complexity and difficulty of the antecedent clause are tightly matched. If the readers have given up processing these sentences, we should not have observed such difference between the Ellipsis conditions and the Pronoun conditions. They should be equally too hard to process and similar effects should be predicted. We conclude that readers indeed had access to the structural information of the antecedent and recovered it when processing the ellipsis site.

	Factor1	Factor2	example
1	CP	sluicing	I wonder who the consultant denied that the new proposal had pleased, but no one knows <u>who</u> , in fact, nobody cares.
2	NP	sluicing	I wonder who the consultant's denial about the new proposal had pleased, but no one knows <u>who</u> , in fact, nobody cares.
3	CP	pronoun	I know who the consultant claimed that the new proposal had pleased, but no one knows <u>about it</u> , in fact, nobody cares.
4	NP	pronoun	I know who the consultant's claim about the new proposal had pleased, but no one knows <u>about it</u> , in fact, nobody cares.

Table1. A sample set of stimuli

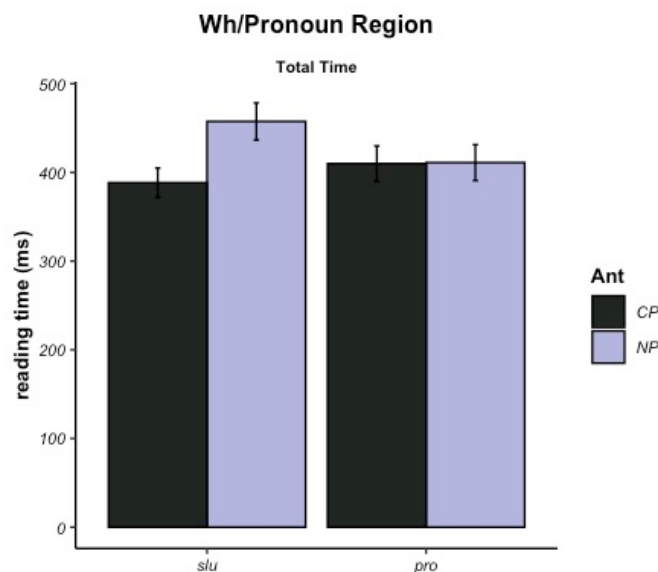


Figure 1. The Total Time Duration (TTD) at the target (wh/pronoun) region

References

- [1] Frazier, L. & Clifton, C. 2000. *Journal of Psycholinguistic Research*.
- [2] Frazier, L. & Clifton, C. 2001. *Syntax*.
- [3] Gibson, E., & Warren, T. 2004. *Syntax*.
- [4] Hankamer, J., & Sag, I. 1976. *Linguistic Inquiry*.
- [5] Kehler, A. 2002. *Coherence, Reference, and the Theory of Grammar*.
- [6] Keine, S. 2015. *University of Massachusetts, Amherst*.
- [7] Kim et al. 2018. *Language Cognition and Neuroscience*.
- [8] Martin, A. E. & McElree, B. 2008. *Journal of Memory and Language*.
- [9] Martin, A. E., & McElree, B. 2011. *Journal of memory and language*
- [10] Murphy, G. 1985. *Journal of Pragmatics*.
- [11] Paape, D., Nicenboim, B., & Vasisht, S. (2017). *Glossa: a journal of general linguistics*

I(interpolated) Maze: High-sensitivity measurement of ungrammatical input processing

Pranali Vani, Ethan Wilcox and Roger Levy

Summary: The Maze task (Forster et al., 2009), in which an experimental participant “navigates” through a text by successively choosing the contextually appropriate target next word over an inappropriate one, is web-deployable with better power and sensitivity for detecting incremental-processing RT effects than self-paced reading (Boyce et al., 2020). In G(rammatical)-Maze, the distractor is a contextually inappropriate word; in L(exical)-Maze, the distractor is a nonce word. G-Maze is the more sensitive of the two, but it has a limitation for sentence processing research: using it to study the processing of ungrammatical input is problematic, since neither the target nor the distractor would be contextually appropriate. Here we introduce Interpolated Maze (I-Maze) to address this limitation. I-Maze mixes G-Maze and L-Maze distractors, with L-Maze distractors for ungrammatical words. We assess the three Maze variants in two English experiments: Wh-Cleft Structures (which tests syntactic category featural match) and Main Verb / Reduced Relative Clause (MVRR) Garden-Path sentences (which tests expectations for parses, rather than true ungrammaticality). We find G-Maze and I-Maze more powerful than L-Maze. We also discover a novel result for MVRR Sentences: a critical-region garden-path disambiguation effect of relative clause reduction not only for ambiguous participles (*brought*), but also a smaller effect for unambiguous participles (*given*). Interestingly, these patterns also appear in surprisals of the GPT-2 neural language model.

I-Maze: I-Maze items were created by interpolating G-Maze and L-Maze distractors. L-Maze distractors were used for the second word of the sentence, all words in critical regions, and ~35% of the remaining words, in groups of two, where the first word appeared in all conditions. L-maze distractor words outside of the critical region provided a baseline estimate for L-Maze distractor times. L-Maze distractors were produced with Wuggy (Keuleers & Brysbaert, 2010).

Experiments: The Wh-Cleft experiment consisted of four conditions (Ex. 1). Slowdowns were expected in the *mismatch* conditions, relative to *match* conditions. The MVRR Garden-paths (Ex. 2) crossed (i) reduction of relative clauses and (ii) ambiguity of RC verb. We expected slowdowns in the reduced RC or ambiguous verb, as well as an interaction between conditions. Each experiment, which was hosted on Ibex Farm, included thirty Wh-Cleft items, twenty MVRR Garden-paths items and twenty fillers. Thirty participants were recruited on Amazon M-Turk.

Results: The results for Wh-Clefts can be seen in Figure 1. We find a significant effect of matching for G and I-Maze ($p < 0.001$) but not for L-Maze. We also found a main effect where NP continuations were read faster ($p < 0.001$, L-Maze $p < 0.01$), likely because the content verb (e.g. “fixed”) sets up stronger expectations for an object than a light verb (e.g. “did”) does for its verbal complement. The results for MVRR Garden-paths can be seen in Figure 2. We find main effects of reduction for all Maze variants ($p < 0.001$), a main effect of ambiguity for G-Maze and I-Maze ($p < 0.01$, $p < 0.001$), and the expected interaction for G-Maze and I-Maze ($p < 0.001$, $p < 0.05$). The increased power of the Maze task reveals a surprising novel effect, which is that unambiguous but reduced RCs produce slower reading times in downstream critical regions than unreduced ambiguous RCs even though both strings are consistent with only one syntactic parse. This effect is significant for G-Maze and I-Maze ($p < 0.05$). Noisy-channel models could account for this behavior, with reduced RCs being sufficiently rare that the processor reserves significant probability for nearby high prior-probability parses. However, we show that the incremental surprisal values (negative log probabilities) of a contemporary neural language model also capture this behavior (Figure 4), suggesting that predictive processing models which do not maintain an incremental stack of parses can capture these effects just as well.

I(terpolated) Maze: High-sensitivity measurement of ungrammatical input processing

Pranali Vani, Ethan Wilcox and Roger Levy

Figure 1: Cleft Sentences. Error bars are 95% confidence intervals

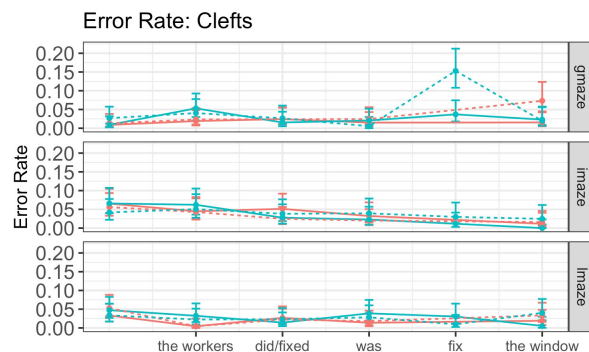


Figure 2: MV/RR Gardenpaths. Error bars are 95% confidence intervals

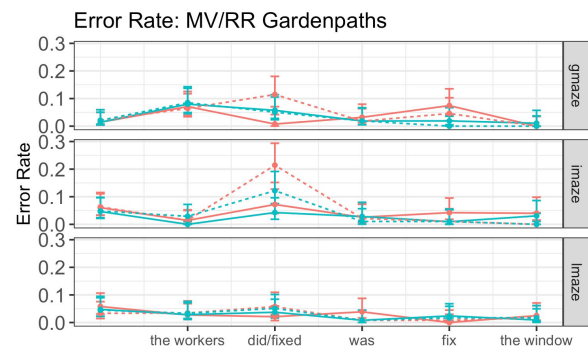
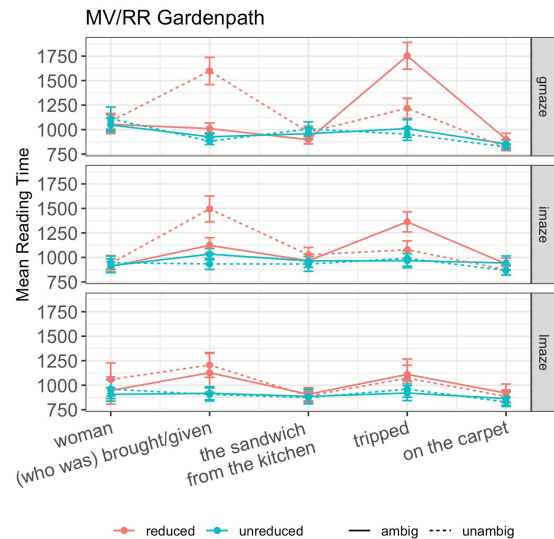


Figure 3: I-Maze and L-Maze critical region error rate in ungrammatical conditions is comparable to the other conditions. G-Maze critical region error rates are higher. This indicates that I-Maze eliminates the speed/accuracy tradeoff that G-maze is prone to, but produces results that are more sensitive to the different conditions than L-Maze.

Ex1: Wh-Cleft Sentences (critical regions are underlined)

What the workers did was repair the window [match/vp]

What the workers fixed was repair the window [mismatch/vp]

What the workers fixed was the window [match/np]

What the workers fixed was repair the window [mismatch/np]

Ex 2: MV/RR Gardenpath Sentences

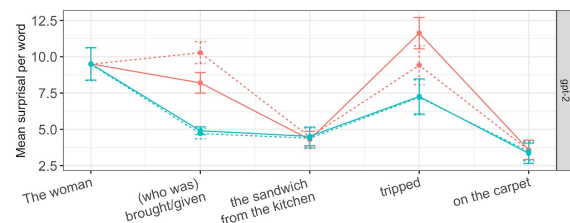
The woman who was given the sandwich tripped. [unreduced / unambiguous]

The woman who was brought the sandwich tripped. [unreduced / unambiguous]

The woman given the sandwich tripped. [reduced / unambiguous]

The woman brought the sandwich tripped. [reduced / ambiguous]

Figure 4: Model results (MVRR Gardenpaths) Y-axis is mean surprisal values ($-\log(\text{word}|\text{context})$) derived from GPT2.



References: Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze Made Easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111, 104082. • Forster, K. I., Guerrero, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1), 163-171. • Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3), 627-633.

Task influences on lexical underspecification: Insights from the Maze and SPR

John Duff, Adrian Brasoveanu, and Amanda Rysling (UC Santa Cruz)

In assigning an incremental interpretation to linguistic input, sometimes a resolution to a temporary ambiguity will prove incompatible with downstream material, leading to costly reanalysis. However, in some cases comprehenders appear to avoid this cost. One approach to such observations is to posit that comprehenders can **underspecify** some input, delaying full commitment and interpretation [2,3,7].

One set of data often taken as evidence for underspecification is [5]’s landmark eyetracking investigation of polysemy, words with multiple meanings which share core features (e.g. *newspaper* as printed object vs. corporate entity) and homonymy, properly ambiguous words with entirely distinct meanings (e.g. *jam* as fruit spread vs. blockage). [5] report that when disambiguation to a less frequent meaning follows a polyseme, reanalysis cost (first-pass RT, probability of regressions out) is less than when such disambiguation follows a homonym. This difference has been argued to follow from an account where decisions among meanings are initially underspecified only when choosing between overlapping meanings [5-7].

From findings of this sort, however, it’s unclear why polyseme specification is delayed. We might consider two hypotheses: first, that underspecification is **utility-based** (effective under typical comprehension strategies); or, underspecification may be **necessary** due to some property of lexical representation and semantic commitment. The present study seeks to address this question by replicating [5] across additional tasks: if underspecification is deployed strategically, it will be sensitive to changes in a participant’s priorities.

In the first experiment, we use [4]’s Maze, in which participants advance word-by-word by making decisions between the correct continuation of a sentence and a foil (see Fig. 1). In particular, we will use the A-Maze of [1], where foils are words with high surprisal in the existing context. If a participant chooses a foil instead of a target, the trial terminates.

Participants in the Maze must engage in eager interpretation to maximize their ability to proceed through the stimulus, making underspecification a less useful strategy than in natural reading. A utility-based account of underspecification then predicts that in the Maze, polysemes may exhibit reanalysis costs similar to homonyms. Alternately, a hypothesis under which underspecification is necessary predicts we should replicate [5], with some interaction in RTs (here, response latencies) for the disambiguation region, such that late disambiguation shows greater costs for homonymy.

Expt. 1 ($n=48$) presented two sets of 32 items featuring polysemy (1) and homonymy (2) in the Maze. Each set crossed disambiguation POSITION (EARLY/LATE) X MEANING (M1/M2), after [5], with dominance established by acceptability norming. Note that EARLY conditions feature cataphoric dependencies. Participants saw items Latin-squared and randomized with 128 fillers.

Log RTs in the disambiguator, residualized over position and length, were analyzed in a Bayesian-fit linear mixed-effects model (Table 1). We observe a Pos main effect we link to the lack of cataphora in LATE, a POS X MEANING interaction indicative of a cost for late disambiguation to M2 for polysemes, and no interaction terms suggesting a difference for homonyms.

Expt. 2 ($n=48$) presented the same items in fixed-window SPR for minimal comparison with the Maze. Analysis reveals a POS X TARGET interaction such that late disambiguation bears larger costs for homonyms than polysemes, consistent with [5]’s underspecification findings.

An account where underspecification is necessary makes the wrong predictions. Instead, we observe task-dependent variation in line with the utility account: polyseme underspecification is avoided in the Maze, a fact we attribute to the task’s demands for eager interpretation. Researchers using the Maze should be aware that it imposes unique task pressures, but awareness of this fact can allow us to see where standard online behavior derives from strategic deployment of the language processing architecture rather than its limitations or requirements.

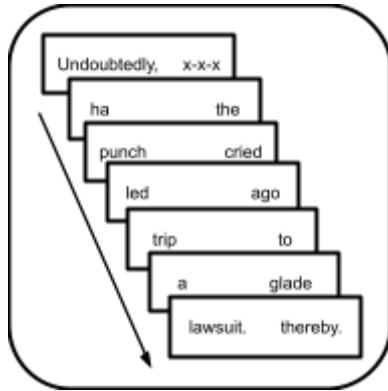


Fig. 1. A depiction of a toy A-Maze trial.

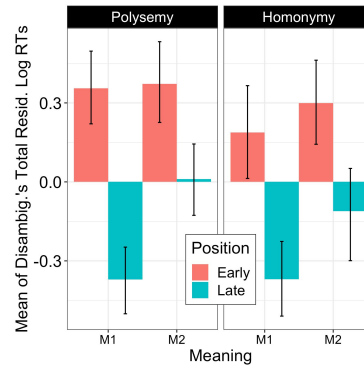


Fig. 2. Mean total residualized log RTs in the disambiguating region (E1).

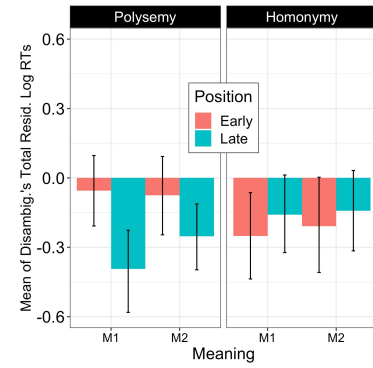


Fig. 3. Mean total residualized log RTs in the disambiguating region (E2).

Fixed Effect	Expt. 1 (Maze)				Expt. 2 (SPR)			
	Mean	SD	95% CI		Mean	SD	95% CI	
POSITION: LATE	-0.74	0.12	-0.97	-0.50 *	-0.33	0.13	-0.58	-0.77 *
TARGET: HOMONYMY	-0.17	0.15	-0.46	0.12	-0.20	0.13	-0.45	0.06
POSITION x MEANING	0.35	0.16	0.05	0.66 *	0.16	0.18	-0.19	0.50
POSITION x TARGET	0.16	0.17	-0.18	0.49	0.44	0.18	0.09	0.79 *
POS x M x TARGET	-0.16	0.23	-0.61	0.29	-0.19	0.25	-0.67	0.30

Table 1. Excerpted mixed-effects models fit to total resid. log RTs in disambiguating region.

(1) **POLYSEMY** (disambiguating region)

- Unfortunately, after it was soaked with rain the newspaper was destroyed. [EARLY,M1]
- Unfortunately, after it lost its advertising profits the newspaper was destroyed. [E.,M2]
(x-x-x intend in job lips discover obtain kid conducted add extension.)
- Unfortunately, the newspaper was destroyed after it was soaked with rain. [LATE,M1]
- Unfortunately, the newspaper was destroyed after it lost its advertising profits. [L.,M2]
(x-x-x kid conducted add extension intend in job lips discover obtain.)

(2) **HOMONYMY** (disambiguating region)

- Reportedly, after it made his toast soggy the jam displeased Tom. [EARLY,M1]
- Reportedly, after it doubled his morning commute the jam displeased Tom. [EARLY,M2]
(x-x-x, come fit detail sir thinks begin kept ours indecision Need.)
- Reportedly, the jam displeased Tom after it made his toast soggy. [LATE,M1]
- Reportedly, the jam displeased Tom after it doubled his morning commute. [LATE,M2]
(x-x-x, kept ours indecision Need come fit detail sir thinks begin.)

References

- [1] Boyce, V., Futrell, R., & Levy, R. P. 2020. *Journal of Memory and Language* 111: 1-13.
- [2] Egg, M. 2010. *Language and Linguistics Compass* 4: 166-181.
- [3] Ferreira, F., & Patson, N. D. 2007. *Language and Linguistics Compass* 1: 71-83.
- [4] Forster, K. I., Guerrero, C., & Elliot, L. 2009. *Behavior Research Methods* 41: 163-171.
- [5] Frazier, L., & Rayner, K. 1990. *Journal of Memory and Language* 29: 181-200.
- [6] Frisson, S., & Frazier, L. 2004. Poster at CUNY 17, University of Maryland.
- [7] Frisson, S. 2009. *Language and Linguistics Compass* 3: 111-127.

The interaction of semantic information and parsing biases: An A-maze investigation

Xinwen Zhang & Jeffrey Witzel (University of Texas Arlington)

This study uses the A-maze task (Boyce et al., 2020) to examine the influence of semantic information on online parsing biases. In the A-maze task, as in all maze task variants (see e.g., Forster et al., 2009), each word in the sentence is presented along with a distractor, and the participant selects the member of the pair that best continues the sentence as quickly and accurately as possible. Boyce et al. (2020) have demonstrated that like other versions of this task, the A-maze produces robust and highly localized indications of incremental processing difficulty. However, the A-maze improves on other versions of this task in that it involves distractors that are automatically generated for each word, which simplifies and systematizes item creation. Crucially, these distractors are generated by a program that selects words that are unlikely (or high in “surprisal”) at each point in the sentence. This means that distractor words are sometimes ungrammatical, sometimes semantically unexpected, and sometimes both. In this way, the A-maze essentially forces incremental syntactic and semantic integration of each word into the developing sentence structure. The present study takes advantage of this task feature to examine the influence of semantic constraints on syntactic parsing biases in sentence types that have yielded somewhat conflicting findings under other online reading paradigms.

The sentence types of interest involved reduced and unreduced relative clauses (RCs) with sentential subjects that were either animate (and good agents for the RC verb) or inanimate (and poor agents/good themes for the RC verb), as in the following examples:

reduced/animate

The defendant examined by the lawyer turned out to be unreliable.

unreduced/animate

The defendant who was examined by the lawyer turned out to be unreliable.

reduced/inanimate

The evidence examined by the lawyer turned out to be unreliable.

unreduced/inanimate

The evidence that was examined by the lawyer turned out to be unreliable.

Many studies have indicated a clear preference for a main-clause interpretation of the RC verb (*examined*) in reduced RC sentences. This is evidenced by processing difficulty (compared to unreduced RC controls) at words that disambiguate the structure of the sentence -- i.e., at and after the RC *by*-phrase. In a now-classic eye-tracking study, however, Trueswell et al. (1994) found that these “garden-path effects” were effectively eliminated under first-pass time in sentences with inanimate subjects. This was taken to indicate that semantic information -- in this case, animacy and semantic fit with the verb -- can override structure-based parsing biases. In a set of follow-up experiments, however, Clifton et al. (2003) found comparable garden-path effects in these sentences, regardless of the animacy of the subject. This was particularly the case under regression path duration, a first-pass reading measure that includes regressive fixations.

The present study ($N=32$) attempted to adjudicate between these somewhat conflicting findings using the A-maze task. The results revealed processing difficulty at the RC verb in reduced/inanimate sentences (see the results table and figure below), indicating that readers detected the semantic mismatch between the inanimate NP and the verb at this point in the sentence (!*The evidence examined...*). Despite this clear indexation of semantic information, however, there were robust garden-path effects for both reduced/inanimate and reduced/animate sentences. These effects were found exclusively at the first word of the disambiguating *by*-phrase (*by*) and were particularly large for reduced/animate sentences. Taken together, these results indicate that in an online reading task that appears to force incremental syntactic and semantic processing, semantic constraints cannot override syntactic parsing biases. Rather, semantic information appears only to facilitate reanalysis when the input is inconsistent with these biases. This study also indicates that the maze task -- and the A-maze task in particular -- provides a useful method for investigating core theoretical questions in sentence processing.

Mean response times (in milliseconds) by condition and region, with standard errors of the mean for repeated measures in parentheses.

	examined	by	the	lawyer	turned
<i>The defendant</i>					
reduced/animate	1175 (19)	917 (30)	608 (9)	981 (16)	886 (21)
unreduced/animate	1334 (27)	680 (13)	598 (6)	926 (17)	834 (19)
<i>The evidence</i>					
reduced/inanimate	1411 (34)	727 (15)	581 (8)	899 (19)	847 (23)
unreduced/inanimate	1141 (27)	638 (18)	578 (8)	932 (14)	853 (23)
Animacy	--	***	*	--	--
RC type	--	***	--	--	--
Animacy* RC type	***	**	--	--	--
animate reduced vs. unreduced	**	***	--	--	--
inanimate reduced vs. unreduced	***	***	--	--	--

*** $p < .001$, ** $p < .01$, * $p < .05$, -- not significant



Mean response time (in milliseconds) by region and condition.

References

- Boyce, V., Futrell, R., & Levy, R. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111.
- Clifton Jr., C., & Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revised. *Journal of Memory and Language*, 49, 317-334.
- Forster, K. I., Guerrero, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41, 163-171.
- Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285-318.

Is the relationship between word probability and processing difficulty linear or logarithmic?

James Michaelov, Megan Bardolph, Seana Coulson, Benjamin Bergen (UC San Diego)

j1michae@ucsd.edu

Research has found that the surprisal of a word—the negative logarithm of the probability of the word being the next word in an utterance—predicts multiple behavioral metrics of processing difficulty such as reading time [8, 10, 3, 6, 12, 17], as well as the N400, a neural index of semantic processing difficulty [7, 5, 1, 11]. However, a recent high-powered study by Brothers & Kuperberg [4] finds a linear relationship between word predictability and two behavioral metrics of processing difficulty (self-paced reading time and picture naming latency), and finds additional evidence of the same relationship in previous eye-tracking literature [13, 14, 15, 16].

One possible explanation is based on the different measures of predictability used between studies. Most of the surprisal studies use corpus-based metrics of word predictability such as trigram probability [17], or the probability estimated by recurrent neural network language models (RNN-LMs) trained on text corpora [7, 1, 11]. By contrast, Brothers & Kuperberg [4] use cloze probability, the probability that participants in a norming study will fill a specific gap in a sentence with a specific word. Thus, these two metrics represent different kinds of predictability—we should not necessarily expect the relationship between them and processing difficulty to be the same.

If part of the language system consists of a neurocognitive implementation of an RNN-LM-like system, that is, a system that predicts the next word in a sequence based on long-term knowledge of the statistics of language and the preceding context, there is no need in principle for its output to be directly proportional to the raw probability output of an RNN-LM. If this output is proportional to the negative log-transformed probability of an RNN-LM, as is supported by the aforementioned behavioural and neural evidence, then we should expect downstream tasks such as the cloze task to use these outputs. Thus, we would expect cloze probability to have a linear relationship with processing difficulty metrics, as was found by Brothers & Kuperberg [4]; and we would expect raw RNN probability to have a logarithmic relationship with both processing difficulty and cloze. In the present study, we investigate whether this is the case with N400 amplitude as our operationalization of processing difficulty.

To test this, we first investigate how well cloze and RNN-LM predicted probability and their log-transformed counterparts (i.e., their surprisals) predict N400 amplitude. To do this, we use a subset of the stimuli and EEG recordings from a previous ERP study [2]. We used the cloze probabilities for the sentence completions collected in the original study, and used a pretrained RNN-LM [9] to calculate corpus-based probability. Each of these was also log-transformed to get surprisal values. We used linear-mixed effects models to predict the by-trial, by-electrode mean amplitudes over the 300-500ms period after stimulus presentation (the canonical N400 time period). As can be seen in Figure 1A, we see that the models using raw cloze probabilities fit N400 amplitude better than those with the log-transformed probabilities, but that the reverse is true for the RNN-LM-derived probabilities. This shows that our initial hypothesis that the two probabilities may differ in this way is supported by the evidence.

To further investigate the hypothesis, we test whether RNN-LM probability or surprisal best predicts cloze probability by using linear mixed-effects models to predict cloze probability based on these two metrics. As can be seen in Figure 1B, we find that RNN-LM surprisal better fits cloze probability than raw RNN-LM probability. Thus, cloze probability more closely reflects corpus-derived probabilities that have been log-transformed than those that have not.

Therefore, if a neurocognitive system that predicts upcoming words based only on the surface-level statistics of previous linguistic input is either used in the cloze task or underlies the N400 response, we provide evidence that its output is closer to RNN-LM surprisal than raw probability.

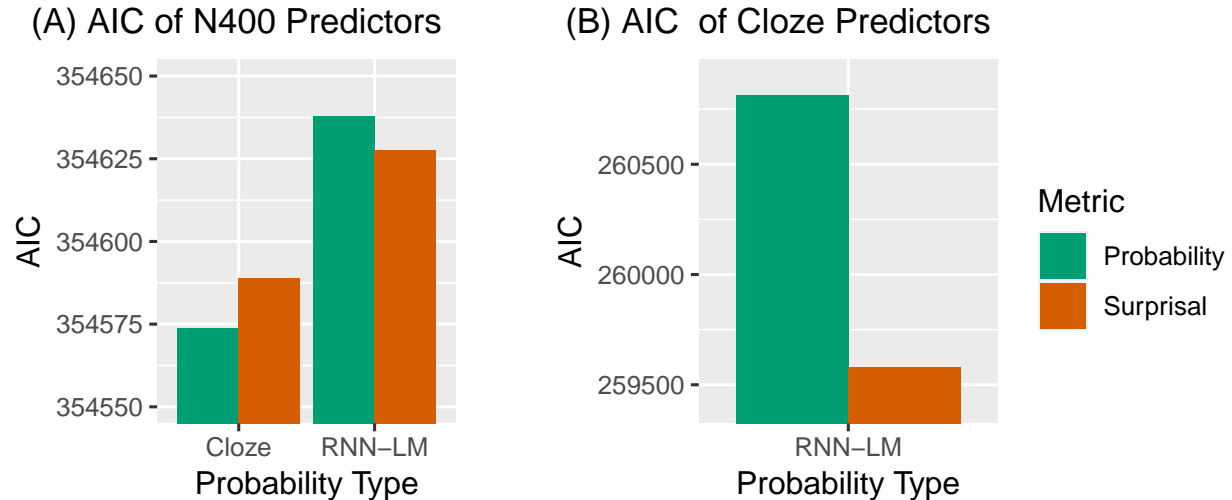


Figure 1: (A) AIC of linear mixed-effects models predicting N400 amplitude by predictor. (B) AIC of linear mixed-effects models predicting Cloze probability by predictor.

References

- [1] Aurnhammer C., Frank S. L., 2019, *Neuropsychologia*, 134, 107198
- [2] Bardolph M., Van Petten C., Coulson S., 2018, in *Twelfth Annual Meeting of the Society for the Neurobiology of Language*. Quebec City, Canada
- [3] Boston M. F., Hale J., Kliegl R., Patil U., Vasishth S., 2008, *Journal of Eye Movement Research*, 2
- [4] Brothers T., Kuperberg G. R., 2021, *Journal of Memory and Language*, 116, 104174
- [5] Delaney-Busch N., Morgan E., Lau E., Kuperberg G., 2017, in *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. London, UK, p. 6
- [6] Demberg V., Keller F., 2008, *Cognition*, 109, 193
- [7] Frank S. L., Otten L. J., Galli G., Vigliocco G., 2015, *Brain and Language*, 140, 1
- [8] Hale J., 2001, in *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01*. Association for Computational Linguistics, Pittsburgh, Pennsylvania, pp 1–8, doi:10.3115/1073336.1073357
- [9] Jozefowicz R., Vinyals O., Schuster M., Shazeer N., Wu Y., 2016, arXiv:1602.02410 [cs]
- [10] Levy R., 2008, *Cognition*, 106, 1126
- [11] Michaelov J. A., Bergen B. K., 2020, in *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics
- [12] Monsalve I. F., Frank S. L., Vigliocco G., 2012, in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. pp 398–408
- [13] Rayner K., Well A. D., 1996, *Psychonomic Bulletin & Review*, 3, 504
- [14] Rayner K., Li X., Juhasz B. J., Yan G., 2005, *Psychonomic Bulletin & Review*, 12, 1089
- [15] Rayner K., Reichle E. D., Stroud M. J., Williams C. C., Pollatsek A., 2006, *Psychology and Aging*, 21, 448
- [16] Sereno S. C., Hand C. J., Shahid A., Yao B., O'Donnell P. J., 2018, *Quarterly Journal of Experimental Psychology*, 71, 302
- [17] Smith N. J., Levy R., 2013, *Cognition*, 128, 302

Bridging language acquisition and processing via the integrated systems hypothesis: Evidence from self-paced reading of newly-learned words within sentential contexts

Laura M. Morett (University of Alabama), Sarah S. Hughes-Berheim (University of Alabama), John F. Shelley-Tremblay (University of South Alabama)

Introduction. The integrated systems hypothesis posits that gesture and speech mutually and obligatorily interact, affecting language processing [1-2]. Despite the absence of gesture at recall, the presence of semantically-congruent (matching) iconic gesture during word learning enhances subsequent memory for newly-learned words, whereas the presence of semantically-incongruent (mismatching) iconic gesture during word learning hinders subsequent memory for newly-learned words [3], indicating that this hypothesis extends to language representation. At present, however, it is unclear whether semantic congruency of gesture and definitions presented during word learning (e.g., *kroosk*—*to sweep* [definition], *to drink* [gesture]) affects subsequent processing of newly-learned words within sentential contexts (e.g., He took the cup to *kroosk*). Moreover, it is unclear how gesture is integrated with orthography, which is processed sequentially with gesture in the visual modality. This work fills these lacunae.

Methods. Via a succession of interleaved word learning and self-paced reading (SPR) blocks, native English speakers ($n=32$) learned 96 pseudowords (English phonotactics; controlled for phonological neighborhood density) in sets of 4 and then read corresponding sets of 4 English critical sentences, each ending with a pseudoword, as quickly as possible. In each word learning trial, participants were presented with a pseudoword as text, then a video of an iconic gesture either matching or mismatching the pseudoword's definition (verified via ratings from a separate sample), and then the pseudoword's English definition, which they were instructed to remember, as text (presented sequentially with gesture to avoid splitting visual attention between gestures and text; Fig. 1A). In each SPR trial, participants pushed a button to read the context sentence wholesale and the critical sentence word-by-word, ending with the pseudoword (sentence pairs normed to elicit English definitions and gesture meanings; Fig. 1B). Both the semantic congruency of gestures that pseudowords were learned with as well as the semantic congruency of definitions that pseudowords were learned with were manipulated relative to critical sentences to examine how they affect pseudoword processing within this context. Word learning and SPR trials were counterbalanced in congruency and order within their respective blocks.

Results. Prior to analysis, pseudoword SPR latencies ≥ 3 SD beyond cell means (15.54%) were trimmed. Remaining latencies were modeled via linear mixed-effect regression using the maximal random effect structures justified. These analyses revealed that gesture-definition match at learning did not affect pseudoword processing within critical sentences. However, definition-sentence and (to a lesser extent) gesture-sentence semantic congruency ($B=5.33$, $t=2.03$, $p=.048$; $B=-5.43$, $t=-1.92$, $p=.06$) affected pseudoword processing within critical sentences, although these effects were non-interactive. Tukey-corrected pairwise comparisons revealed that SPR latencies were higher for pseudowords with definitions incongruent and gestures congruent with critical sentences than for pseudowords with definitions congruent and gestures incongruent with critical sentences ($p=.02$; see Fig. 2). No other comparisons reached significance.

Discussion. The results demonstrate that pseudowords learned with mismatching iconic gestures are processed less efficiently within sentential contexts with which their definitions are semantically-incongruent and their gestures are semantically-congruent than vice versa for sentential contexts. Moreover, the results demonstrate that pseudowords learned with matching iconic gestures are processed with similar efficiency regardless of whether their definitions and gestures are semantically-congruent or -incongruent with sentential contexts. Both of these findings contradict the prediction of the integrated systems hypothesis that semantic (in)congruency of gestures and definitions should affect processing of newly-learned words similarly, suggesting that it may fail to bridge language acquisition and processing within sentential contexts.

References. [1] Kelly, Ozyurek, & Maris (2010). *Psych. Sci.* Kelly, Creigh, & Bartolotti (2010). *J Cog. Neuro.* [3] Kelly, McDevitt, & Esch (2009). *Lang. & Cog. Processes.*

Figure 1. (A) Word learning trial with iconic gesture mismatching definition of pseudoword. Pseudoword and definition duration: 2000 ms; gesture duration: ~2000 ms; ISI duration: 1000 ms. (B) SPR trial featuring pseudoword learned with semantically-incongruent definition and semantically-congruent gesture. All stimuli presented until button pressed to proceed.

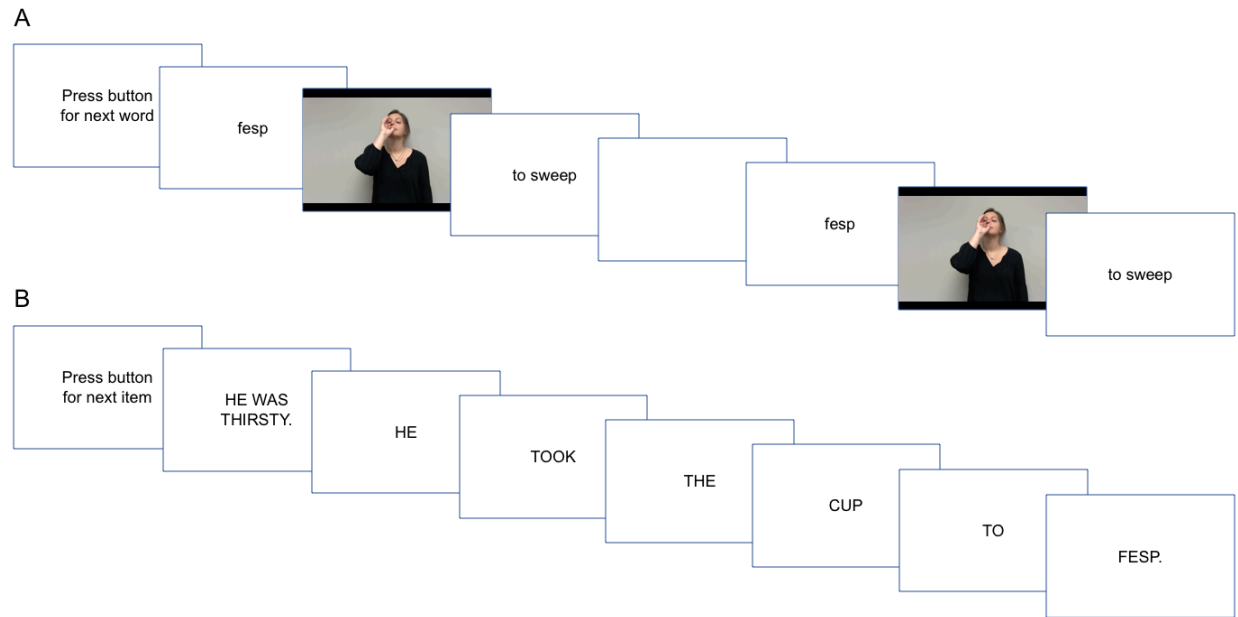
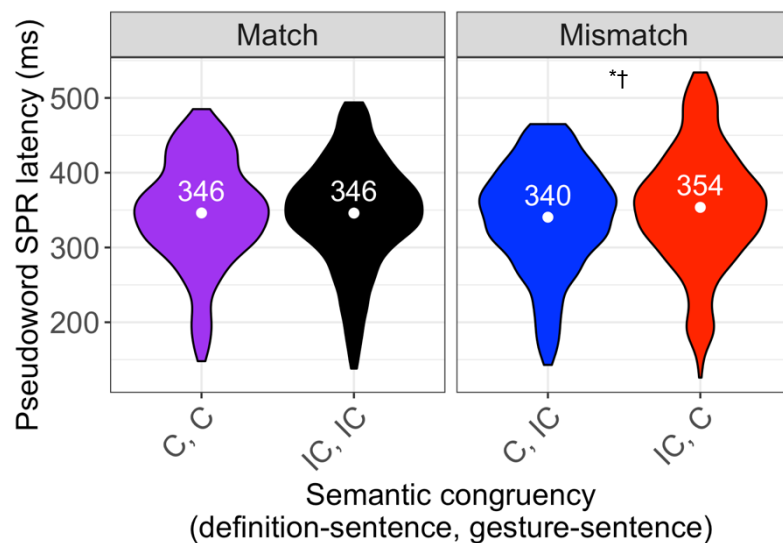


Figure 2. Pseudoword SPR latency by gesture-definition match at learning and definition-sentence and gesture-sentence semantic congruency. White dots and values represent cell means. ^{*} $p < .05$, definition-sentence congruency; [†] $p < .10$, gesture-sentence congruency



Lexical access in sentence reading is mediated by domain-general processing speed

Naomi Sellers, Shannon McKnight, Phillip Gilley, Albert Kim (University of Colorado Boulder)

Understanding the operations of word recognition and their timecourse is crucial for sentence processing theory because the rapid syntactic and semantic commitments during sentence processing are rooted in word recognition [1-2]. Psycholinguistic theories have made significant progress characterizing the timecourse of major perceptual and cognitive operations that yield word recognition in the modal brain [3-4]. A major missing piece in standard models, however, is an account of individual variation in the timecourse of word recognition. We investigated the timecourse of word recognition in individuals by examining speed of lexical access, which we indexed as the latency of the earliest effects of lexical frequency on brain activity in individual experimental participants during sentence-embedded word recognition. **We asked whether and how this stage of word recognition during sentence processing is shaped by individual differences in language knowledge and general cognitive ability.**

We examined the timing of lexical frequency effects on EEG activity collected from 205 participants who read plausible, grammatical sentences in RSVP format (120 sentences, ~600 target content words per participant). We used lexical frequency effects to index lexical access on the assumption that effects of lexical frequency on brain activity must reflect the evaluation of a visual stimulus at the level of lexical representation, as opposed to a lower, sublexical form-based or morphological representation [1, 5]. In a separate lab visit, as part of a larger investigation of individual variation during sentence processing, we also measured individual cognitive abilities, including breadth of language experience (print exposure and vocabulary measures) and perceptual speed (timed button press responses to visual stimuli).

As a baseline measure of lexical frequency effects, we evaluated whether the group-averaged left-occipital N170 ERP peak amplitude between 150-220ms post-word onset was predicted by lexical frequency or length. N170 peak amplitude was positively related to lexical frequency (Fig1; $\beta = 0.040$ $p = 0.015$), suggesting that lexical access has occurred by this latency in general.

We hypothesized that individuals with a wider experience of the English language would exhibit faster lexical access, controlling for processing speed. To test this hypothesis, we measured the amplitude of left-occipital EEG elicited in each 10 ms window between 100 and 260ms. For each participant, at each 10ms measurement, we tested whether amplitude was predicted by lexical frequency while controlling for word length. **We then identified the earliest lexical frequency effects in each participant. Multiple regression modeling tested whether individual lexical frequency effect latencies were predicted by three factors: language experience, verbal speed, and nonverbal speed.**

The latency of participants' earliest lexical frequency effects was not predicted by individual language experience levels ($p > 0.1$), providing no support for our hypothesis that greater language experience leads to faster lexical access. Instead we find the onset of lexical frequency's effect on amplitude was significantly predicted by nonverbal speed ($\beta = -4.932$, $p = 0.025$) and moderately predicted by verbal speed ($\beta = 3.938$, $p = 0.068$).

Although there is growing evidence that individual differences in language experience affects multiple aspects of sentence processing [6-8] language experience did not influence the speed of lexical access here. **Overall, we find that measures of perceptual speed predict the onset of lexical access during sentence processing. This suggests that faster minds access lexical information faster.** Our ongoing work examines whether other measures of lexical access might reveal effects of language experience during downstream sentence processes.

REFS: [1] MacDonald, Perlmuter & Seidenberg (1994) *Psych Rev.* [2] Trueswell & Tannenhaus (1994) *Perspectives on sentence processing* [3] Grainger & Holcomb (2009), *LangLingCompass* [4] Reichle, Rayner & Pollastek, (2003) *Behavioral & Brain Science.* [5] Hauk & Pulvermüller (2004) *Clinical Neurophys.* [6] McKnight, Miyake, Bell-Souder & Kim (2018) 31st *CUNY Conference* [7] Pakulak & Neville (2010) *JCoN* [8] Freed, Hamilton, Long (2017) *JML*

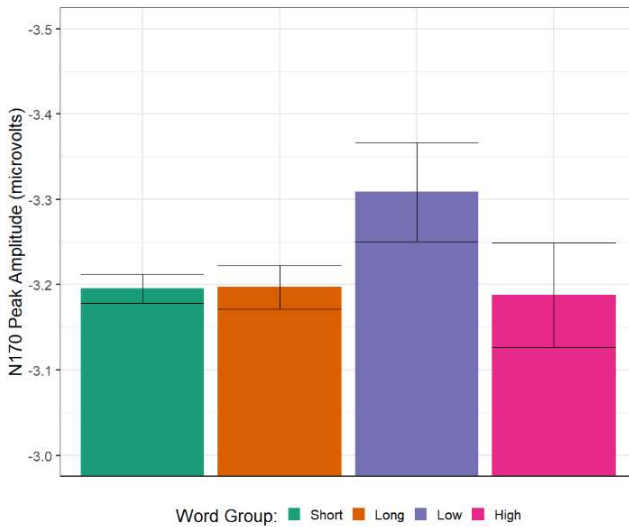


Figure 1: N170 Peak Amplitude is mediated by word frequency. Peak amplitude as a function of word length (Short/Long) and lexical frequency (Low/High). Low frequency words elicit significantly larger N170 peaks than high frequency words. Length does not appear to significantly affect the N170 amplitude. This suggests that, for most people, the N170 reflects a stage of word recognition where lexical representations are available.

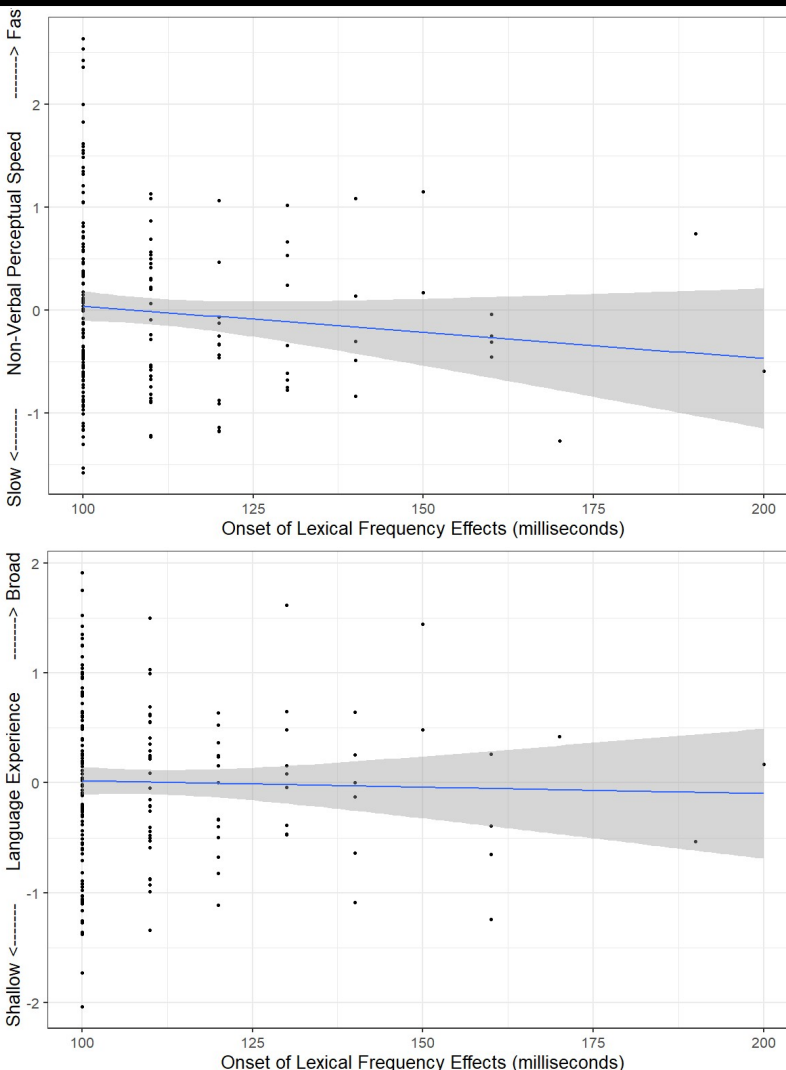


Figure 2: Frequency sensitivity is partially mediated by general processing speed.

Non-verbal speed was quantified by measuring reaction times during simple speeded decision tasks (eg. “is the arrow pointing left or right?”). Language experience was quantified as a composite measure of performance on a vocabulary test, author-magazine recognition task, spelling-from-audio task, and an irregular word reading task. We find that the onset of frequency effects are significantly predicted by speed and not experience.

Perceptual connectivity influences toddlers' attention to known objects and subsequent label processing

Ryan E Peters (UT Austin), Justin B Kueser, & Arielle Borovsky (Purdue University)
ryan.peters@austin.utexas.edu

Does a toddler's knowledge of and attention to subcomponents of word meanings impact lexico-semantic processing? Adult research suggests the answer is "yes". Different aspects of word meanings influence a range of adult psycholinguistic processes, including categorization, word/concept learning, and semantic priming. Meanwhile, recent work exploring effects of portions of word meanings on toddlers' normative vocabulary development have revealed not all types of meaning are equally important: perceptual features matter most (Engelthaler & Hills, 2017; Peters & Borovsky, 2019). However, less is known about whether and how perceptual aspects of word meanings influence lexical processing in early language development.

One possibility is increased perceptual connectivity (i.e., having more words in a child's lexicon that share perceptual features with the item) facilitates processing for familiar words. This option is supported by evidence that 2-year-old toddlers' processing is facilitated for words that are members of categories they are more knowledgeable of (Borovsky et al., 2016), and likely to share many perceptual features with (Hills et al., 2009). A second possibility is that increased perceptual connectivity facilitates attention to objects pre-labeling, with cascading effects on subsequent label processing. This option is supported by evidence linking children's speed of visual object processing and sensitivity to holistic shape to word knowledge (Pereira & Smith, 2009; Smith, 2003) and experience with same category items (Quinn, 2004).

Thus, in the current study, we explored whether and how building a lexicon with perceptual connectivity supports either *pre-labeling* attention to and/or *post-labeling* recognition of word meanings. We explored this question in 24-30-month-olds (N=60) in relation to other individual differences, including age, vocabulary size, and temperamental tendencies to maintain focused attention. Participants' looking to item pairs with high vs low perceptual connectivity (Figure 1A) was measured before and after target item labeling (Figure 1B).

Results from a permutation cluster analysis revealed pre-labeling attention to novel items is biased to both high and low connectivity items: first to high, and second, but more robustly to low connectivity items (Figure 1C). Exploratory analyses of first looks showed the initial bias towards high-connectivity items mainly resulted from a greater likelihood for first looks to land on high connectivity items according to an exact binomial test (*probability* = 0.47, *95% confidence interval* = [0.44, 0.5], *p* = .029), while the later bias towards low connectivity items was driven by longer durations for first looks that landed on low-connectivity items according to a linear mixed effects model (*coefficient* = -0.123, *95%CI*=[-0.238, -0.008], *p*=.035).

Subsequent object-label processing was also marginally facilitated for high connectivity items (Figure 1D), and connectivity significantly interacted with temperamental tendency to maintain focused attention according to a linear mixed effects model (*coefficient*=0.098, *95%CI*=[0.02, 0.176], *p*=.013), even while considering significant cascading effects of pre-labeling attentional biases (*coefficient*=0.08, *95%CI*=[0.04, 0.122], *p*<.001). This result suggests that a tendency for maintaining focused attention during learning opportunities may provide crucial support for the recognition of shared perceptual features between objects.

This work provides the *first empirical evidence* that patterns of shared perceptual features within children's known vocabularies influences both visual and lexical processing, highlighting the potential for a newfound set of developmental dependencies based on the perceptual/sensory structure of early vocabularies.

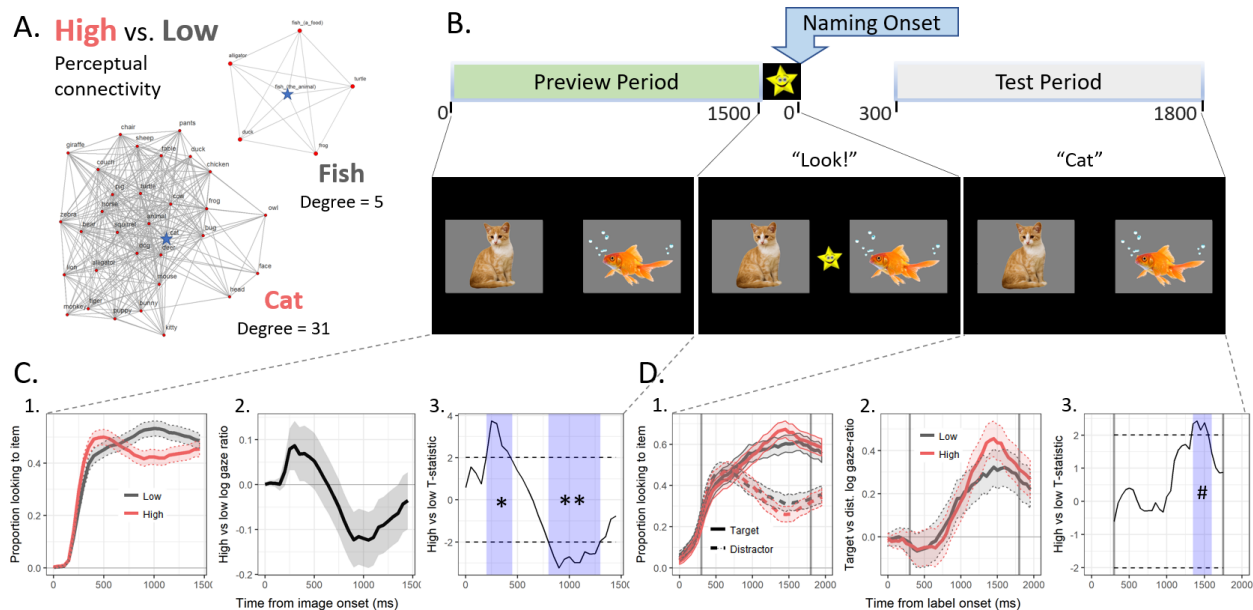


Figure 1. (A) Networks showing words connected to CAT and FISH via shared perceptual features according to a recently developed extension of the McRae et al. (2005) feature production norms, (B) an example of visual stimuli in an experimental trial, and timecourse plots—for both the (C) pre-labeling Preview Period and (D) post-labeling Test Period—of the 1. proportion of fixations, 2. log gaze-ratios, and 3. pointwise comparisons with periods of consecutive significant differences identified by cluster analyses (light blue).

** $p < .01$. * $p < .05$. # $p < .1$.

References

- Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Development*, 87(6), 1893-1908.
- Engelthaler, T., & Hills, T. T. (2016). Feature biases in early word learning: network distinctiveness predicts age of acquisition. *Cognitive science*.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009b). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, 112(3), 381-396.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4), 547-559.
- Pereira, A. F., & Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental science*, 12(1), 67-80.
- Peters, R., & Borovsky, A. (2019). Modeling early lexico-semantic network development: Perceptual features matter most. *Journal of Experimental Psychology: General*, 148(4), 763.
- Quinn, P. C. (2004). Is the asymmetry in young infants' categorization of humans versus nonhuman animals based on head, body, or global gestalt information?. *Psychonomic Bulletin & Review*, 11(1), 92-97.
- Smith, L. B. (2003). Learning to recognize objects. *Psychological Science*, 14(3), 244-250.

Virtual-World eye-tracking: The efficacy of replicating word processing effects remotely

Zoe Ovans, Jared Novick & Yi Ting Huang (University of Maryland)

Psycholinguistic research has generated detailed models of moment-to-moment language processing, and has increasingly turned to virtual methods that recruit diverse participants, yield large sample sizes, and remain pandemic-proof. However, there is substantial uncertainty about the feasibility and sensitivity of measurements from remote settings. While methods such as self-paced reading and acceptability judgements replicate well online [1], it is unknown whether fine-grained effects (e.g., within word recognition) will be observable. Recent attempts using visual-world eye-tracking have relied on automatic gaze-detection (e.g., [2,3]), but this requires calibration and can have limited accuracy. To validate the efficacy of remote eye-tracking for word processing, the present study employed a novel webcam paradigm (via *PCibex* [4]) to semi-replicate Experiment 1 in Allopenna et al., 1998 [5]. This landmark study (cited over 1600 times) revealed listeners' incremental activation of phonemic competitors during spoken-word recognition. It is an ideal candidate for validating remote testing, since real-time fixations track the extent to which subtle acoustic changes incrementally alter predictions of word identity. It has been replicated in laboratory settings (e.g., [6]), but, to our knowledge, not remotely.

Compared to eye-tracking in the lab, webcam eye-tracking introduces additional variability, including participants' screen size, camera quality, internet bandwidth, and environmental distractions. It was our aim to determine whether these factors limit sensitivity to the time-course of word recognition. While data collection is ongoing, 34 participants have been collected from Amazon Mechanical Turk and the university study pool. Some participants had hardware difficulties or did not yield suitable data, but our overall data-retention rate was 79%. We showed listeners an image of a spoken target (e.g., "beaker"), phonological cohort competitors (e.g., "beetle"), rhyme competitors (e.g., "speaker"), and unrelated distractors (e.g., "carriage"). If incremental word recognition is observable in this format, we expect to see looks to the target and cohort-competitor images after word onset, and to a lesser extent, to the rhyme-competitor after word offset. To increase the feasibility of virtual testing, we included only partial-set trials (e.g., with two unrelated objects, target and cohort-competitor) in a Latin square design, reducing the trial number from 96 to 18. This ensured that cohort and rhyme competitor looks were independent, encouraged participants to stay engaged, and reduced video upload time. Looks were recorded through participants' webcams and hand-coded frame-by-frame [7].

As Fig.1 shows, looks to the target increased following disambiguation, confirming that participants successfully link the audio to our visual displays. Looks to the target object began 400ms after word onset, about 200ms slower than lab-based eye-tracking [5]. To examine the extent of competition, we averaged fixations in a 1000ms time-window after word onset, and compared competitor fixations to unrelated controls. As predicted, participants looked to cohort and, to a lesser extent, rhyme competitors after target word offset (Fig. 1). Mixed-effects models reveal more looks to cohort than rhyme and unrelated competitors ($p < .01$), though looks to rhymes did not differ from unrelated items ($p = .29$). Next, we calculated the relative target and competitor frequency and included this as a fixed effect (Fig. 2). Consistent with [8], we found an interaction between frequency and rhyme looks. Participants looked to rhymes more than unrelated controls when rhymes were more frequent than targets ($p = .02$). Together, this shows that incremental word processing and subtle frequency effects are observable in virtual testing. We conclude that webcam eye-tracking produces similar results to in-lab testing, but eye-movements are slower, and subtle effects like rhyme competition may be harder to detect. Even so, the presence of cohort competition and frequency modulation provides evidence for this method's sensitivity to incremental processing, and provides validation for a new, virtual avenue for visual-world sentence processing research for closely time-locked effects.

Figure 1: Proportion of looks to items surrounding target word onset

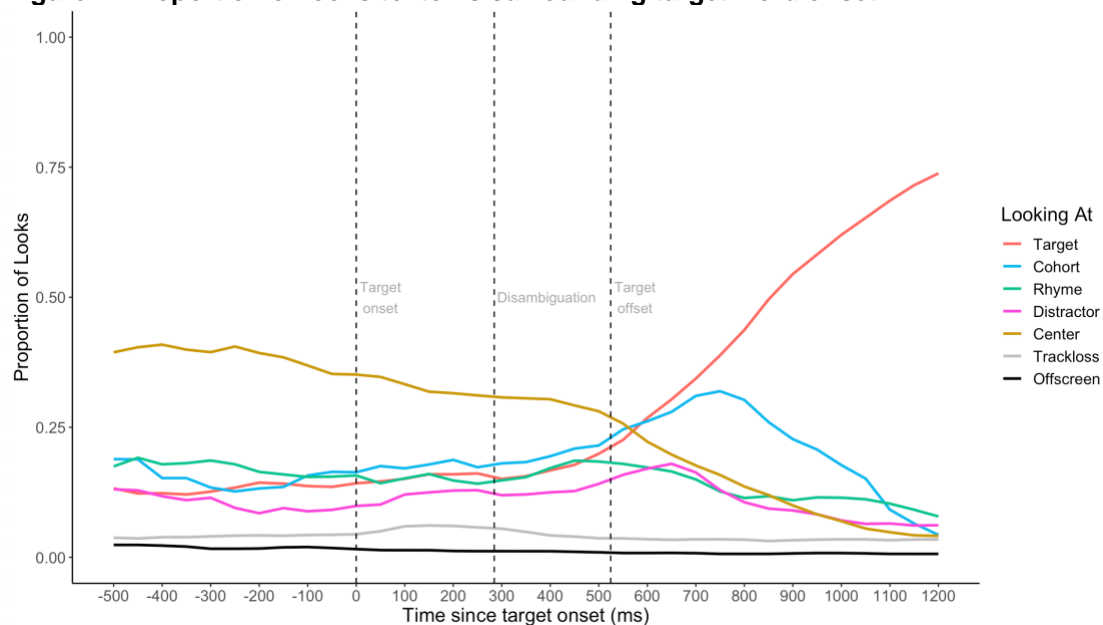
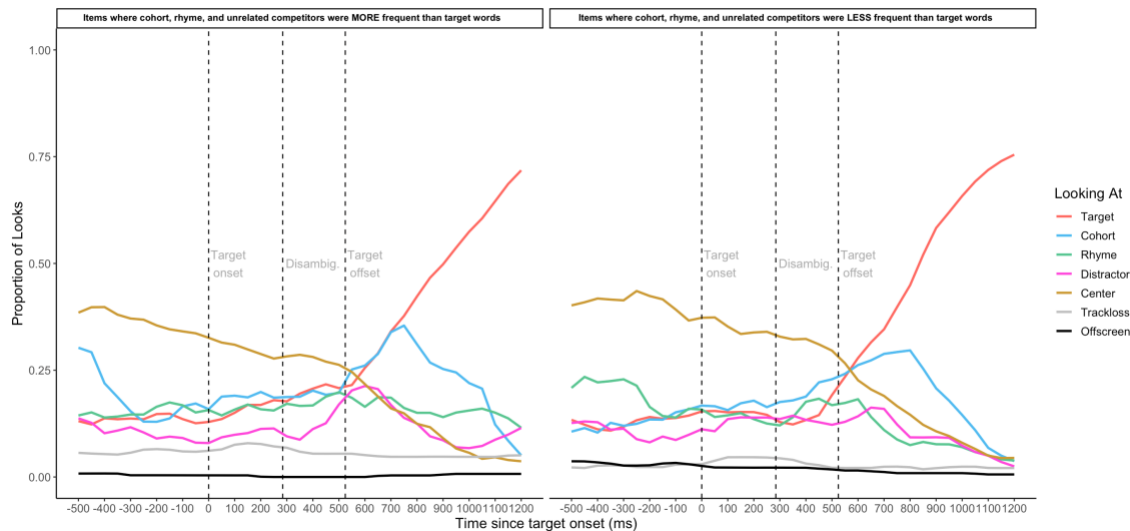


Figure 2a: Items where cohort, rhyme, and unrelated competitors were MORE frequent words than target words

Figure 2b: Items where cohort, rhyme, and unrelated competitors were LESS frequent words than target words



References:

- [1] Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1), 155-167.
- [2] Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.
- [3] Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, 50(2), 451-465.
- [4] Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX).
- [5] Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, 38(4), 419-439.
- [6] Farris-Trimble, A., & McMurray, B. (2013). Test-retest reliability of eye tracking in the Visual World Paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research*.
- [7] Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive psychology*, 49(3), 238-299.
- [8] Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive psychology*, 42(4), 317-367.

Complex syntax and conversational turn-taking during toddler-adult picture book reading

Anastasia Stoops, Jane Hwang, Mengqian Wu, Jessica Montag (University of Illinois at U-C)
A characteristic of skilled adult language use is the ability to produce and comprehend complex utterances. Increasing evidence suggests that experience with these complex structures contributes ease of processing (Real & Christiansen, 2007; Dabrowska, 2012), and such experience may disproportionately come from written language (Roland, Dick & Elman, 2007; Montag & MacDonald, 2015). What is the role of spoken **and written** language input at the earliest stages of language development?

Talk generated during picture book reading differs from typical child-directed speech in a number of ways. Book reading elicits rich caregiver-child conversation (Muhinyi, et al., 2020; Whitehurst et al., 1988) and more parent speech and conversational turns than free-play (Gilkerson et al., 2017; Sosa, 2016). While extra-textual talk is studied extensively, the book text has received less attention. Picture book text is more lexically diverse (Montag et al., 2015) and syntactically complex (Cameron-Faulkner et al., 2013; Montag, 2019) than typical child-directed speech. A key mechanism by which book reading may affect language outcomes is by exposing children to complex language, including complex syntax, that might otherwise be rare. To hypothesize plausible pathways for the *causal* book reading contribution to language outcomes, we need a clearer description of the talk generated during picture book reading, including the degree to which the complex syntax in book text becomes part of the language environment.

To observe consistency/variation across families and the effect of various book features on the generated talk, we provided families with 4 novel picture books that varied in length and syntactic complexity. Families recorded themselves reading the books at home as they normally would. **First**, we examine how much of the complex syntax in picture books caregivers say. **Second**, we examine how book length and syntactic complexity affects the book reading talk.

Method

12 families, children aged 24-37 months (7 girls) recorded 6-12 reading sessions (total=58). Book length and number/type of complex constructions are shown in Table 1. Target syntactic constructions are defined in Table 2. A team of research assistants transcribed adult and child speech and marked utterance boundaries using the ELAN software. The corpus will be available to other researchers upon publication of this work.

Results

Overall, families spent more time reading the longer books, but there was considerable variability between families (Figure 3). Both features of the books and family individual differences contribute to overall reading times, but time spent reading a book can vary wildly.

Crucially, we find that the complex sentence constructions in the books were indeed produced by caregivers (Figure 1). Out of 181 target constructions approximately 82% (149) were read from the book without any modification (described in Table 3). Most modifications were additions before or after the target construction, so the complex construction was produced intact 97% of the time. Picture books may be an important source of complex syntax for children because adult caregivers indeed read the complex language in the book text aloud.

Finally, the turn-taking counts were the highest for short and simple book with little variability among the others (Figure 2). The longer books and the books with the most complex syntax were *not* the books that promoted the most parent-child conversation.

Discussion and Conclusion

We demonstrate that picture books may be an important source of complex syntax for young children. As we see for adults, written and spoken language, even in early childhood, may provide different types of language input. However, we find that a different kind of rich language, caregiver-child turn-taking was more frequent in the shorter, syntactically simple books. This dissociation suggests that interventions that aim to identify the “best” books may be misguided. Books of different lengths or books with more or less complex syntax may provide *different* linguistic input for children, all of which may be important for language development.

Table 1. Book classification summary

Book Title N=58 (≈9 hours (Reading session count)	Book Length-Syntactic Complexity (Word count)	Counts of Syntactic Construction (SC) Types				SC counts per book
		SRC	ORC	Oblique	Passive	
That is not a good idea (21)	Short-Simple SS (125)	0	0	0	0	0
When dinosaurs came with everything (17)	Medium-Simple MS (1018)	0	1	1	0	2
Stellaluna (11)	Long-Simple LS (1211)	2	1	0	0	3
Oh, the places you'll go! (9)	Medium-Complex MC (939)	5	4	4	2	15

Table 2. SC summary

Syntactic Construction	Example	Count (N=181)
Subject Relative Clause (SRC)	More bats gathered around to see the strange young bat who behaved like a bird ("Stellaluna")	45
Object Relative Clause (ORC)	The next thing I knew , she had him cleaning the gutters ("When dinosaurs came with everything")	64
Oblique Relative Clause (Oblique)	The places you'll go! ("Oh the places you'll go!")	54
Passive Main Clause (Passive)	You'll be left in a Lurch ("Oh the places you'll go!")	18

Table 3. SC modification summary

SC modification type	Example	%(N) 100%=32
Addition/omission before SC	And you may not find any you'll want to go down ORC	12.50(4)
Addition after SC	You'll be left in a Lurch Passive Parent: Oh... his poor balloon got caught up in a tree	65.62(21)
Addition within SC	Stellaluna was terribly hungry – but not for the crawly things that Mama Bird brought ORC	9.38(3)
SC repetition	Parent: You can steer yourself any direction you choose ORC You can steer yourself any direction you choose ORC	6.25(2)
SC omission	The places you'll go Oblique	6.25(2)

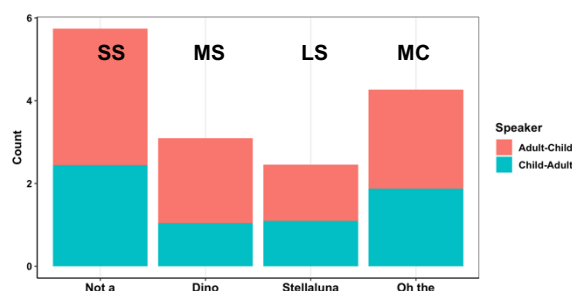


Figure 2. Turn-taking per-minute by book: Adult vs Child initiated

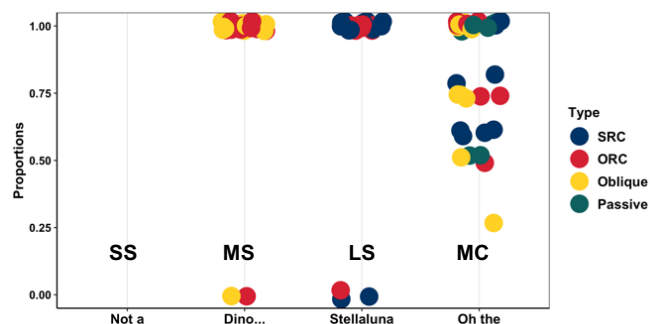


Figure 1. Proportion of SC uttered unchanged by book; 1 dot = SC of interest in one reading session

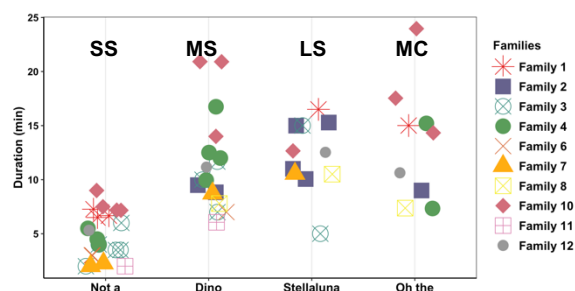


Figure 3. Reading session duration by book and family; 1 point = one reading session

References

- Cameron-Faulkner, T., & Noble, C. (2013). A comparison of book text and child directed speech. *First Language*, 33(3), 268-279.
- Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2, 219-253.
- Gilkerson, J., Richards, J. A., & Topping, K. J. (2017). The impact of book reading in the early years on parent-child language interaction. *Journal of Early Childhood Literacy*, 17, 92-110.
- Muhinyi, A., Hesketh, A., Stewart, A. J., & Rowland, C. F. (2020). Story choice matters for caregiver extra-textual talk during shared reading with preschoolers. *Journal of Child Language*, 47, 633-654.
- Montag, J. L. (2019). Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language*, 39, 527-546.
- Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, 144, 447.
- Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, 26, 1489-1496.
- Real, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *JML*, 57, 1-23.
- Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *JML*, 57, 348-379.
- Sosa, A. V. (2016). Association of the type of toy used during play with the quantity and quality of parent-infant communication. *JAMA Pediatrics*, 170, 132-137.
- Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology*, 24, 552.

Preferences for communicative efficiency in miniature languages are independent of learners' L1s

Lucy Hall Hartley, Masha Fedzechkina (University of Arizona)

Using miniature language learning methodology, researchers have claimed to have uncovered a variety of abstract cognitive biases giving rise to cross-linguistically frequent patterns (also known as language universals) [1–3]. However, since participants in these experiments are typically adults proficient in at least one language (their L1), it raises an important question of whether some aspects of learners' performance observed in miniature languages can be better explained by L1 influences rather than by more general pre-L1 biases. Indeed, recent work has shown that learners' preference for suffixing over prefixing, previously attributed to processing constraints, is better explained by the L1 [4]. Here, we ask whether learners' biases in communicative efficiency can be explained by L1 influences as well. Consider work by [5,6], where English speakers exposed to either fixed or flexible constituent order languages with optional case marking maintained case in their productions when it was informative about grammatical function assignment (flexible order) and dropped case when it was redundant (fixed order). This preference is consistent with a bias to efficiently trade off the effort required to produce case against message uncertainty as claimed by [5,6]. However, using less case in the fixed order language is also consistent with L1 influence: Learners of the fixed order language may have dropped case to bring the language closer to their L1 (English), which has fixed order and no case. We ask whether the preference for communicative efficiency holds across structurally different L1s. Specifically, we ask whether speakers of English (fixed order, no case), German (flexible order, 4 case categories), and Russian (flexible order, 6-7 case categories) restructure miniature language input to use more case where it is informative (suggesting a general bias at work) or show different preferences in using case and constituent order (suggesting an L1 influence).

Method: English, German and Russian L1 speakers (20 per L1/minature language) learned a miniature language in 2 online sessions over 2 consecutive days. Both input languages had optional case marking on the object only (67% present). The languages had either flexible (VSO/VOS 50/50%) or fixed (VSO 100%) constituent order. Participants first learned alien character-name pairings and then learned the grammar by watching videos of transitive actions accompanied by miniature language descriptions. At the end of each session, participants described previously unseen transitive action videos using the miniature language. We assessed the use of constituent order and case in production.

Results: We analyzed learners' VSO and case use using generalized linear mixed effects models (with maximal converging random effects structure). All three L1 groups matched the input proportion of VSO in the fixed and flexible order languages (p 's > 0.11; Fig.1), replicating the behavior of English speakers in [5,6]. There were L1 differences in the overall amount of case used by learners: German speakers used the same amount of case as English speakers ($\hat{\beta}$ =0.46, z =0.94, p =0.34), but Russian speakers used less case than English speakers ($\hat{\beta}$ =-0.19, z =-2.43, p =0.01). Across all L1s, learners of the flexible order language used significantly more case compared to the learners of the fixed order language ($\hat{\beta}$ =1.28, z =6.34, p <0.001), suggesting a preference to use more case when it is informative. L1 did not interact with constituent order flexibility (p 's > 0.34), crucially suggesting that the preference to use more case in the flexible order language did not depend on the L1 (Fig.2). Thus, all three L1 groups restructured the input in the same way: They matched the input constituent order and, following a bias for communicative efficiency, used more case in the flexible order language compared to the fixed order language.

Conclusion: Our findings suggest that learners restructure miniature language input in a communicatively efficient way regardless of how case and constituent order are used in their L1. We add to a growing body of work investigating L1 influences in miniature language learning and show that by collecting crosslinguistic data, we can begin to understand precise circumstances of L1 influence and its interactions with more general universal biases in the paradigm.

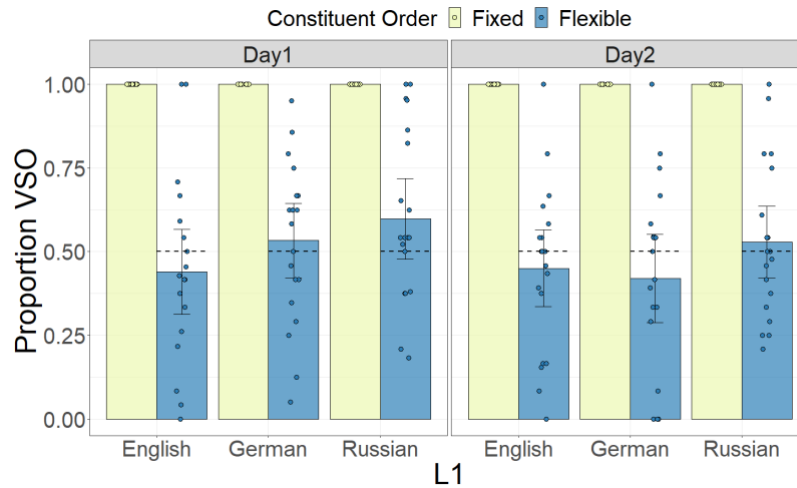


Figure 1: VSO use in production by day of training and L1 background. The dashed line represents the input proportion for the flexible order language (VSO input for the fixed order language is 1.0). Dots are individual participants' means. Error bars are bootstrapped 95% confidence intervals.

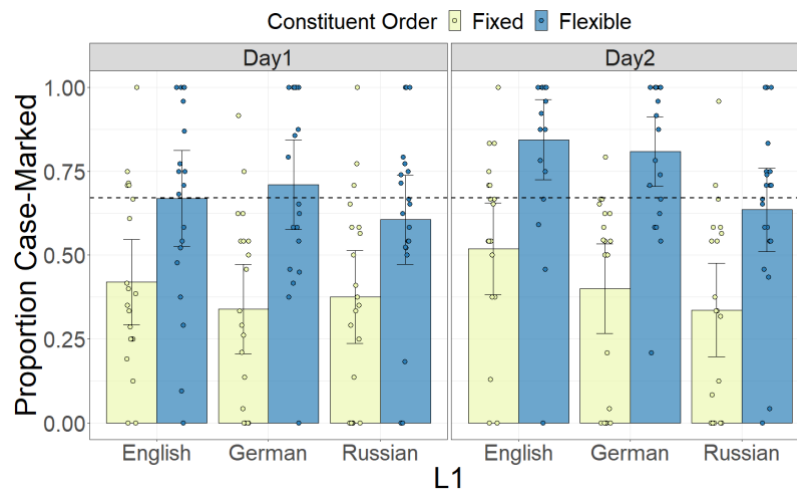


Figure 2: Case marker use in production by day of training and L1 background. The dashed line represents the input proportion (same across fixed and flexible order languages). Dots are individual participants' means. Error bars are bootstrapped 95% confidence intervals.

References

- [1] J. Culbertson, P. Smolensky, and G. Legendre, "Learning biases predict a word order universal," *Cognition*, vol. 122, no. 3, pp. 306–329, 2012.
- [2] M. C. St. Clair, P. Monaghan, and M. Ramscar, "Relationships between language structure and language learning: The suffixing preference and grammatical categorization," *Cogn. Sci.*, vol. 33, no. 7, pp. 1317–1329, 2009.
- [3] J. Kanwal, K. Smith, J. Culbertson, and S. Kirby, "Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication," *Cognition*, vol. 165, pp. 45–52, 2017.
- [4] A. Martin and J. Culbertson, "Revisiting the Suffixing Preference: Native-Language Affixation Patterns Influence Perception of Sequences," *Psychol. Sci.*, 2020.
- [5] M. Fedzechkina, E. L. Newport, and T. F. Jaeger, "Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case Marking," *Cogn. Sci.*, vol. 41, no. 2, pp. 416–446, 2017.
- [6] M. Fedzechkina and T. F. Jaeger, "Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission," *Cognition*, vol. 196, p. 104115, 2020.

The role of L1 and L2 frequency in cross-linguistic structural priming: An artificial language learning study

Merel Muylle (Ghent University), Sarah Bernolet (University of Antwerp), Robert J. Hartsuiker (Ghent University)

Hartsuiker and Bernolet's (2017) developmental account of shared syntactic representations postulates that, during second-language (L2) acquisition, the L2 representations evolve gradually from being item-specific to more abstract, and finally become shared with the native language (L1). Such sharing may be reflected in the emergence of structural priming between two sentences. The account assumes faster development of syntactic representations for frequent vs. infrequent L2 structures. If this is true, there may be earlier and stronger cross-linguistic priming for more frequent L2 structures. In addition, less frequent structures are often found to elicit more priming than more frequent ones (i.e., the so-called inverse frequency effect) and it has been shown that frequency of a structure in one language might affect priming in the other language. Still, it remains unclear how L1 and L2 frequency effects contribute to the acquisition of syntax in early stages of L2 acquisition.

In the current study, we investigated frequency effects at the onset of L2 learning using an artificial language (AL) learning paradigm (Muylle et al., 2020; see Table 1, Figure 1). L1 Dutch speakers ($N = 96$) learned an AL that either had a prepositional-object (PO) dative bias (i.e., PO datives appeared three times as often as double-object datives, or DO datives) or a DO dative bias (i.e., DOs appeared three times as often as POs). Priming was assessed from the AL to Dutch (that has a strong PO bias). We put forward three contrasting hypotheses on how AL frequency modulates the sharing of syntax across languages: 1) the most frequent AL structure is shared before the less frequent one, 2) there is no sharing for either structure early on in the acquisition (and hence no frequency modulation yet), or 3) both structures are shared or at least connected between languages by the end of the training session, and priming effects will be modulated by both AL and L1 frequency effects in an additive way.

We analyzed the results (see Figure 2) using generalized linear mixed effects models with *PO answer* (binomial) as dependent variable and the interaction *Bias* (PO vs. DO) * *Prime Structure* (PO vs. DO vs. baseline) as fixed effects (N of observations = 2913). This analysis showed that there was a main effect of *Prime Structure*, with marginally significant priming for DOs, but not for POs compared to a baseline condition with a transitive or intransitive prime. However, the difference between DO and PO priming was not significant. Importantly, the priming effect was similar across both bias conditions (i.e., no *Bias* * *Prime Structure* interaction), which suggests that L1, but not AL frequency influenced immediate priming (i.e., when the prime is immediately followed by the target). Interestingly, participants in the DO bias group produced significantly more DO targets (10%) in Dutch than participants in the PO bias group, showing that AL frequency exerted cumulative priming effects on L1 productions.

Our findings suggest that both structures are shared, in line with the third hypothesis, but in contrast to our predictions, immediate priming effects seemed to be modulated by L1 frequency only (i.e., the less frequent L1 structure, DO, could be primed more easily from the AL). Importantly, cumulative priming effects indicated that AL frequency did exert an effect on L1 structural choices in general (i.e., the overall proportion of PO vs. DO responses was different for both bias groups). This pattern of results did not provide evidence for or against the hypothesis (based on Hartsuiker & Bernolet's developmental theory) that the representations of frequent L2 structures are shared with L1 before less frequent ones, but can be partly explained in terms of implicit learning accounts of structural priming.

Hartsuiker, R. J., & Bernolet, S. (2017). The development of shared syntax in second language learning. *Bilingualism: Language and Cognition*, 20, 219–234.

Muylle, M., Bernolet, S., & Hartsuiker, R. J. (2020). The role of case marking and word order in cross-linguistic structural priming in late L2 acquisition. *Language Learning*, 70, 194–220.

Table 1. Examples of sentences in the AL & Dutch.

	AL	Dutch
Intransitive	Fuipam jaltsi <i>Cook waves</i>	De kok zwaait <i>The cook is waving</i>
Active	Fuipam zwifsi dettus <i>Cook kisses clown</i>	De kok kust de clown <i>The cook is kissing the clown</i>
Passive	Dettus nast zwifo ka fuipam <i>Clown is kissed by cook</i>	De clown wordt gekust door de kok <i>The clown is being kissed by the cook</i>
DO	Fuipam stiesi dettus sifuul <i>Cook shows clown hat</i>	De kok toont de clown de hoed <i>The cook is showing the clown the hat</i>
PO	Fuipam stiesi sifuul bo dettus <i>Cook shows hat to clown</i>	De kok toont de hoed aan de clown <i>The cook is showing the hat to the clown</i>

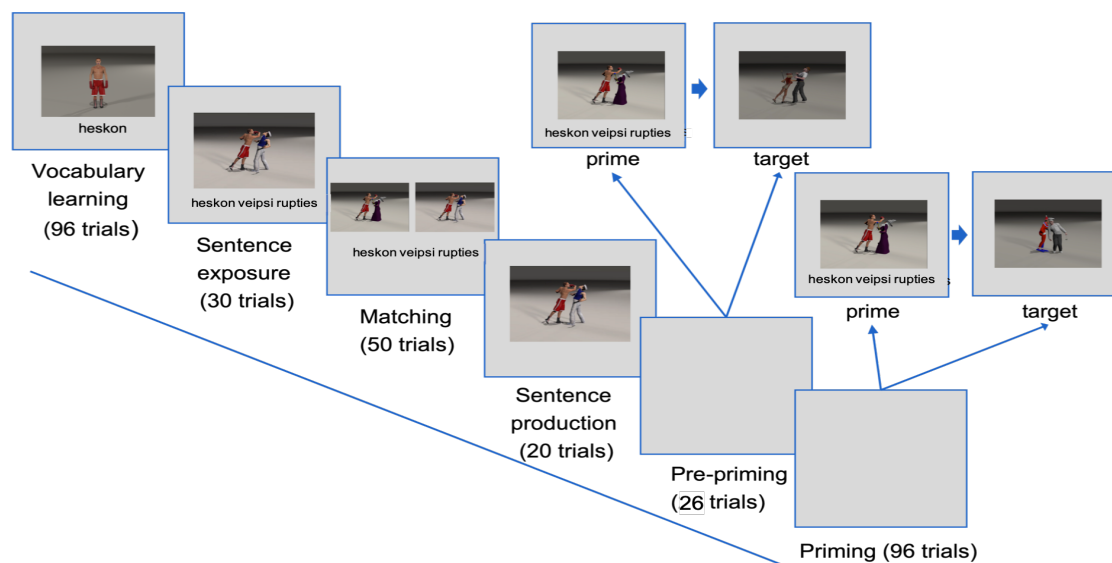


Figure 1. Sequence of the different experimental blocks in the AL learning paradigm.

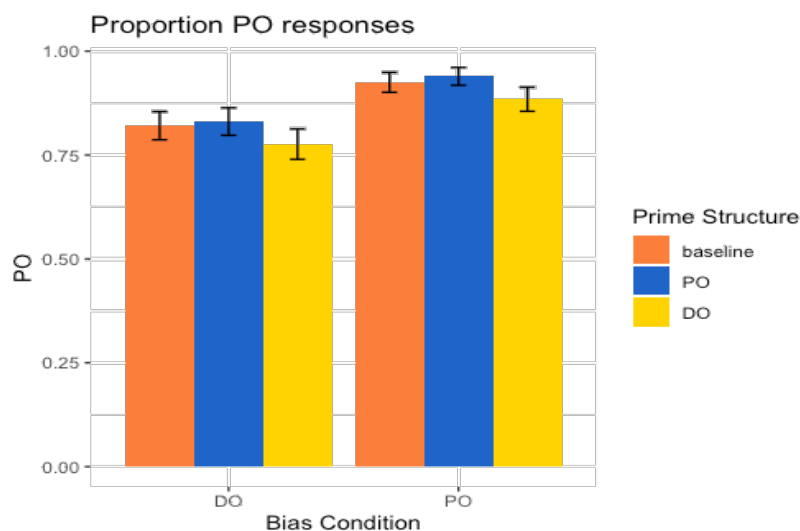


Figure 2. Proportion of PO responses in Dutch for each priming condition and bias (with 95% confidence intervals).

Probability matching vs. regularization in contact-induced syntactic change

Ming Xiang¹, Christine Gu¹, Yixue Quan², Weijie Xu¹, Suiping Wang²

¹The University of Chicago; ²South China Normal University

In statistical learning, both probability matching and (over-)regularization have been found in human behavior [1]. In the former, humans reproduce the probability distribution in the input; whereas in the latter, the frequent pattern in the input is produced even more frequently than its input frequency. Statistical learning has been suggested as a possible mechanism for language change, but the conclusions are often based on artificial language learning tasks. The current study looks at syntactic change due to language contact in multilingual communities. Using a picture-description production task, we investigated the usage of ditransitive verbs across multiple generations of Cantonese speakers from Guangzhou, China. Cantonese is the major local language spoken in Guangzhou, but its usage is in decline amid intensive contact with Mandarin Chinese. The current study, being one of the first to quantitatively evaluate syntactic change in Cantonese, revealed that the younger generation of Cantonese speakers, instead of shifting to a direction that probability-matches the distribution of Mandarin (the dominant contact language), actually over-regularized the originally preferred pattern within Cantonese.

Procedure: Two main groups of participants were tested on the same set of stimuli (Table 1). The **target Group 1** are native Cantonese speakers (18-70 years old) that were born and raised in Guangzhou and currently live there. The second **control Group 2** are native Beijing Mandarin speakers (18-60 years old) that were born and raised in Beijing and currently live there. In a picture description task, participants used a verb provided to them in their respective native language to describe a picture that depicts a ditransitive event. The critical trials (n=21) all have verbs that can be used ditransitively. There are an additional 20 filler trials.

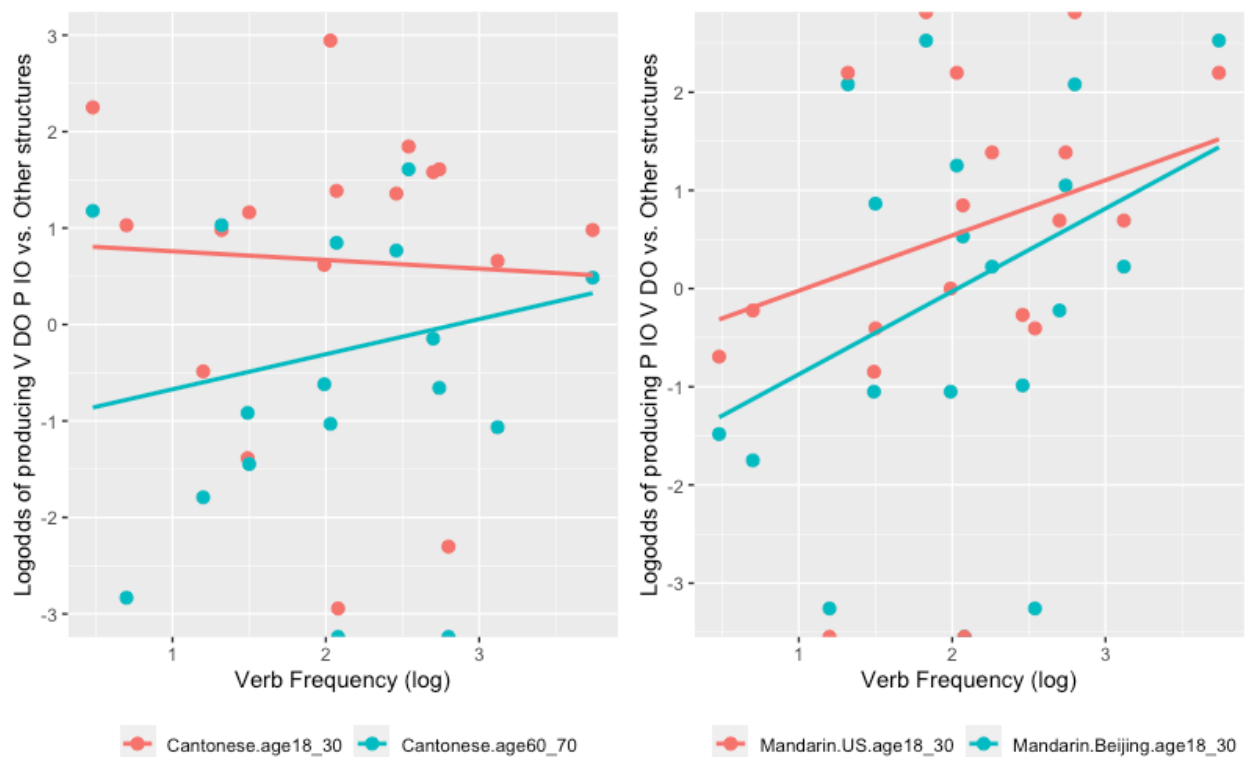
Results: Although the range of possible syntactic frames produced by the two groups of speakers are largely identical, there is a sharp contrast in the distribution of the patterns. The most frequently produced Cantonese structure is V DO P IO (55% on average, *sent some apples to the friends*), whereas for Mandarin it is P IO V DO (46%, *to the friends sent some apples*). As shown in Table 1, for Cantonese, the production frequency of the dominant V DO P IO order gradually increased from older to younger generations, and there is no change in the P IO V DO frequency, showing no assimilation to Mandarin. A mixed-effects logistic regression model, with age as a continuous variable, confirmed that older adults produced fewer V DO P IO structures (Est=-0.027, SE=0.005, z=-5.5, p<.0001). Younger Cantonese speakers therefore have **over-regularized** the originally preferred V DO P IO pattern. For each generation of Cantonese speakers, we also calculated an entropy measure based on the frequency (aggregated by participants and items) of each syntactic frame produced. We found entropy reduction from older to younger generations (Table 1), consistent with the observation of over-regularization. When parallel analyses were carried out for the group of Beijing Mandarin speakers, there is no evidence for any change at all. The over-regularization in Cantonese is therefore not the result of a universal diachronic process. It appears to have taken place because Cantonese is under the pressure of being in contact with another dominant language. To explore whether contact-induced over-regularization is a more general phenomenon, we conducted a pilot study on a third group of young Beijing Mandarin speakers, between 18-30 years old, who were born and raised in Beijing but moved to Chicago in their late adolescence or early adulthood. Compared to the age-matched Mandarin speakers living in Beijing (18-30 years), the Mandarin speakers in the US showed clear over-regularization, producing significantly more instances of the P IO V DO pattern (Est=0.75, SE=0.34, z= 2.1, p<.05). Regularization has been argued to be frequency-dependent in some previous studies [2,3]. To understand the individual item effect, we correlated each verb's **lexical frequency** with its log-odds of being used in the most dominant syntactic frame (Figure 1). We did not find an interaction between lexical frequency and old/young Cantonese groups (p>.4), nor did we find an interaction between lexical frequency and US/Beijing Mandarin groups (p>.7).

Conclusion: In an intensive multilingual environment, the weaker language does not necessarily assimilate to the dominant language. Instead we observe contact-induced (over-)regularization, suggesting a potential relationship between regularization and cognitive load pressure [4].

Table 1:

	Age group	Mandarin speakers				Cantonese speakers from Guangzhou			
		# of participants	Frequency of V DO P IO	Frequency of P IO V DO	Syntactic frame entropy	# of participants	Frequency of V DO P IO	Frequency of P IO V DO	Syntactic frame entropy
Beijing	60-70	NA				20	0.4	0.06	2.68
	50-60	14	0.04	0.45	2.48	23	0.58	0.07	2.23
	40-50	17	0.05	0.47	2.45	23	0.62	0.03	2.07
	30-40	22	0.11	0.47	2.42	16	0.63	0.02	1.98
	18-30	27	0.05	0.47	2.52	22	0.67	0.08	1.94
US	18-30	10	0.03	0.57	2.12	NA			

Figure 1: Relationship between lexical verb frequency and the log-odds of producing the dominant syntactic frame over other structures. **Cantonese (Left):** oldest and youngest generations. **Mandarin (Right):** speakers living in the US and their age-matched counterparts living in Beijing.



References:

- [1] Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development*, 1(2), 151-195.
- [2] Morgan, E., & Levy, R. (2016). Frequency-Dependent Regularization in Iterated Learning. In *The Evolution of Language: Proceedings of the 11th International Conference*.
- [3] Liu, Zoey and Morgan, Emily (2020) Frequency-dependent Regularization in Constituent Ordering Preferences. *The 42nd Annual Conference of the Cognitive Science Society*.
- [4] Ferdinand, V., Kirby, S. & Smith, K. (2019) The cognitive roots of regularization in language. *Cognition*.

Title: When animacy overshadows word order in sentence comprehension: The case of late first-language acquisition

Unlike hearing individuals who always have full access to language from birth, deaf individuals often have impoverished early language experience in childhood. Under some extreme circumstances, deaf individuals may acquire American Sign Language (ASL) as their first language (L1) after late childhood, resulting in poor language outcomes (Mayberry, 1993; Mayberry et al 2002; Frejan Ramirez et al, 2013; Cheng & Mayberry, 2019). This population provides a rare opportunity to investigate the sensitive period for language. One unanswered question is what strategies late L1 signers use to comprehend simple transitive sentences in ASL.

When comprehending transitive sentences, young children often use heuristic strategies and rely on non-linguistic cues, such as animacy and event plausibility, before they can fully rely on word order (Dodson & Tomasello, 1998; Strohner & Nelson 1974). Cheng and Mayberry (2020) found that late L1 signers of ASL also predominantly rely on event plausibility rather than word order (SVO in ASL) or subject animacy when interpreting implausible sentences like BANANA BITE BOY or DUCK CARRY CLOWN. However, the animacy features of the participating nouns can be confounded with event plausibility. Also, given that event plausibility is a strong cue in this population, it may overshadow the use of both animacy and word order. Therefore, it is not clear if late L1 signers of ASL can make use of either animacy or word order when comprehending implausible and non-reversible transitive events in ASL. On the other hand, all late L1 signers in this study had an extremely late ASL onset (after 9 years of age). It is crucial to also examine individuals with less severe delays in L1 ASL, in order to understanding the role of ASL onset on the acquisition and use of basic linguistic cues such as word order.

In the current study, we conduct two experiments to explicitly test the roles of animacy and word order during sentence comprehension with deaf late L1 signers with extremely late language onset (Exp. 1), and with deaf signers with various ASL onsets (Exp. 2). We adopted a sentence-picture matching task and crossed the two nouns in SVO ASL sentences (subject, object) with noun animacy (animate, inanimate), yielding four sentence conditions (Figure 1). All transitive sentences consisted of two nouns and one plain verb indicating implausible events and involved no human characters. We also included 4 filler conditions including both plausible and implausible intransitive events and spatial relations. Each condition includes 15 items, yielding 60 target items and 60 filler items. Experiment 1 was conducted in person with 5 deaf late L1 signers and 5 deaf native L1 signers. All deaf late L1 signers were born profoundly deaf, did not use hearing devices, and had minimal spoken/written language proficiency based on self-report. They all had an extremely late onset to ASL, ranging from 11 to 25 years of age; they all had at least 3 years of ASL exposure by the time of testing, ranging from 3 to 42 years. Experiment 2 will be conducted online, and we are currently recruiting participants. We plan to include 4 groups of L1 ASL signers with at least 9 years of ASL experience: Native Signers (NS, N=10, ASL onset 0-2yo); Early Signers (ES, N=10, ASL onset 3-5yo); Late Signers (LS, N=10, ASL onset 6-8yo); Severely Late Signers (SLS, N=10, ASL onset >9yo). In addition to the online comprehension task, we will also gather the following information: a) detailed language background information using a questionnaire; b) English reading comprehension skills using Woodcock Reading Mastery Tests; c) non-verbal cognitive skills using a group of standardized cognitive tests.

Results from Exp. 1 (Figure 2) show consistent use of word order with little interference from noun animacy for the native signers. In contrast, the Late L1 Signer group performed above chance when there is no animacy conflict (animate-animate, $z=4.45^{***}$; animate-inanimate, $z=5.12^{***}$; inanimate-inanimate, $z=3.25^{**}$), but only at chance level when the subject was inanimate and object was animate ($z=0.34$). These results indicate that 1) late L1 signers make use of both word order and animacy when an event plausibility cue is available; and 2) animacy plays a more salient role when the two cues conflict with each other. When animacy conflicts with the syntactic role, late L1 signers are less likely to rely on word order. In Exp. 2,

we expect to see increasing reliance on word order with earlier ASL age onset (NS>ES>LS>SLS). Alternatively, there may be a cut-off age of language onset such that word order is robust when acquired before a certain age (e.g. NS=ES=LS>SLS).

The current findings confirm previous findings, showing that when early language is impoverished, even basic linguistic cues appear to be less accessible to the learner. This incomplete learning may affect subsequent learning mechanisms, such as syntactic bootstrapping, impeding the further development of more complex sentence structures.

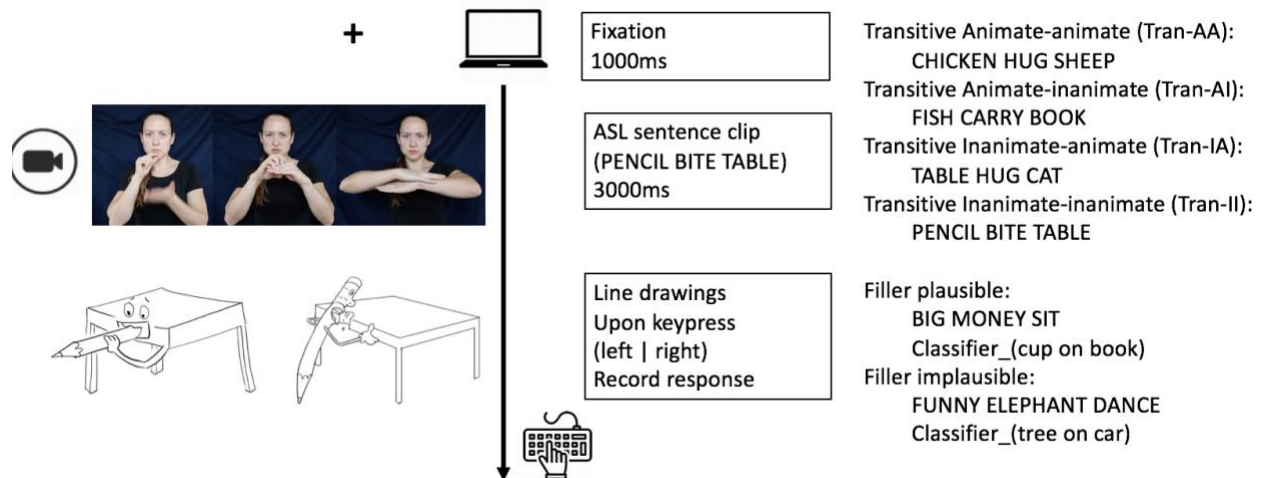


Figure 1: Experimental paradigm and conditions (with ASL gloss examples in upper case English)

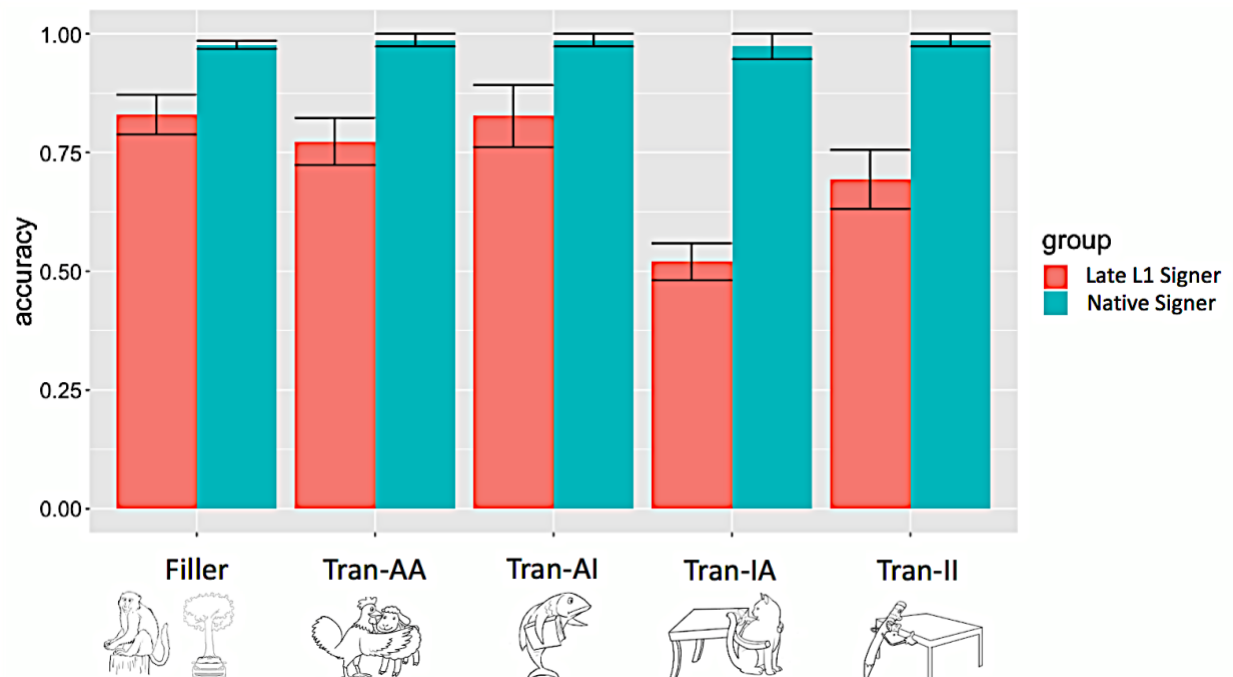


Figure 2: Group accuracy results of filler and target conditions (with matched picture examples)

References:

- Cheng, Q., & Mayberry, R. I. (2019). Acquiring a first language in adolescence: the case of basic word order in American Sign Language. *J Child Lang*, 46(2), 214-240.
- Cheng, Q., & Mayberry, R. I. (2020). When event knowledge overrides word order in sentence comprehension: Learning a first language after childhood. *Developmental Science*, e13073.
- Dodson, K., & Tomasello, M. (1998). Acquiring the transitive construction in English: The role of animacy and pronouns. *Journal of child language*, 25(3), 605-622.
- Ferjan Ramirez, N., Lieberman, A. M., & Mayberry, R. I. (2013). The initial stages of first-language acquisition begun in adolescence: When late looks early. *J Child Lang*, 40(2), 391-414
- Mayberry, R. I. (1993). First language acquisition after childhood differs from second language acquisition: The case of American Sign Language. *Journal of Speech and Hearing Research*, 36(6), 1258-1270.
- Mayberry, R. I., Lock, E., & Kazmi, H. (2002). Linguistic ability and early language exposure. *Nature*, 417(6884), 38-38.
- Strohner, H., & Nelson, K. E. (1974). The young child's development of sentence comprehension: Influence of event probability, nonverbal context, syntactic form, and strategies. *Child Development*, 567-576.

Distributed Morphology feature geometries crosslinguistically: Acquiring the copula
Shiloh Drake
Bucknell University

In Distributed Morphology (DM), it is assumed that words, phrases, and sentences are made up of the same hierarchical relationships: that is, elements of sentences and elements of words can be diagrammed in constituent structures, and morphemes are not simply the result of morphophonological processes (Harley & Noyer, 1999). Morphemes are made up of three elements that, when combined, result in a structure that contains grammatical features, semantic features, and the phonology necessary to utter the word or phrase in question. This paper tests the acquisition of the feature bundles that comprise the grammatical features of a word—for instance, person, number, clusivity, case, and other morphosyntactic features that differentiate between the functions of words.

As an offline model, DM must specify how morpheme selection occurs. One proposed method is through a feature hierarchy or geometry, similar to those proposed as typological universals (Drake, 2020; Harley & Ritter, 2002; Hanson, 2000). In acquisition, the feature hierarchy would predict that less marked morphemes are acquired before more marked morphemes, and morphemes that express more agreement features are acquired later than morphemes that express fewer agreement features. Harley and Ritter (2002) analyzed typologically distinct languages as well as acquisition data from Hanson (2000), and showed that this type of feature hierarchy correctly predicts the acquisition and distribution of pronouns agreeing in person and number. Drake (2020) showed that the acquisition of English copula followed a similar pattern, where the stages of acquisition follow a default 3.pres.sing *is* at around 2;0 (years;months) and more complex forms of agreement like the use of the past participle *been* occur much later at 3;1.

To further test the assumptions set forth by the previous studies, additional corpora of child speech from CHILDES (MacWhinney, 2000) in English, French, Irish, Japanese, Sesotho, and Welsh were analyzed for the occurrence of copulas, with children ranging from 0;11 to 7;00 in age. According to previously proposed feature geometries, 1st and 3rd person present singular forms should occur at earlier ages, followed by 2nd person present singular. Forms with more distinguishing features, such as number, tense and aspect, should appear later.

This hypothesis is borne out after a preliminary analysis of the corpora. In each of the corpora, the first instance of a copula occurred at roughly 1;0 and was a “default” non-second, non-past singular form. Forms specified for additional features, such as past, plural, and aspect, occurred much later, as found in Drake’s (2020) previous study. Overall, 1st and 3rd person singular forms are most often used, with the present tense observed slightly more often than the past tense in the children’s speech.

This analysis provides further support to the feature geometry proposed by Harley and Ritter (2002) and Drake (2020), and also provides support for DM’s potential usefulness as a model of on-line language processing (Pfau, 2008; Gwilliams & Marantz, 2015; Drake, 2018; *inter alia*). Analyzing the longitudinal naturalistic speech of typically developing children who speak many different languages provides a measure of on-line grammatical processing that is difficult to obtain in a setting other than in a child’s home, but also provides rich data to aid in validating theories and models of language. As children seem to acquire morphemes in an orderly fashion (e.g., Brown, 1973) regardless of the language that they speak, acquisition data can only enhance the models and frameworks that pay it heed—especially given the frequently cited divide between linguistic competence and linguistic performance.

Effects of word order on L1 and L2 semantic prediction

Carrie N. Jackson (Penn State University), Holger Hopp (TU-Braunschweig), Theres Grüter (University of Hawai'i)

Previous research shows that adult L2 speakers use semantic cues to predict upcoming input during language comprehension (e.g., Chambers & Cooke, 2009; Dijkgraaf et al., 2017; Ito et al., 2018). However, this research has relied on subject-first (SVO) sentences and no studies have investigated whether L2 speakers also use semantic cues predictively when embedded in a different word order that poses difficulties and is used less frequently in L2 production compared to L1 production (e.g., Jackson & Ruf, 2017; O'Brien & Féry, 2015). Here, we investigate how syntactic structure, i.e. word order differences, affects the timing and magnitude of semantic prediction, especially when L1 and L2 word orders differ, to investigate whether and how syntax constrains L2 semantic prediction, as compared to L1 semantic prediction.

In a visual-world experiment, 32 L1 English-L2 German speakers and 32 L1 German speakers listened to subject-first (SVO) and adverb-first (AdvVS) sentences. For subject-initial sentences, English and German share SVO surface order (1), while non-subject initial sentences have V3 order in English (AdvSV), but V2 order in German (2). We tracked participants' eye-movements to image displays (Fig. 1) and measured if they used semantic information from the lexical verb predictively to anticipate the upcoming noun (constraining-verb; 1a/2a). Looks to the target in sentences using modal verbs (neutral-verb; 1b/2b), in which the lexical verb appears at the end of the sentence, served as a baseline (see Dahen & Tanenhaus, 2004, for L1 Dutch).

- (1a) Simone_{SUB} füttert_V täglich [den Hund]_{OBJ} im Garten. (SVO; constraining-verb)
Simone feeds daily the dog in the garden
- (1b) Simone_{SUB} soll_{Vmod} täglich [den Hund]_{OBJ} im Garten füttern_V. (SVO; neutral-verb)
Simone should daily the dog in the garden feed
"Simone feeds/should feed the dog daily in the garden."
- (2a) Im Sommer springt_V täglich [der Frosch]_{SUB} ins Wasser. (AdvVS; constraining-verb)
In summer jumps daily the frog into the water
- (2b) Im Sommer wird_{Vmod} täglich [der Frosch]_{SUB} ins Wasser springen_V. (AdvVS; neutral-vb)
In summer will daily the frog to the water jump
"In summer the frog will jump/jumps into the water daily."

Data were analyzed using a bootstrapping procedure with confidence intervals (Stone & Lago, 2020) to identify the time point at which looks to the target diverged in constraining-verb versus neutral-verb sentences. An analysis of looks time-locked to verb onset revealed more looks to the target in constraining-verb versus neutral-verb sentences prior to the onset of the target noun for both SVO and AdvVS sentences among both L1 and L2 speakers, though prediction was generally delayed for L2 speakers. For L1 speakers, the divergence point for SVO sentences (806ms [CI: 782, 850]) and AdvVS sentences (894ms [CI: 833, 1020]) were similar, with overlapping CIs (Fig. 2). For L2 speakers the divergence point for SVO sentences (1169ms [CI: 1071, 1343]) and AdvVS sentences (1317ms [CI: 1224, 1479]) were also similar, with overlapping CIs (Fig. 2). A second analysis to examine effects of L2 proficiency revealed that among the L2 speaker group, higher proficiency was associated with more looks overall to the target noun (from verb to noun onset) in both word orders, but that L2 speakers engaged in predictive processing regardless of proficiency level and word order. These results demonstrate that adult L2 speakers engage in semantic prediction across syntactic contexts, including contexts not present in the L1, suggesting that any modulations in L2 semantic prediction based on syntax may be quantitative, not qualitative, in nature (Kaan, 2014).

Figures

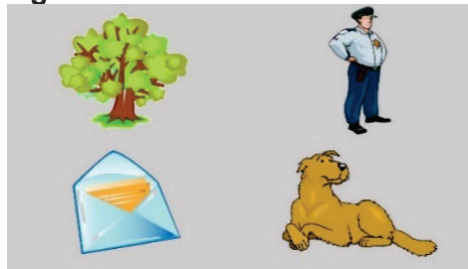


Figure 1. Image display (for 1a/1b)

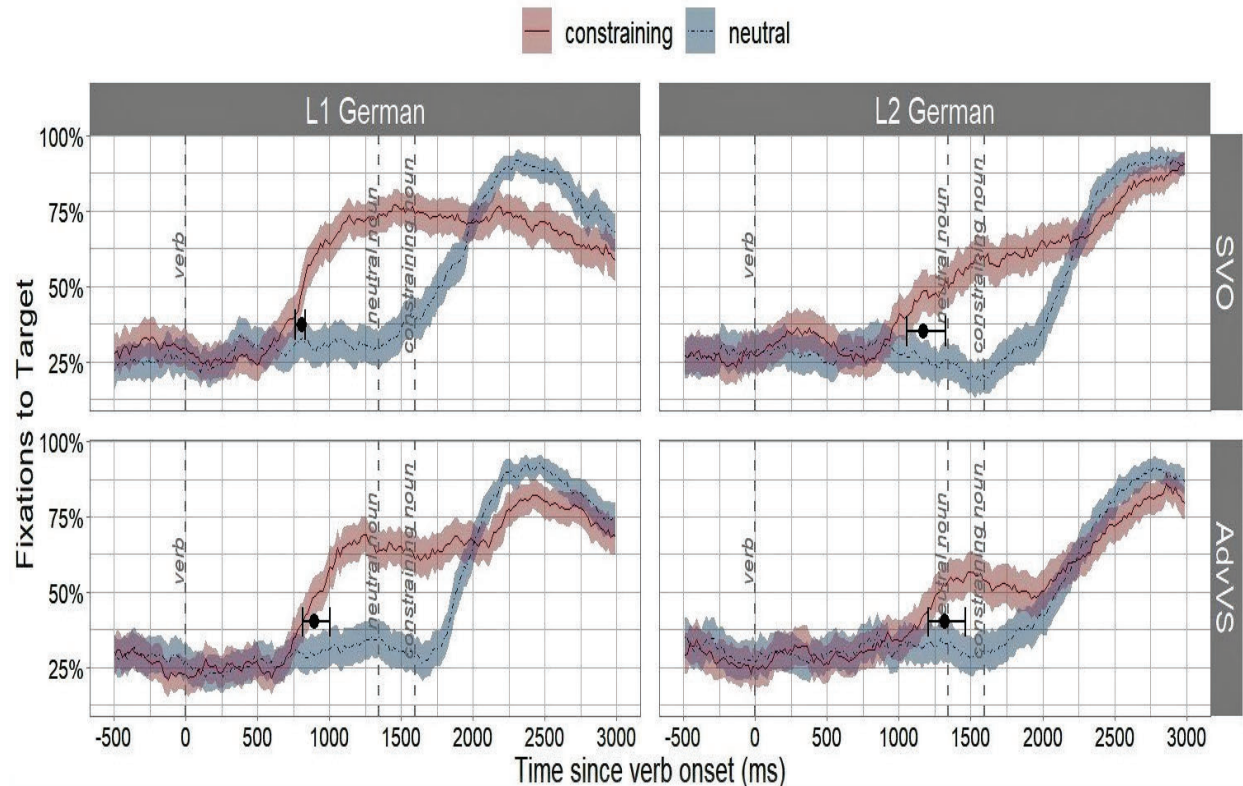


Figure 2. L1 and L2 speaker fixations to target noun in neutral-verb vs. constraining-verb sentences. Divergence points (with bootstrapped 95% confidence intervals) in black.

References

- Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: Sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 1029-1040.
- Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 498-513.
- Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, 20(5), 917-930.
- Jackson, C. N., & Ruf, H. T. (2017). The priming of word order in second language German. *Applied Psycholinguistics*, 38(2), 315-345.
- Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?. *Linguistic Approaches to Bilingualism*, 4(2), 257-282.
- Stone, K., Lago, S., & Schad, D. J. (2020). Divergence point analyses of visual world data: applications to bilingual research. *Bilingualism: Language and Cognition*, 1-9.
- O'Brien, M. G., & Fery, C. (2015). Dynamic localization in second language English and German. *Bilingualism*, 18(3), 400-418.

Model-based estimates of predictability reveal brain's robust sensitivity to variation in semantic fit even among unexpected words.

Jakub M. Szewczyk, Kara D. Federmeier (University of Illinois at Urbana-Champaign)

The brain's graded response to words that vary in their predictability in a sentence has been well-established: There is a monotonic relationship between the amplitude of the N400 ERP component and human production norms (cloze probability: CP), such that more predictable words elicit smaller N400s. However, this discriminability approaches a limit for words with CPs near zero because of limited variance in CP values and noisy estimation of CPs for weakly predictable items. Moreover, even comparisons between plausible but unexpected and wholly anomalous words yield only small N400 differences (e.g., Kuperberg et al., 2020). It is unclear whether this pattern is revealing of the mechanism underlying contextual facilitation – for example, that the language processor predicts only a small set of specific lexical items (e.g., Van Petten & Luka, 2012) – or simply because the CP metric is at floor and thus is unable to pull out variance that is actually in the signal. Two views provide opposing predictions: 1) the N400 is sensitive to differences in predictability of all words and it is related to the predictability on the log scale (the surprisal theory, Levy, 2008; Kuperberg & Jaeger, 2016); 2) the N400 is sensitive to predictability only in the range measurable with CP tests and the relationship is linear (Brothers & Kuperberg, 2021).

To adjudicate between these possibilities, in this study, in this study, we reanalyzed data from an ERP experiment in which 32 participants saw 282 simple English sentences that were completed by expected and unexpected (but plausible) words. We quantified the predictability of the sentence endings using GPT2-xl, a state-of-the-art machine learning model of language, which assigns a probability distribution across all possible sentence continuations. We first tested how model-derived predictability compares with predictability estimated by classic CP tests in explaining N400 amplitudes to expected endings, in which CP varied in the range 0.09-1.00 (mean CP=0.56). The mixed-effects regression revealed that both sources of predictability estimates are excellent predictors of the N400 amplitude to the sentence endings ($t=4.9$ for the GPT-2 model, $t=5.1$ for CPs), although, as revealed by model comparison, CP explained N400 variance over and beyond GPT2 ($\Delta\log\text{Lik} = 4$, $p < .01$) but not the other way around ($\Delta\log\text{Lik} = 1$, $p = .17$). Overall, both models explained N400 amplitude variance in the range of 5 μV (see Figure 1, left panel). Additional GAMM models showed that the relationship is linear (both with predictability estimated by CP tests and by the GPT2 model).

Next, we analyzed the response to unexpected endings. Here, CP could not explain N400 amplitude as all unexpected words had CP=0. However, a mixed-effects regression using the GPT2-based index of predictability revealed that the N400 was robustly sensitive to predictability even in this range ($t = 5.9$) and even though all the unexpected endings were fully plausible. Indeed, variance in N400 amplitude to unexpected endings was surprisingly large, exceeding variance observed to expected endings (see Figures 1 & 2). Additional GAMM models showed that the relationship between the N400 and predictability in this range is logarithmic. We replicated these findings using two other ERP experiments using similar items, involving 42 participants and 8602 data-points in total.

Because different functions related predictability and the N400 to expected and unexpected words, we made a final model using a single function that could jointly fit both types of words: $\beta_1 * p + \beta_2 * \log(p)$ (see Figure 2). We propose that the logarithmic component ($\log(p)$) reflects updating of conceptual representations, in line with the surprisal theory (Levy, 2008), while the linear component (p) corresponds to the degree to which lexical representation of the word was hierarchically preactivated by the representation of the context.

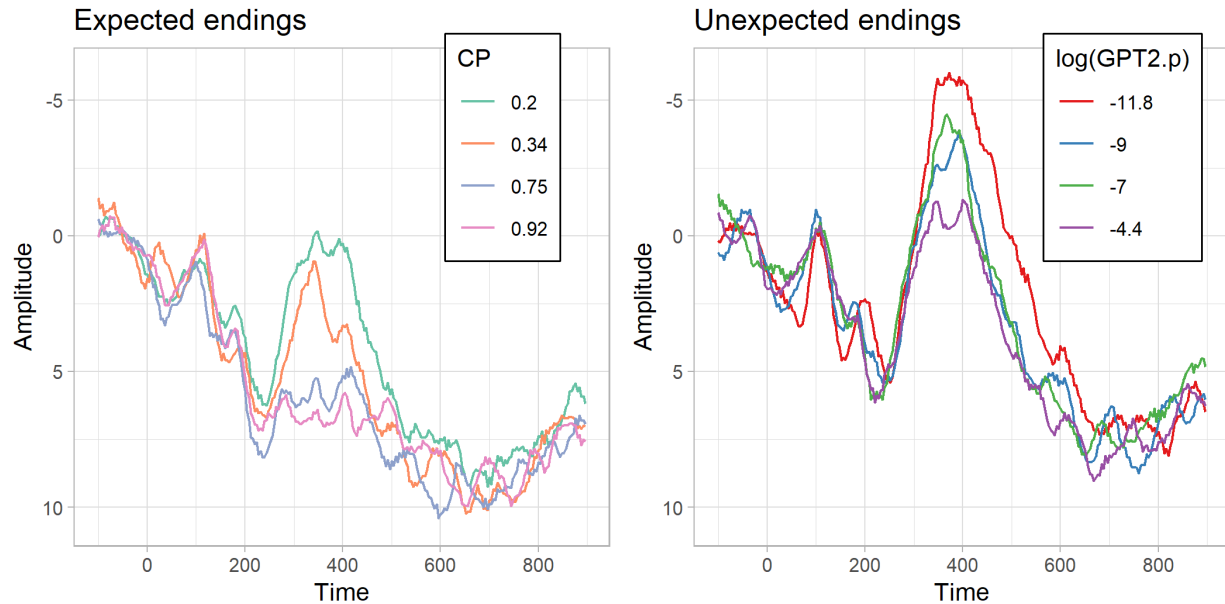


Figure 1. ERPs to expected (left panel) and unexpected (right panel) sentence endings, broken down by their predictability estimated by cloze probability tests (left panel, linear scale) or the GPT2 model (right panel; log scale). The bins were set to have an equal number of items. Values in the legend correspond to the mean predictability in each bin.

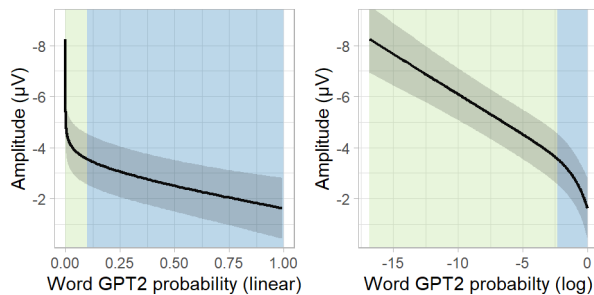


Figure 2. Predictions of the model of the N400 amplitude that includes predictors of word's probability both on the linear and logarithmic scale. Left panel: probability on the linear scale; right panel: probability on the logarithmic scale. The contrast between two colors of background corresponds to a threshold (arbitrarily set at $p = .1$) separating regions where the relationship between word probability and N400 is more linear (blue) and more logarithmic (green).

References:

- Brothers & Kuperberg (2021). JML, v116, 104174
 Levy, R. (2008). Cognition, v106, 1126–1177
 Kuperberg, G. R., & Jaeger, T. F. (2016). LCN, v31, 32–59
 Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). JoCN, v32, 1-35
 Kutas, M., & Hillyard, S. A. (1984). Nature, v307, 161–163
 Van Petten, C., & Luka, B. J. (2012). IJoP, v83, 176–190

Online cloze evidence for rapid use of lexical and grammatical cues

Masato Nakamura & Colin Phillips (University of Maryland)

Predictions about upcoming input are standardly measured via facilitated processing of explicitly presented words (fixation times, N400 amplitudes) or anticipatory looks in scenes (e.g. [1, 2]). In this study we examine predictions via a spoken, speeded cloze task in Japanese. We use information from spoken responses to understand how and when contextual cues are used to generate predictions, revealing effects obscured in EEG studies.

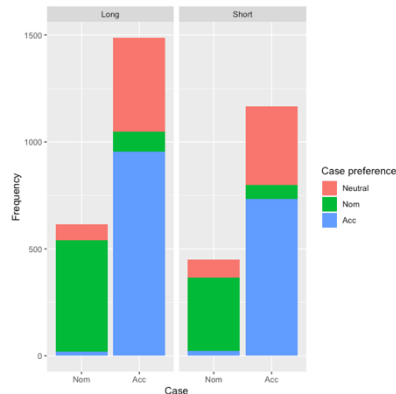
Situations where online measures of prediction diverge from corpus probabilities or late measures are particularly informative about how predictions arise. A useful test case is argument role reversals, in which an anomalous verb is processed as if it is more expected than it should be (e.g. [3,4]). For example, in *The customer that the waitress served* vs. *The waitress that the customer served* the verb *serve* differs in offline cloze probability, but EEG studies in many languages have found that it elicits identical N400 amplitudes. Additional time between the arguments and the verb yields an N400 contrast [5]. These findings motivated the claim that early predictions reflect lexical associations, with role-specific predictions emerging only after a delay. However, the explicit presentation of anomalous verbs in these studies might bias the estimate of how expected those anomalous words were.

We examined the timing of use of argument role and lexical cues in a Japanese speeded cloze task, using materials from a previous EEG study [6]. We presented minimal contexts of a noun and a case marker, which participants completed with a verb. The cloze task measures predictions via speakers' own productions. Instead of measuring the degree of convergence of open-ended predictions (i.e. cloze probability), we used the full set of productions to test predictions at specific times by (i) using simple contexts, to control lexical and grammatical content of cues, (ii) limiting the response time windows [7] and (iii) using a simple NLP measure to assess the relationship between contexts and produced items. This was possible by gathering spoken responses via the internet.

80 speakers [40 analyzed so far] each completed 160 visually presented fragments. In a 'long' block responses had to start after 1.6-2.8s, and by 1.2s in a 'short' block. The timeline and the stimuli matched an existing study that found identical N400s at the verb, regardless of case. For each of the 5389 produced noun-verb pairings we measured speech onset latency, noun-verb similarity using Japanese word2vec [8], and whether the pairing would be more plausible with nominative or accusative case-marking, e.g., *thief-acc arrest* is more plausible than *thief-nom arrest*. Pairings featuring the dispreferred case were coded as reversals.

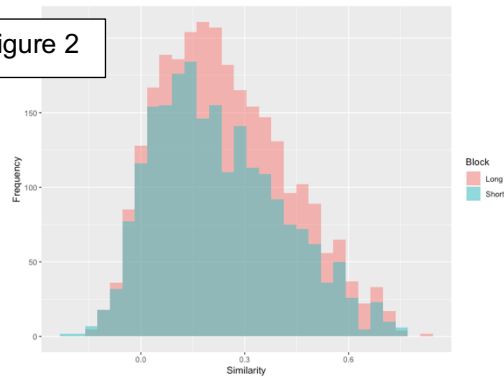
Argument roles clearly affected productions. Case-compatible productions were far more common than reversals, comprising 94.4% of trials in the short condition (Fig. 1). Verb transitivity clearly matched the case marking. Noun-verb similarity was higher in the long condition, suggesting more specific expectations with more time (Fig. 2). The verbs produced in reversed responses tend to have high cloze probabilities in the other case markings, suggesting that role-independent lexical associations serve as lures (Fig 3.). Speech onset latencies were shorter for more similar pairings. Overall, the speeded cloze results show that both argument roles and lexical association shape early predictions [cf. 9]. The discrepancy with prior EEG results could reflect a monitoring process that filters (most) role-incompatible productions in the cloze task, or a biasing effect of explicitly presented lures in EEG studies.

Figure1



The frequency of transitive verb productions. The x-axis represents the case marking of the context noun, and the color represents whether that verb is more plausible if that noun is in nominative or accusative case, or is neutral. Very few noun-verb productions involved the dispreferred case marker

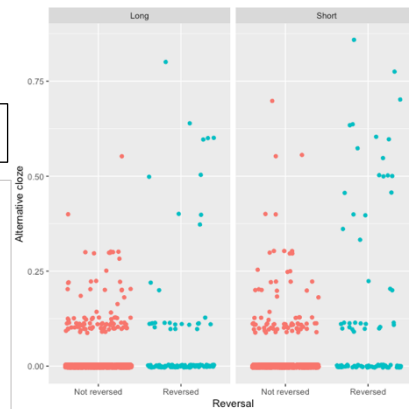
Figure 2



The word2vec similarity of the noun-verb pairing produced in each trial.

Figure 3

Lure strength: for each noun-verb production, each dot represents the cloze probability (in the current experiment) of the same verb in trials where the noun had the alternative case marker, e.g., for *thief-nom arrest*, the figure shows the cloze probability of *thief-acc arrest*. Elevated values indicate strong lures.



References

- [1] Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161–163.
- [2] Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- [3] Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1), 117–129.
- [4] Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2016). A “bag-of-arguments” mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5), 577–596.
- [5] Chow, W.-Y., Lau, E., Wang, S., & Phillips, C. (2018). Wait a second! Delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7), 803–828.
- [6] Momma, S. M. (2016). *Parsing, Generation and Grammar*.
- [7] Chow, W.Y. & Kurenkov, I., Buffinton, J., Kraut, B., & Phillips, C. (2015). How predictions change over time: evidence from an online cloze paradigm. Poster presented at the 28th annual CUNY Human Sentence Processing Conference, Los Angeles, CA.
- [8] Matsuno, S., Mizuki, S., & Sakaki, T. (2019). 日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築 [Construction of a distributed word representation model using large-scale SNS + web-based Japanese corpora]. 人工知能学会全国大会論文集 [Proceedings of the Annual Conference of the Japanese Society of Artificial Intelligence], JSAI2019, 4Rin113-4Rin113.
- [9] Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5), 602–616.

Children with Hearing Loss Use Semantic and Syntactic Cues for Prediction in Sentence Comprehension

Rebecca Holt (Macquarie University), Benjamin Davies (Macquarie University), Laurence Bruggeman (Western Sydney University), Katherine Demuth (Macquarie University)

Prediction of upcoming words benefits listeners' spoken language processing. Predictable words can be identified with less acoustic information [1], can be accessed earlier [2], and require less effort to process [3]. Prediction may thus be particularly advantageous to those for whom speech input is degraded and for whom language processing is slow and effortful, such as children with pre-lingual hearing loss (HL) [e.g., 4]. Prediction has not yet been examined among children with HL, though they may struggle to employ contextual information [e.g., 1]. This suggests that their ability to predict based on context may be less efficient than their normal-hearing (NH) peers.

Children with NH as young as 2 years can predict based on a range of linguistic cues, including semantic context [5] and subject-verb syntactic agreement [6]. These different types of prediction may pose different challenges for children with HL. While semantic prediction is predominantly based on content words, which are highly salient in speech, agreement-based syntactic prediction depends on function words and affixes, which are often less salient and less accessible to those with HL. Syntactic prediction can also be inconsistent; NH children demonstrate better prediction using plural subject-verb agreement than singular [6, 7]. We therefore hypothesised that children with HL would predict less than their NH peers, if at all. However, if children with HL did predict, we expected this in the more perceptually-salient semantic context, rather than in the syntactic.

In Experiment 1, 25 English-speaking children with HL (hearing aid and/or cochlear implant users; $M_{age} = 10;2$) and 25 with NH ($M_{age} = 9;6$) participated in a visual world paradigm eye-tracking task [8]. They heard sentences in which the object noun was semantically related (predictable) or unrelated (unpredictable) to the subject noun and verb while viewing four images on screen: the object noun and three distractors. Experiment 2 included two additional children with HL ($N = 27$; $M_{age} = 10;2$), and six additional children with NH ($N = 31$; $M_{age} = 9;9$). Children heard sentences (Table 1) with (predictable) or without (unpredictable) copula number agreement with the target noun while viewing two images: a single animal and a group of animals. Logistic curves were fit to the proportion of looks to the target for each participant and condition in both experiments. The crossover points of each curve, reflecting the timing of looks to the target, were analysed using linear mixed-effects models. Fixed factors were Predictability and Group, plus Number (i.e., singular/plural target; for Experiment 2 only). Models had maximal random effects.

In Experiment 1, participants looked to the target earlier in the predictable than the unpredictable condition ($\beta = 23.28$, $SE = 3.40$, $p < .001$), demonstrating semantic prediction. In Experiment 2, there was a significant interaction between Predictability and Number ($\beta = -32.48$, $SE = 7.76$, $p < .001$). Participants looked to the target earlier in the predictable than the unpredictable condition, but only for plural targets. Agreement was thus used for prediction, but only for *are*, not *is*, similar to [6, 7]. No significant differences between groups were found in either experiment. Thus, in contrast to our hypotheses, and previous findings of limited use of context among children with HL [e.g., 1], children with HL were able to predict on par with their NH peers based on both more- and less-salient auditory information. Note that our participants typically received earlier and more comprehensive intervention than those in these earlier studies. Our findings suggest that these relatively recent advances in HL intervention may have been successful in allowing children with HL to achieve more NH-like spoken sentence processing, and that interventions relying on prediction may be beneficial for children with HL.

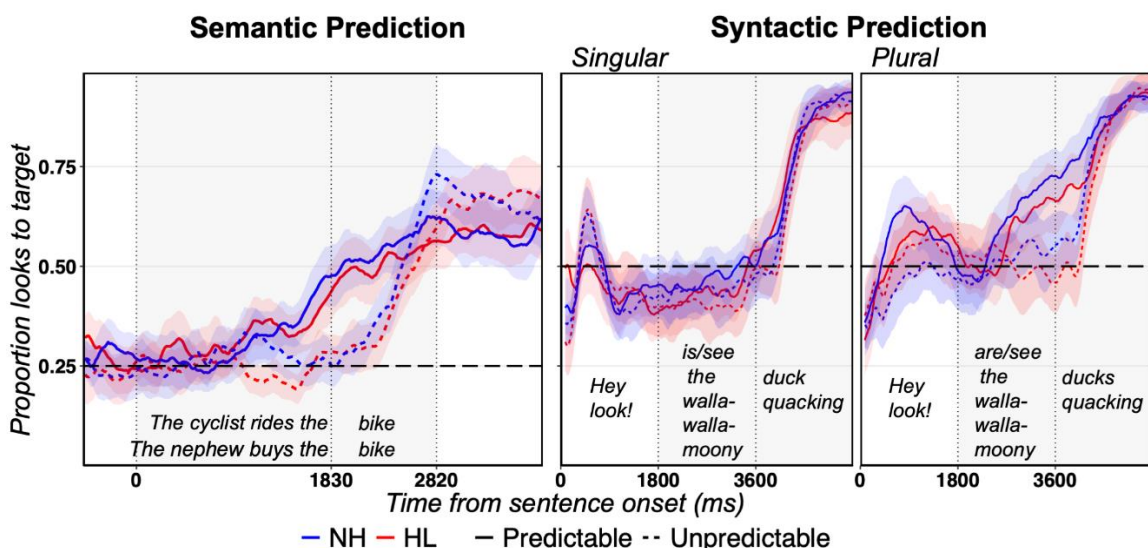
References

1. Conway CM, Deocampo JA, Walk AM, Anaya EM, Pisoni DB. Deaf children with cochlear implants do not appear to use sentence context to help recognize spoken words. *J Speech Lang Hear R.* 2014 Dec; 57:2174-2190.
2. DeLong KA, Urbach TP, Kutas M. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat Neurosci.* 2005 July; 8:1117-1121.
3. Winn, MB. Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends Hear.* 2016 Sep; 20:1-17.
4. McGarrigle R, Gustafson SJ, Hornsby BWY, Bess FH. Behavioral measures of listening effort in school-age children: examining the effects of signal-to-noise ratio, hearing loss, and amplification. *Ear Hear.* 2019 Mar; 40:381-392.
5. Mani N, Huettig F. Prediction during language processing is a piece of cake – But only for skilled producers. *J Exp Psychol Human.* 2012 Jul; 38:843-847.
6. Lukyanenko C, Fisher C. Where are the cookies? Two- and three-year-olds use number-marked verbs to anticipate upcoming nouns. *Cognition.* 2016 Jan; 146:349-370.
7. Davies B, Xu Rattanasone N, Demuth K. Comprehension of the copula: preschoolers (and sometimes adults) ignore subject-verb agreement during sentence processing. *J Child Lang.* 2020 May; 47:695-708.
8. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. *Science.* 1995 Jun; 268:1632-1634.

Table 1 – Sample stimulus sentences. The novel adjective ‘wallawallamoony’ occurred in all sentences in Exp. 2 to delay the onset of the target noun, allowing time for anticipatory looks.

Experiment	Predictable sentence	Unpredictable sentence
Exp. 1: Semantic context	The cyclist rides the bike.	The nephew buys the bike.
Exp. 2: Subject-verb agreement	Hey look! Are the wallawallamoony ducks quacking?	Hey look! See the wallawallamoony ducks quacking.

Figure 1 – Mean proportion of looks to the target image. Exp. 1 on left, Exp. 2 on right. Horizontal dashed line shows chance.



Does reading unexpected words lead to engagement of cognitive control?

Suzanne R. Jongman (sjongman@illinois.edu), Yaqi Xu, and Kara D. Federmeier
University of Illinois at Urbana-Champaign

When a sentence ends unexpectedly, readers must make adjustments to successfully integrate the unexpected word in the previous sentence context. A previous self-paced reading ERP study by Payne and Federmeier [1] suggested that readers have two mechanisms available to cope with expectancy violations. For highly constraining sentences ending unexpectedly, they found a late anterior positivity (LPC) previously argued to reflect suppression of the anticipated word and/or the revision of the sentence message [2]. Importantly, they found the LPC only for fast final-word reading trials. For slow trials, they instead found an anterior N2, previously linked to domain-general cognitive control [3]. The authors argued that the N2 acts to inhibit the prepotent motor response to move forward, giving readers time to resolve the conflict between the expected and presented word. This suggests that reading relies on cognitive control and that slow trials are actually trials of successful employment of control. In situations where readers are too late to exert cognitive control and move on quickly, they instead have to rely on a late semantic revision process as reflected by the LPC.

To test the hypothesis that reading unexpected words may rely on cognitive control, we used a cross-task paradigm interleaving self-paced reading trials with cognitive control trials, a paradigm used successfully to show ambiguity resolution engages cognitive control [4]. We presented, word-by-word, 136 highly constraining sentences from [5], half ending expectedly and the other half unexpectedly. Each sentence was followed by a Flanker trial. Adler et al. [6] showed that Flanker performance was modulated by prior reading of a cognitively demanding code-switch sentence: subjects were faster on incongruent Flanker trials that followed a code-switch compared to a non-switch sentence, but no prior sentence effect was found for congruent trials. This reflects conflict adaptation: cognitive control engagement facilitates subsequent conflict resolution [the Gratton effect, 7]. If reading an unexpected word engages cognitive control, we should see better performance on a subsequent incongruent Flanker trial.

The reading-Flanker task was performed online. To ensure participants read the sentences, a block of 34 trials was followed by 6 old/new memory questions. Only individuals with memory performance above 70% were included (48 out of 61). We used a linear mixed effects model for Flanker RTs and mixed effects logistic regression for Flanker accuracy, as [6]. Both models included prior sentence ending, current Flanker trial, and their interaction as fixed effects and subject as a random intercept. For a second analysis, we sorted Flanker responses into four separate bins based on final-word reading times, separately for each participant and condition [1]. We tested if including the three-way interaction Expectancy x Congruency x Bin improved model fit to investigate if reading speed influences control adjustments.

Results indicated a typical Flanker effect both in RTs and accuracy (Table 1): responses were faster and more accurate overall on congruent trials than incongruent trials. Did the prior sentence ending modulate this pattern? We found no such evidence as there was no significant interaction for RT nor accuracy (Table 2). Performance on incongruent trials was not enhanced after unexpected endings compared to expected endings. Instead, neither congruent nor incongruent trials were influenced by the previous final word. Including the three-way interaction with bin did not improve model fit (RT: $\chi^2(3) = 4.46$, $p = .22$; ACC: $\chi^2(3) = 2.24$, $p = .52$). Thus, there was no evidence that slow trials in particular exhibited enhanced cognitive control (Fig. 1).

To conclude, we found no evidence for cognitive control adjustments when readers encountered an unexpected word. Employment of control, previously evidenced by an N2 for slow reading trials [1], did not appear to sustain long enough to impact a subsequent Flanker trial. Whereas ambiguity resolution or code-switching [4,6] may require continued control spanning several words, reading an unexpected word may instead engage control only briefly to slow down reading for the current word, with control lifted instantly to resume normal reading.

Table 1. Response time and accuracy performance on Flanker trials*, dependent on the previous sentence ending type (as determined by cloze probability ratings).

Previous Sentence Ending	Current Flanker Trial Type	Flanker RT** (ms)	Flanker Accuracy (%)
Expected	Congruent	514 (<i>SD</i> = 166)	98.6 (<i>SD</i> = 11.7)
Unexpected	Congruent	510 (<i>SD</i> = 172)	98.7 (<i>SD</i> = 11.1)
Expected	Incongruent	689 (<i>SD</i> = 220)	91.9 (<i>SD</i> = 27.3)
Unexpected	Incongruent	694 (<i>SD</i> = 237)	93.0 (<i>SD</i> = 25.4)

* Excludes trials with final-word reading times at 99.7th percentile within person and within condition (1.5%) [1], and with Flanker RTs beyond 2.5SDs from the overall mean (1.4%) [5].

** Excludes incorrect trials (4.2%).

Table 2. Results of mixed model analyses for Flanker RT and Flanker accuracy.

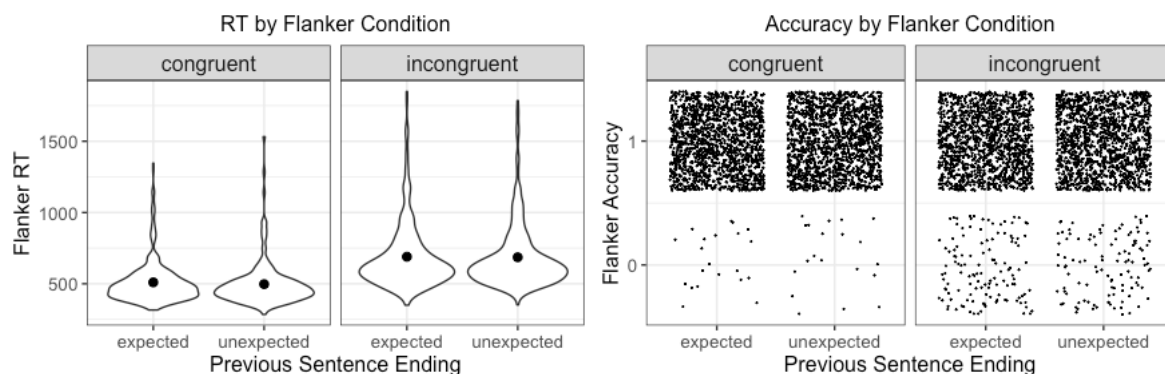
	Fixed effects*	Estimate	Std. Error	T/Z value	P value**
RT***	Intercept	6.350	0.028	224.33	<.0001
	Expectancy	0.005	0.005	1.05	0.29
	Congruency	-0.293	0.005	-58.14	<.0001
	Exp * Con	0.013	0.010	1.31	0.19
Accuracy	Intercept	4.75	0.289	16.44	<.0001
	Expectancy	-0.178	0.181	-0.98	0.33
	Congruency	2.328	0.193	12.05	<.0001
	Exp * Con	0.135	0.361	0.37	0.71

* Sum-to-zero contrast coding.

** Computed using Satterthwaite's approximation for denominator degrees of freedom.

*** RTs were log-transformed to correct for non-normal distribution.

Figure 1. Flanker RTs and Flanker Accuracy for slow reading trials (bin 4) only, separated by congruency type. Previous sentence type* did not appear to influence either Flanker trial type.



*There was no main effect of sentence ending (i.e., similar RTs for expected and unexpected words). This however does not entail that unexpected words do not require control: [6] found no main effect for code-switching, yet did find an influence on subsequent incongruent flankers.

References

[1] Payne & Federmeier, 2017, *J Cogn Neurosci*, 29:5; [2] Van Petten & Luka, 2012, *Int J Psychophysiol*, 83:2; [3] Folstein & Van Petten, 2008, *Psychophysiology*, 45:1; [4] Kan et al., 2013, *Cognition*, 129:3; [5] Federmeier et al., 2007, *Brain Res*, 1146; [6] Adler et al., 2020; *J Exp Psychol Learn Mem Cogn*, 46:4; [7] Gratton et al., *J Exp Psychol*, 121:4.

Prediction accuracy facilitates processing of visual word form.

Yang Agnes Gao, Tamara Y. Swaab, Matthew J. Traxler (University of California Davis)

In sentence processing, word retrieval is facilitated in predictable contexts, as is evidenced from faster response times in naming and lexical decision, fewer and shorter fixations during natural reading, anticipatory eye movements in visual world experiments, and reduced N400s in ERP experiments (Schwanenflugel & Shoben, 1985; Altmann & Kamide, 1999; Federmeier & Kutas, 1999). ERP evidence has shown that predictive effects are separate from and precede contextual integration (Brother et al., 2015). The precise nature of the processes that generate predictions and the types of linguistic representations that are affected by prediction remain unclear. We designed this study to produce evidence regarding the types of representations that may be affected by prediction during language processing. In particular, we tested whether anticipatory processes pre-activate word form information in a priming paradigm.

To test whether word form information is pre-activated by anticipatory processes, we asked participants (N=198) to predict target words in a priming study followed by a lexical decision task (adapted from Brothers et al, 2015; Dave et al, 2018). Participants read lists of words comprising prime and target pairs. They were asked to actively predict the upcoming target after reading the prime word, and to perform a lexical decision task on the target. On related trials, the prime and target words had a forward association strength of .5 (*circus* - *CLOWN*; *trim* - *CUT*). On unrelated trials, the forward association strength was 0 (*trim* - *CLOWN*; *circus* - *CUT*). Each participant read 480 sets of word-word pairs, and 125 filler sets of word-non word pairs (*cartoon*-*CRECKED*; *detail* - *NELB*; to generate "no" responses) None of the words were repeated within subjects, but the same target words occurred in both related and unrelated conditions across different lists. Participants completed the experiment online via PCI Ibex. Readers were presented with the first word, followed by a 1800ms delay, during which they were asked to generate a prediction of the second word based on the meaning of the first word. Then, the second word or non-word target appeared. Subsequently, readers were asked to perform two consecutive tasks: 1) speeded lexical decision: indicate whether the target is a real word in English or not; 2) prediction: indicate whether their prediction matched the second word they saw. We compared the lexical decision RTs to the target words based on prediction accuracy and relatedness (accurately predicted related vs. unpredicted related, vs unrelated - unpredicted related words).

We subjected the RT data to linear mixed-effects models with RT as the dependent measure and fixed effects of condition. We found a significant effect of prediction accuracy (Figure 1). Lexical decisions were faster for related words when they were accurately predicted than when they were unpredicted. However, there was no difference in RT latency between the unpredicted related and the unpredicted unrelated target words, suggesting that there was no effect of semantic matching when the words were not accurately predicted.

If lexical form information is pre-activated as a consequence of successful prediction, we should observe smaller effects of lexical variables such as length for accurately predicted words compared to words that were unpredicted or unrelated. To test this, we included word length along with prediction success vs. failure in an additional LMER. We found a significant interaction of prediction accuracy and length ($b=4$, $SE=1.45$, $\chi^2(1) = 8.24$, $p < .05$). When words were not predicted, we found a main effect of length (driven by very long words > 9chars; mean target length = 5 chars); this effect was not present for successfully predicted words (Figure 2).

In conclusion, we found that prediction success led to faster lexical decision times. Importantly, accurate prediction eliminated word length effects, indicating that the actual word form had been pre-activated prior to presentation of the target words. Hence, anticipation of words during language processing may operate in a similar fashion to word identification and lexical access during reading of words.

Supplemental materials

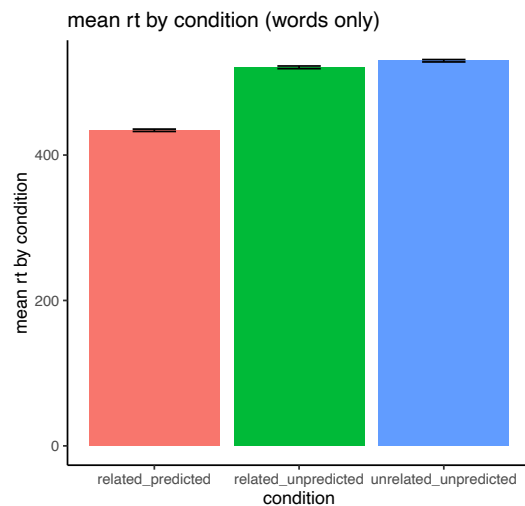


Figure 2. Mean response time as a function of condition.

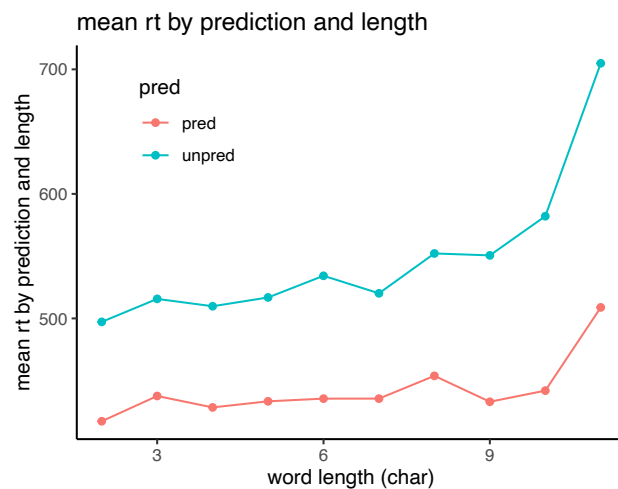


Figure 1. Mean response times as a function of word length for predicted and unpredicted words.

References

- Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, 136, 135–149.
- Dave, S., Brothers, T. A., Traxler, M. J., Ferreira, F., Henderson, J. M., & Swaab, T. Y. (2018). Electrophysiological evidence for preserved primacy of lexical prediction in aging. *Neuropsychologia*, 117, 135–147.
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, 41(4), 469–495.
- Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, 24(2), 232–252.

Putting the pieces together: Two-year-olds hearing an unfamiliar accent recognize known words and learn new words, but do not use known words to learn new words

Alexander LaTourrette (University of Pennsylvania), Cynthia Blanco (Northwestern University), Sandra Waxman (Northwestern University)

By their second birthdays, infants process speech efficiently in their native language and are expert word-learners. They also successfully recognize familiar words and learn new words spoken in unfamiliar accents (Best et al., 2009; Schmale et al., 2011). Here, we ask whether subtler difficulties in processing remain and can affect word learning. Prior work reveals that in native-accented speech, 2-year-olds use known words to infer the meaning of novel words: if infants hear “The dax is sleeping,” they infer “dax” refers to an animal (Ferguson et al., 2018). However, if processing unfamiliar accents remains a challenge for 2-year-olds, as for older children (Bent, 2014), infants may struggle to infer meanings solely from linguistic context.

To test this question, we adopted the eye-tracking paradigm established by Ferguson et al. (2018) with native-accented speech. However, we presented sentences in Spanish-accented speech, an unfamiliar accent for our participants. See Figure 1. Infants ($n=48$) heard dialogues between two speakers featuring familiar nouns (6 trials) and then novel nouns (6 trials). No referents were shown during dialogues. For novel nouns, we varied whether they were presented in an Informative linguistic context with an animacy-restricted verb (e.g., “The dax is sleeping”) or a Neutral linguistic context (“The dax is clean”). At test, infants viewed an animate and inanimate object and were prompted to look to the target noun’s referent. If infants used the Informative verb’s selectional restrictions to infer the referent of the novel noun, they should look more to the animate referent. If the unfamiliar accent posed too great a processing challenge, then performance should resemble the Neutral condition.

In Experiment 1, 24-month-olds ($M=23.79$ mo, $SD=.71$) successfully identified the referents of familiar nouns in Spanish-accented speech, $p<.01$. However, they did not use familiar verbs to learn novel nouns. A cluster-based permutation test revealed that 24-month-olds failed to use the Informative context to learn novel nouns: unlike previous native-accent conditions, looking patterns in the Informative and Neutral conditions did not significantly diverge, $p>.5$ (Figure 2).

To assess whether infants’ difficulties with the unfamiliar accent truly stemmed from the challenge of using the linguistic context, not simply learning words, we conducted Experiment 2. The task was identical to Experiment 1’s Informative condition, except the referent for each novel noun was present during the dialogue. Two-year-olds ($n=24$) in this Co-present Referent condition successfully learned novel words: performance diverged from the Neutral control condition, $p=.01$, with infants looking more to the animate referent 550ms to 1250ms after noun onset. Thus, infants learned novel words in an unfamiliar accent when a co-present referent was available. Infants’ performance in the Co-Present Referent condition was also predicted by their preference for the target on familiar noun trials, $r(20)=.46$, $p=.032$. Success in comprehending familiar words across accents is thus associated with success in learning new ones.

These findings reveal a nuanced developmental trajectory for processing unfamiliar accents. While 2-year-olds both recognize and learn words in unfamiliar accents—and these skills are inter-related—they still struggle in using known words to learn new ones. This may reflect difficulties in online sentence processing in unfamiliar accents or limits on infants’ willingness to make semantic inferences from unfamiliar accented speech. These findings also cohere well with older children and adults’ continued difficulties in processing unfamiliar accents.

References

- Bent, T. (2014). Children's perception of foreign-accented words. *J. Child Lang*, 41(6), 1334-1355.
- Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of phonological constancy: Toddlers' perception of native-and Jamaican-accented words. *Psychol. Sci.*, 20(5), 539-542.
- Ferguson, B., Graf, E., & Waxman, S. R. (2018). When veps cry: Two-year-olds efficiently learn novel words from linguistic contexts alone. *Lang. Learn. Dev.*, 14(1), 1-12.
- Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Dev. Sci.*, 15(6), 732-738.

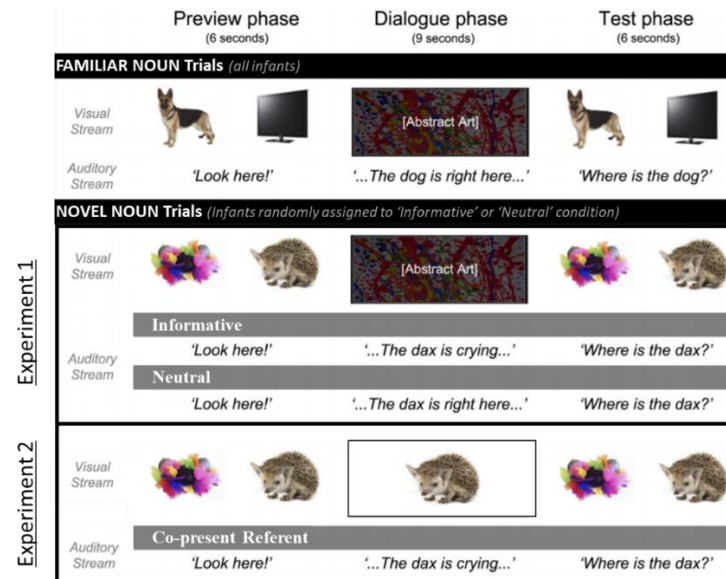


Figure 1. Experimental design. All infants began with 6 Familiar Noun trials, featuring known objects and words. This also provided 2 minutes of exposure to the unfamiliar accent. Next, infants saw 6 Novel Noun trials, with the learning context determined by condition. The dependent variable was the proportion of looking directed to the target object during test.

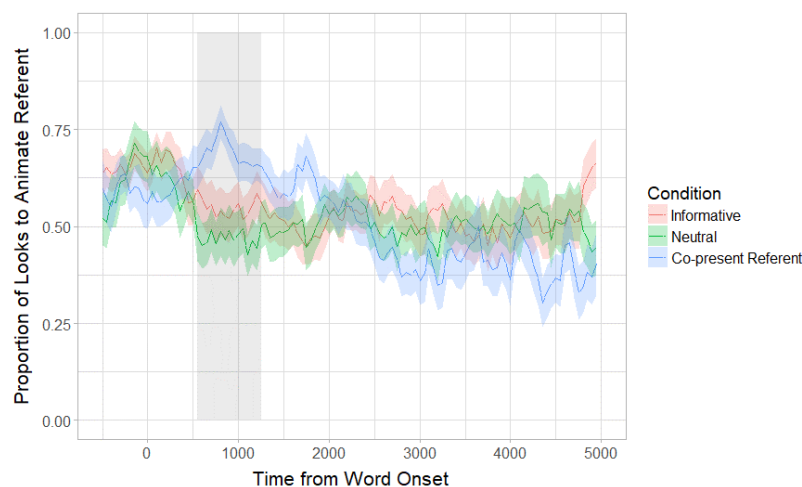


Figure 2. Test timecourse. In the Informative and Neutral conditions, looking patterns did not significantly differ, $t_{\text{cumulative}} < 5$, $p > .5$. However, the Co-Present Referent condition significantly diverged from the Neutral condition 550ms to 1250ms after word onset, $t_{\text{cumulative}} = 38.05$, $p = .01$.

Inside the *wug*-test: phonological well-formedness and processing costs

Canaan Breiss (University of California, Los Angeles)

Introduction: Recent phonological research has focused on the role of lexical storage as a way to explain unexpected morpheme-specific deviations from grammar-wide phonological principles (Zuraw 2000, 2007, 2015; Moore-Cantwell & Pater 2016; Moore-Cantwell & Smith 2017; Zymet 2018, 2019). This implies a feed-forward relationship between grammar and lexicon in production: the phonological forms of morphemes are retrieved, along with optional item-specific information, and then the phonological grammar combines the morphemes subject to a set of general well-formedness principles, overridden only by lexically-specific information. This paper presents evidence for a bidirectional relationship between lexicon and phonological grammar, focusing on a phenomenon known as Lexical Conservatism (Steriade 1997). Lexical Conservatism describes scenarios in which a novel form (the Derivative (D), ex., *compensable*) unexpectedly undergoes a phonologically-motivated (markedness-improving) change to the Local Base (B_L) which would not otherwise be possible (ex., rightward stress shift, as in *cómpensate* + *-able* → *compénsabe*, **cómpensable*, while *ínundate* + *-able* → *ínundable*, **ínúndable*). Steriade argues that this behavior depends on the presence of a phonologically-advantageous morphologically-related word (the Remote Base (B_R); here the final-stressed root allomorph in *compéns-atory* exists but **ínúnd-X* does not). This theoretical explanation makes strong psycholinguistic claims about the relationship between lexicon and grammar, suggesting the phonology can “recruit” related forms from the lexicon in real time.

Exp. 1 replicated and extended Steriade’s original survey. 31 subjects were asked to read aloud 120 sentences where a B_L was presented alongside a D formed by attaching one of the affixes *-able*, *-ity*, and *-ism* (as in figure 1). Half the B_Ls had phonologically advantageous B_Rs. Afterwards, subjects completed a *knowledge check* where they were asked to read aloud and indicate whether they knew each of the B_Ls they had seen, as well as the B_Rs for the half of B_Ls which had them. The dependent variable was stress placement in the D relative to that subject’s production of B_L and B_R. Analysis was carried out using Bayesian hierarchical logistic regression; here I discuss findings for which there is greater than 95% certainty of a true effect.

Results: The effect of an individual subject knowing the relevant B_R increased the likelihood that a D had stress placement mismatching B_L. We also observe phonological determinants of stress placement (figure 2). Exp. 1 supports Steriade’s informal survey results and demonstrates that the form of the D is causally related to the presence of the B_R, but the effect is probabilistic, and interacts with purely-phonological principles of stress placement.

Exp. 2 extends Exp. 1 and incorporates a priming manipulation. If the findings of Exp. 1 are due to the presence of B_Rs in individual speakers’ lexicons, we might expect the strength of the effect to be moderated by lexical characteristics of the B_R such as frequency and semantic similarity between B_L and B_R, and the influence of the B_R should be able to be increased by making it more salient to the speaker before they create the D from the B_L. 30 new subjects participated in an experiment with a similar design as Exp. 1 which included 40 B_Ls, half with B_Rs, fully crossed with affixes *-able* and *-ic*. Procedure followed Exp. 1, except that the *knowledge check* for half of the B_Rs (counterbalanced across subjects) preceded the D formation task, thus priming the B_R for when its B_L was encountered during the experiment. Data annotation and modeling followed Exp. 1. **Results:** As in Exp. 1, both lexical (knowing the B_R) and phonological (syllable weight, secondary stress) factors influenced D stress placement. Focusing on those B_Ls for which the B_R was known, we observe that a primed B_R exerted a greater effect, and this interacted with semantic similarity (figure 3). These facts suggest an architecture where the phonological grammar can “recruit” non-local phonological allomorphs (B_Rs) in real time, implying a dynamic trading relationship between processing effort in retrieving a second non-local form and potential gain in phonological well-formedness by doing so. This is not compatible with strictly feed-forward assumptions, since the data show effects of optimizing both for lexical and phonological factors, but is integrable with Levelt (1993)’s production model.

“An ideology centered on *illustrating* could be called illustrism”

Figure 1: Example of a carrier sentence used in Exp. 1. The B_L is italicized, and the D is underlined.

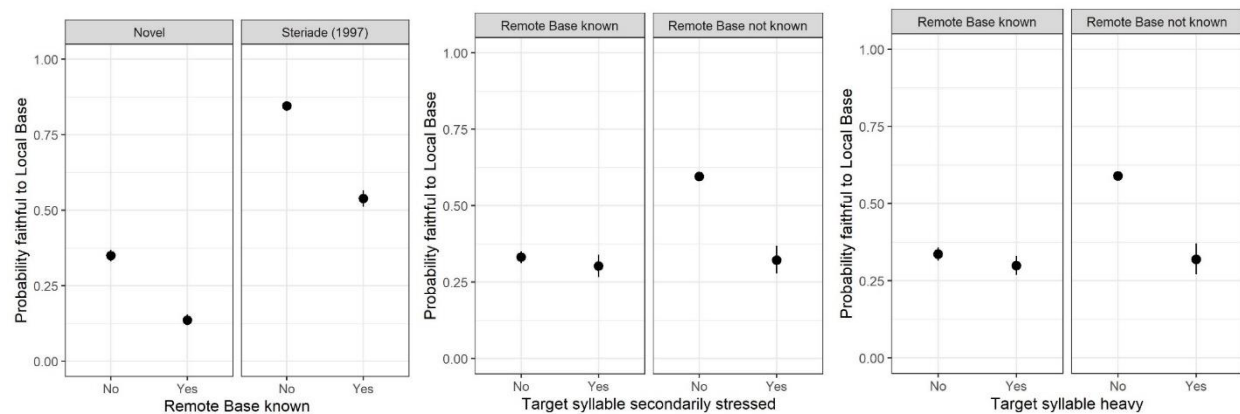


Figure 2: Partial results of Experiment 1, mean and standard error in each plot. The leftmost panel plots the probability of Derivative stress matching B_L stress as a function of whether the B_L was from Steriade (1997)’s original study, or novel for Experiment 1. The center panel plots the intersection of whether the B_R was known to an individual subject with whether the target syllable bore secondary stress (*no* as in *métá* vs. *yes* as in *ínsect*). The rightmost panel plots the intersection of whether the B_R was known to an individual subject with whether the target syllable was heavy (*no* as in *drama* vs. *yes* as in *ballast*).

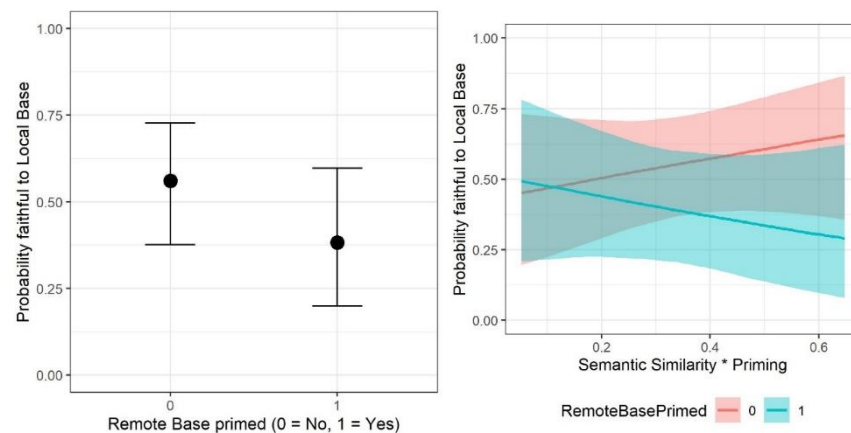


Figure 3: Marginal means and 95% Credible Intervals from the Bayesian hierarchical regression model in Exp. 2. Left panel indicates that Derivatives with primed B_Rs are more likely to be unfaithful in stress placement to their B_L. Right panel plots the interaction of priming with the semantic similarity between B_L and B_R, estimated by using the cosine similarity of their word embeddings in a Word2Vec neural network, normalized to the 0 (less similar) -1 (more similar) interval.

Selected References

Levelt, W. J. (1993). Speaking: From intention to articulation (Vol. 1). MIT press. **Steriade, D. (1997).** Lexical conservatism. *Linguistics in the morning calm*, 157-179. **Zymet, J. (2018).** Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning (Doctoral dissertation, UCLA).

Talking, like, a Valley Girl? Online Processing of Sociolinguistic Cues

Daisy Leigh, Judith Degen, Robert J. Podesva (Stanford University)

Listeners form impressions about a speaker's social persona both from *what* they say and *how* they say it [7,8]. How and when do we do this? Sociolinguistic work has long shown that the social meanings associated with specific phonetic variants ('cues') contribute to listeners' perceptions of speaker persona in offline judgments: for example, speakers may sound more 'casual' and 'unprofessional' when they use *-in'* rather than *-ing* (e.g., *talkin'* vs. *talking*) [4,9]; or more 'excitable' or 'like a Valley Girl' when they use High Rising Terminals (HRT, aka 'uptalk') rather than declarative prosody [13]. But the effects of these cues are not fixed or absolute; *-in'* might indicate an 'unprofessional' persona in some voices but not others, for example [3,10]. This suggests that sociophonetic cues compete with other information in the speech signal; listeners must integrate the meaning contribution of sociophonetic cues with all the other social impressions that arise when hearing someone talk. As yet, little is known about how or when this happens: while most psycholinguistic work investigating the online processing of linguistic cues has focused on listeners' inferences about upcoming linguistic *material* (e.g., [1,6,11]), very little attention has been paid to listeners' unfolding inferences about the *speaker* (though see e.g., [2,7,14]). We conducted two eye-tracking studies to investigate listeners' online and offline uptake of two sociophonetic cues: *in'* (Exp.1) and HRT (Exp.2).

Design: In a 2AFC visual world paradigm, participants heard a stimulus and selected the speaker they thought produced it. The two speakers were representations of different personas: a Tough and a Valley Girl (Fig.1). We measured participants' persona selections and eye movements. On critical trials, participants in Exp.1 (N=160) heard *-in'* and *-ing* cues; those in Exp.2 (N=152), HRT and declaratives. Participants heard stimuli produced by four different voices in each experiment. Experiments were conducted online on Prolific, using Webgazer.js [10] to capture gaze data. **Predictions:** The Tough and Valley Girl images were normed to ensure they captured similar meanings to those reported for *-in'* (e.g. 'chill, unprofessional') vs. HRT (e.g., 'excited, feminine'). We therefore expected listeners to look towards, and select, Toughs more often after hearing *-in'* (vs. *-ing*) in Exp.1, and less after hearing HRT (vs. declaratives) in Exp.2. **Results:** Both cues modulated participants' offline judgements of speaker persona in the predicted directions, regardless of how Tough or Valley Girl the four different voices sounded overall (Fig.3 and 5). They also modulated online behavior: from the 800-900ms window after cue onset onwards, participants were more likely to look at the Tough image if they had heard *-in'* rather than *-ing*. In Exp.2, participants were significantly *less* likely to look at the Tough image after hearing HRT rather than a declarative, from the 1000-1100ms window onwards. *Overall* Tough/Valley bias for each voice was also reflected in looking patterns: e.g., in Exp.1, participants initially looked more to the Valley Girl persona when listening to Voice 4, but the presence of the *-in'* cue biased them towards the Tough interpretation (Fig. 4). **Discussion:** Our results suggest that participants processed both cues probabilistically by weighing the meaning contributions of each against their existing expectations about the speaker. Online cue uptake was observed much later than the 200ms typically allocated to executing signal-driven eye movements [1]. Given the sparsity of existing work, we can only speculate on the reasons: it is possible that listeners simply take longer to process phonetic cues' social meanings than their purely referential ones. Alternatively, these cues may be weak or less reliable cues to social identity; stronger/more reliable cues may result in faster online integration. Despite the delayed online effects, our results qualitatively (Figs.3-6) point to HRT having stronger biasing effects on interpretation than *-in'*, indicating that these cues vary with respect to their relative social informativity. These considerations raise exciting questions regarding the role of cue strength, reliability, and timing in the online integration of social and denotational information. We consider the current findings a promising starting point for future empirical work examining the online processing of sociophonetic cues.

Fig.1: Example Persona Images



Fig.2: Example stimuli

Exp.1		Exp.2	
-in'	I'm talkin' about the beam.	HRT	I'm talking about the beam
-ing	I'm talking about the beam.	Declarative	I'm talking about the beam.

All critical items took the form 'I'm talking about the x', where x is a monosyllabic word. Participants heard 16 different items: 8 -in' items, 8 -ing in Exp.1, 8 HRT and 8 Declarative items in Exp.2. Stimuli were identical across conditions other than the differences outlined in Fig.2, and the same stimuli were used for the -ing and Declarative conditions in Exps. 1 and 2. We used existing utterances from the NSP corpus [5] and manipulated them to include the cues of interest.

Statistical details

For persona selections, we fit logistic regression models predicting log-odds of selecting a Tough persona given the presence of an -in' (Exp.1) or HRT cue (Exp.2). For both models, there was a main effect of cue ($\beta=0.57$, $p<0.001$ in Exp.1, $\beta=-0.80$, $p<0.001$ in Exp.2). For eye-tracking data, we fit logistic regression models predicting log-odds of looking at the Tough (vs. Valley Girl) in each 100ms window after cue onset, given the cue heard. (We took 200ms either side of cue onset as a baseline period against which looks in subsequent windows were compared). In Exp.1, the earliest window that the presence of -in (vs. ing) predicted a significant increase in Tough looks was 800-900ms ($\beta=0.29$, $p<0.05$). In Exp.2, the earliest window where HRT predicted a significant decrease in Tough looks was 1000-1100ms ($\beta=-0.34$, $p<0.05$). For all models, we included the maximal random effects structure justified by the data.

Figs.3-6 are ordered by 'Toughest' to most 'Valley Girl' sounding speaker; in both experiments, the voices were heard in random order. Error bars represent bootstrapped 95% confidence intervals. In Fig.4 and 6, black vertical line represent cue onset, and pink lines, audio offset.

Fig.3: Exp.1, Persona selections

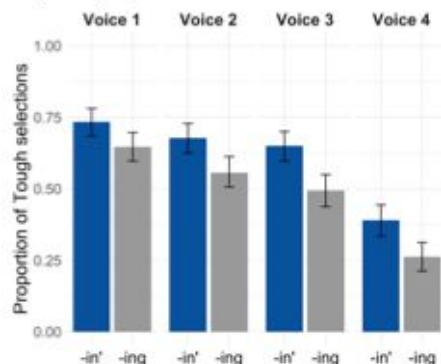


Fig.4: Exp.1, Eye-tracking data

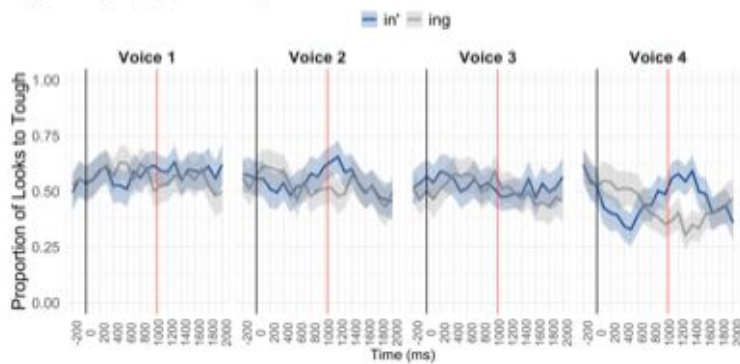


Fig.5: Exp.2, Persona selections

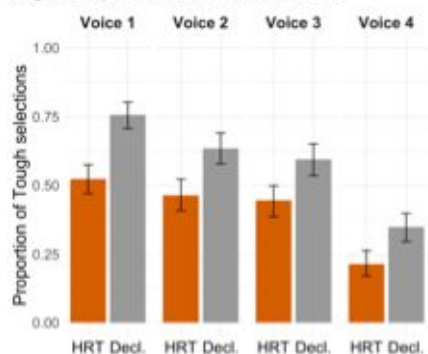
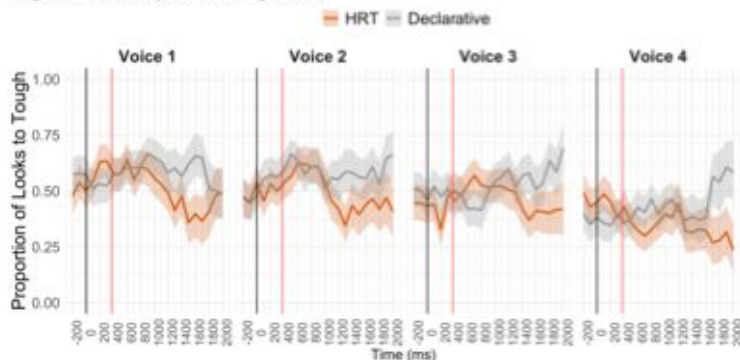


Fig.6: Exp.2, Eye-tracking data



References: [1] Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*. [2] Austen, M. & Campbell-Kibler, K. (2020). Eye-tracking for sociolinguistic perception. Poster at *Annual Meeting of the Linguistic Society of America*. [3] Campbell-Kibler, K. (2008). I'll be the judge of that: Diversity in social perceptions of ING. *Language in Society*. [4] Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change*. [5] Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*. [6] Degen, J. & Tanenhaus, M.K. (2016). Availability of Alternatives and the Processing of Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science*. [7] D'Onofrio, A. (2018). Controlled and automatic perceptions of a sociolinguistic marker. *Language Variation and Change*. [8] D'Onofrio, A. (2018). Personae and phonetic detail in sociolinguistic signs. *Language in Society*. [9] Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M. & Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics*. [10] Levon, E. & Fox, S. (2014). Social Salience and the Sociolinguistic Monitor: A Case Study of ING and TH- fronting in Britain. *Journal of English Linguistics*. [11] McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomics Bulletin Review*. [12] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. [13] Tyler, J.C. (2015). Expanding and Mapping the Indexical Field: Rising Pitch, the Uptalk Stereotype, and Perceptual Variation. *Journal of English Linguistics*. [14] Weissler, R. E. & Brennan, J. R. (2020) How do Listeners Form Grammatical Expectations to African American Language?, *University of Pennsylvania Working Papers in Linguistics*.

Structural Priming and Non-Native Language Processing

Douglas J. Getty, Scott H. Fraundorf

Contemporary theories generally hold that comprehension is fast and accurate because it draws on experience with the linguistic environment [e.g., 1, 2]. But, variability in spoken language can pose a challenge by rendering prior experience unhelpful or even misleading [3]. To study how comprehenders overcome this challenge, we examined differences in listening to native (Nat) vs. non-native (NN) speakers. NN speakers typically do not achieve Nat-like proficiency [4], resulting in accented speech and higher rates of syntactic errors. Past findings suggest that comprehension of NN speech may be underspecified relative to Nat speech [5,6,7]. We contrast two accounts of why this may occur: (a) An expectation account whereby listeners expect lower linguistic competence in NN speakers, and thus adaptively rely less on the literal speech input and more on top-down methods of comprehension [5]. This account predicts that reduced proficiency should *always* lead to less reliance on bottom-up input. (b) A “good enough” account assumes that listeners optimize comprehension resources to the task goal [8]. This account predicts that listeners *may* use heuristic-based processing to comprehend NN speech but *can* process it deeply if prompted (e.g. if a low proficiency NN speaker requires more resources).

Method. We adapted the **classic syntactic-priming paradigm** [9] to the online Qualtrics platform. In each of 48 critical trials, participants first heard a dative prime sentence in either the prepositional-object structure (PO, 1a in Table 1) or double-object structure (DO, 1b). Immediately after, they typed a description of an unrelated dative-eliciting image. All prime-target pairs were pseudo-randomly embedded within a list of 144 unrelated filler sentences and images. Picture descriptions were coded as either a PO, DO, or OTHER and analyzed using linear mixed-effects regression models separately for each block (see Table 2). Comprehension was assessed on 48 filler sentences to ensure attention remained on-task.

Results Exp 1 (N=128). Speaker was manipulated within-subjects in a blocked design, such that half participants heard 24 primes spoken by a NN speaker (L1 Mandarin) in Block 1, followed by 24 primes spoken by a Nat speaker (Block 2), or the reverse counterbalanced order. In each block, we **replicated the classic syntactic-priming effect** (main effect of Prime, increased probability of producing POs after PO primes). Critically, in Block 1, we also observed a negative interaction between Speaker and Prime, driven by a **reduced priming effect in the NN speaker condition** (Table 2; Figure 1a). The reduction in priming does not appear to be driven by reduced attention to the task (*M* comprehension accuracy = 99% for both Nat and NN conditions), but rather a mode of processing that relies less on structural information, consistent with both the “expectation” and “good enough” accounts. Interestingly, this reduced priming **carried over into Block 2** such that participants who were primed less by the NN speaker in Block 1 were also primed less by Nat speaker in Block 2, and vice versa (Figure 1b).

Results Exp 2 (N=114). Exp. 2 used the same design, except both talkers were NN (L1 Mandarin), and one was manipulated to be less proficient by introducing ungrammaticalities on 30% of the filler sentences (e.g. Table 1, 2a-2b). In Block 1, there was a main effect of Prime and a positive interaction, suggesting **the less proficient speaker elicited more priming** (Table 2; Figure 2a). By Block 2, the main effect remained, but there was no difference in priming between the less proficient and more proficient conditions (Table 2; Figure 2b).

Discussion. We found that **NN speaker status and NN speaker proficiency influence priming**. In Exp. 1, decreased priming for the NN speaker in Block 1 carried over into subsequent processing of the Nat speaker in Block 2, while those exposed to the Nat speaker first showed consistent priming throughout. While both accounts predict the reduced priming effect, the carryover effect is explained by the “good enough” principle that processing adapts to the task. In Exp. 2, proficiency does influence priming; however, given the direction of the interaction, the results are again more consistent with the “good enough” account that allows for increased resources to processing if the task demands. We suggest that listeners’ comprehension of NN speech reflects **contextually optimized processing strategies rather than an intrinsic reliance on top-down comprehension** when processing NN speech.

References

[1] Levy (2008). *Cognition*. [2] MacDonald et al. (1994). [3] Liberman et al., (1967). [4] Birdsong & Molis (2001). *JML*. [5] Lev-Ari (2015). *Front. Psych.* [6] Hanulíková et al. (2012). *JCN*. [7] Gibson et al. (2017). *Psych Sci*. [8] Karimi & Ferreira, (2016). *QJEP*. [9] Bock (1986). *Cog Psych*.

Table 1. Example sentences. Primes like 1a-1b were used in both experiments, while ungrammatical sentences like 2a-2b were only used in Experiment 2.

Prime Sentences (Exp 1 and Exp2)

(1a) Dative {PO}: The man gave the toy to his daughter.

(1b) Dative {DO}: The man gave his daughter the toy.

Ungrammatical Fillers (Exp 2)

(2a) The janitor **is clean a floors** daily.

(2b) The union leader **is assist a workers** in **organize** the strike.

Table 2. Fixed-effects test of priming in each model. A near-maximal random-effects structure was used in each model. Estimates reflect the probability (in logits) of producing a PO in each condition (descriptions coded as OTHER were dropped for analysis).

		<i>Experiment 1</i>				<i>Experiment 2</i>			
		Beta	SE	z	p	Beta	SE	z	p
Block 1	(Intercept)	0.24	0.22	1.11	.27	(Intercept)	-0.19	0.23	.400
	Prime	0.47	0.11	4.25	<.001	Prime	0.63	0.12	<.001
	Speaker	0.25	0.32	0.80	.42	SpeakProf	-0.10	0.27	.71
	Prime:Speaker	-0.67	0.22	-3.01	<.01	Prime:SpeakProf	0.78	0.24	<.01
Block 2	(Intercept)	0.04	0.19	0.23	.82	(Intercept)	-0.69	0.21	<.01
	Prime	0.57	0.12	4.82	<.001	Prime	0.66	0.12	<.001
	Speaker	-0.41	0.31	-1.32	.19	SpeakProf	-0.21	0.36	0.57
	Prime:Speaker	0.72	0.24	3.06	<.01	Prime:SpeakProf	-0.35	0.23	0.13

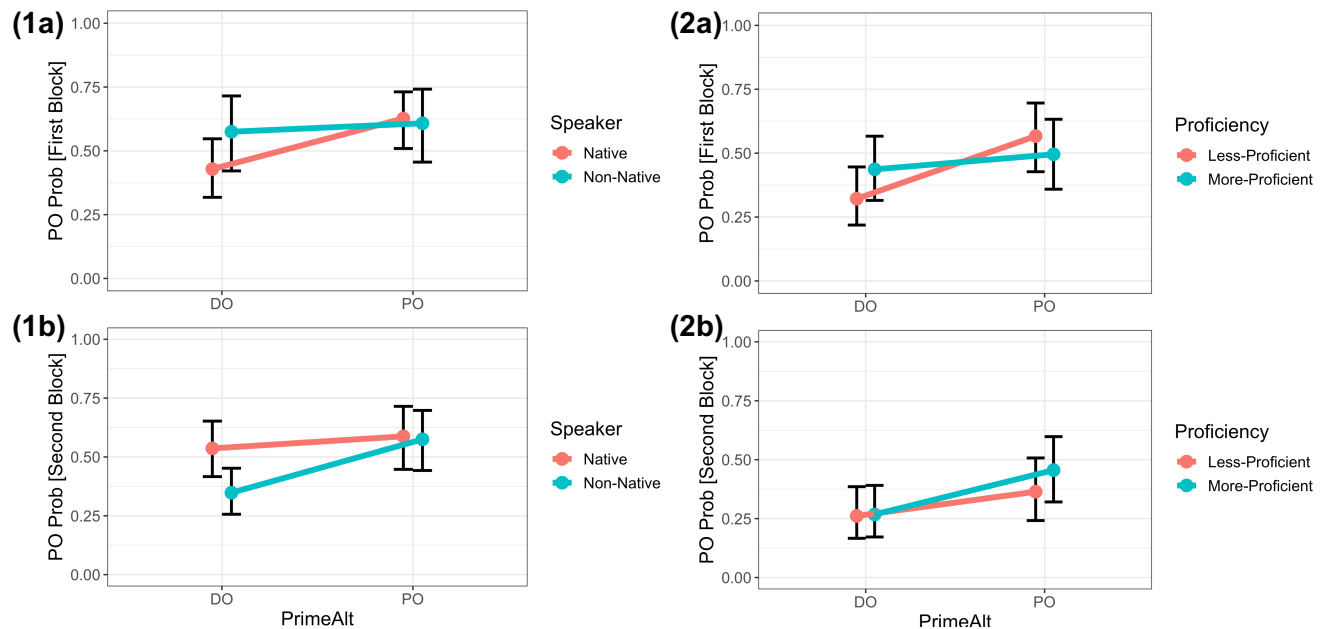


Figure 1a-2b. Points represent model-estimated marginal means, transformed from logits to probability. Errors bars represent 95% confidence intervals.

Pre-schoolers process word onsets and codas similarly: A time-course analysis

Rosanne Abrahamse, Nan Xu Rattanasone, Katherine Demuth & Titia Benders (Macquarie University).

Words can contain minimal phonemic differences in onset (*bin-pin*) or coda positions (*map-mat*). Being able to recognise these words and distinguish them from alternatives is critical for understanding other people. Infants and toddlers process word onsets quickly, recognising the first part (e.g. /beɪ:/) of a word as quickly and accurately as the whole word (e.g. /beɪ:bi/ (*baby*)) [1], and resolving the onset of the word before the coda of the word [2]. Less is known about children's processing of codas. In production, codas are typically acquired later than onsets [3]. In perception, adults confuse newly learned coda minimal pairs more frequently than onset pairs [4]. These findings suggest that children may process codas less efficiently than onsets. However, children (by 1;6), as well as adults, can detect mispronunciations in both onsets and codas with a roughly comparable speed [5]. Although this suggests similar processing for onsets versus codas, the processing time-courses of onsets and codas were not directly compared. The aim of the present study is, thus, to directly compare children's processing of onsets and codas in real words, and to investigate how this relates to adult processing.

Processing was compared between onset and coda minimal pairs (e.g. *bin-pin* vs. *map-mat*). A direct comparison was made by aligning the processing time-courses for onsets and codas at comparable disambiguation points in the acoustic signal (i.e., onset vs. coda burst). If children's later acquisition of codas in production and adults' less accurate learning of coda words are linked to slower processing of codas, we would expect slower processing time for codas than onsets. However, if children and adults process codas as rapidly as onsets, we expect no differences in processing time. Adults are expected to be faster overall than children [2, 5].

Seven Australian English (AuE) speaking adults ($M_{age} = 31$ years; range 20-41; 4 males) and 28 AuE speaking children ($M_{age} = 4.6$ years; range 3.2-5.8; 16 males) participated in an eye-tracking study with a Looking-While-Listening paradigm [6]. The stimuli consisted of 30 minimal pairs (18 onset trials, 12 coda trials), with voicing and place of articulation contrasts (Table 1). The session began with a picture naming task to familiarise participants with the stimuli. Then, during the eye-tracking task, participants were shown two pictures for 2000 ms, representing a minimal pair. They heard 'Look at the X', after which the pictures remained on screen for a further 4000ms. Onset and Coda trials were blocked with order of presentation counterbalanced across participants. We calculated proportion of looks to the target. Differences between looking curves for Onsets vs. Codas (burst-aligned) were analysed across groups (Adults vs. Children) using cluster-based permutation tests [7]–[9] (Figure 1). Analyses were performed over a -500 to +2000ms window to take into account transitional cues in the preceding vowel.

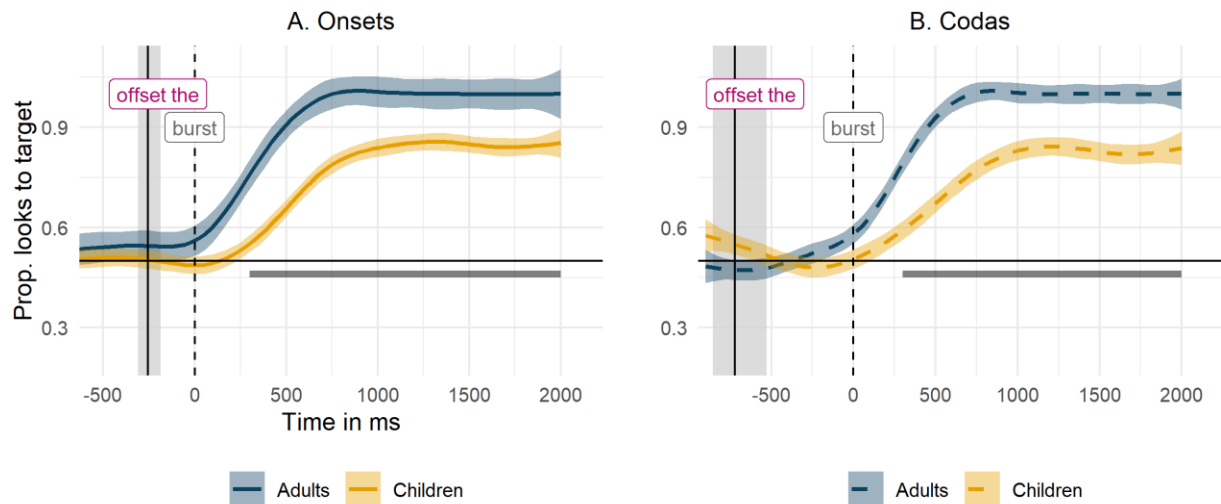
The results did not support the hypothesis that children (or adults) would process onsets faster than codas, as no significant differences were found when comparing Onset vs. Coda curves for either group. However, significant differences were found between Adult vs. Child curves for both Onsets (time-window: 300-2000ms, Monte Carlo $p < 0.001$) and Codas (time-window: 300-2000ms, Monte Carlo $p < 0.001$). This indicates that adults looked significantly more towards the target than children from 300ms until 2000ms after the burst.

In line with [5], these results provide direct evidence that pre-schoolers process codas as fast as onsets, albeit more slowly than adults. This suggests that even though word processing speed increases with age, the mechanisms to process the beginnings and ends of words rapidly are in place early in development. Overall, these findings provide an important baseline to test the word processing speed of children with slow language processing, e.g., those with hearing loss or developmental language disorders.

Table 1. Example phonetic contrasts (number) used for onset and coda minimal pairs. PoA = Place of Articulation

	Onsets	Codas	Examples
PoA (n=20)	b/g(2) p/k(2) b/d(2) p/t(2) d/g(1) t/k(3)	b/g(1) d/g(1) p/k(1) p/t(2) t/k(3)	boat-goat, pea-key, bow-dough rub-rug, mud-mug, cape-cake pen-ten, date-gate, tape-cape map-mat, net-neck
Voicing (n=10)	b/p(2) d/t(2) g/k(2)	d/t(2) g/k(2)	bin-pin, deer-tear, goat-coat seed-seat, log-lock

Figure 1. Time-course of looks to target for Onsets (A) and Codas (B), aligned at the start of the respective stop burst. Curves smoothed using general additive model curve fitting (with 95% confidence intervals). ‘Offset the’ indicates mean, minimum and maximum offset of ‘the’ in carrier sentence ‘look at the X’ Grey horizontal bars mark statistically significant time-windows.



References

- [1] A. Fernald, D. Swingley, and J. P. Pinto, “When Half a Word Is Enough: Infants Can Recognize Spoken Words Using Partial Phonetic Information,” *Child Dev.*, vol. 72, no. 4, pp. 1003–1015, 2001.
- [2] D. Swingley, J. P. Pinto, and A. Fernald, “Continuous processing in word recognition at 24 months,” *Cognition*, vol. 71, no. 2, pp. 73–108, 1999.
- [3] K. Demuth, J. Culbertson, and J. Alter, “Word-minimality, epenthesis and coda licensing in the early acquisition of english,” *Lang. Speech*, vol. 49, no. 2, pp. 137–174, 2006.
- [4] S. C. Creel, R. N. Aslin, and M. K. Tanenhaus, “Acquiring an artificial lexicon: Segment type and order information in early lexical entries,” *J. Mem. Lang.*, vol. 54, no. 1, pp. 1–19, 2006.
- [5] D. Swingley, “Onsets and codas in 1.5-year-olds’ word recognition,” *J. Mem. Lang.*, vol. 60, no. 2, pp. 252–269, 2009.
- [6] A. Fernald, R. Zangl, A. L. Portillo, and V. A. Marchman, “Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children,” *Dev. Psycholinguist. On-line methods Child. Lang. Process.*, no. 2008, pp. 97–135, 2008.
- [7] E. Maris and R. Oostenveld, “Nonparametric statistical testing of EEG- and MEG-data,” *J. Neurosci. Methods*, vol. 164, no. 1, pp. 177–190, 2007.
- [8] B. Ferguson and J. Dink, “Package ‘eyetrackingR,’” 2018.
- [9] K. Tamási, C. McKean, A. Gafos, and B. Höhle, “Children’s gradient sensitivity to phonological mismatch: Considering the dynamics of looking behavior and pupil dilation,” *J. Child Lang.*, vol. 46, no. 1, pp. 1–23, 2019.

Having to predict a (native or non-native) partner's utterance increases adaptation in L2

Theres Grüter (U. of Hawai'i), Yanxin (Alice) Zhu (U. of Hawai'i), Carrie N. Jackson (Penn State U.)

Effects of structural priming and adaptation have been argued to arise as a result of the computation of prediction error (ChangEtAl2006, Jaeger&Snider2013). Top-down factors such as explicit instructions to predict (BrothersEtAl2017) and social characteristics of the interlocutor (WeatherholtzEtAl2014) have been shown to modulate the size of prediction and priming effects. Within the context of second language (L2) acquisition, the view of priming as an (implicit) learning mechanism has led to the exploration of structural priming as a tool for L2 learning (McDonoughEtAl2015) and offered a potential theoretical framework for more unified study of L2 processing and learning (Jackson&Hopp 2020). Yet while of immediate relevance to applied and theoretical goals in L2 acquisition, the modulating roles of top-down factors such as explicit prediction and speaker characteristics on L2 priming and adaptation remain largely unexplored. We present evidence from two written production priming experiments with Korean L2 learners of English, focusing on double-object datives, to address the following questions:

RQ1: Do task instructions to predict a partner's utterance increase effects of (i) immediate priming, and (ii) longer-term adaptation as measured by change from baseline to posttest?

RQ2: Do the partner's social and linguistic status as a native or non-native speaker affect the size of (i) immediate priming, and (ii) longer-term adaptation?

Method. In both experiments, participants in the 'guessing-game' (GG) group (Exp1: $n=18$, Exp2: $n=27$) had to predict how a virtual partner would describe a picture prior to seeing the actual prime sentence, which they then evaluated as the same or different from their initial guess (Fig1). This manipulation was intended to explicitly induce prediction and computation of prediction error. Participants in the control group (CC; Exp1: $n=17$, Exp2: $n=26$) only re-typed the prime sentence in a standard repetition priming procedure (Fig2). The virtual partner consistently used double-object datives (DOs: *The girl fed the squirrel a nut*) with ditransitives, thus priming and adaptation should manifest in terms of increased use of DOs compared to prepositional datives (POs: *The girl fed a nut to the squirrel*), the strongly preferred construction for Korean learners (Kaan&Chun2018). The partner was presented as a native speaker of English ('Jessica') in Exp1 and as a Korean learner of English ('Soo-Min') in Exp2. In a baseline-priming-posttest design (Table1), participants alternated between repeating(CC)/guessing(GG) the partner's picture descriptions (primes) and describing pictures themselves (targets).

Results. Mixed logit models showed increases in DO production from baseline to priming phase in both experiments ($bs>2$, $ps<.001$; Fig3). While effects appeared numerically larger in GG vs CC groups, interactions with group were not robust (Exp1: $b=1.32$, $p=.06$; Exp2: $b=.52$, $p=.3$). Yet group significantly modulated change from baseline to posttest (Exp1: $b=1.62$, $p=.03$; Exp2: $b=1.31$, $p=.006$), with GG participants continuing to produce DOs more frequently than CC participants. While priming effects were numerically smaller in Exp2 than Exp1, experiment did not emerge as a robust modulator in a combined analysis of data from both experiments.

Discussion. In both experiments, explicit instructions to predict a partner's utterance (**RQ1**) led to greater adaptation in terms of change from baseline to posttest. Notably, the effect of this manipulation (GG/CC) only became robust in the posttest, suggesting it affected longer-term adaptation, or learning, more strongly than short-term activation of a primed structure. Future studies including delayed posttests will need to examine the longevity of this effect, yet this finding presents preliminary evidence to suggest that applied approaches seeking to use priming as a tool for L2 learning may benefit from incorporating a forced prediction or guessing component. Meanwhile, no clear evidence for modulation of L2 priming by social factors (**RQ2**) emerged. This is unexpected in light of findings showing native speakers adapt more to talkers using a more standard variety (WeatherholtzEtAl2014), but aligns with the only previous study of social factors in L2 structural priming (Chun&Kaan2020), which suggested such effects may be more complex than predicted by models based on data from native language processing.

References

- Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, 93, 203–216.
- Chang, F., Dell, G., & Bock, K. (2006) Becoming syntactic. *Psychological Review*, 113, 234–72.
- Chun, E., & Kaan, E. (2020). The effects of speaker accent on syntactic priming in second-language speakers. *Second Language Research*. <https://doi.org/10.1177/0267658320926563>
- Jackson, C. N., & Hopp, H. (2020). Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism*, 24, 895–911.
- Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation. *Cognition*, 127, 57–83.
- Kaan, E., & Chun, E. (2018). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition*, 21, 228–242.
- McDonough, K., Neumann, H., & Trofimovich, P. (2015). Eliciting production of L2 target structures through priming activities. *Canadian Modern Language Review*, 71, 75–95.
- Weatherholtz, K., Campbell-Kibler, K., & Jaeger, T. F. (2014). Socially-mediated syntactic alignment. *Language Variation and Change*, 26, 387–420.

Table 1. Experiment design. (NB: no lexical boost)

Phase	Experimental items
	# and structure of prime-target pairs
Baseline	6 prime: (in)transitive target: ditransitive
Priming	8 prime: ditransitive: DO target: ditransitive
Posttest	6 prime: (in)transitive target: ditransitive

Figure 2. Prime trial, CC condition (Exp1)



Figure 1. Prime trial, GG condition (Exp1); sample participant responses in blue

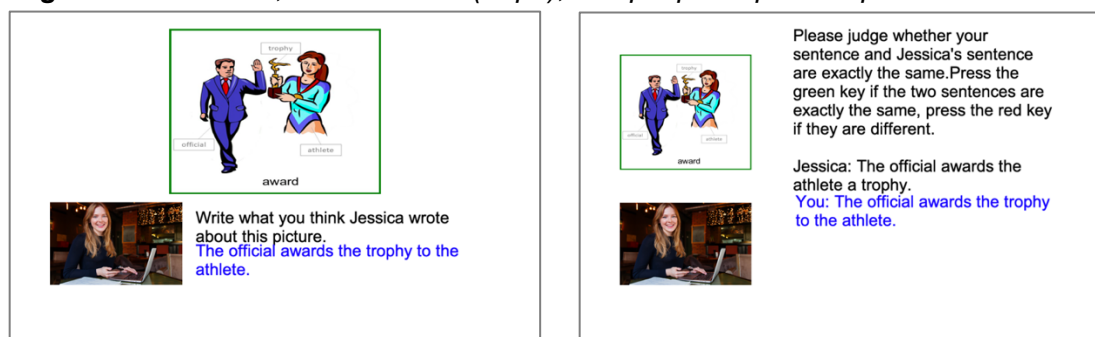
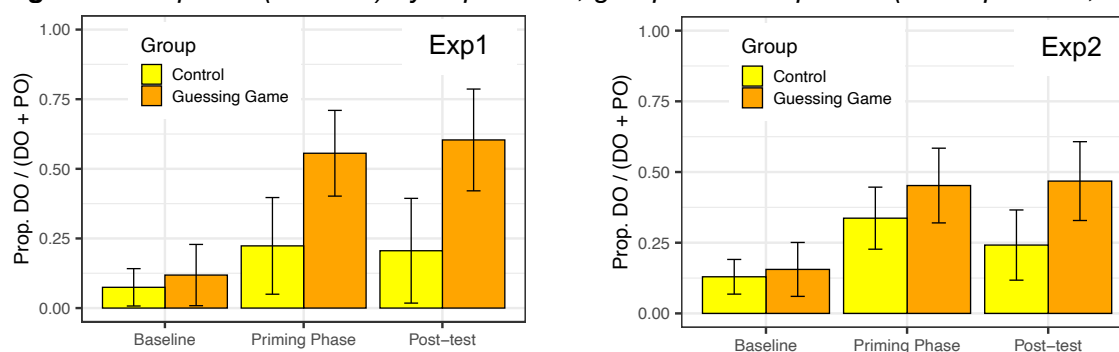


Figure 3. Prop. DO/(DO+PO) by experiment, group and task phase. (Participant Ms, 95% CIs)



Does bilingual inhibitory control operate over structural representations?

Andrea Seañez, Alejandra Fanith, Iva Ivanova (University of Texas at El Paso)

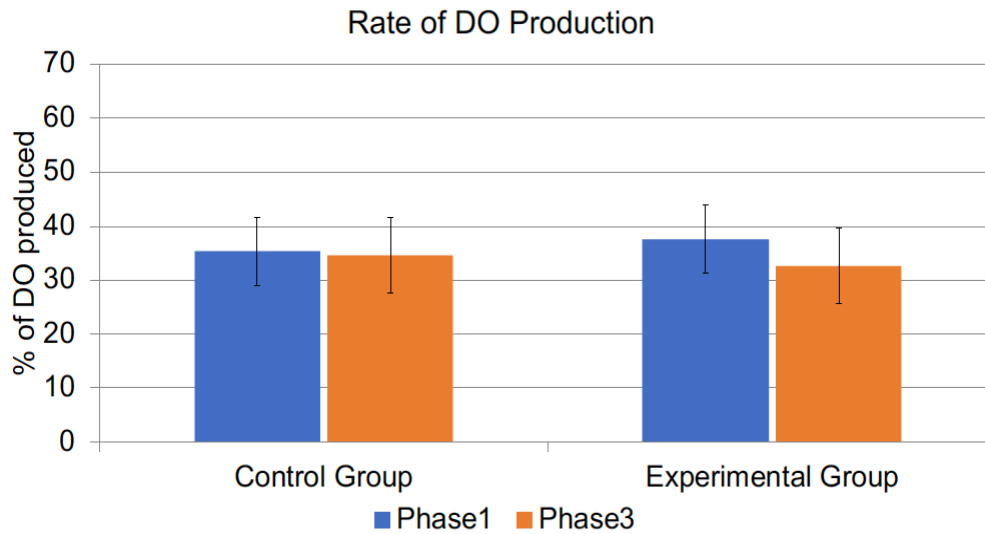
imivanova@utep.edu

Bilinguals rarely make wrong-language intrusions when their other language would not be understood. In the widely accepted Inhibitory Control Model, this is because they inhibit the non-target language to avoid interference (Green, 1998). Such inhibition can happen at two levels: *local inhibition*, when language task schemas (e.g., “Production in Language X”) inhibit outputs of the lexico-semantic system belonging to the non-target language, and *global inhibition*, when they inhibit whole non-target-language schemas (e.g., “Production in Language Y”). But the model does not specify if syntactic representations also get inhibited. It may be that they do not, and inhibition operates only over lexical representations. It may also be that syntactic representations are inhibited too, although how such inhibition would operate depends on the architecture of the bilingual syntactic system. For example, the language task schema would not be able to send inhibition to all syntactic representations in a language if the syntactic system is shared across bilinguals’ languages (Hartsuiker et al., 2004) and instead may operate only over language-tagged non-shared representations. Establishing if inhibition operates at the structural level will thus help constrain both accounts of bilingual language control and of bilingual structural representation.

The existence of structural inhibition was tested in a picture-description experiment with Spanish-English bilinguals dominant in English. The study is based on the fact that a transfer event can be expressed with both a prepositional dative (e.g., *The nun is giving a book to the pirate*) and a double object sentence in English (e.g., *The nun is giving the pirate a book*), but only with a prepositional dative in Spanish since Spanish lacks the double object structure. We compared differences in bilinguals’ double-object production rates in English before and after speaking Spanish, to those of another group of bilinguals who spoke only English throughout.

The experiment was administered online using Qualtrics. In Phase 1, all bilinguals gave typed descriptions of a set of 24 dative pictures in English (containing six written verbs repeated across four pictures), and an intermixed set of 36 intransitive fillers. In Phase 2, an Experimental group described in Spanish a set of monotransitive pictures (e.g., a waitress eating a cake), half of which had animate and half inanimate objects. A Control group described the same monotransitive pictures in English. In Phase 3, all bilinguals described a different set of dative pictures using the same six verbs as in Phase 1 (mixed with another set of 36 fillers). If structures get inhibited, speaking Spanish in Phase 2 should induce global inhibition of English, affecting in the very least the English structures that are not shared with Spanish, among them double objects. If so, upon returning to English, double objects should have reduced accessibility because of the prior inhibition. The Experimental group should thus produce fewer double-objects in Phase 3 than in Phase 1, while for the Control group there should be no change. To ensure sufficiently high baseline double-object production, the experiment began with a phase priming double objects, and target verbs were the six verbs that elicited highest rates of double objects in prior norming with the same population. Bilinguals’ English and Spanish proficiency and language history was assessed with a language history questionnaire (summarized in Table 1).

Preliminary results (Figure 1) showed no significant effects. Of most interest, bilinguals in the Experimental group ($N = 29/48$) were not differentially affected by Phase type than participants in the Control group ($N = 33/48$; Phase type X Group interaction in the LMER model: $p = .57$). Thus, tentatively, so far we have failed to detect any evidence for global structural inhibition. Experiment 2 will further test for effects of local structural inhibition: It may be that the non-existent-in-Spanish double object structure needs to be inhibited especially or only during production of Spanish prepositional datives, but not during production of monotransitives, with which it does not compete. New in addition to repeated dative verbs in Phase 3 will further test if effects are lexically driven; if so, double object production should decrease in Phase 3 for the Experimental group only for repeated but not for new verbs.



*Error bars represent standard error.

Figure 1. Percentage of double object (DO) production for the Control and Experimental groups in Phases 1 and 3.

Table 1. Language history characteristics of bilinguals in the Control and Experimental groups. The groups did not differ on any characteristic (all $ps > .32$).

	Control Condition	Experimental Condition
Age of first exposure in years		
English	2.46 (3.39)	3.67 (4.37)
Spanish	0.64 (1.57)	1.41 (4.67)
Other language(s)	10.85 (5.74) N=13	15.25 (8.71) N=12
Age of Acquisition in years		
English	4.54 (4.15)	4.72 (4.72)
Spanish	3.66 (3.50)	2.87 (4.89)
Other language(s)	Not collected	Not collected
% daily use		
English	67.76% (22.70%)	63% (24.61%)
Spanish	32% (22%)	34% (24%)
Other language(s)	3% (7.43%) N=13	1% (3.11%) N=12
Self-rated proficiency		
English	9.5 (0.75)	9.19 (1.21)
Spanish	7.71 (2.42)	8.26 (2.25)
Other language(s)	2.23 (1.64) N=13	3.33 (2.57) N=12

Note: Standard deviations are provided in parentheses.

References:

- Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1(2), 67-81.
- Hartsuiker, R. J., Pickering, M. J., & Velkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, 15(6), 409-414.

Bilingual language control in connected speech

Kyle Wolff and Iva Ivanova (University of Texas at El Paso)

Bilinguals make wrong-language intrusions extremely rarely in situations when their other language will not be understood. In the most established theory of this phenomenon, bilinguals inhibit the non-target language to prevent interference during target-language production (Inhibitory Control Model, Green, 1998). Such inhibition can act at the level of individual lexical representations (*local inhibition*) or at the level of the whole language (*global inhibition*). The most robust behavioral index of inhibitory control is a naming delay of previously inhibited words from the non-target language when this language becomes target, attributed to recovery from inhibition (and such recovery may last for at least ten minutes: Christoffels et al., 2016). This effect is more pronounced or only present for bilinguals' dominant language, consistent with the Inhibitory Control Model's feature that inhibition – and hence recovery from it – is proportional to the strength of the language it acts on (Calabria et al., 2012; Meuter & Allport, 1999). Such a slow-down in dominant-after-non-dominant picture naming is extremely robust, but it is unknown how bilingual inhibitory control dynamics affect connected speech.

In connected speech, lexical retrieval delays (assumed to reflect retrieval difficulties) should be manifest in a reduced speech rate, more filled (*uhs* and *uhms*) and unfilled pauses (Hartsuiker & Notebaert, 2009), fewer words overall, and/or an increased use of cognates (words with the same meaning and a similar form across two languages), which may be less affected by inhibition. More speculatively, a greater use of easier-to-retrieve words such as higher-frequency and more generic words (expected in the face of lexical retrieval difficulties, e.g. in AD: Ostrand & Gunstad, 2020) would be inconsistent with the implication of the Inhibitory Control Model that more robust representations are more strongly inhibited.

Method (Fig. 1). Eighty-six English-dominant Spanish-English bilinguals viewed two 8-min. videos (Tom-and-Jerry-type cartoons with no language) and after each viewing orally explained the video contents. Participants in the Changed-language group explained the first video in Spanish and participants in the Same-language group explained it in English (Phase 1). All participants explained the second video in English (Phase 2). Of interest was how the speech rate, fluency and quality during dominant English production in Phase 2 would be affected in the Changed-language group relative to the Same-language group. Also, half of the participants in each group explained the same two videos in Phases 1 and 2 (to target local inhibition), while the other half explained different videos (to target global inhibition). Bilinguals' English and Spanish proficiency (Table 1) was assessed with tests of productive vocabulary (MINT, Gollan et al., 2012) and grammar knowledge (MELICETⁱ and DELEⁱⁱ), and a language history questionnaire.

The data were analyzed with 2 (Phase 1 language) x 2 (Video Identity) ANOVAs. Contrary to the Inhibitory Control Model predictions, the Phase 2 English speech of the Changed-language group showed no significant differences from that of the Same-language group in speech rate, unfilled pauses and filled pauses. However, the Changed-language group produced fewer words overall ($p = .04$), fewer unique content words ($p = .04$), and words of higher overall frequency ($p = .04$) than the Same-language group (Figs 2-4). Video identity across phases had no effects except for unique content word frequency (Fig. 5). The remaining analyses will target a continuous measure of cognate status and, more exploratory, mean utterance length and number of clauses.

In conclusion, connected speech in bilinguals' *dominant* language showed clear effects of language control induced by previously speaking the non-dominant language. However, these effects were only partially consistent with strong predictions of the Inhibitory Control Model, and there was little support for a division of inhibition into local and global. Instead, our results may suggest that bilinguals possess compensatory measures to recover from adverse language-control effects on the dominant language to maintain speech fluency and quality – instead of being more disfluent or speaking more slowly, they used fewer and easier words.

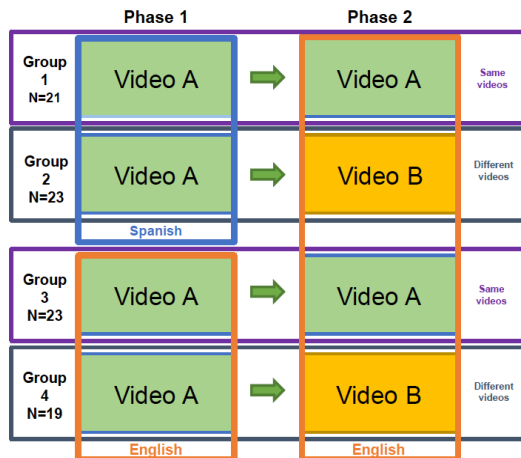


Figure 1: Design. The videos were counterbalanced across phases.

	Mean (SD)
Age of first exposure in years	
English	^a 4.4 (3.7)
Spanish	^a 2.5 (3.0)
Other	15.1 (4.3), N = 43
% daily use now	
English	^a 71% (19%)
Spanish	^a 29% (18%)
Other	9% (15%), N = 6
% daily use as a child	
English	^a 56% (26%)
Spanish	^a 44% (26%)
Other	40%, N = 1
Self-rated proficiency	
English	^a 9.6 (0.6)
Spanish	^a 6.9 (2.1)
Other	2.0 (1.3), N = 32
Productive vocabulary (MINT, of 68)	
English	60.2 (3.6)
Spanish	41.0 (12.5)
Grammar knowledge	
English (MELICET Adapted, of 50)	39.0 (7.8)
Spanish (DELE Adapted, of 50)	22.4 (6.5)

^aN = 80 (the language history questionnaires of six participants could not be uniquely identified).

Table 1. Participant language history

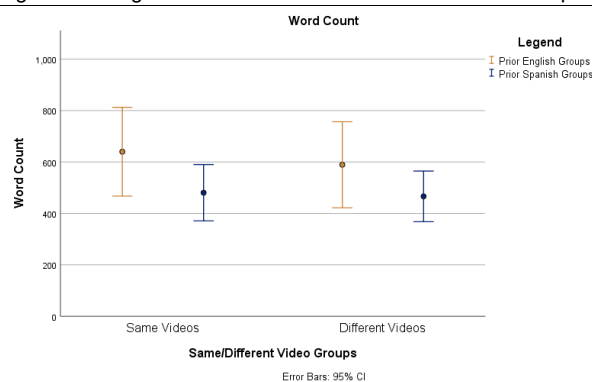


Figure 2: Total number of words

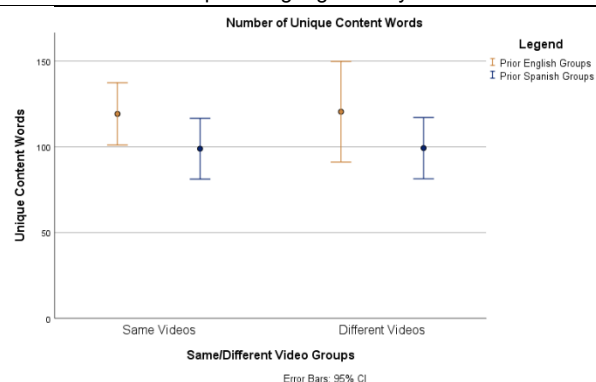


Figure 3: Number of unique content words

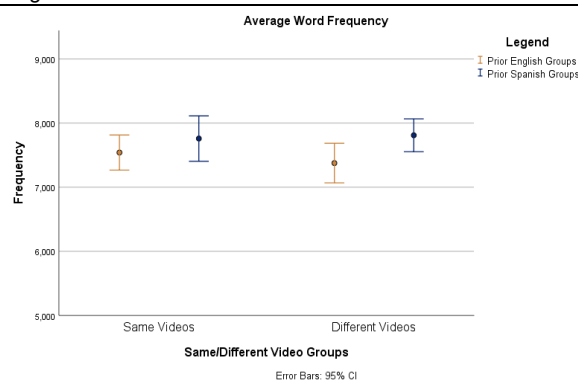


Figure 4: Average overall word frequency

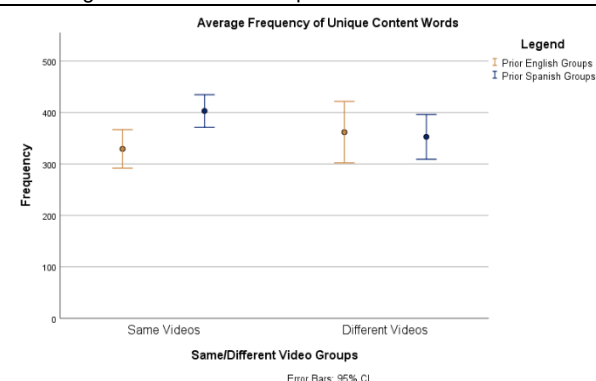


Figure 5: Frequency of unique content words

References:

- Calabria, M., Hernández, M., Branzi, F. M., & Costa, A. (2012). Qualitative differences between bilingual language control and executive control: Evidence from task-switching. *Frontiers in psychology*, 2, 399.
- Christoffels, I., Ganushchak, L., & La Heij, W. (2016). When L1 suffers. *Cognitive Control and Consequences of Multilingualism*, 2, 171.
- Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of memory and Language*, 50(4), 491-511.
- English Language Institute (2001) MELICET — GCVR user's manual. Ann Arbor, MI: English Language Institute, The University of Michigan.
- Gollan, T. H., Weissberger, G. H., Runqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition*, 15(3), 594-615.
- Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and cognition*, 1(2), 67-81.
- Hartsuiker, R. J., & Notebaert, L. (2009). Lexical Access Problems Lead to Disfluencies in Speech. *Experimental Psychology*.
- Meuter, R. F., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of memory and language*, 40(1), 25-40.
- Ministry of Education, Culture, and Sport of Spain. Diploma de español como lengua extranjera (Diploma of Spanish as a Second Language). 2006.
- Ostrand, R., & Gunstad, J. (2020). Using Automatic Assessment of Speech Production to Predict Current and Future Cognitive Function in Older Adults. *Journal of Geriatric Psychiatry and Neurology*, 0891988720933358. Advance online publication.

¹ Michigan English Language Institute College Entrance Test, English Language Institute. (2001). MELICET—GCVR user's manual.

² Diplomas de Español como Lengua Extranjera, Ministry of Education, Culture, and Sport of Spain.

Gap-filler dependencies are sensitive to islands: The case of Japanese relative clauses

Maho Takahashi, Grant Goodall (University of California, San Diego)

mtakahas@ucsd.edu

In filler-gap dependencies, gaps within certain structural environments (known as “islands”) are severely degraded. Does the same phenomenon arise in gap-filler dependencies, which are common in head-final languages? Here we address this question by examining relative clauses (RCs) in Japanese. RCs are known to be islands in many languages [1]. For instance, relativization out of another RC in English (i.e., a filler-gap dependency across an RC boundary) is not allowed (=1b).

- (1) a. The professor that [RC _ wrote a novel] is very proud. *filler-gap*
b. *This is the novel that [RC2 the professor that [RC1 _ wrote _]] is very proud.

RCs in Japanese are head-final, as shown schematically in (2a), thus exemplifying a gap-filler dependency. If this dependency is sensitive to islands, further relativization out of the RC, as in (2b), should not be possible (cf. 1b).

- (2) a. [RC _ a novel wrote] the professor is very proud. *gap-filler*
b. This is [RC2 [RC1 _ _ wrote] the professor is very proud] the novel.

Such structures have been often thought to be grammatical [2, 3], but here we explore this rigorously by means of an acceptability experiment using a factorial design, looking for the super-additivity that signals the presence of an island effect [4].

Experiment 1: 36 native speakers of Japanese participated in an online sentence acceptability experiment using a 7-point scale. The experiment had a 2x2 design, crossing EMBEDDED CLAUSE (RC vs. non-island) and EXTRACTION (relativization) out of the embedded clause (+ vs. -). The non-island clause is headed by *koto* ‘the fact (that),’ known not to induce an island effect [5, 6]. Participants saw 5 tokens of each condition (20 in total), together with 40 filler items of widely varying acceptability. Each of the 4 lists was fully counterbalanced and pseudorandomized. Sample stimuli are displayed in (3).

Results/Discussion: A linear mixed-effect model with random effects of subject and item reveals a significant main effect of EXTRACTION ($p < 0.001$), and importantly, a significant interaction between EMBEDDED CLAUSE and EXTRACTION ($p = 0.002$) (Figure 1). This interaction shows the super-additivity that defines an island effect, thus suggesting that gap-filler dependencies are indeed sensitive to islands. However, is the effect here specific to gap-filler dependencies, or could it occur with any “backwards” dependency? Exp. 2 explores the latter scenario with an anaphor that can precede its referent.

Experiment 2: A new group of 36 speakers participated in an online experiment consisting of the same number of stimuli as Exp. 1 (20 critical + 40 fillers = 60 total) and a similar 2x2 design crossing EMBEDDED CLAUSE and ANAPHOR DEPENDENCY (+ vs. -), the latter replacing the gap-filler dependency (EXTRACTION) of Exp. 1. The anaphor *zibun* ‘self’ was used, forming a backwards dependency with its referent *gakusha* ‘professor.’

Results/Discussion: A linear mixed-effect model with random effects of subject and item shows a significant main effect of EXTRACTION ($p < 0.001$), but the interaction effect between EMBEDDED CLAUSE and ANAPHOR DEPENDENCY is not significant ($p = 0.78$) (Figure 2). The absence of an interaction here suggests that the island effect observed in Exp. 1 is specific to gap-filler dependencies and is not a property of backward dependencies in general.

Conclusions: On a par with filler-gap dependencies, then, gap-filler dependencies seem to be sensitive to islands (though the relatively high acceptability of the island violation suggests this may be a “subliminal island” effect [7]). Our results are in accord with the general findings in the literature that the processing of head-initial and head-final structures is much more similar than one might expect [8, 9, 10]. The source of island effects in filler-gap dependencies has of course long been hotly contested, but the current results suggest that any account that attributes the effect solely to the rightward search for a gap would appear to be incorrect.

(3) **Sample items:** Exp.1 with a sentence-initial gap, Exp.2 with the anaphor *zibun*

a. [-RC] [-extraction] (Exp.1) / [-anaphor] (Exp.2)

[_{koto} Gakusha-ga SF-shousetsu-o kai-ta-koto-ga saikin shoten-de
professor-NOM Sci-Fi novel-ACC write-PST-fact-NOM recently bookstore-at
tokusyu-sa-re-ta.

feature-do-PASS-PST

“The fact [_{koto} that a professor wrote a sci-fi novel] was recently featured in a bookstore.”

b. [-RC] [+extraction] (Exp.1) / [+anaphor] (Exp.2)

[_{RC} [_{koto} ___ / *Zibun_i-ga* SF-shousetsu-o kai-ta-koto-ga saikin
(self-NOM) Sci-Fi novel-ACC write-PST fact-NOM recently

shoten-de tokusyu-sa-re-ta] *gakusha_i-wa* hokorashige-da.

bookstore-at feature-do-PASS-PST professor-TOP looks.proud-COP

“The professor_i [_{RC} who the fact [_{koto} that ___ / *self_i* wrote a sci-fi novel] was featured in a bookstore] looks proud.”

c. [+RC] [-extraction] (Exp.1) / [-anaphor] (Exp.2)

[_{RC} Gakusha-ga ___ kai-ta] SF-shousetsu-ga saikin shoten-de
professor-NOM write-PST Sci-Fi novel-NOM recently bookstore-at

tokusyu-sa-re-ta.

feature-do-PASS-PST

“The sci-fi novel [_{RC} that the professor wrote ___] was featured in a bookstore.”

d. [+RC] [+extraction] (Exp.1) / [+anaphor] (Exp.2)

[_{RC2} [_{RC1} ___ / *Zibun_i-ga* ___ kai-ta] SF-shousetsu-ga saikin shoten-de
(self-NOM) write-PST Sci-Fi novel-NOM] recently bookstore-at

tokusyu-sa-re-ta] *gakusha_i-wa* hokorashige-da.

feature-do-PASS-PST professor-TOP looks.proud-COP

“The professor_i [_{RC2} who the sci-fi novel_j [_{RC1} that ___ / *self_i* wrote ___] was recently featured in a bookstore] looks proud.”

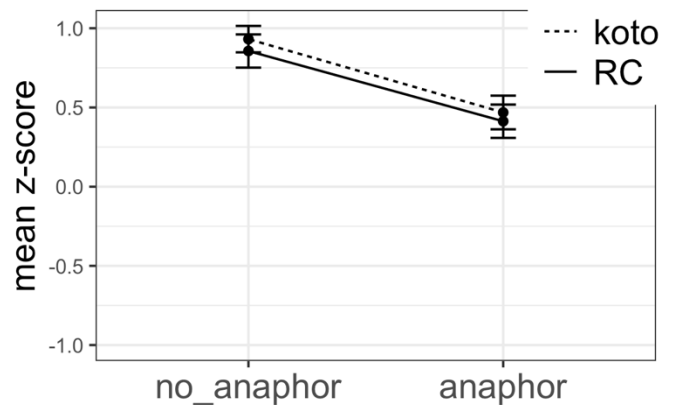
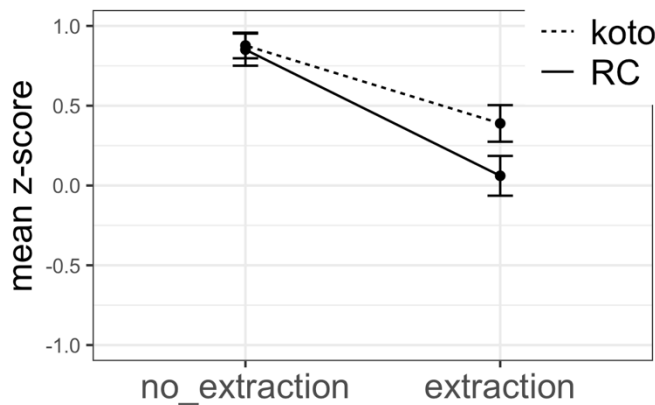


Figure 1: Mean acceptability from Exp. 1 (in z-score). **Figure 2:** Mean acceptability from Exp. 2 (in z-score).

References: [1] Ross (1967). *Constraints on variables in syntax*. [2] Ishizuka (2009). CNPC Violations and Possessor Raising in Japanese. ICEAL 2. [3] Nakamura & Miyamoto (2013). The object before subject bias and the processing of double-gap relative clauses in Japanese. *L&CP*. [4] Sprouse et al. (2011). Reverse island effects. *Syntax*. [5] Fukuda & Sprouse (2019). *Islandhood of Japanese Complex NPs and the Factorial Definition of Island Effects*. [6] Omaki et al. (2020). Subextraction in Japanese and subject-object symmetry. *NLLT*. [7] Almeida (2014). Subliminal wh-islands in Brazilian Portuguese. *RdA* 13. [8] Kahraman et al. (2011). Incremental processing of gap-filler dependencies. *TCP* 12. [9] Aoshima et al. (2004). Processing filler-gap dependencies in a head-final language. *JML*. [10] Omaki et al. (2015). Hyper-active gap filling. *Frontiers*.

Verb Metaphoric Extension during Sentence Processing

Daniel King and Dedre Gentner | Northwestern University

How does one understand a sentence like *the lantern limped*? Metaphoric uses of verbs are frequent in everyday language (Krennmayr, 2011). Yet the vast majority of research on metaphor processing has focused on noun-noun metaphors (e.g., *my job is a jail*, *my lawyer is a shark*) (Glucksberg et al., 1997). Comparatively little experimental work has focused on verb metaphors (but see Cardillo et al., 2012; Stamenković et al., 2019; Torrealano et al., 2005). This dearth of research on verb metaphor is problematic, as verb metaphor may in fact be more common than noun metaphor (Krennmayr, 2011; Jamrozik et al., 2013).

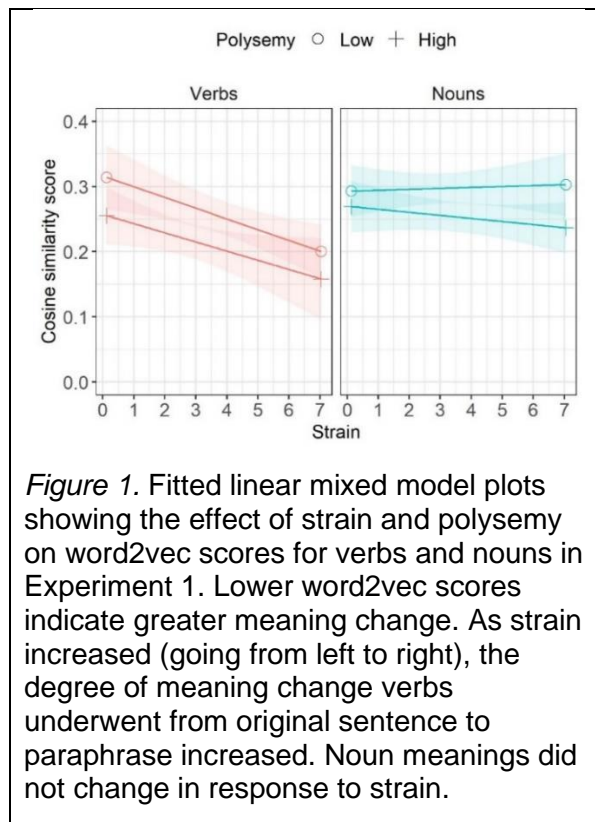
The objective of this research is to characterize the processes underlying verb metaphoric extension in sentence processing. We base our work on prior findings of a *verb mutability effect* in sentence processing (Gentner, 1981; Gentner & France, 1988; Kersten & Earles, 2004; Reyna, 1980), which reported that verbs have a greater propensity to alter their meaning in semantically strained contexts than do nouns. For example, Gentner and France (1988) found that, when asked to paraphrase strained (nonliteral) sentences like *the lantern limped*, people tend to alter the verb's meaning rather than the noun's, producing (for example) *the light flickered*. Here, we test two possible process accounts of verb mutability: online adjustment vs sense selection. The *online adjustment* account is that verb mutability is due to online adaptation processes that alter the verb's meaning to fit the noun's. The *sense selection* account is that verb mutability is a matter of selecting an appropriate meaning from the verb's existing senses. Past work on verb mutability (e.g., Gentner and France, 1988; Kersten & Earles, 2004) used verbs that were significantly more polysemous than the nouns, leaving open the possibility that sense selection can explain these findings.

In Study 1, we compared the sense selection account with the online adjustment account. As in Gentner & France's (1988) paradigm, participants paraphrased a mix of unstrained (literal) sentences (e.g., *the professor complained*) and strained (nonliteral) sentences (e.g., *the box complained*). Sentences were generated by combining 6 nouns and 6 verbs factorially for a total of 36 intransitive sentences, constructed such that half the nouns and verbs used were low-polysemy and half were high-polysemy. We asked new participants to paraphrase these and assessed the degree of noun and verb meaning change in the paraphrases using word2vec (Mikolov et al., 2013). The results supported the online adjustment account: both low- and high-polysemy verbs changed meaning in response to strain to an equal extent, while both low- and high-polysemy noun meanings remained equally stable (see Figure 1).

In Study 2, we tested the *minimal subtraction hypothesis* (Gentner and France, 1988), which states that verbs extend metaphorically in a graded manner, with domain-specific aspects being altered before more abstract relational structure. Using the same paraphrase paradigm, we selected 18 new nouns (6 human, 6 artifact, and 6 abstract) and 54 new verbs from 3 different classes (manner of motion, communication, and bodily process). Strain was increased systematically by varying the noun subject type (e.g., *the woman limped*, *the wagon limped*, *the fantasy limped*). The results replicated Study 1 and were consistent with minimal subtraction: word2vec scores indicated that verb (but not noun) meaning change increased in a graded manner as a function of strain. Regardless of type (human, artifact, or abstract), noun meanings remained stable across strain.

In our current research (Study 3), we directly investigate which aspects of the verb's meaning are altered as strain increases. The paraphrases from Study 2 were given to a new group of participants, who judged which components of the original verb's meaning were retained in each paraphrase. The results so far are consistent with minimal subtraction: (1) verb meanings changed in a graded manner; (2) domain-specific aspects of the verb's meaning changed before more domain-general aspects (see Figure 2).

Figures



Noun Type	Physical Motion Involving Legs	Physical motion not involving legs	Metaphoric motion	No physical or metaphoric motion	Unsure
Human	75	8	5	7	14
Artifact	6	79	6	14	6
Abstract	2	2	54	44	13

Figure 2. Current results of Experiment 3 (ongoing) for manner of motion verbs. Numbers represent response frequencies for verb meaning components retained across paraphrases. Rows correspond to the semantic strain of the stimuli sentences, increasing from top to bottom. Columns represent the dependent measure: level of verb abstraction for a given sentence. Yellow cells indicate responses consistent with minimal subtraction. For example, *the wagon limped* was paraphrased as *the damaged cart creaked along*, resulting in a code of *physical motion not involving legs*. *the fantasy limped* was paraphrased as *the story moved along slowly*, and was coded as *metaphoric motion*.

References

- Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *NeuroImage*, 59(4), 3212–3221.
- Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, 4, 161–178.
- Gentner, D., & France, I. M. (1988). Chapter 14 - The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical Ambiguity Resolution* (pp. 343–382). Morgan Kaufmann.
- Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language*, 36(1), 50–67.
- Jamrozik, A., Sagi, E., Goldwater, M., & Gentner, D. (2013). Relational words have high metaphoric potential. *Proceedings of the First Workshop on Metaphor in NLP*, 21–26.
- Kersten, A. W., & Earles, J. L. (2004). Semantic context influences memory for verbs more than memory for nouns. *Memory & Cognition*, 32(2), 198–211.
- Krennmayr, T. (2011). *Metaphor in newspapers*: LOT, Utrecht, Netherlands.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*.
- Reyna, V. (1980). When words collide: Interpretation of selectionally opposed nouns and verbs. *Symposium on Metaphor and Thought*.
- Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, 105, 108–118.
- Torreano, L. A., Cacciari, C., & Glucksberg, S. (2005). When Dogs Can Fly: Level of Abstraction as a Cue to Metaphorical Use of Verbs. *Metaphor and Symbol*, 20(4), 259–274.

34th Annual CUNY Conference on Human Sentence Processing

Saturday March 6, 2021

Hour	Session	Time	Title	Authors
Hour 1	1	12:30	Do faces speak volumes? A methodological perspective on social biases in speech comprehension and evaluation across three age groups	Adriana Hanulíková
Hour 1	1	12:30	Recognition of minimal pairs in (un)predictive sentence contexts in two types of noise	Marjolein van Os, Jutta Kray and Vera Demberg
Hour 1	1	12:30	★ Parents speak more about Object Features when children engage in Sustained Attention	Ryan Peters and Chen Yu
Hour 1	1	12:30	Facilitating the processing of foreign accent reduces bias against nonnative speakers	Shiri Lev-Ari and Katarzyna Grabka
Hour 1	1	12:30	Languages spoken by more people are more sound-symbolic	Shiri Lev-Ari, Ivet Kancheva, Louise Marston, Hannah Morris, Teah Swingler and Madina Zaynudinova
Hour 1	1	12:30	Individual differences in accent adaptation	Xin Xie and T. Florian Jaeger
Hour 1	2	12:30	Multiverse analysis of eye-tracking data: Reexamining the ambiguity advantage effect	Caren Rotello, Brian Dillon and Caroline Andrews
Hour 1	2	12:30	Computational Estimation of Lexical Semantic Norms: A New Framework	Bryor Sneffjella and Idan Blank
Hour 1	2	12:30	Objective ages of acquisition for 3300+ simplified Chinese characters	Zhenguang Cai, Shuting Huang, Zebo Xu and Nan Zhao
Hour 1	2	12:30	Visual recognition of morphologically complex words by second language learners: A masked priming study	Mariia Baltais and Anna Jessen
Hour 1	2	12:30	Online Processing of Derived and Inflected Words in L1 Turkish: A Masked Priming Experiment	Refika Cimen and Filiz Cele
Hour 1	3	12:30	★ ERP responses to lexical-semantic processing differentiate toddlers at high clinical risk for autism and language disorder	Chiara Cantiani, Valentina Riva, Chiara Dondena, Elena Maria Riboldi, Maria Luisa Lorusso and Massimo Molteni
Hour 1	3	12:30	Individual differences in language ability: Quantifying the relationships between linguistic experience, general cognitive skills and linguistic processing skills	Florian Hintz, Cesko C. Voeten, Christina Isakoglou, James M. McQueen and Antje S. Meyer

Hour	Session	Time	Title	Authors
Hour 1	3	12:30	★ Exposure to Plurals Can Help or Hurt Plural Production	Justin Kueser, Ryan Peters, Pat Deevy and Laurence Leonard
Hour 1	3	12:30	★ Recovery from semantic prediction violations during sentence processing in preschoolers with Developmental Language Disorder	Michelle Indarjit, Mariel Schroeder, Patricia Deevy, Laurence Leonard and Arielle Borovsky
Hour 1	3	12:30	Individual Differences in the Perception of Foreign-Accented Irony	Veranika Puhacheuskaya and Juhani Järvikivi
Hour 1	3	12:30	The impact of structural and functional lesions in the ventral stream on online semantic integration	Noelle Abbott, Niloofar Akhavan, Michelle Gravier and Tracy Love
Hour 1	4	12:30	Is reanalysis selective when regressions are manually controlled?	Dario Paape and Shravan Vasishth
Hour 1	4	12:30	Prediction of successful reanalysis based on eye-blink rate and reading times	Lola Karsenti and Aya Meltzer-Asscher
Hour 1	4	12:30	Reanalysis difficulty modulates cumulative structural priming effects in sentence comprehension	Ming Xiang and Weijie Xu
Hour 1	4	12:30	Interaction between local coherence and garden path effects supports a nonlinear dynamical model of sentence processing	Roeland Hancock and Whitney Tabor
Hour 1	4	12:30	Coordination ambiguity resolution in native and non-native language comprehension	Yesi Cheng, Hiroki Fujita and Ian Cunnings
Hour 1	4	12:30	Back to the Future: Do Influential Results from 1980s Psycholinguistics Replicate?	Fernanda Ferreira, Gwendolyn Rehrig, Madison Barker, Eleonora Beier, Suphasiree Chantavarin, Beverly Cotter, Zhuang Qiu, Matthew Lowder and Hossein Karimi
Hour 1	5	12:30	Presupposition projection from disjunction is symmetric	Alexandros Kalomoiros and Florian Schwarz
Hour 1	5	12:30	★ Pragmatic inference facilitates word retention in school-aged children	Katherine Trice, Marina Hernandez Santana, Dionysia Saratsli, Leah Heisler and Zhenghan Qi
Hour 1	5	12:30	A corpus-based study of (non-)exhaustivity in wh-questions	Morgan Moyer and Judith Degen
Hour 1	5	12:30	`At least' as a scalar modifier: Scalar diversity and ignorance inferences	Stavroula Alexandropoulou

Hour	Session	Time	Title	Authors
Hour 1	5	12:30	Priming pragmatic reasoning in the verification and evaluation of comparisons	Vishakha Shukla, Madeleine Long, Vrinda Bathia and Paula Rubio-Fernandez
Hour 1	6	12:30	Self-reported inner speech salience moderates implicit prosody effects	Mara Breen and Evelina Fedorenko
Hour 1	6	12:30	Guiding Implicit Prosody with Delexicalized Melodies: Evidence from a Match/Mismatch Task	Nicholas Van Handel, Matthew Wagers and Amanda Rysling
Hour 1	6	12:30	Using eye movements to predict performance on reading comprehension tests	Diane Meziere, Lili Yu, Erik Reichle, Titus von der Malsburg and Genevieve McArthur
Hour 1	6	12:30	Reading Minds, Reading Stories: Social-Cognitive Abilities are Related to Linguistic Processing of Narrative Viewpoint	Lynn S. Eekhof, Kobie van Krieken and Roel M. Willems
Hour 2	7	13:30	The interaction between grammaticality congruence and register-situation formality congruence in German sentence processing: an eye-tracking-reading pilot study	Camilo Rodriguez Ronderos, Katja Maquate and Pia Knoeferle
Hour 2	7	13:30	★ Event Completion, Not Ongoingness, Is Language Dependent: Crosslinguistic Evidence from ERPs in English and Russian	Anna Katikhina and Vicky Lai
Hour 2	7	13:30	Case marking influences the apprehension of briefly exposed events	Arrate Isasi-Isasmendi, Caroline Andrews, Sebastian Sauppe, Monique Flecken, Moritz Daum, Itziar Laka, Martin Meyer and Balthasar Bickel
Hour 2	7	13:30	Conceptualisation and formulation of motion event sentences in L2.	Matias Morales, Martin Pickering and Holly Branigan
Hour 2	7	13:30	★ Patterns of motion expression in children with or without a language disorder	Samantha Emerson, Karla McGregor and Şeyda Özçalışkan
Hour 2	7	13:30	The role of prior discourse in the context of action: Insights from pronoun resolution	Tiana Simovic and Craig Chambers
Hour 2	7	13:30	The social cost of maxims violation: Pragmatic behavior informs speaker evaluation	Andrea Beltrama and Anna Papafragou
Hour 2	8	13:30	Perceptual contrast as a visual heuristic in the formulation of referential expressions	Madeleine Long, Isabelle Moore, Francis Mollica and Paula Rubio-Fernandez
Hour 2	8	13:30	But what can I do with it?: Speakers name interactable objects earlier in scene descriptions	Madison Barker, Gwendolyn Rehrig and Fernanda Ferreira

Hour	Session	Time	Title	Authors
Hour 2	8	13:30	Culture, collectivism, and second language use affect perspective taking in language production	Max Dunn, Zhenguang G. Cai, Zebo Xu, Holly Branigan and Martin Pickering
Hour 2	8	13:30	The Role of Relatedness on Sentence Production	Jacqueline Erens and Jessica Montag
Hour 2	8	13:30	Speech Rate Convergence in Spontaneous Conversation	Maya Ricketts, Benjamin Schultz and Duane Watson
Hour 2	8	13:30	A cross-cultural study of the use and comprehension of color words: English vs Mandinka	Paula Rubio-Fernandez and Jara-Ettinger Julian
Hour 2	9	13:30	★ Recall and production of singular they/them pronouns	Bethany Gardner and Sarah Brown-Schmidt
Hour 2	9	13:30	★ Gender-inclusivity in English pronoun selection by L1 English and Spanish speakers	Cara Walker and Lauren Ackerman
Hour 2	9	13:30	Bias against "she" pronouns can be rapidly overcome by changing event expectations	Till Poppels, Veronica Boyce, Chelsea Ajunwa, Titus von der Malsburg and Roger Levy
Hour 2	9	13:30	The online application of structural and semantic biases during pronoun resolution	Markus Bader and Yvonne Portele
Hour 2	9	13:30	Singular vs. Plural Themselves: Evidence from the Ambiguity Advantage	Nicholas Van Handel, Lalitha Balachandran, Stephanie Rich and Amanda Rysling
Hour 2	9	13:30	Singular they in transition: ERP evidence and individual differences	Peiyao Chen, Olivia Leventhal, Sadie Camilliere, Amanda Izes and Daniel Grodner
Hour 2	10	13:30	Mismatches in Subject-Verb Agreement: The Processing of Numeral Quantifiers in Turkish	Ayşe Gül Özay-Demircioğlu
Hour 2	10	13:30	Regional constructions still need learned after adaptation	Emily Atkinson and Julie Boland
Hour 2	10	13:30	Understanding center embedding sentences: Can agreement and resumption help?	Hila Davidovitch, Maayan Keshev and Aya Meltzer-Asscher
Hour 2	10	13:30	When singular morphology meets notional plurality: another puzzle for agreement	Martina Abbondanza and Francesca Foppolo
Hour 2	10	13:30	Distribution matters: change in relative frequency affects syntactic processing	Valerie Langlois and Jennifer Arnold

Hour	Session	Time	Title	Authors
Hour 2	10	13:30	Cognitive Control and Ambiguity Resolution: Beyond Conflict Resolution	Varvara Kuz, Keyue Chen, Clement Veall and Andrea Santi
Hour 2	11	13:30	★ Six-month-old infants' abilities to represent regularities in speech	Irene de la Cruz-Pavía and Judit Gervain
Hour 2	11	13:30	★ The newborns' brain detects utterance-level prosodic contours	Anna Martinez-Alvarez, Silvia Benavides-Varela and Judit Gervain
Hour 2	11	13:30	★ Distributional learning as a driver of robust speech processing	Xin Xie, Andrés Buxó-Lugo and Chigusa Kurumada
Hour 2	11	13:30	★ The Identifiability of Consonants and of Syllable Boundaries in Infant-Directed English	Daniel Swingley
Hour 2	11	13:30	★ The presence of background noise reduces interlingual phonological competition during non-native speech recognition	Florian Hintz, Cesko C. Voeten and Odette Scharenborg
Hour 2	12	13:30	An investigation of the time-course of syntactic and semantic interference in online sentence comprehension	Daniela Mertzen, Brian W. Dillon, Ralf Engbert and Shravan Vasishth
Hour 2	12	13:30	A cue-based approach to processing adjuncts	Ethan Myers and Masaya Yoshida
Hour 2	12	13:30	Retrieval interference in the processing of RCs: Evidence from the visual-world paradigm	Gwynna Ryan and Matthew Lowder
Hour 2	12	13:30	Longer encoding times facilitate subsequent retrieval during sentence processing	Hossein Karimi, Michele Diaz and Eva Wittenberg
Hour 2	12	13:30	Cue-based retrieval model of parsing	Jakub Dotlacil
Hour 2	12	13:30	Competing Effects of Syntax and Animacy in Priming of Relative Clause Attachment	Melodie Yen, Idan Blank and Kyle Mahowald
Hour 2	13	13:30	Language modeling using a neural network shows effects on N400 beyond just surprisal	Don Bell-Souder, Shannon McKnight, Vladimir Zhdanov, Sean Mullen, Akira Miyake, Phillip Gilley and Albert Kim
Hour 2	13	13:30	The Posterior P600 reflects Reanalysis but not Repair	Edward Alexander, Trevor Brothers and Gina Kuperberg
Hour 2	13	13:30	Modeling subcategorical information maintenance in spoken word recognition	Wednesday Bushong and T. Florian Jaeger

Hour	Session	Time	Title	Authors
Hour 2	13	13:30	Interpreting implausible sentences: The role of phonological similarity	Jianyue Bai and Zhenguang Cai
Hour 2	13	13:30	The Stability of Individual ERP Response Dominance Within and Across Conditions	Tamarae Hildebrandt and Jonathan R. Brennan

Do faces speak volumes? A methodological perspective on social biases in speech comprehension and evaluation across three age groups

Adriana Hanulíková (University of Freiburg)

An unresolved issue in social perception concerns the effect of perceived ethnicity on speech processing. Bias-based accounts assume that listeners activate stereotypes in the case of a talker classification as nonnative (Rubin, 1992; Kang & Rubin, 2009), resulting in conscious misunderstanding and negative evaluation of speech. In contrast, expectation/exemplar-based accounts suggest that correct anticipation of a talker's accent facilitates processing (Babel & Russell, 2015; McGowan, 2015). Driven by theoretical and methodological differences in previous research, this study seeks to establish the extent to which effects of perceived ethnicity on speech processing depend on three sources of variability: experimental method, speech context, and age group. Life-long experiences with certain speakers and their language use shapes the distributional knowledge and can contribute to differences across age group. To this end, sentence recall (assessing speech intelligibility) and accent ratings from three white European non-university populations (72 teens, mean age 14.1; 50 younger adults, mean age 36; 50 older adults, mean age 77.6; all native speakers of German) were examined. Participants were primed with photographs of young Asian and white European women and asked to repeat utterances spoken in standard German, Korean-accented German, and a regional German variety, all embedded in speech-shaped noise. Each ethnicity was presented across all three levels of the accent factor (i.e. there were three photographs for each ethnicity). After the recall task, participants were asked to provide accent ratings for each speaker on a scale from 1-5 (5 = strong accent). A linear mixed effect logistic regression model for binary responses was fitted to the recall data, and a cumulative link mixed model was used for the accent ratings. Sentence recall accuracy increased in the foreign-accented speech for the Asian prime compared to the white European prime, in line with expectation/exemplar-based accounts. However, this matching expectation effect varied during the course of the experiment across the accents and groups (first vs. second half, see Figure 1) and was most pronounced in the group of teens in the foreign accent. In contrast, speech presented along Asian primes received the most negative accent ratings (see Figure 2) irrespective of the speech context, consistent with a bias-based view. The effect was stronger in the group of older adults than in the other groups. Younger adults showed weak or no effects of ethnicity in either task. Taken together, both methods show in part successful integration of social information, but the conclusions diverge. A disconnect between linguistic measures and a non-equivalence of sentence recall and indexical judgments like accent ratings became apparent. Clearly, they seem not to tap into the same underlying construct. The malleability of ethnicity effects shows the importance of a substantial scrutiny of the methodological disparities used by theoretical accounts. While the present findings show that each theory has its share, they also suggest that theoretical contradictions are a consequence of methodological choices that tap into distinct aspects of social information processing. Importantly, predictive abilities and strategies vary across the age groups, underlining the importance of the inclusion of underrepresented populations in future research.

References

- Babel, M., Russell, J. (2015). Expectations and speech intelligibility. *J. of Acoustical Society of America*, 137(5), 2823–2833.
- Kang, O., Rubin, D.L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *J. of Language and Social Psychology*, 28, 441–456.
- McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech*, 58(4), 502–521.
- Rubin, D.L. (1992). Nonlanguage factors affecting undergraduates' judgements of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33(4), 511–531.

Figure 1: Proportion of correctly repeated words in each speech context and listener group, for the first and second half of the experiment. Black dots represent the overall means and the colored dots show the individual participant means. The violin plots depict probability density. Error bars represent 95% confidence intervals.

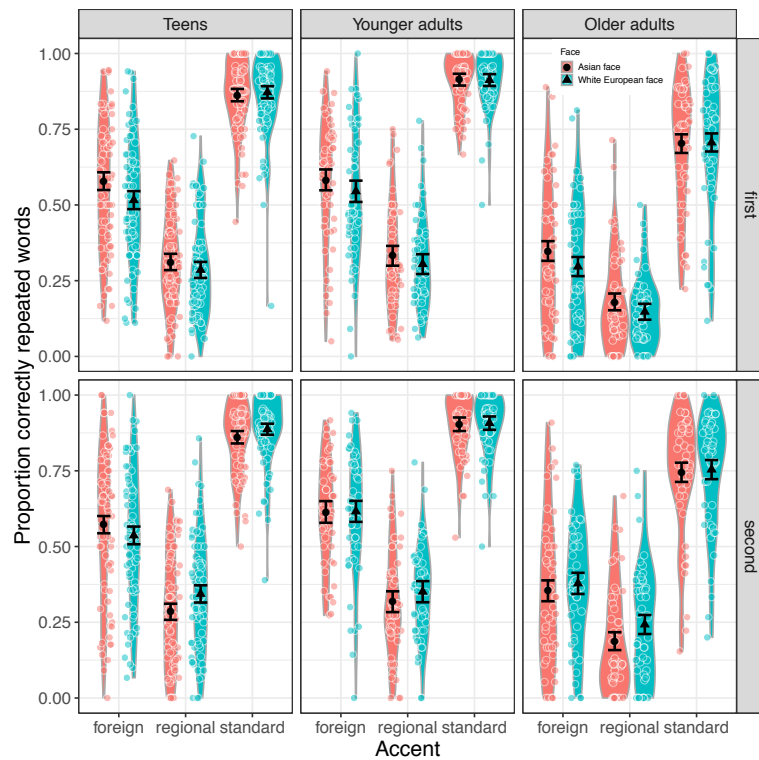
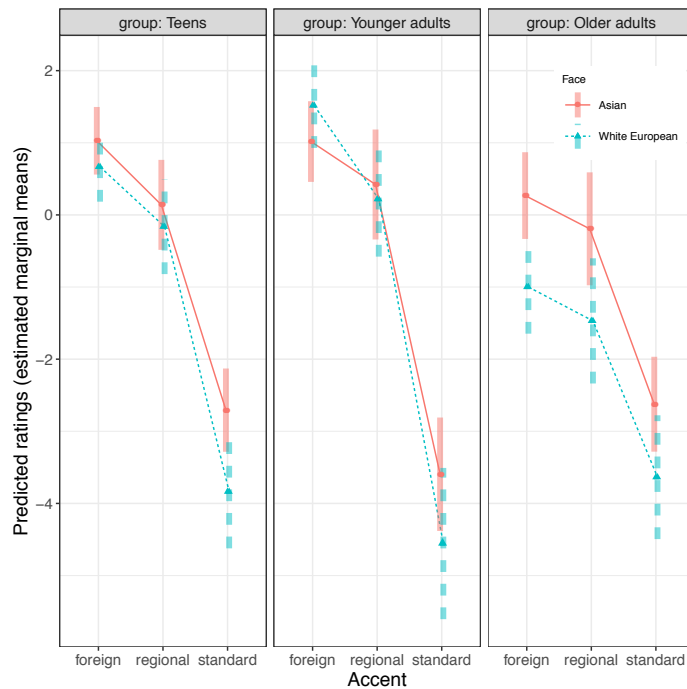


Figure 2: Estimated marginal means for ratings based on the clmm model. Error bars represent 95% confidence intervals.



Recognition of Minimal Pairs in (un)predictive Sentence Contexts in two Types of Noise

Marjolein van Os, Jutta Kray, Vera Demberg (Saarland University)

Language understanding is facilitated by highly predictive contexts even in noisy conditions (Dubno, Horwitz, & Ahlstrom, 2000; Sommers & Danielson, 1999). Here we investigated whether the type of noise influences speech comprehension, that is, the recognition of minimal pairs, differently. While multi-speaker babble noise approximates the average long-term spectrum of the speech of an adult male speaker, white noise has a flat spectral density with the same amplitude throughout the audible frequency range. Both types of noise lead to energetic masking of the target speech, as both the speech signal and the noise have energy in the same spectral frequency bands (Brungart, 2001). However, as babble noise shows more overlap with the spectral information of a single speaker, it may lead to greater energetic masking than the more spread out energy of white noise. While previous studies have compared babble and white noise (Lecumberri & Cooke, 2006; Taitelbaum-Swead & Fostick, 2016), no studies have so far directly compared the effect of noise type on mishearing in different noise contexts. We expected that babble noise reduces recognition performance more than white noise and that this effect is more pronounced when sentence endings are unpredictable target words.

To examine this, participants listened to recordings of sentences embedded in babble and white noise at -5 dB SNR and in quiet. They typed in the last word of the sentence they had heard. Each sentence was presented visually up to the target word to provide a predictive context that was understandable regardless of background noise. The final target word either fit the sentence semantically (mean cloze 0.72, high predictability condition, HP), or was unpredictable based on the preceding context (mean cloze 0, low predictability condition, LP). Example stimuli can be found in Table 1. The target words formed minimal pairs differing in one phonetic feature in medial position, and were swapped with the respective sentence frames to create LP items. This allowed us to investigate whether listeners can rely on small acoustic cues for word recognition, even in noise, while keeping sentence contexts equal across conditions.

Responses from 48 participants (31 males, mean age = 24 years) were coded on whether they matched the auditorily presented word (e.g., in example 1A in Table 1 “Liege” / “lounge”, *target*), the similar sounding *distractor* (e.g., in 1A “Liebe” / “love”), or were a different word entirely (e.g., in 1A “Platz” / “space”, *wrong*). Using a General Linear Mixed Model with fixed effects of Noise and Predictability as well as the interaction, and Trial No, and random intercepts for Subject and Item, with random intercepts for Noise and Predictability for both, we find that both noise conditions lead to fewer correctly identified target responses than quiet ($\beta = -5.30$, $SE = 0.85$, $z = -6.21$, $p < .001$ for babble and $\beta = -4.50$, $SE = 0.82$, $z = -5.51$, $p < .001$ for white noise). The rate of correctly identified targets does not differ significantly between the two noise conditions ($p = .09$). Regarding the beneficial effect of predictability, we find that participants correctly identify the target more often for HP compared to LP ($\beta = -6.01$, $SE = 0.91$, $z = -6.58$, $p < .001$; see Figure 1). On the subset of unpredictable items, we next tested whether the types of errors (wrong vs. distractor) differ between the noise conditions, see Figure 2. Here the distractor fit the context and most of the acoustic signal. The wrong response did not fit both. We ran the model with fixed effects of Noise and Trial No, and random intercepts for Subject and Item. We find more wrong responses in babble noise compared to quiet ($\beta = -1.23$, $SE = 0.34$, $z = -3.61$, $p < .001$), as well as a to white noise ($\beta = 0.73$, $SE = 0.24$, $z = 3.01$, $p < .01$). The wrong responses in the babble condition cannot have been caused by competing speech in the noise: due to the high number of speakers, specific speech streams were unintelligible.

The results suggest that noise hamper speech comprehension irrespective of sentence predictability. The type of noise induced different errors indicating that white noise is indeed an easier condition than babble. Analyses of semantic fit and phonetic distance in the wrong responses will shed more light on this.

Table 1. Example Stimuli

1A	Am Pool im Hotel gab es nur noch eine freie Liege . <i>At the pool in the hotel there was only one free lounge left.</i>	HP
1B	Nach vier Jahren heiratete Paul seine große Liebe . <i>After four years, Paul married his big love.</i>	HP
1C	Am Pool im Hotel gab es nur noch eine freie Liebe . <i>At the pool in the hotel there was only one free love left.</i>	LP
1D	Nach vier Jahren heiratete Paul seine große Liege . <i>After four years, Paul married his big lounge.</i>	LP

Note. Highly predictable sentences (HP) were made based on minimal pairs (*Liebe / Liege*) in 1A and 1B), then sentence-final target words were swapped to make low predictability items (LP) with the sentence frames of 1A and 1B, resulting in 1C and 1D. English translations have been given in *italics*.

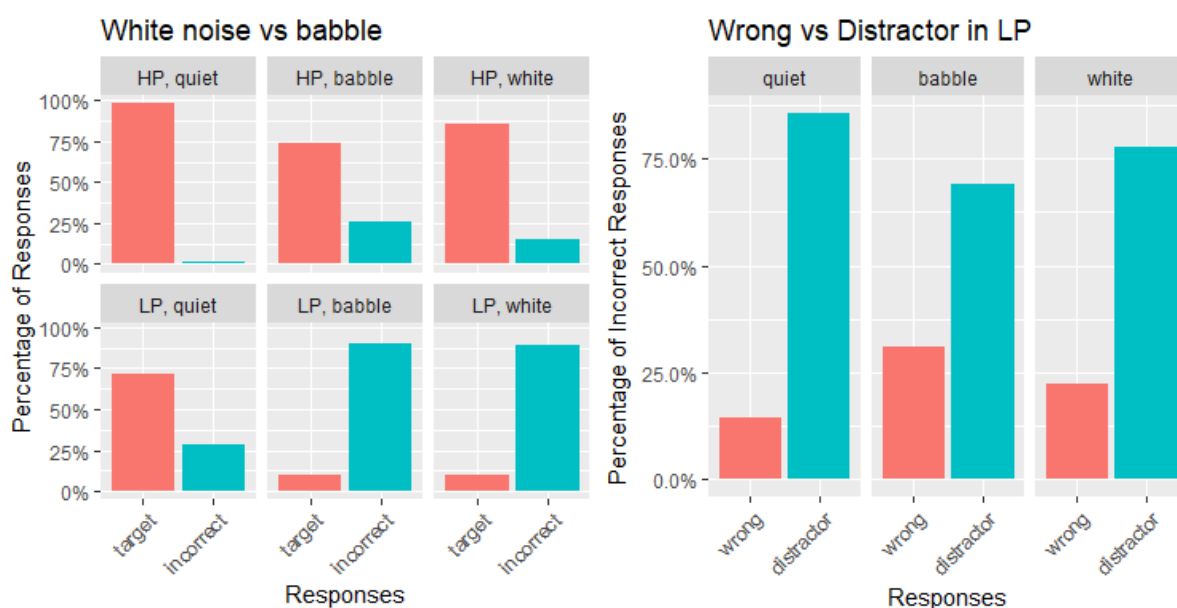


Figure 1: % of target and incorrect responses for high (HP) and low predictability condition (LP) in quiet, babble noise, and white noise. Figure 2: % of wrong and distractor responses for the low predictability condition (LP) in quiet, babble noise, and white noise.

References

- Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3), 1101-1109.
- Dubno, J. R., Ahlstrom, J. B., & Horwitz, A. R. (2000). Use of context by young and aged adults with normal hearing. *The Journal of the Acoustical Society of America*, 107(1), 538-546.
- Lecumberri, M. G., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, 119(4), 2445-2454.
- Sommers, M. S., & Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging*, 14(3), 458-472.
- Taitelbaum-Swead, R., & Fostick, L. (2016). The effect of age and type of noise on speech perception under conditions of changing context and noise levels. *Folia Phoniatrica et Logopaedica*, 68(1), 16-21.

Parents speak more about Object Features when children engage in Sustained Attention

Ryan E Peters & Chen Yu (UT Austin)

ryan.peters@austin.utexas.edu

Do parents prioritize certain types of information when their children are engaged in Sustained Attention (SA) to an object? Here, guided by work on parent responsiveness on the one hand (Tamis-LeMonda et al., 2001) and work highlighting the importance of children's visual experience during naming on the other (Yu et al., 2019), we hypothesize the topical content of parent speech relates to children's visual SA (defined as episodes of attention longer than 3 s) to speech targets. But this hypothesis is complicated by research separately indicating the content of parent utterances (Chang & Deák, 2019) and children's patterns of SA (Suanda et al. 2016) are both tied to larger scale structure in parent discourse. Thus, here we explore relations between the content of parent utterances and coinciding patterns of child SA while considering interrelated influences of discourse structure.

To address this aim, we recruited parent-child dyads into the lab to participate in a free toy play session—during which we recorded parent speech and collected gaze data via head-mounted eye-trackers (Figure 1A&B). We coded parent speech using a novel coding scheme of mutually exclusive speech content types that builds upon the framework created by Chang and Deák (2019; Figure 1C).

We asked: How do the (IV.1) topical content of referential parent speech, (IV.2) whether it is in a discourse and (IV.3) timing relative to first (consecutive) target reference relate to the (DV) temporal patterning of coinciding child SA to speech referents? We explored relations between these IVs and SA by conducting analyses of paired events consisting of parent speech overlapping with either preceding or following episodes of SA using linear mixed effect models. Models included the three IVs as fixed effects and subjects and items as random effects. Given limited space, here we focus on the subset of results showing pairwise differences between the estimated marginal means for (IV.1) topical content types output from the models.

First, results showed parent speech conveying information about Object Features is significantly more likely to overlap with preceding episodes of SA than the other content types (Figure 1D). Second, we found speech about Object Features occurs significantly later in preceding SA episodes than other speech types (Figure 1E). Finally, analyses revealed episodes of SA overlapping with speech about Object Features are significantly longer than for episodes of SA overlapping with speech about Actions and Activities or object Labels (Figure 1F). Crucially, while not discussed here, all of these results hold true while considering the interrelated influences of discourse structure.

This work is the first systematic exploration of relations between the topical content of parent speech and child attention in a naturalistic environment that takes into account larger scale structure in parent speech. The findings show parents prioritize conveying information about the features of objects during optimal learning opportunities: namely, when the child has been engaged in an episode of SA to the referent object. While the degree to which such prioritization generalizes across contexts remains unexplored, the potential impacts on early lexico-semantic development are wide ranging. For example, recent network modeling work showing that labels for objects with more perceptual features are learned earlier (Peters & Borovsky, 2019) could in part be explained by the current findings. Thus, this work highlights how recently proposed developmental dependencies between the perceptual/sensory structure of early vocabularies and lexical development may, in part, be driven by patterns of what parents say to their children and when in the course of daily interactions.

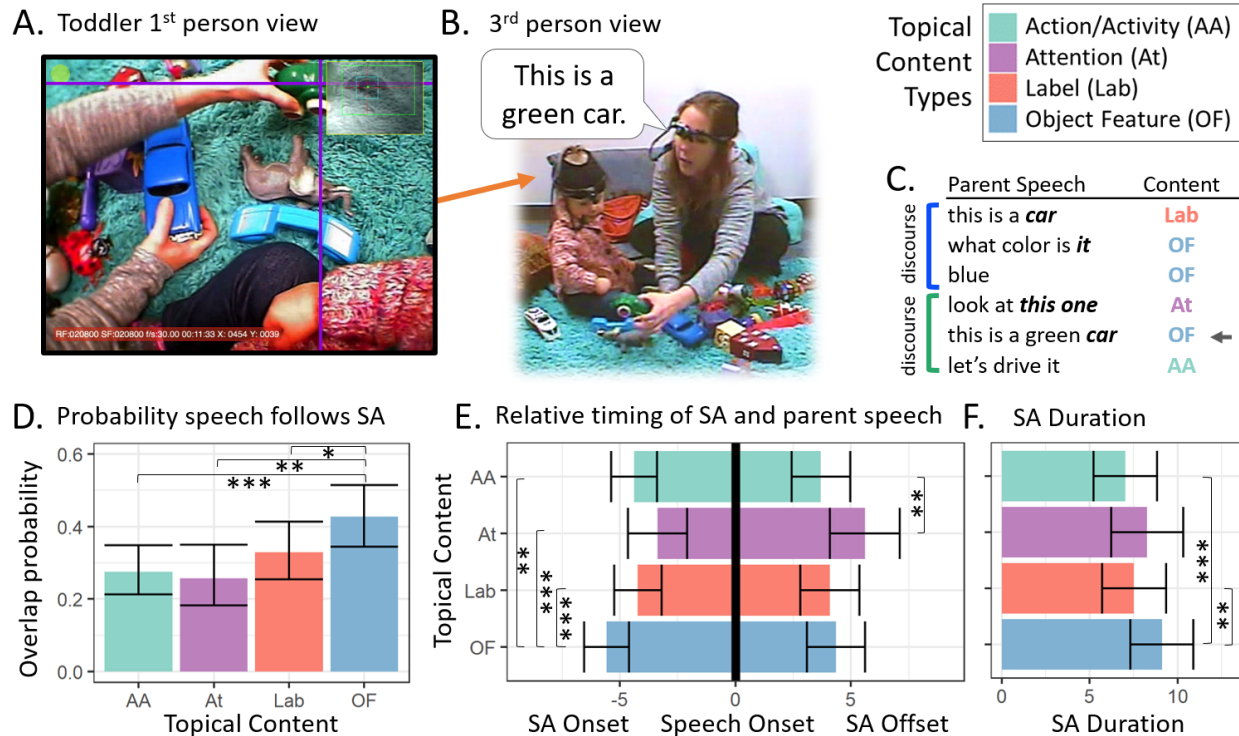


Figure 1. Row 1: Experiment setup showing A) first-person toddler view, B) third-person view and C) corresponding section of speech transcript. Row 2: Comparisons between content types of estimated marginal means output from linear mixed effects models of the D) probability parent speech overlaps with preceding episodes of Sustained Attention (SA), E) timing of SA onsets and offsets relative to speech onsets and F) durations of overlapping episodes of SA. *** $p < .001$. ** $p < .01$. * $p < .05$.

References

- Chang, L. M., & Deák, G. O. (2019). Maternal discourse continuity and infants' actions organize 12-month-olds' language exposure during object play. *Developmental science*, 22(3), e12770.
- Peters, R., & Borovsky, A. (2019). Modeling early lexico-semantic network development: Perceptual features matter most. *Journal of Experimental Psychology: General*, 148(4), 763.
- Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, 83(5), 1762-1774.
- Suanda, S. H., Smith, L. B., & Yu, C. (2016). More than Words: The Many Ways Extended Discourse Facilitates Word Learning. In *CogSci*.
- Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child development*, 72(3), 748-767.
- Yu, C., Suanda, S. H., & Smith, L. B. (2019). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental science*, 22(1), e12735.

Facilitating the processing of foreign accent reduces bias against nonnative speakers

Katarzyna Grabka and Shiri Lev-Ari (Royal Holloway, University of London, Egham, UK)

Shiri.Lev-Ari@rhul.ac.uk

People are more likely to believe things that are easier to process (McGlone & Tofigbakhsh, 2000; Reber & Schwarz, 1999). For example, people are more likely to believe trivia statements when they are written in clearer color contrast (Reber & Schwarz, 1999). This has implications for interactions between native and non-native speakers since foreign-accented speech is harder to process than native speech even when it is fully understood (Munro & Derwing, 1995). Indeed, prior research has shown that people believe information less when it is delivered in a foreign accent rather than a native accent (Lev-Ari & Keysar, 2010), presumably because of the greater processing difficulty. The latter finding, however, has not always been replicated (Stocker, 2017). Furthermore, there is no direct evidence that the bias against non-native speakers stems from greater processing difficulty. Here we show that by exposing people to Polish-accented English we reduce their tendency to distrust information delivered in Polish-accented English. Furthermore, we show that the reduction in bias is fully mediated by improvement in comprehension of Polish-accented speech.

While foreign-accented speech is harder to process (Munro & Derwing, 1995), this difficulty can be alleviated by exposure (Clarke & Garrett, 2004). Furthermore, exposure to several foreign-accented speakers can improve comprehension of other speakers with the same accent (Bradlow & Bent, 2008) and even of speakers with similar accents (Baese-Berk et al., 2013). This suggests that exposing listeners to foreign-accented speech should facilitate their comprehension of that accent, and consequently, reduce their bias against non-native speakers with that accent.

To test whether exposure to accent can reduce the bias against its speakers, two-hundred and twenty native speakers of British English participated in the study. First, participants listened to 8 short stories and answered simple comprehension questions about them. Critically, the stories were told by either native speakers of British English or Polish-accented speakers of English. Next all participants listened to 50 trivia statements (e.g., *An ostrich's eye is bigger than its brain*) and estimated their truth value on a continuous 100-point scale ranging from False to True. Each trivia statement had two versions, one recorded by a native speaker and one by a Polish-accented speaker, but each participant heard only one version of each statement. Statements were presented in random order and each participant heard half of the statements in native accent and half in foreign accent. At the end of the experiment, all participants were tested on their comprehension of Polish-accented English by transcribing a few sentences produced by the Polish-accented speakers.

Participants' truth-ratings were analyzed with a mixed effects regression analysis. Results showed that participants believed statements more when they were produced in a native rather than a foreign accent ($\beta=22.26$, $SE=2.40$, $t=9.29$), but that this effect interacted with Exposure ($\beta=-7.71$, $SE=3.21$, $t=-2.40$), such that participants who were exposed to Polish accent in the Exposure phase showed significantly smaller bias (See Figure 1). Furthermore, participants exposed to Polish accent performed better in the accent comprehension task ($\beta=2.58$, $SE=0.21$, $t=12.36$, $p<0.001$; See Figure 2), and a mediation test using the mediation package in R (Tingley et al., 2014) revealed that the improvement in accent comprehension accounted for 90.1% of the effect of Exposure on truth judgment.

The results of this study show that the relative difficulty of understanding foreign-accented speech leads to believing it less, but that this bias can be reduced by exposure to that foreign accent. This study is the first to provide direct evidence for the role of processing difficulty in the bias against non-native speakers. It thus illustrates how cognitive processes involved in language processing as well as language-based interventions can have social consequences.

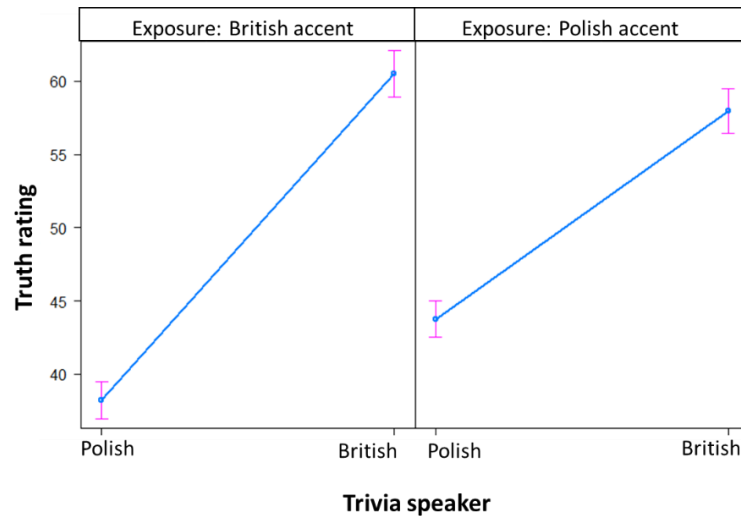


Figure 1

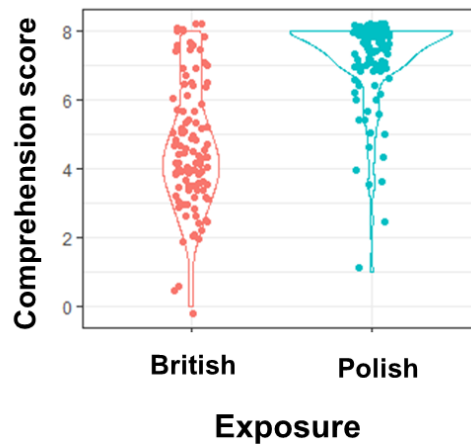


Figure 2

References

- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America*, 133, 3, EL174-EL180.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106, 2, 707-729.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116, 3647-3658.
- Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46, 6, 1093-1096.
- McGlone, M. S., & Tofiqbakhsh, J. (2000). Birds of a feather flock conjointly(?): Rhyme as reason in aphorisms. *Psychological Science*, 11, 424-428.
- Munro, M. J., & Derwing, T. G. (1995). Processing time, accent and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38, 289-306.
- Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8, 338-342.
- Stocker, L. (2017). The Impact of Foreign Accent on Credibility : An Analysis of Cognitive Statement Ratings in a Swiss Context. *Journal of Psycholinguistic Research*, 46, 617-628.
- Tingley D., Yamamoto T., Hirose K., Keele L., & Imai K. (2014). Mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59, 5, 1-38.

Languages spoken by more people are more sound-symbolic

Shiri Lev-Ari, Ivet Kancheva, Louise Marston, Hannah Morris, Teah Swingler & Madina Zaynudinova (Royal Holloway, University of London, Egham, UK)

shiri.lev-ari@rhul.ac.uk

Languages are spoken in different social environments. These environments impose different communicative challenges. For example, larger communities might have less shared knowledge and greater difficulty of converging on a shared system. Recent research shows that languages adapt to their social environment [1], and in particular, that languages spoken by larger communities are less morphologically complex [1-2]. These findings suggest that languages spoken by more people might evolve to have features that make them more robust for learning and communication. This study tests this hypothesis by examining whether languages spoken by more people are more sound symbolic. This question is important not only because it sheds light on how linguistic features are shaped by communicative pressures, but also because it can address the question of the role of sound symbolism in language.

Sound symbolism has been shown to facilitate language learning and processing [e.g., 3-4]. Therefore, to the degree that languages spoken by more people should be more robust, they might rely more on sound symbolism. This hypothesis is in line with recent research on facial expressions that shows that more heterogeneous communities, which also face greater communicative challenges, use more exaggerated facial expressions, and these are better understood by non-community members [5].

To test whether languages spoken by more people are more sound symbolic, we selected 20 languages spoken by millions of people (Median=81.7m; range: 24.5m-1.1billion) and 20 languages spoken by only hundreds or thousands of people (Median=3,750; range: 200-314,000). Next, we generated recordings of the words 'large' and 'small' in those languages using a text-to-speech synthesizer. We selected the words 'large' and 'small' as there is an established link between high front vowels such as 'i' and 'e' and small size and low and back vowels such as 'a', 'o' and 'u' and large size [e.g., 6]. 128 participants heard the words in a random order, and for each word, they guessed whether it means 'small' or 'large'. If they were familiar with the word, they indicated that they knew the word and did not provide a guess.

A logistic mixed effects regression revealed that participants were better at guessing word meanings in languages spoken by many vs few people ($\beta=-0.3$, $SE=0.15$, $z=-2$, $p<0.05$; see Fig 1a). An exploratory analysis using the (log-transformed) number of speakers rather than a categorical predictor suggests greater influence of community size when communities are small ($\beta=0.3$, $SE=0.01$, $z=2.1$, $p<0.04$; see Fig 1b). Next, we examined whether participants relied on vowels to make their judgments. Participants exhibited the established sound-symbolic patterns: they were more likely to guess that a word means "large" the more back vs front vowels it had ($\beta=0.16$, $SE=0.05$, $z=3.2$, $p<0.01$; see Fig 2). Interestingly, widely-spoken languages were not more likely than less common languages to have front/back vowels to indicate small/large size ($p>0.1$). This suggests that widely-spoken languages relied on other sound-symbolic cues.

This study shows that languages adapt to their social environment, and in particular, that widely-spoken languages are more sound symbolic than languages spoken by few people. We propose that this is driven by the need to overcome the greater communicative challenges involved in interaction in larger communities. Some research suggests that language lose their iconicity with time [e.g., 7 for sign languages]. This study suggests that having a larger community of speakers can lead to maintenance of this iconicity. Interestingly, results also showed that even though widely-spoken language were more sound symbolic, they were less likely to follow vowel-size congruency. It might therefore be the case that community size influences not only the degree to which languages are sound symbolic but the type of sound symbolism that languages develop.

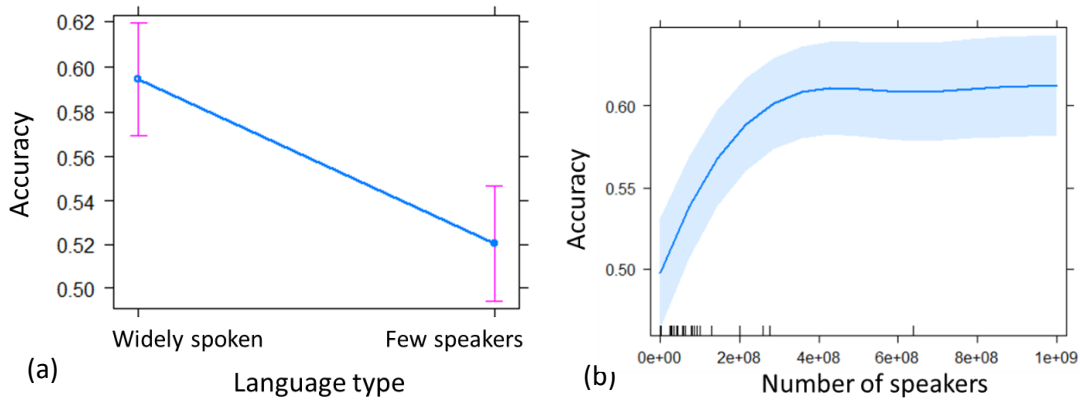


Figure 1

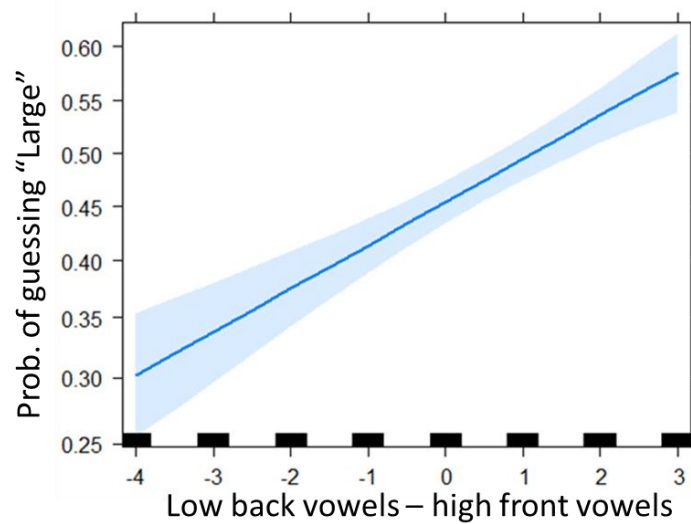


Figure 2

References

- [1] Lupyan, G., & Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends in cognitive sciences*, 20, 9, 649-660.
- [2] Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286, 1907, 20191262.
- [3] Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109, 54-65.
- [4] Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia*, 76, 264-275.
- [5] Wood, A., Rychlowska, M., & Niedenthal, P. M. (2016). Heterogeneity of long-history migration predicts emotion recognition accuracy. *Emotion*, 16, 4, 413.
- [6] Peña, M., Mehler, J., & Nespors, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological Science*, 22, 1419-1421.
- [7] Frishberg, N. (1975). Arbitrariness and iconicity: historical change in American Sign Language. *Language*, 696-719.

Individual differences in accent adaptation

Xin Xie & T. Florian Jaeger (University of Rochester)

Exposure to a talker with atypical pronunciations changes how this talker is perceived subsequently, often improving comprehension of that talker [1-2]. One line of research focuses on how listeners make adjustments when exposed to isolated words with atypical pronunciations of a particular contrast (e.g., /s/-/sh/; [3-4]). In these studies, as few as 10 words containing the sound can elicit significant adaptation for that contrast; post-exposure tests often involve a 2 Alternative Forced Choice (2AFC) task on minimal pairs and as such, listeners' attention is explicitly directed towards a particular sound/contrast. In contrast, research on the perception of globally accented L2 speech sometimes assumes that significantly more exposure is required for successful learning (but see [5-6]). These studies typically use transcription tasks to probe comprehension enhancements [1,7]. It is thus possible that adaptation proceeds equally rapidly in both paradigms but is left undetected in the transcription task. We report two experiments that directly compare listeners' adaptation to an unfamiliar L2 accent when assessed by a 2AFC or a transcription task. In addition, we ask whether the particular adaptation pattern is dependent on talker-specific properties by examining two test talkers (late L2 learners of English with intermediate intelligibility).

Methods. In two MTurk-based experiments (N = 47, 56), we assessed perception of word-final stop voicing in Mandarin-accented English, among native English listeners. Both experiments use the same stimuli and design (Fig.1), differing only in the task (2AFC vs. transcription). Between participants, both experiments manipulated whether exposure presented isolated spoken words from the L2-accented test talker (L2-accented exposure) or the same words from an L1 speaker (L1-accented exposure). This manipulation was crossed with the L2-accented test talker: half of the participants in each exposure condition heard test talker M4 and half heard M15. On each exposure trial, participants heard a word and had to choose which of two words on the screen they heard (2AFC) or had to transcribe the word (transcription). Either way, participants received immediate feedback about the correct response after each exposure trial. At test, participants completed the same task as during exposure but without feedback and on a new set of words. Critical trials (<50% in both exposure and test) involved minimal pairs with a word-final stop (e.g., 'seed' vs. 'seat'). Response accuracy was measured; transcriptions were considered as accurate if the voicing (voiced vs. voiceless) was correctly recognized.

Results and discussion. Data from each experiment was analyzed with logistic mixed-effects regression (accuracy ~ Test Talker * Condition * Voicing + maximal converging random effect structure; see Table 1). For the transcription task, our results showed a *Test Talker* effect (M15 > M4), a *Test Talker X Voicing* interaction, and critically, a three-way *Test Talker X Condition X Voicing* interaction. The three-way interaction was driven by a significant *Condition X Voicing* interaction for both talkers but in opposite directions: for M15, the experimental group had reduced accuracy for voiced tokens and increased accuracy for voiceless tokens, indicating a bias shifted towards voiceless tokens; for M4, there was an opposite bias shift towards voiced tokens. Further simple effect analyses were shown in Fig.2A. A similar pattern was observed for the 2AFC task, although the effects were overall smaller (Fig.2B). Pulling data from both experiments, there was an overall three-way interaction as observed separately for each task, with no *Test Talker X Condition X Voicing X Task* interaction. Taken together, the effects of accent exposure (~10 mins of exposure; replicating prior work using sentence stimuli [3-4]) were highly consistent across tasks but exhibited strikingly distinct patterns for two test talkers of the same L2 accent. Our finding offers two critical insights for future work on L2 speech perception. First, researchers should consider examining learning effects even within a short paradigm. Second and more importantly, the large by-talker differences in adaptation reveals an underestimated role of talker variability, even among talkers who are assumed to be extremely similar. Therefore, research on

accent adaptation—in particular, work focusing on cross-talker generalization—should be cautious drawing conclusions from just one test talker (cf. [1,2]).

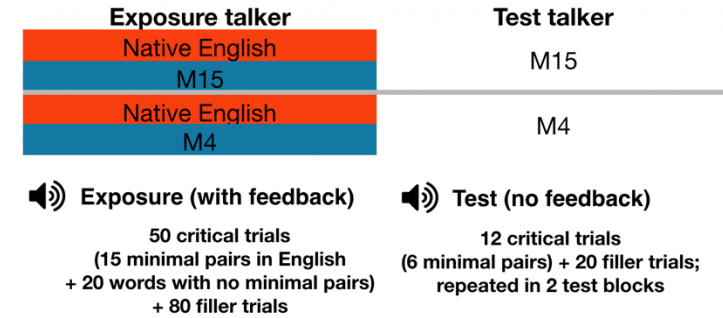


Figure 1.
Experimental design. In each experiment, two test talkers (M15 and M4) were employed. For each test talker, there were two exposure talker conditions (Native English. Vs. M15).

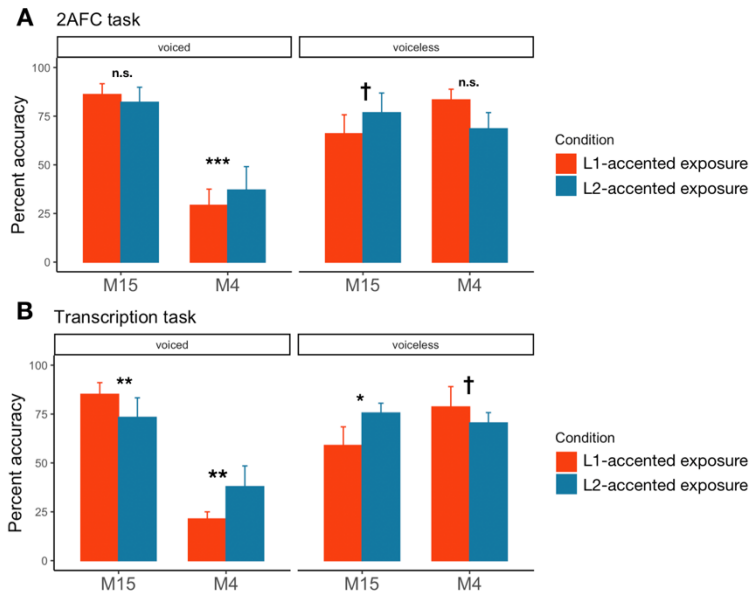


Figure 2.
Human responses. The x-axis shows performance for the two L2-accented test talkers.

* represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < .0001$ and † represents $p < 0.1$.

2AFC Task				
Levels	Coefficient $\hat{\beta}$	SE	z	p
Intercept	-0.591	0.204	-2.897	0.004 **
Talker	1.116	0.145	7.695	0.000 ***
Condition	0.376	0.153	2.451	0.014 *
Voicing	-0.795	0.181	-4.390	0.000 ***
Talker X Condition	-0.351	0.145	-2.429	0.015 *
Talker X Voicing	1.122	0.110	10.168	< 2.00E-16 ***
Condition X Voicing	0.247	0.121	2.043	0.041 *
Talker X Condition X Voicing	-0.525	0.110	-4.779	0.000 ***

Transcription Task				
Levels	Coefficient $\hat{\beta}$	SE	z	p
Intercept	0.723	0.271	2.674	0.008 **
Talker	0.715	0.112	6.396	< 1.60E-10 ***
Condition	0.011	0.125	0.091	0.928
Voicing	-0.409	0.260	-1.575	0.115
Talker X Condition	-0.164	0.111	-1.477	0.140
Talker X Voicing	0.938	0.084	11.158	< 2.00E-16 ***
Condition X Voicing	-0.013	0.100	-0.128	0.898
Talker X Condition X Voicing	-0.539	0.083	-6.481	< 9.09E-11 ***

Table 1. Mixed-effects model results.

* represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < .0001$ and † represents $p < 0.1$.

Multiverse analysis of eye-tracking data: Reexamining the ambiguity advantage effect

Caren Rotello & Brian Dillon (UMass Amherst), Caroline Andrews (University of Zürich)

Statistical analysis of eye-tracking-while-reading data involves many decisions. For example, researchers may analyze different dependent measures (e.g. regression path duration, or total reading time). These choices create multiple-comparisons issues in eye-tracking research, leading to unacceptably high Type I error rates [1]. However, similar multiple comparisons issues implicitly arise whenever a researcher faces choice points in constructing her dataset, such as when semi-arbitrary subject exclusion criteria are set [2]. Counterintuitively, the existence of these alternative datasets (Gelman's *garden of forking paths*) can create a multiple comparisons problem even if only a single dataset is analyzed, and only a single statistical test is ever performed [2]. This reality cannot be remedied using familiar corrections. A new strategy to manage this is *multiverse analysis*, which involves enumerating all plausible alternative datasets that could be used in statistical analysis (i.e., all reasonable choices a researcher might make), analyzing all possible datasets at once, and evaluating how robust the results are to different choice points in the analysis [3].

We apply the multiverse approach to eye-tracking-while-reading data. Eye-tracking data are a good candidate for multiverse analysis because there is often uncertainty about where (target or spillover) or in what measure post-lexical effects will be seen [4]. We investigate the *ambiguity advantage effect*, the finding that some globally ambiguous sentences are read more quickly than unambiguous counterparts [e.g., 5]. Participants ($N_{subj} = 84$) read sentences like (1) in either ambiguous (**AMBIG**) or unambiguous (**HIGH ATTACH, LOW ATTACH**) variants ($N_{item} = 27$, verb number counterbalanced across items). We identified 7 different choice points, such as the measure and region of interest (ROI) to use; see (2) for a summary. ROIs were generated by considering 1-4 word spans centered on the critical disambiguating word *was* and two spillover words. Taking all possible combinations of the decision points yielded 2,880 possible datasets. For each dataset, we fit a linear mixed-effects regression model to the data to estimate the effect of *AMBIGUITY* on RT; Random-effects structure for each model was determined using the parsimonious approach of [6]. We obtained the *p*-value for *AMBIGUITY* using the Satterthwaite approximation [7].

Figure 1 plots the distribution of *p*-values for the effect of *AMBIGUITY* across datasets. Overall, two analysis choice points substantially shift the distribution of *p*-values across datasets: the eye-tracking measure used, and the ROI. Figure 1 suggests limited evidence for the ambiguity advantage effect in first fixation measures, with a somewhat uniform distribution of *p*-values across datasets. In contrast the distribution of *p*-values is more concentrated below 0.05 in first pass, go-past, and especially total time measures. These trends interact with ROI: 'spillover only' ROIs that did not include the critical disambiguating word (*was*) were overall less likely to have *p*-values less than 0.05, but multiword ROIs including the disambiguating word revealed the opposite tendency. The other choice points considered largely did not systematically shift the distribution of *p*-values across datasets. Overall, our analysis yields evidence for the ambiguity advantage effect, but it does not appear in all combinations of ROI and eye-tracking measure. We suggest that multiverse analyses may profitably serve as guides for strong pre-registered studies on eye-tracking while reading.

(1) Edwin has been reading about...

AMBIG: the sister of the actor who was visiting the resort...

HIGH ATTACH: the sister of the actors who was visiting the resort...

LOW ATTACH: the sisters of the actor who was visiting the resort...

(2) Analysis choice points:

Eye-tracking measure: {first fixation, first pass, go-past time, total times}

Duration scale: {raw RT (ms), log-transformed RT (log ms)}

ROI: {was, was visiting, was visiting the, who was, who was visiting, who was visiting the, the, visiting, visiting the}

Exclude subjects by accuracy: {No cut off, >50% accuracy, > 60%, > 70%, > 80%}

Exclude subjects with excessive track loss: {Yes, No}

Exclude trials with first pass regression: {Yes, No}

Exclusion of fixations < 80ms or > 1000ms: {Yes, No}

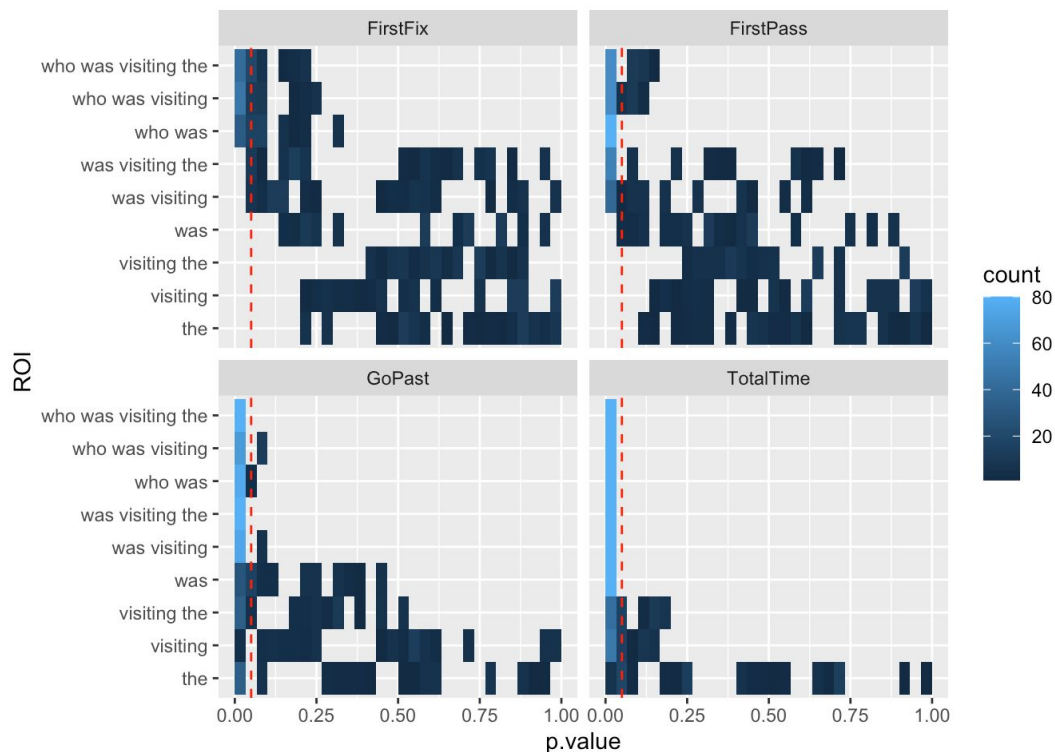


Figure 1: Distribution of p -values for effect of *AMBIGUITY* across datasets. Panels represent different eye-tracking measures. Different ROI are represented in the rows; each row has 80 total datasets.

[1] von der Malsburg & Angele (2017). *JML*. [2] Gelman & Loken (2014). *American Scientist*. [3]. Steegen et al. (2016). *Persp on Psych Science* [4] Clifton, Staub & Rayner. (2007) *Eye movements and reading*. [5] van Gompel et al. (2005). *JML*. [6] Matuschek et al. (2017). *JML*. [7] Kuznetsova et al. (2017). *J of Stat Software*.

Computational Estimation of Lexical Semantic Norms: A New Framework

Bryor Sneffjella & Idan Blank (UCLA)

The meanings stored in our mental lexicon are vast and varied, and include many dimensions (or “semantic features”) such as a word’s emotional attributes (e.g., valence, arousal), functional properties (e.g., usefulness), sensorimotor attributes (e.g., size, color), etc. (Binder et al., 2016; Lynott et al., 2019; Warriner et al., 2013). Over the past 60 years, the space of semantic features has been steadily increasing, yet the study of meaning has struggled with data sparsity throughout. Whereas English speakers know approximately 40,000 words, most semantic features have available behavioural ratings (“semantic norms”) for merely 1,000–10,000 words (Fig. 1), despite massive online crowdsourcing efforts at considerable cost. The default methodological solution is to limit statistical analyses only to the subset of words for which all semantic features of interest are available; any words with partial data are simply excluded. This precarious practice, known as listwise deletion or complete case analysis, is known to damage statistical power and can bias data analysis (Rubin, 2004).

A recent alternative to complete case analysis in the field of lexical semantics replaces expensive survey methods with “efficient” computational methods which have been shown to predict semantic norms with high accuracy (c.f. Hollis et al., 2017). This task is performed in two steps. First, a representation of words as high-dimensional vectors (“word embeddings”) is automatically generated from corpus co-occurrence data; then, the vector features are used as predictors in a machine learning algorithm that is trained on a small set of words for which norms have been empirically collected. This model then predicts the missing semantic norms based on those words’ embeddings. Such “extrapolated semantic norms” are now publicly shared and their use in statistical inference, in place of empirical norms, is an emerging practice.

Herein, we argue that both complete case analyses and norm extrapolation are statistically problematic. First, we show that words lacking empirical semantic norms are a non-random selection from the lexicon, making complete case analysis an unwise default practice. This problem has gone unacknowledged when semantic norms are used to predict behavior (e.g., lexical decision times) in megastudies, so the semantic effects discovered therein may have yielded biased results. Second, we claim that while norm extrapolation has been construed as a *prediction* problem, it should be conceived of as a *missing data* problem. To demonstrate the far-reaching statistical implications of this reframing, we draw upon principles of analysis of partially observed data, simulations, and empirical data.

Given the pattern of missing data and the misguided framing of the statistical problem at hand, deficiencies in current semantic norm extrapolation methods include (1) overconfidence, due to “forgetting” of the uncertainty in the imputation model; (2) biased statistical inference, particularly when testing hypotheses involving nonlinearities or interactions; and (3) inefficiency, due to a failure to take into account all relevant sources of information, and not accounting for missing data in variables other than semantic norms (namely, dependent variables in analyses, such as reading times). Practical solutions to these issues are offered by the technique of multiple imputation (Rubin, 2004). Our specific analysis pipeline uses a combination of LASSO variable selection (Tibshirani, 1996) and a model-based multiple imputation method (SMCFCS, Bartlett and Morris, 2015) embedded within multiple imputation by chained equations (MICE, van Buuren and Groothuis-Oudshoorn, 2011). We use simulation evidence to show these methods in concert can accommodate high-dimensional imputation with an analysis model potentially involving nonlinearities and interactions, and restore unbiased estimation with close to nominal confidence interval coverage. We also revisit theorized effects of words’ connotations of danger and usefulness (Wurm, 2007) in lexical decision, where our method yields qualitatively different results (Fig. 2) than the existing, naive extrapolation methods. Surprisingly, our results further indicate that given the particular nature of missing data, a proper implementation of semantic norm extrapolation via multiple imputation should in fact be preferred over the de-facto default use of complete case analysis in lexical semantics.

Figure 1

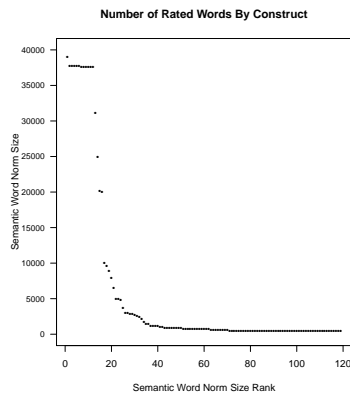
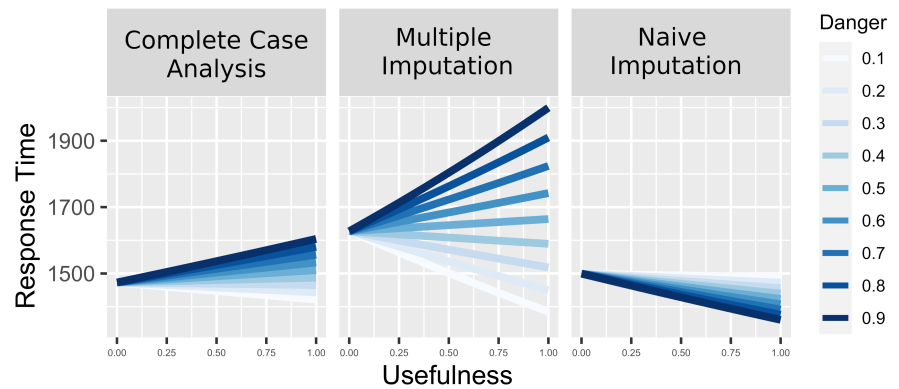


Figure 2

Usefulness by Danger Interaction



Leftmost Panel (Fig. 1): Number of words normed for 118 semantic features in the lexical semantics literature, ranked by number of words normed. Only a handful of semantic features have measurements matching the size of an average English speaker's lexicon.

Right Panels (Fig. 2): Interaction of word danger and usefulness on lexical decision response times in the English Crowdsourcing Project, as analyzed by a complete case analysis (left panel), after multiple imputation of danger and usefulness norms (middle panel), and after a naive imputation of danger and usefulness norms (right panel). The multiple imputation shows the predicted usefulness by danger interaction with correct functional shape, where high danger, high usefulness words yield slowed responses, but low danger, high usefulness words speed responses. This interaction is flipped and insignificant when danger and usefulness norms are imputed naively. A complete case analysis using empirical danger and usefulness norms shows an insignificant interaction of reduced magnitude.

References

- Bartlett, J. W., & Morris, T. P. (2015). Multiple imputation of covariates by substantive-model compatible fully conditional specification. *The Stata Journal*, 15(2), 437–456.
- Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4), 130–174.
- Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, 70(8), 1603–1619.
- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 1–21.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 1–67. <https://www.jstatsoft.org/v45/i03/>
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4), 1191–1207.
- Wurm, L. H. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review*, 14(6), 1218–1225.

Objective ages of acquisition for 3300+ simplified Chinese characters

Zhenguang G. Cai (The Chinese University of Hong Kong), Shuting Huang (The Chinese University of Hong Kong), Zebo Xu (The Chinese University of Hong Kong), Nan Zhao (Hong Kong Baptist University)

Age of acquisition (AoA) of a word refers to the age in which people first learn the word. Words that are acquired earlier in life, compare to late-acquired words, show processing advantages for participant in word recognition (e.g., Bylund, Abrahamsson, Hyltenstam, & Norrman, 2019) and spoken/handwritten word production (e.g., Chalard, Bonin, Méot, Boyer, & Fayol, 2003; Yum & Law, 2019). Besides, neuroimaging studies also shown that early-acquired words elicit greater activation in semantically related brain areas than late-acquired words (e.g., Fiebach, Friederici, Müller, von Cramon, & Hernandez, 2003). To facilitate research on AoA effects in Chinese, an AoA norms are required.

Following Liu, Shu and Li (2007) and Shu, Chen, Anderson, Wu and Xuan (2003), we constructed two AoA norms, one norm based on 18 textbooks of Chinese (corresponding to 18 terms from Grade 1 to 9) published by the People's Education Press in response to the 2001 national curriculum, and the other based on the 18 textbooks of Chinese by the same publisher in light of the 2011 national curriculum. One term is equated to 0.5 years. As children start Grade 1 at 6 years old, a character learned in the Term 1 then has an AoA of 6.5 years. There are AoAs for 3358 characters in the 2001 norm and 3395 characters in the 2011 norm (with 3013 characters available in both norms).

Descriptive results show that the current norms have significantly larger coverage of characters than previous norms developed by Shu et al. (2003) and Liu et al. (2007). We then compare these norms in terms of predicting behavioural indices in four large-scale psycholinguistic databases of simplified Chinese character processing: Sze, Rickard Liow, & Yap (2014; a character decision database), Tsang et al. (2018; a character decision database), Liu et al. (2007; a character naming database), and Wang, Huang, Zhou, & Cai (2020; a character handwriting database). As can be seen in Table 1, the current norms outperformed previous norms in explaining the behavioural data (e.g., the new norms had the two largest adjusted R^2 s in 6 out of the 8 comparisons). To further quantify such explanatory power differences, we also used the Bayes factor to assess how good a model is compare to another basing on the common characters. As is shown in Table 2, models using the 2001 AoA norm outperformed models using the Shu or Liu norm in 12 out of the 16 comparisons. Models using the 2011 norm outperformed their alternative models in 15 out of 16 comparisons (see Table 3). We also conducted a comparison between the two current norms, results shown that models with the 2011 norm outperformed models with the 2001 norm in 5 of the 8 comparisons and was outperformed in 1 comparison, with evidence being not clear in 2 other comparisons (see Table 4).

The developed objective AoAs are available at Open Science Framework (<https://osf.io/j587y/>) and can be used for subsequent research on Chinese character recognition or production.

Table 1: Regression results using different AoA norms on the four databases. All p -values for the t -tests are $< .001$; the largest adjusted R^2 is in **bold** and the second largest in ***italic bold***.

	Accuracy			Reaction time		
	β	t	R^2_{adj}	β	t	R^2_{adj}
<i>Sze et al. (2014)</i>						
N2001	-0.011	-14.14	0.086	0.088	28.28	0.273
N2011	-0.011	-15.89	0.103	0.080	29.14	0.279
Shu	-0.006	-7.26	0.028	0.069	18.30	0.159
Liu	-0.006	-11.23	0.073	0.58	21.37	0.224
<i>Tsang et al. (2018)</i>						
N2001	-2.200	-13.31	0.219	0.103	18.06	0.341
N2011	-2.050	-13.49	0.221	0.103	18.23	0.341
Shu	-1.649	-6.82	0.085	0.082	9.80	0.163
Liu	-1.538	-10.36	0.176	0.074	14.94	0.309
<i>Liu et al. (2007)</i>						
N2001				16.549	25.93	0.244
N2011				15.651	26.48	0.246
Shu				15.510	15.68	0.121
Liu				12.670	30.44	0.279
<i>Wang et al. (2020), accuracy and latency</i>						
N2001	-0.049	-21.43	0.237	0.190	25.72	0.309
N2011	-0.049	-21.41	0.234	0.190	25.12	0.296
Shu	-0.029	-10.66	0.082	0.146	16.36	0.175
Liu	-0.036	-16.81	0.187	0.128	18.59	0.219
<i>Wang et al. (2020), duration</i>						
N2001	0.138	12.70	0.098			
N2011	0.127	11.71	0.083			
Shu	0.148	10.96	0.087			
Liu	0.100	10.08	0.076			

Visual recognition of morphologically complex words by second language learners: A masked priming study

Mariia Baltais (Ghent University), Anna Jessen (University of Potsdam)

Many recent studies examined early visual recognition of morphologically complex words in native (L1) speakers by conducting masked priming experiments, in which prime words are presented so briefly that they are typically not consciously visible to the readers. Lexical decisions to the target words preceded by morphologically related primes (e.g., *walked* – *walk*) tend to be significantly shorter than those obtained in the unrelated condition (e.g., *brush* – *walk*). This facilitation is attributed to automatic decomposition of the primes into their constituent morphemes (*walk* + *-ed*). However, it is not clear whether this mechanism applies to irregular inflections (e.g., *taught*), or they are stored and processed as indecomposable whole forms. Research on non-native (L2) visual recognition of regular and irregular inflections also provides controversial results (Neubauer & Clahsen, 2009; Clahsen & Jessen, 2020). According to the Shallow Structure Hypothesis, L2 processing tends to rely on “shallow” parsing strategies, which make use of surface information (Clahsen & Felser, 2006). Some masked priming studies found that in contrast to native speakers, L2 speakers can show effects of purely orthographic relatedness (Feldman et al., 2010).

To examine early visual processing of inflected words in L2 speakers, we tested 63 highly proficient Russian learners of German (mean age of acquisition: 13 years, SD: 5.93, range: 5-30 years) and compared their results to a control group of 32 German native speakers. Experimental materials come from the study by Clahsen and Jessen (2020). In the morphological item set, three types of German past participles were used as related primes: regular, irregular without stem allomorphy, and irregular with stem allomorphy (see Table 1). The corresponding verbs in the first person singular of the present tense acted as targets. Purely orthographically and purely semantically related primes and targets formed two control item sets. All related and unrelated prime-target pairs were distributed over two counter-balanced lists, so that each participant would see each target only once. Each list consisted of 150 experimental and 210 filler pairs; in half of the trials, targets were non-words. Prime words were preceded by a hash mask and appeared on the screen for 50 ms (see Figure 1).

Lexical decision times were analyzed by fitting mixed-effect linear regression models. The L1 group showed genuine morphological priming for all types of inflections, indicative of their morphological decomposition. Similar priming effects, including numerical patterns, were observed for the Russian L2 group (see Figure 2). However, statistical models could not reliably distinguish these latter effects from facilitation caused by purely orthographic overlap between control primes and targets: there were no significant interactions of prime type (related, unrelated) and relatedness type (morphological, orthographic). It suggests that orthographic relatedness could have played a role in the L2 speakers' early recognition process of inflected words.

Our results can be directly compared to those obtained by Clahsen and Jessen (2020) for Turkish-German bilinguals, who demonstrated genuine morphological priming for regular inflections and no purely orthographic facilitation. One of the possible explanations for greater importance of orthographic relatedness for the Russian speakers might be the difference in L1 and L2 scripts: Russian uses a Cyrillic-script alphabet while both Turkish and German use Latin-script alphabets. Feldman et al. (2010) also tested a group of L2 speakers whose L1 had a Cyrillic-script alphabet (Serbian learners of English): orthographic and irregular primes produced statistically undistinguishable patterns, although regular primes yielded genuine morphological facilitation in more proficient L2 speakers. However, the present study demonstrates that even highly proficient L2 speakers can rely on orthographic information during the processing of regular inflections, which goes in line with the Shallow Structure Hypothesis.

Condition	Target	Related prime	Unrelated prime	No of targets
Regular (-t)	<i>lande</i> 'I land'	<i>gelandet</i> 'landed'	<i>furchtbar</i> 'awful'	30 items
Irregular (-n, no stem change)	<i>falle</i> 'I fall'	<i>gefallen</i> 'fallen'	<i>klüger</i> 'smarter'	30 items
(-n, stem change)	<i>finde</i> 'I find'	<i>gefunden</i> 'found'	<i>herrlich</i> 'splendid'	30 items
Orthographic	<i>Lasche</i> 'strap'	<i>Flasche</i> 'bottle'	<i>Herbst</i> 'autumn'	30 items
Semantic	<i>Arzt</i> 'physician'	<i>Doktor</i> 'doctor'	<i>Presse</i> 'press'	30 items

Table 1. Stimulus examples (based on Clahsen & Jessen, 2020).

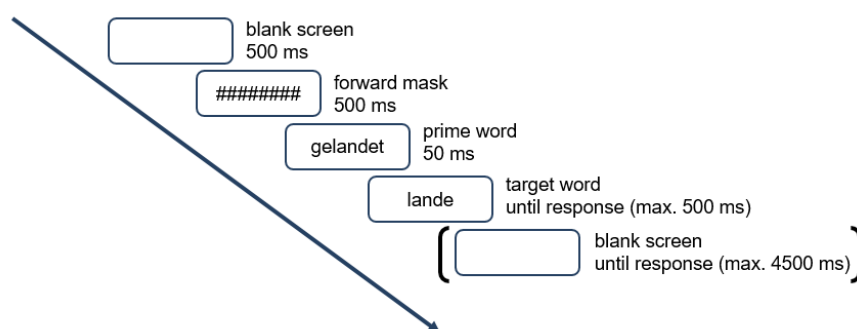


Figure 1. Structure of a trial (based on Clahsen & Jessen, 2020). The number of hashes in the forward mask was equal to the number of letters of the prime word.

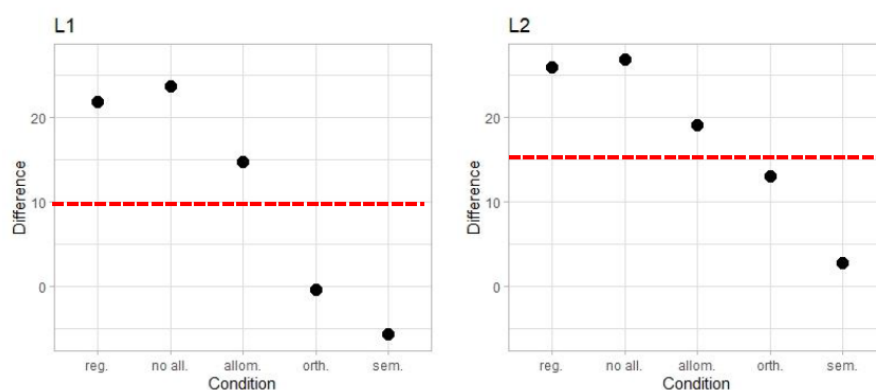


Figure 2. Priming effects (in milliseconds) obtained for each group (L1, L2) and condition (regular participles, irregular participles without allomorphy, irregular participles with allomorphy, orthographic, semantic). The dashed line represents the significance level ($p < .05$).

References

- Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, 27(1), 3–42. <https://doi.org/10.1017/S0142716406060024>
- Clahsen, H., & Jessen, A. (2020). Variability and its limits in bilingual word recognition: A morphological-priming study. *The Mental Lexicon*, 15, 292–326. <https://doi.org/10.1075/ml.20013.cla>
- Feldman, L. B., Kostić, A., Basnight-Brown, D. M., Đurđević, D. F., & Pastizzo, M. J. (2010). Morphological facilitation for regular and irregular verb formations in native and non-native speakers: Little evidence for two distinct mechanisms. *Bilingualism: Language and Cognition*, 13(2), 119–135. <https://doi.org/10.1017/S1366728909990459>
- Neubauer, K., & Clahsen, H. (2009). Decomposition of Inflected Words in a Second Language: An Experimental Study of German Participles. *Studies in Second Language Acquisition*, 31(3), 403–435. <https://doi.org/10.1017/S0272263109090354>

Online Processing of Derived and Inflected Words in L1 Turkish: A Masked Priming Experiment

Refika Cimen, Filiz Cele (Istanbul Aydın University)

In recent decades, a considerable amount of psycholinguistic research has been driven by the question of whether morphologically-complex word forms are decomposed or stored as full-forms in the mental lexicon. In general, earlier studies investigated complex morphological processes, especially inflection, in verbal forms and suggested a decompositional pattern for processing these forms in L1. Yet, the results of a limited number of studies which examined inflectional processing in nominal forms raises the question of whether word category can be a determining factor for the preferred morphological processing route [1-2] since they show that L1 speakers may make less use of rule-based decomposition on the processing of inflected nouns. Due to a lack of studies focusing on the role of word category in the investigation of morphological processing, the present study aimed to provide a broader picture of L1 morphological processing by investigating the inflectional and derivational paradigms in both nominal and verbal forms via a masked priming experiment with 24 adult L1 speakers of Turkish, an agglutinative language with rich morphology. The experimental stimuli consisted of a nominal list (Fig. 1) and a verbal list (Fig.2), both of which were designed to form six different conditions (i.e., Identity, Derivation, Inflection, Semantic, Orthography, and Unrelated) with the same targets to achieve a direct comparison between different types of primes. The orthographically- and semantically-related primes were included in order to determine whether any priming effects had a morphological nature or not. Each list involved 36 experimental items and 180 fillers. The experiment started with the presentation of a forward mask (#####) on the screen as a fixation point for 500 ms., which was immediately followed by the prime word, which was presented in lowercase letters on the screen for 50 ms. The target word appeared on the screen right after the prime word in uppercase letters, which was meant to prevent visual priming by minimizing any orthographic overlap. The target word remained on the screen for 5000 ms, during which the participants pressed 'yes' or 'no' buttons on the keyboard to indicate whether the target word was a word or a non-word. A repeated-measures ANOVA was conducted on the participants' responses to experimental stimuli and a significant interaction was found between prime type and response time (RT) ($p < .05$). Pairwise comparisons on the verbal stimuli revealed repetition priming effects in Identity, while the absence of a statistically significant difference in the mean RT between Identity and the two morphological conditions, i.e., Derivation and Inflection, ($p = 1.000$) was indicative of a full-priming pattern (Fig. 3). Yet, the mean RTs in Inflection and Derivation differed significantly from the mean RTs in Semantic and Orthography ($p < .05$), indicating that the full-priming effects were not due to an orthographic overlap or a semantic relationship between the prime-target pairs. The nominal stimuli, on the other hand, yielded a different result with respect to Inflection (Fig. 4). While pairwise comparisons revealed repetition priming in Identity and full-priming in Derivation, no priming effects were obtained in Inflection, suggesting that the participants showed lower sensitivity to the morphological structure of inflected nouns. On the other hand, the mean RTs in Semantic and Orthography differed from the mean RT in Derivation significantly ($p < .05$), indicating that the full-priming effects yielded by the derived primes were purely morphological in nature. Our findings show that morphologically-complex verbs, whether derived or inflected, are decomposed in L1 Turkish, whereas morphologically-complex nouns are only decomposed when they are derivational. This suggests that rule-based processing for accusative nominal marker could be a time-costly process, which could be attributed to the frequency of this marker as a nominal ending. Therefore, we conclude that word category can be a determining factor in the processing route of morphology for L1 Turkish speakers.

Keywords: morphological processing, masked priming, Turkish, inflection, derivation, decomposition, full-listing, L1 speakers

Figure 1. A sample set of noun stimuli

Target	Prime Type					
	Identity	Inflection -(y)l	Derivation (-CI)	Orthography	Semantic	Unrelated
BÜYÜ 'spell'	büyü 'spell'	büyüyü 'spell-ACC'	büyücü 'wizard'	büyük 'big'	sihir 'magic'	şeker 'sugar'
KİRA 'rent'	kira 'rent'	kirayı 'rent-ACC'	kiracı 'tenant'	kiraz 'cherry'	ev 'home'	duygu 'emotion'

Figure 2. A sample set of verb stimuli

Target	Prime Type					
	Identity	Inflection (-SA)	Derivation (-IM)	Orthography	Semantic	Unrelated
BAKMAK 'to look'	bakmak 'to look'	baksa 'if he/she looked'	bakım 'care'	bakkal 'grocery store'	gör 'see'	tüket 'consume'
SATMAK 'to sell'	satmak 'to sell'	satsa 'if he/she sold'	satım 'sale'	saten 'satin'	ödemek 'to pay'	çevir 'turn'

Figure 3. Results of the verbal stimuli

	Identity	Inflection	Derivation	Orthography	Semantic	Unrelated
RTs	631.96	658.29	650.29	727.90	709.26	710.37
(SDs)	(113.41)	(110.88)	(124.89)	(95.61)	(106.07)	(114.94)
Error Rate (%)	2.08	0.69	1.39	0.69	5.56	4.86
Priming Effect	78.41	52.08	60.08		1.11	

Figure 4. Results of the nominal stimuli

	Identity	Inflection	Derivation	Orthography	Semantic	Unrelated
RTs	605.70	684.65	621.70	716.54	703.48	692.36
(SDs)	(105.99)	(102.42)	(106.28)	(97.78)	(98.50)	(96.56)
Error Rate (%)	0.69	0	2.78	0.69	3.47	2.78
Priming Effect	86.66	17.71	60.66			

References:

[1] Ahn, H. D., Cho, Y., Hwang, J. B., Jeon, M., Jeong, K., & Kim, J. (2014). L2 morphological processing of Korean nominal marker -ka: Evidence from masked and cross-modal priming with advanced Chinese learners. *Linguistic Research*. [2] Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*.

ERP responses to lexical-semantic processing differentiate toddlers at high clinical risk for autism and language disorder

Chiara Cantiani, Valentina Riva, Chiara Dondena, Elena Maria Riboldi, Maria Luisa Lorusso, Massimo Molteni (Child Psychopathology Unit, Scientific Institute, IRCCS Eugenio Medea, Bosisio Parini, Lecco, Italy)

Delays in early expressive vocabulary are among the most common reasons motivating diagnostic evaluation (Morgan et al., 2020). These symptoms might be the first signs of the presence of a neurodevelopmental disorder, including developmental language disorder or Autism Spectrum Disorder (ASD). Although many typical symptoms can differentiate children with early signs of ASD from those who specifically show delays in language development, the two populations show striking similarity when focusing only on the observed language impairment (Paul et al., 2008). It seems important to determine how the brains of these young children in the early phases of language acquisition work. We expected differences between groups suggesting different mechanisms in processing words in meaningful contexts.

Here, we directly compared two groups of 19-month-old toddlers identified via clinical assessment as being at risk for such neurodevelopmental disorders, i.e., children characterized by low expressive vocabulary (LDS score $\leq 15^{\circ}$ percentile; Late Talkers, LT, $N=18$), children with early symptoms of ASD (ADOS-2 total score ≥ 6 ; ASD, $N=18$), and a group of typically developing children (TD, $N=28$). Specifically, we investigated the electrophysiological underpinnings of the (dis)ability to establish the first lexical–semantic representations during the critical phase of lexical acquisition, with the aim of identifying similarities and specificities among these groups in lexical-semantic processing. Event-Related Potentials (ERPs) elicited by words (either congruous or incongruous with the previous picture context; Match or Mismatch, M vs. MM) and pseudo-words (PW) are investigated within a picture-word matching paradigm (Cantiani et al., 2017) in the three groups, considering the three specific ERP components reported in toddlers for similar tasks (i.e., phonological-lexical priming effect; N400; Late Positive Component), and investigating longitudinal intra-group associations with language and socio-communications skills at age 24 months.

As expected, we found differences between the groups that might underlie specificities, but also similarities. Whereas no or subtle differences emerged across groups concerning the N400 component, the two clinical groups differed significantly from the typically developing group in the other two ERP components. On the one side, the LT group differed from the other two groups in the phonological-lexical priming effect, reflecting detection of the correspondence between the heard word and the lexical representation pre-activated by the picture ($F(4,103) = 2.998$, $p = .029$, $\eta^2 = .089$). Specifically, they showed no evidence for this effect (see Figure 1), suggesting that they miss the early automatic recognition of incongruencies between what is heard and what is expected (e.g., Torkildsen et al., 2009). On the other side, the ASD group differed from the other two groups in the Late Positive Component, reflecting the effortful semantic reanalysis following a violation ($F(2,61) = 4.306$, $p = .018$, $\eta^2 = .124$). Specifically, they were characterized by a complete lack of such component (see Figure 2), indicating that higher-level processing mechanisms of re-analysis of the violation are missing in this population (e.g., DiStefano et al., 2019). The functional interpretation of the two components is corroborated by significant correlations suggesting that the early component is associated with later socio-communication skills whereas the late component is associated with linguistic skills.

The results point in the direction of differential impaired mechanisms in the two populations (i.e., impaired automatic detection of the incongruencies in LT vs. absence of high-level reanalysis of such incongruencies in the children with early signs of ASD). The differential impaired mechanisms emerged in the present study could inform the definition of early interventions for populations at high risk for neurodevelopmental disorders because showing the very first clinical signs.

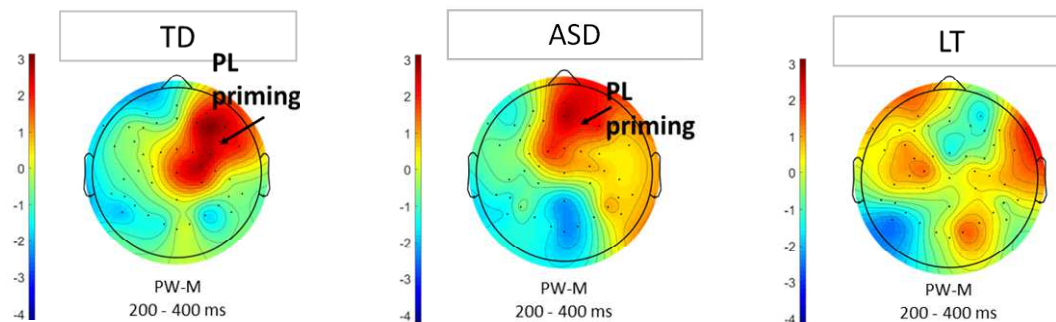


Figure 1. Topographical maps relative to the Phonological-Lexical priming effect. Distribution of difference waveforms (PW-M) is shown for the selected Time-Windows (200-400 msec).

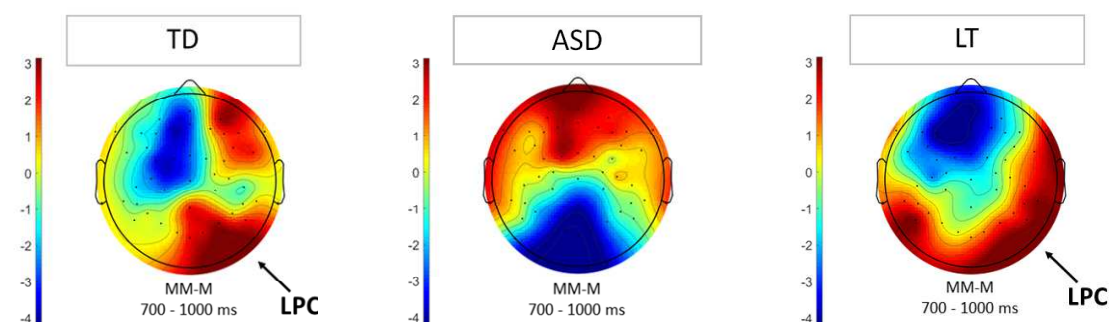


Figure 2. Topographical maps relative to the LPC. Distribution of difference waveforms (MM-M) is shown for the selected Time-Windows (700-1000 msec).

Selected references:

- Cantiani, C., Riva, V., Piazza, C., Melesi, G., Mornati, G., Bettoni, R., Marino, C., & Molteni, M. (2017). ERP responses to lexical-semantic processing in typically developing toddlers, in adults, and in toddlers at risk for language and learning impairment. *Neuropsychologia*, 103, 115–130.
- DiStefano, C., Senturk, D., & Jeste, S. S. (2019). ERP evidence of semantic processing in children with ASD. *Developmental Cognitive Neuroscience*, 36, 100640.
- Lord, C., DiLavore, P., Gotham, K., Guthrie, W., Luyster, R. J., Risi, S., & Rutter, M. (2012). *Autism Diagnostic Observation schedule: ADOS-2*. Western Psychological Services.
- Matson, J. L., Mahan, S., Kozlowski, A. M., & Shoemaker, M. (2010). Developmental milestones in toddlers with autistic disorder, pervasive developmental disorder not otherwise specified and atypical development. *Developmental Neuropsychology*, 13(4), 239–247.
- Paul, R., Chawarska, K., & Volkmar, F. (2008). Differentiating ASD From DLD in Toddlers. *Perspectives on Language Learning and Education*, 15(3), 101–111.
- Rescorla, L., & Alley, A. (2001). Validation of the Language Development Survey (LDS). *Journal of Speech, Language, and Hearing Research*, 44(2), 434–445.
- Torkildsen, J. V. K., Friis Hansen, H., Svagstuen, J. M., Smith, L., Simonsen, H. G., Moen, I., & Lindgren, M. (2009). Brain dynamics of word familiarization in 20-month-olds: Effects of productive vocabulary size. *Brain and Language*, 108(2), 73–88.

Individual differences in language ability: Quantifying the relationships between linguistic experience, general cognitive skills and linguistic processing skills

Florian Hintz (Max Planck Institute for Psycholinguistics), Cesko C. Voeten (Leiden University), Christina Isakoglou (Radboud University), James M. McQueen (Radboud University), Antje S. Meyer (Max Planck Institute for Psycholinguistics)

Most people acquire their native language effortlessly, yet individuals differ greatly in how they use it. Language ability strongly influences people's functioning in society and is an important predictor of professional success. Although the question 'what makes someone a good language user?' has intrigued scholars for a long time, little is known about the principal dimensions of language skills. Most studies adopted qualitative approaches (i.e., asking 'is X involved in Y?') and focused on the involvement of just one variable in linguistic processing skills (e.g., working memory in sentence comprehension). Such approaches ignore the contribution of other potentially relevant variables and do not allow for a quantification of the relationships between multiple potentially relevant variables. Therefore, the conclusions that may be drawn about the principal dimensions of language skills are limited.

In the present study, we took a first step towards a comprehensive characterization of individual differences in language skills. We tested 112 young adults (aged between 18 and 29 years) in a lab-based setting on a recently developed behavioral test battery (Hintz et al., 2020). The battery included 33 tests designed to assess nine key constructs reflecting language skills and skills assumed to be involved in linguistic processing: Word production, Word comprehension, Sentence production, Sentence comprehension, Linguistic experience, Non-verbal processing speed, Working memory, Inhibition, and Non-verbal intelligence. Except for non-verbal intelligence, we included multiple tests per psychological construct to address task impurity.

Using principal component analysis (PCA, number of expected components was unconstrained) we first assessed how strongly each individual test loaded on the construct it was assumed to measure. The results showed that the majority of tests loaded strongly on their respective construct, none of the PCAs yielded more than one component (Table 1, for an overview). Then one score for each of the nine constructs was extracted for each participant. These scores were submitted to a correlation analysis. The correlations among the nine scores are presented in the heatmap in Figure 1 (Panel A). Finally, the correlation matrix from Panel A was converted into a distance matrix and then submitted to a hierarchical clustering analysis. Panel B plots the outcome of this analysis as a dendrogram. Correlation and hierarchical clustering analyses revealed strong correlation/similarity between non-verbal processing speed and language comprehension, especially word comprehension. Moreover, we observed a strong correlation between linguistic experience and language production, especially word production. While word-level and sentence-level skills *within* a domain were related, the hierarchical clustering analysis yielded separate clusters for comprehension and production. In line with previous research, working memory, non-verbal intelligence, and to a lesser extent inhibition clustered together. These general cognitive skills correlated weakly to moderately with linguistic processing skills and formed a separate cluster in the hierarchical clustering analysis.

In sum, the present study constitutes a first step towards a comprehensive, quantitative characterization of individual differences in language skills. Our results extend previous research by demonstrating a strong influence of general cognitive skills (i.e., processing speed) on comprehension and of linguistic experience on production. The present data further suggest that production and comprehension skills are less related than one might have thought.

We are currently testing a larger sample of participants with diverse educational backgrounds using versions of the tests presented here that can be run via the internet. Moreover, next to charting the variability in language skills at the behavioral level, we will investigate its neurobiological and genetic underpinnings.

Table 1: The table presents the loadings of the individual tests on the construct they were assumed to measure as well as the amount of variance explained, established using PCA.

Word production (42% variance explained)	Word comprehension (43% variance explained)	Sentence production (61% variance explained)	Sentence comprehension (55% variance explained)
Picture naming -.53	Non-word monitoring noise -.44	Phrase generation .78	Gender cue activation .91
Antonym production -.54	Word monitoring noise -.27	Sentence generation .78	Verb semantics activation .91
Verbal fluency (Sem.) -.71	Meaning monitoring noise -.42		Monitoring noise .02
Verbal fluency (Phon.) -.72	Rhyme judgment -.85		
Maximal speech rate -.48	Auditory lexical decision -.85		
One minute test -.76	Semantic categorization -.83		
Klepel test -.72			
Linguistic experience (58% variance explained)	Non-verbal processing speed (53% variance explained)	Working memory (48% variance explained)	Inhibition (56% variance explained)
Peabody test .84	Auditory simple RT .74	Digit span (forward) .80	Eriksen flanker task .75
Spelling test .75	Auditory choice RT .85	Digit span (backward) .77	Antisaccade task .75
ART .82	Letter comparison .47	Corsi block (forward) .61	
Idiom recognition .54	Visual simple RT .73	Corsi block (backward) .56	
Prescriptive grammar .83	Visual choice RT .82		

Note: Non-verbal intelligence is not listed as it was measured using a single test.

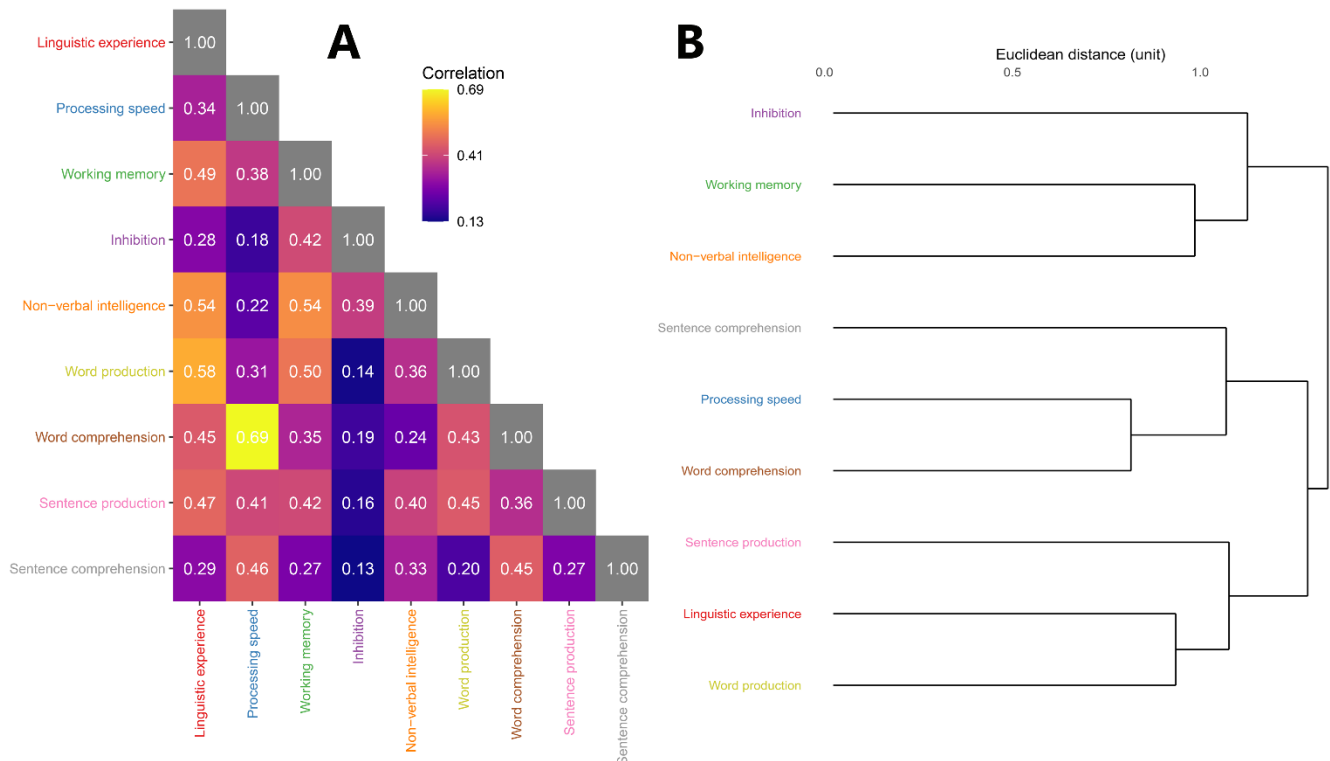


Figure 1: Panel A presents a correlation matrix based on the nine PCA-derived scores. The scale ranges from the weakest to the strongest correlation between any two scores in the set. Panel B presents a dendrogram as outcome of the hierarchical clustering analysis, based on the correlations in Panel A. In the dendrogram, scores that are similar (i.e., closer) cluster together.

Reference

Hintz, F., Dijkhuis, M., Van 't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020). A behavioural dataset for studying individual differences in language skills. *Scientific Data*, 7: 429.

Exposure to Plurals Can Help or Hurt Plural Production

Justin B. Kueser^a, Ryan Peters^b, Pat Deevy^a, and Laurence B. Leonard^a

^a Department of Speech, Language, and Hearing Sciences, Purdue University

^b Department of Psychology, University of Texas at Austin

Research Question

Child language researchers have long sought to explain the “U-shaped” pattern of children’s grammatical development. For example, after a brief period of using irregular plurals such as *mice*, many children go through a period of changing their production to *mouses*, demonstrating over-application of the regular plural -s rule. Recent work from a discriminative learning perspective has emphasized how the learning of inflectional morphology (and exceptions) depends on both positive and negative evidence (Baayen et al., 2011). Consistent with this perspective, Ramscar et al. (2013) found that massed exposure to regular plurals caused counterintuitive changes in the production of irregular plurals by children with typical development (TD). Children with TD early in the acquisition of irregular plurals learned from the exposure in a *positive* way, demonstrating an *increase* in overregularization of irregular plurals compared to a pretest. Older children with TD learned from the exposure in a *negative* way, demonstrating a *decrease* in overregularization. Importantly, both groups received the same exposure to regular plurals, suggesting that a child’s learning from plurals in the input depends crucially on what the child currently knows.

This has important implications for children with developmental language disorder (DLD), who demonstrate much difficulty learning irregular plurals (e.g., Oetting & Rice, 1993). In the current study, we examined how exposure to plural nouns affects the production of irregular, zero, and regular plurals (e.g., *mice*, *deer*, *cats*) in children with DLD and with TD.

Method

Our participants included 20 four-to-five-year-old children with DLD and 20 age-matched children with TD. Children completed a pretest in which they named six images each of irregular, zero, and regular plurals. The children were then exposed to 96 images of regular plurals; half had appeared on the pretest. As a cover task, the children said whether those things were on the pretest but the words were not spoken. After this intervention, the children received a posttest identical to the pretest to examine the influence of the intervention on plural production.

We predicted that children with DLD would demonstrate less accuracy than children with TD. We therefore expected that the effect of the intervention for children with DLD would differ from that for children with TD because the effect of the intervention was shown to be dependent on the accuracy of production by Ramscar et al. (2013). We also expected that the intervention would have a positive effect on regular plurals and an effect on irregular and zero plurals that was dependent of children’s knowledge of these plurals.

Analysis and Results

The accuracy of the children’s production of plural nouns was analyzed in a mixed-effects logistic regression model. The results of the model are presented in Table 1 and Figure 1. We found better posttest than pretest performance for irregular and zero plurals despite the fact that the intervention consisted of regular plurals only, though the groups did not demonstrate different effects of the intervention as predicted. We also found that regular plurals demonstrated a decrease in accuracy from pre- to posttest for both groups.

Implications

The results were in part consistent with prior work in that posttest performance improved compared to pretest for irregular and zero plurals. However, the finding that regular plurals demonstrated a decrease in performance suggested that children were not just responding to the intervention but also to the words on the tests. This suggests that plural acquisition and processing depend on a complex mix of factors: the specific stimuli heard and seen, the

frequency of the stimuli in contrast to their frequency in the language as a whole, and the state of children's knowledge. These results also have important implications for the assessment and teaching of plurals in children with DLD.

Table 1. Statistical model results.

Effect	<i>df</i>	χ^2	<i>p</i>
Group	1	4.83	.028 *
Test	1	0.17	.682
Type	2	33.04	<.001 ***
Freq	1	10.42	.001 ***
Group x Test	1	0.26	.608
Group x Type	2	18.43	<.001 ***
Test x Type	2	7.45	.024 *
Group x Test x Type	2	0.42	.813

Note. Random intercepts for participant and item were included. *p* values are based on likelihood ratio tests. Test = pre- or post-test; Type = irregular, zero, or regular plurals; Freq = log frequency of lemma in CHILDES; Group = TD or DLD.

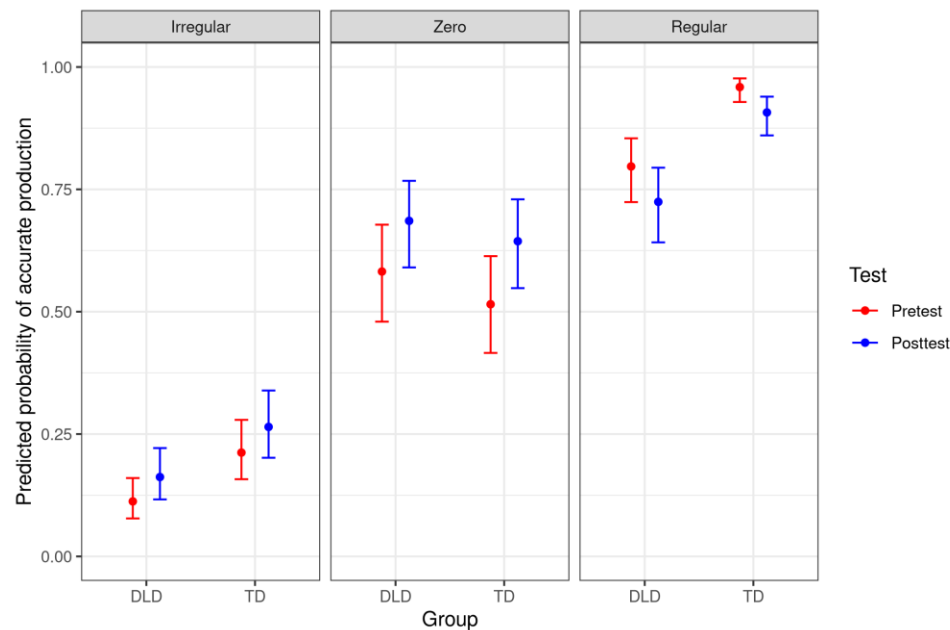


Figure 1. Model predictions for accurate production during pre- and posttest for regular, irregular, and zero plurals. Error bars are standard errors. The likelihood of correct production of each type of plural was significantly different from the others. The TD group demonstrated significantly more accuracy than the DLD group for all plural types except for zero plurals, where there was no significant group difference. The effect of test was significantly different between irregular and regular plurals and between zero and regular plurals, but not between irregular and zero plurals. Pretest accuracy was poorer than posttest accuracy for irregular and zero plurals but better for regular plurals. The effect of test did not significantly differ between groups.

References

- Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3), 438–481. <https://doi.org/10.1037/a0023851>
- Oetting, J. B., & Rice, M. L. (1993). Plural acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 36, 1236–1248. <https://doi.org/10.1044/jshr.3606.1236>
- Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mice in adult speech. *Language*, 89(4), 760–793. <https://doi.org/10.1353/lan.2013.0068>

Recovery from semantic prediction violations during sentence processing in preschoolers with Developmental Language Disorder

Michelle Indarjit, Mariel Schroeder, Patricia Deevy, Laurence Leonard & Arielle Borovsky (Purdue University)

Prediction is one mechanism that is thought to promote rapid spoken language comprehension from at least age 2 (Mani & Huettig, 2012). Importantly, listeners generate graded lexical predictions for a range of expected less-expected but semantically-related items (Federmeier & Kutas, 1999). Some early evidence suggests these graded mechanisms of prediction might vary in children with Developmental Language Disorder (DLD). For example, typically-developing (TD) children and adults predictively fixate to verb-related items that are not highly expected given the entire sentence context (e.g. in a sentence like *The pirate chases the...* listeners will fixate largely to a highly expected item *SHIP*, and to a lesser degree, towards a less-expected, but chase-able *CAT*). However, adolescents with DLD show robust prediction for highly-expected sentence objects, but do not activate (i.e. fixate towards) less-expected objects (Borovsky, Burns, Elman & Evans, 2013). Moreover, children with DLD show lexico-semantic deficits (Sheng & McGregor, 2010) and slower speed of processing in off-line sentence comprehension tasks compared to TD peers (Montgomery, 2000). Here, we ask whether and how these differences in mechanisms of semantic activation during online sentence processing in DLD affect comprehension in unexpected sentence contexts.

Preschoolers (aged 4;0-6;0) with DLD ($n=19$) and TD ($n=23$) completed an eye-tracked sentence recognition task that sought to explore how quickly children in each group recovered from their initial predictions for a highly expected item when the sentence ended with a less-expected object. Children were asked to select images from 4-picture arrays that matched with spoken SVO sentences containing an informative agent and verb (*The pirate chases the...*), followed by object endings in two conditions: (1) unexpected action-related objects (UAR; *CAT*) or (2) unexpected, action-unrelated (UAU) objects, where the ending did not coordinate with the verb (*BONES*). Images from both conditions were present on screen, and children also saw filler sentence trials with expected sentence endings. We compared fixations towards the named target object in the two unexpected conditions in both groups.

Results highlight differences in looks to the named object in the UAR and UAU conditions between conditions and groups (Figures 1 & 2). Children with DLD looked more towards the target in the UAR condition ($M = 0.57$, $SD = 0.13$) compared to UAU condition ($M = 0.47$, $SD = 0.09$), $t(18) = 3.10$, $p = 0.006$, while for TD children there was a marginally significant difference towards the target in the UAR condition ($M = 0.63$, $SD = 0.14$) compared to UAU condition ($M = 0.56$, $SD = 0.13$), $t(22) = 2.05$, $p = 0.052$. TD children looked more toward the target in the UAU condition, $t(38.55) = -2.63$, $p = 0.012$, compared to DLD peers. There was not a significant difference for the UAR condition between groups, $t(39.38) = 1.44$, $p = 0.159$.

These findings suggested that children with DLD were especially slower (vs. TD peers) to recognize the most unexpected (UAU) items in the task. These patterns suggest that preschoolers with DLD show graded activation for UAR items compared to UAU items, contrary to prior results in older adolescents with DLD.

Our results yield novel insights into the dynamics of sentence processing with a range of language learning skills. Specifically, the results highlight that TD preschoolers generate graded predictions even in less-predictable linguistic contexts. Additionally, although children with DLD (in other work) do not generate robust predictive activation for semantically-related sentence outcomes, they more effectively process unexpected outcomes that have a semantic connection to their prior context. Together these findings suggest that lexical activation mechanisms supporting linguistic prediction and recovery in semantically-related and entirely unexpected contexts may not be identical, and suggest avenues for further study.

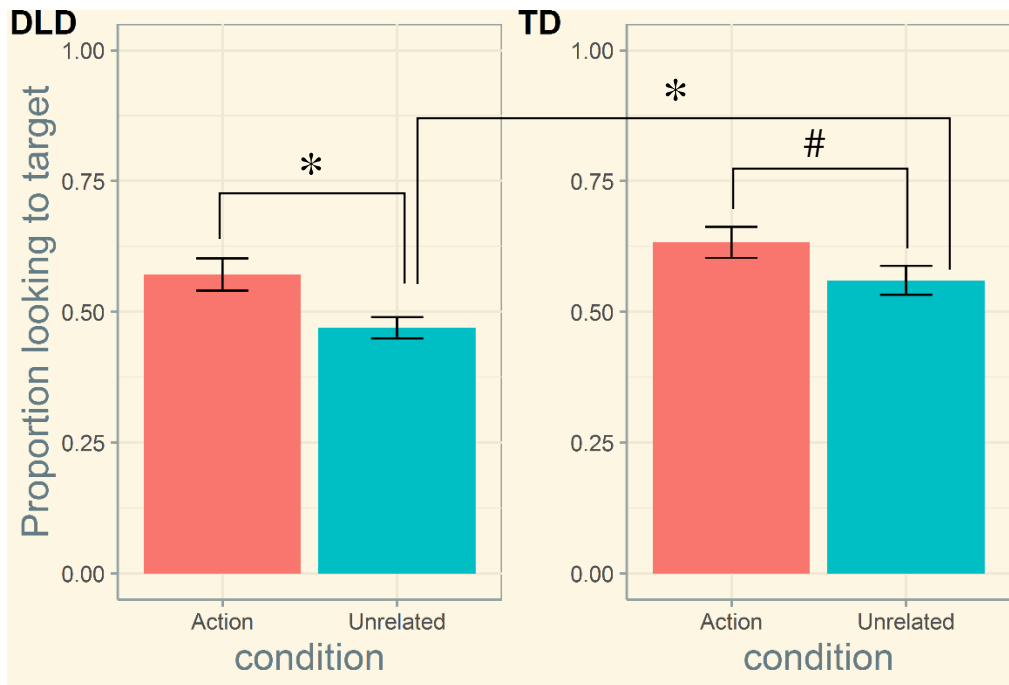


Figure 1: Mean proportion of looking to the Target object shown by group and condition (“Action” = UAR, “Unrelated” = UAU). Mean proportion of looks was calculated by time looking to Target / (Target + Distractor) object, from 300 ms post Target (object) onset to Target offset.

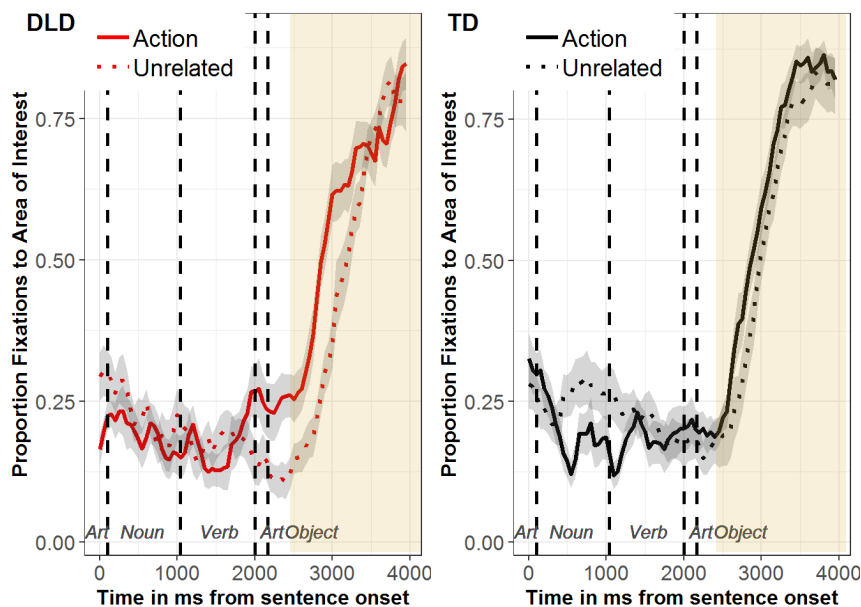


Figure 2: Time-course of fixating on target interest areas in Action-Related (UAR) and Unrelated (UAU) conditions across the entire sentence within each group of participants. Error ribbons represent SEM. Accuracy between groups was measured across a time-period starting 300 ms (shaded) after onset of the sentential object and continued until object offset.

Individual Differences in the Perception of Native and Foreign-Accented Irony

Veranika Puhacheuskaya
Juhani Järvikivi
(University of Alberta)

Research has shown that a foreign accent triggers speaker-specific expectations which alter core language processing mechanisms, such as lexical access, semantic integration, reanalysis, and depth of processing [1,2]. However, psycholinguistic investigations into the processing of foreign accents are relatively recent, and virtually all existing research is on literal language. Our study attempted to fill this gap by examining whether making inferences from non-literal speech is also permeable to such factors as the speaker's accent.

To answer this question, we tasked 96 native speakers of English with listening to short dialogues between native Canadian and foreign-accented (Chinese) speakers and rating them for irony, appropriateness, offensiveness, and one's certainty in the speaker's intent. The two speakers conversed as peers equal in social status. There were 24 experimental dialogues, and each dialogue belonged to one of eight conditions: *native/foreign ironic/literal criticism/praise* (Fig 1). Additionally, we collected the participants' political views, empathy scores, and ambiguity intolerance scores. We hypothesized that, since political views are a robust predictor of anti-immigrant prejudices [3], more conservative listeners may invest less effort in understanding foreign-accented speakers and thus miss the ironic intent of the message more often. High empathy, on the contrary, should facilitate identification of the speaker's ironic intent due to better mentalizing abilities. Finally, high ambiguity intolerance may lead to overconfidence in the achieved interpretation in an attempt to reach the cognitive closure and lessen the load added by ambiguity.

Using generalized additive modelling, we found that foreign-accented irony was indeed considered less ironic than native irony ($p < .001$), and that was true for both criticism and praise (Fig 2). In line with the previous research, ironic praise in general was rated less ironic and less appropriate than criticism, which might be attributed to its surface form violating conversational etiquette. The participants' certainty in the correct interpretation of foreign-accented speech was lower for every condition save literal praise, and the difference between accents was bigger in the ironic conditions. Further, person-based factors significantly affected the ratings and interacted with the type of irony. More conservative participants were worse at detecting irony than their liberal peers but this effect was stronger and more linear for a rarer irony type. In contrast, high empathy facilitated irony detection. We offer several explanations for our findings.

All in all, the results of this study demonstrate that interpersonal variation needs to be accounted for when examining the processing of foreign-accented speech and building the models of speech perception.

References

1. Lev-Ari, S. (2015). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, 5.
2. Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience*, 9.
3. Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society*, 31(3), 419–441.

	Ironic	Literal
Praise	<ul style="list-style-type: none"> - I saw on Facebook that you had a cello concert in Vienna. - Yes, they called me for an encore three times, and the applause was so loud. I stayed there till midnight. - I always knew you were hopeless at playing the cello 	<ul style="list-style-type: none"> - I saw on Facebook that you had a cello concert in Vienna. - Yes, they called me for an encore three times, and the applause was so loud. I stayed there till midnight. - I always knew you were gifted at playing the cello
Criticism	<ul style="list-style-type: none"> - How are your cello lessons going? Are you getting better? - It's torture. Last time the teacher kicked me out. She said she could give me a badge for being her worst student ever. - I always knew you were gifted at playing the cello 	<ul style="list-style-type: none"> - How are your cello lessons going? Are you getting better? - It's torture. Last time the teacher kicked me out. She said she could give me a badge for being her worst student ever. - I always knew you were hopeless at playing the cello

Fig 1. Example materials. Every dialogue in this table was recorded twice, with the native and foreign-accented speakers swapping roles.

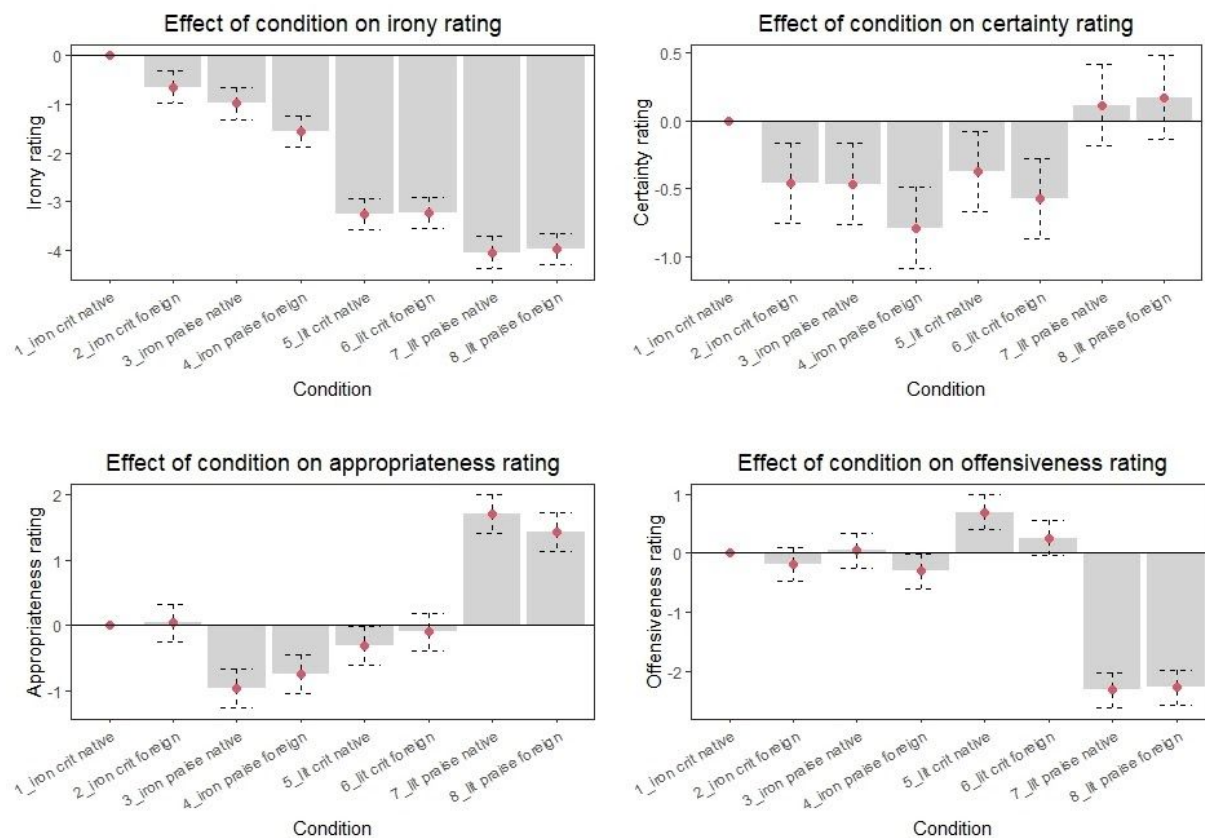


Fig 2. The parametric effect plots for all rating types (irony, certainty in the speaker's intent, appropriateness, and offensiveness).

Impact of structural/functional lesions in the ventral stream on online semantic integration

Noelle Abbott (*), Niloofar Akhavan (*), Michelle Gravier (CSU Easy Bay), & Tracy Love (*)

* SDSU/UCSD Joint PhD Program in Language and Communicative Disorders

Background: Semantic integration (SI), the ability to combine the meaning of words to form more complex representations, is central to rapid, auditory processing of sentences. Prior neuroimaging research has suggested that SI involves widespread left hemisphere activation of cortical regions within the ventral stream (VS) including the anterior middle temporal gyrus (ATL) and the angular gyrus (AG)^{1,2,3,4,5}. However, the necessity of each of these areas within the network to support SI is not clear. One way to investigate their functional role is to identify how damage (structural or functional) to any of these regions impacts the SI process. Post-stroke individuals with chronic aphasia (IWA; a language impairment that typically results from damage to the language dominant hemisphere of the brain) can provide insight into this issue. Structural lesion information is commonly used to map out the association between structural damage and resulting behavior. However, structural damage may not capture underlying alterations to brain function. Following a stroke, cerebral blood flow (CBF) may be hypoperfused (i.e., reduced) in regions of the brain that otherwise appear structurally intact, which can lead to language impairments that would not be predicted by location of structural brain damage alone⁶. One way to capture these functional impairments is through the use of perfusion imaging, which measures CBF of neural tissue⁷.

Current Study: This study investigates the role of regions within the left VS network that have been implicated in SI. We present preliminary evidence for the role of ATL and AG. For this study, we grouped IWA into two groups based on their structural and functional lesion characteristics, those with VS damage (vs-IWA) and those without VS damage (nvs-IWA). We then examined SI using an eye-tracking while listening paradigm (ETL). We predicted that only those with functional or structural damage to ATL or AG within the VS network will exhibit impaired SI abilities.

Participants: 11 neurotypical age-matched controls (AMC) and 11 chronic IWA (>1-year post-stroke) participated in the ETL study. Thus far, 5/11 IWA contributed CBF data in this within-subjects study and are used for analysis to determine compromised brain regions (see Table 1).

Behavioral Task: Using ETL, we tested SI during real-time sentence processing in a group of AMC and our two groups of IWA with CBF data (vs-IWA and nvs-IWA). Here, we operationalized SI as a process by which information from a semantic cue facilitates access of an upcoming noun before it is heard (i.e., anticipation)^{1,8}. In the experimental sentences, semantically biased adjectives (“venomous”) were uniquely associated with the target noun (“snake”), whereas unbiased adjectives (“voracious”) in the control sentences were not (Fig. 1[a]).

Neuroimaging: Using a 3T GE MRI scanner, we investigated both structural and functional brain damage; using structural MRI to determine size and location of lesioned tissue and perfusion MRI to determine the extent of neural integrity in our regions of interest (ATL, AG).

Results: Fig. 1[b] shows the time course of proportion of gazes to the target noun in biased and unbiased conditions for AMC and IWA. Separate multilevel group analyses were conducted to show which participants demonstrated SI, as indexed by rate of lexical access in the biased versus unbiased conditions. Results (Fig. 2) revealed that AMC and IWA were able to access the target lexical item, but IWA demonstrated different anticipatory gaze patterns. The nvs-IWA participants used the semantically biased adjective to anticipate the upcoming noun, whereas the vs-IWA participants did not.

Conclusion: Preliminary results thus far suggest that the ATL and AG play a functional role in SI, by facilitating the use of semantic cues for on time lexical access. When either of these areas become impaired (as measured by structural or functional lesions), semantic cues may no longer be efficiently integrated into the ongoing auditory sentence stream. As will be discussed in the presentation, these effects could be linked to reports of delayed lexical access in aphasia¹¹ and underscore the importance of considering functional and structural brain damage in IWA when mapping the association between brain and behavior.

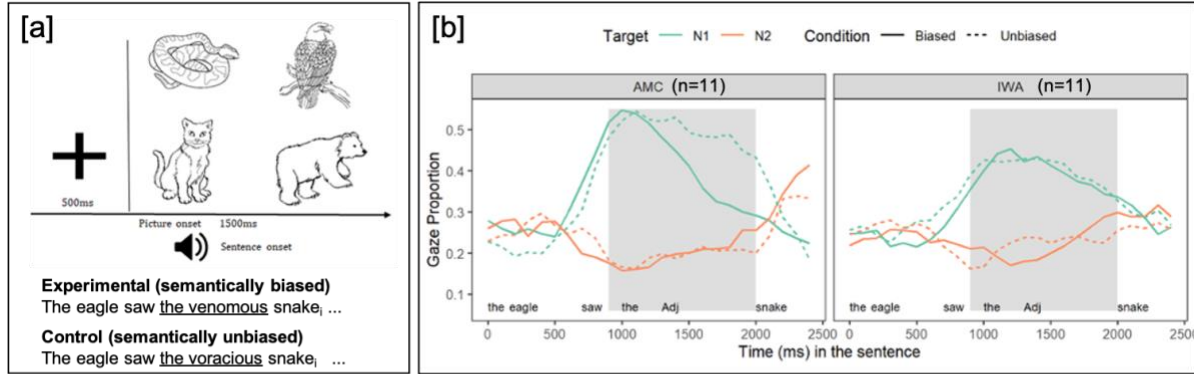


Figure 1. [a] Visual world eye-tracking paradigm. Adjectives (adj) were matched for syllable length and lexical frequency. The time window of interest (underlined) begins at the average onset of the second determiner until the end of the adj (across all items). Follow-up analysis included an extended time window to the end of the second noun (*snake*). [b] Time course depiction of looks to the first and second noun (N1 and N2) as the sentence is heard. Looks to distractor items are excluded from this plot. The shaded region represents the time window of interest for statistical analysis, which captures looks to N2 after hearing the adj.

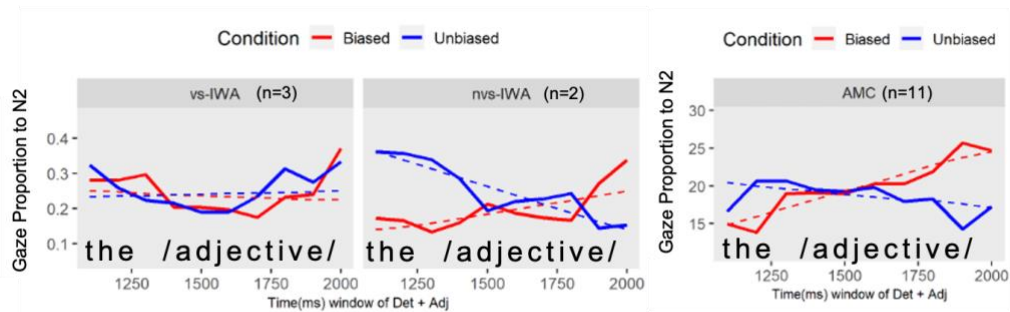


Figure 2. Time course analysis of gaze proportion to N2 in the window of interest [the + Adj] for both IWA groups and AMC across conditions. The dashed line represents the model fit and the solid lines represent the raw data for each condition. Increase in gazes to N2 in the biased versus unbiased condition reflect anticipation of the target noun. The AMC group (rightmost panel) used the biased adjective to access the target noun ($ES = -0.2$, $SE = 0.09$, $t = -2.6$, $p = 0.008$). Similarly, the nvs-IWA participants showed anticipatory access to N2 in the biased versus unbiased condition. This contrasts with the vs-IWA participants ($ES = -0.7$, $SE = 0.3$, $t = -2.7$, $p = 0.007$) indicating that individuals with VS damage do not integrate the semantic properties of the adj to anticipate N2.

Group	IWA	Sex	Years Post-Stroke	Age at Testing (Years)	Education (Years)	Aphasia Subtype	BDAE-3 Severity ^a	WAB-AQ ¹⁰	Lesion Location	% Damage in L ATL	CBF in L ATL	% Damage in L AG	CBF in L AG
vs-IWA	009	M	15	55	17	Mixed non-fluent	2	67.7	Large L lesion, IFG (BA 44/BA45) w/ posterior extension	56%	18.73	5%	20.65
vs-IWA	169	M	4	59	12	Broca	2	28.2	L posterior IFG (BA 44) w/ posterior extension	0%	32.38*	33%	49.95
vs-IWA	190	F	6	76	12	Broca	3	88.2	Left superior temporal lobe	0%	22.64*	9%	68.15
nvs-IWA	101	M	9	67	20	Broca	3	82.6	Large L lesion posterior IFG (BA 44) w/ posterior extension	0%	43.55	2%	79.13
nvs-IWA	151	F	7	65	16	Anomic	4	95.8	L MCA infarct with subcortical extension	0%	39.30	0%	56.83
AMC	Ages 57-66 years (mean = "61.9); 7 females, 4 males; Education 14-18 years (mean = "15.7)*						--	--	--	--	--	--	--

Table 1. Participant Demographics. IWA = individual with aphasia; vs-IWA = ventral stream damage; nvs-IWA = no ventral stream damage; AMC = Age-matched neurotypical controls; BDAE-3 = Boston Diagnostic Aphasia Examination v 3 (1=Severe, 5=Mild); WAB-AQ = Western Aphasia Battery - Aphasia Quotient (<50=severe, 51-70=moderate, >71=mild); L ATL = left anterior temporal lobe; CBF = cerebral blood flow (mL/100g/sec); L AG = left angular gyrus; *Hypoperfusion in the left hemisphere was based on CBF values that were at least 2 standard deviations below each participants right hemisphere mean. Bolded values represent functional lesions (hypoperfused regions). *Missing education data for four AMC

References

- Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9), 416-423.
- Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1), 42.
- Boylan, C., Trueswell, J. C., & Thompson-Schill, S. L. (2015). Compositionality and the angular gyrus: A multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia*, 78, 130-141.
- Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex*, 96, 105-120.
- Graessner, A., Zaccarella, E., & Hartwigsen, G. (2020). Differential contributions of left-hemispheric language regions to basic semantic composition. *bioRxiv*.
- Love, T., Swinney, D., Wong, E., & Buxton, R. (2002). Perfusion imaging and stroke: A more sensitive measure of the brain bases of cognitive deficits. *Aphasiology*, 16(9), 873-883.
- Sorensen, A. G., & Reimer, P. (2000). Cerebral MR perfusion imaging: principles and current applications.
- Nozari, N., Mirman, D., & Thompson-Schill, S. L. (2016). The ventrolateral prefrontal cortex facilitates processing of sentential context to locate referents. *Brain and language*, 157, 1-13.
- Goodglass, H., Kaplan, E., & Barresi, B. (2001). *BDAE-3: Boston Diagnostic Aphasia Examination-Third Edition*. Philadelphia, PA: Lippincott Williams & Wilkins.
- Kertesz, A. (2007). *Western aphasia battery: Revised*. Pearson.
- Love, T., Swinney, D., Walenski, M., & Zurif, E. (2008). How left inferior frontal cortex participates in syntactic processing: Evidence from aphasia. *Brain and Language*, 107(3), 203-219.

Is reanalysis selective when regressions are manually controlled?

Dario Paape & Shravan Vasishth (University of Potsdam)

To what extent is rereading in syntactically complex sentences under conscious control? It has been argued by, e.g., [1] that in cases of syntactic misanalysis, the problematic part of the sentence is selectively reread (“selective reanalysis”). However, it is not clear to what extent readers consciously decide to reread earlier material. It is possible that selective reanalysis is, to some extent, consciously triggered when the reader notices that misanalysis has occurred. Conscious awareness of garden-pathing may occur in extremely difficult examples (*The horse raced past the barn fell*) but may also stochastically arise in milder cases.

Due to the SARS-CoV-2 pandemic, it is currently difficult to conduct lab-based eye-tracking studies. We thus switched to a web-based, modified version of self-paced reading that allows rereading (“bidirectional SPR”): readers press the right arrow key to move forward through the sentence and the left arrow key to move backward (“manual regression”). Readers can also return to the beginning of the sentence or move directly to the comprehension question with specific keys. Sentence presentation is masked and non-cumulative. While this method has obvious drawbacks, such as the inability to skip words, and is quite dissimilar to eye tracking, it also has advantages: oculomotor noise is eliminated as a source of fixation errors and regressions require conscious deliberation.

We designed an experiment in German with two conditions (early versus late disambiguation) and two types on ambiguity: the NP/S coordination ambiguity [2] and the German SVO/OVS ambiguity [3] (see examples on page 2). We recruited 100 participants through Prolific, each of whom read 32 critical sentences plus 52 fillers. In addition to a baseline monetary compensation, participants earned bonus compensation by reading both quickly and accurately. Points were awarded based on the time taken to complete the study and the percentage of correct answers to comprehension questions. Detailed comprehension questions were asked after each trial.

Overall, rereading was quite common in our study: at least one manual regression occurred in about 51% of trials for critical sentences. We analyzed first-pass reading times and rereading times by region, analogously to eye-tracking studies. First-pass reading times showed a divergence between coordination and SVO/OVS sentences in the early disambiguation region (see Figure 1, left), in that coordination sentences showed an ambiguity slowdown (95% CrI: [6 ms, 65 ms]) while SVO/OVS sentences showed a disambiguation slowdown (CrI: [7 ms, 69 ms]). Both sentence types showed a garden-path effect in the late disambiguation region, both in first-pass reading times (CrI: [−2 ms, 35 ms]) and in rereading times (CrI: [12 ms, 127 ms]).

We used multidimensional scaling and model-based clustering to identify clusters of scanpaths across both sentence types [4]. Scanpaths clustered along two dimensions: amount of rereading and location of regressions (see Figure 1, right). Across conditions, participants who scored high on the speed/accuracy measure showed more rereading. Garden-pathing also led to higher values in this dimension. Correspondingly, garden-pathing decreased the probability of scanpath membership in cluster 2 (CrI: [−9%, 0%]), in which there are almost no regressions (see Figure 2), and increased the probability of membership in cluster 3 (CrI: [0%, 6%]), in which large portions of the sentence are reread. Membership in other, smaller clusters, such as cluster 5 with short regressions to the disambiguating region, showed no indication of a difference between conditions.

Our results suggest that manually-controlled rereading as a response to syntactic misanalysis is overall relatively unselective, though there was some increased focus on the disambiguating region. To the extent that these findings are transferable to eye tracking, it may be that conscious disruption of the reading flow results in a global rereading strategy, while more subtle disruptions result in more locally selective strategies.

Coordination, early disamb.

... und der Schauspieler (NOM)
... and the actor

Coordination, late disamb. (garden path)

... und die Schauspielerin (NOM/ACC)

SVO/OVS, early disamb.

... den Forscher (ACC)
... the researcher.SG

SVO/OVS, late disamb. (garden path)

... die Forscherin (NOM/ACC)

The make-up artists powdered the singers ...

mit blauen Augen wurde parfümiert, ...
with blue eyes was perfumed ...

The make-up artists powdered the singers ...

mit blauen Augen wurde parfümiert ...

im Dschungel stachen die Moskitos, ...
in.the jungle bit.PL the mosquitoes ...

im Dschungel stachen die Moskitos, ...

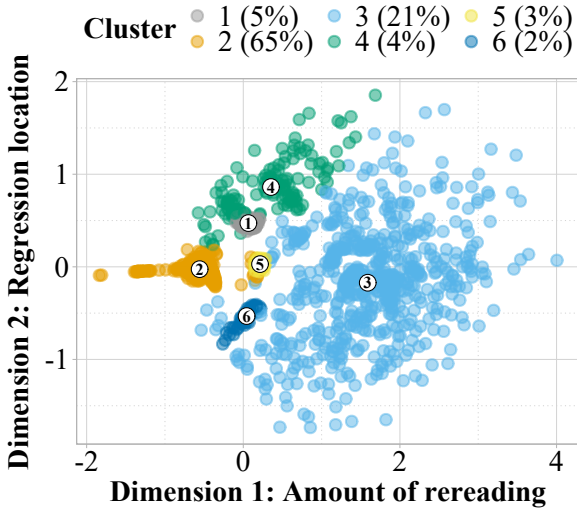
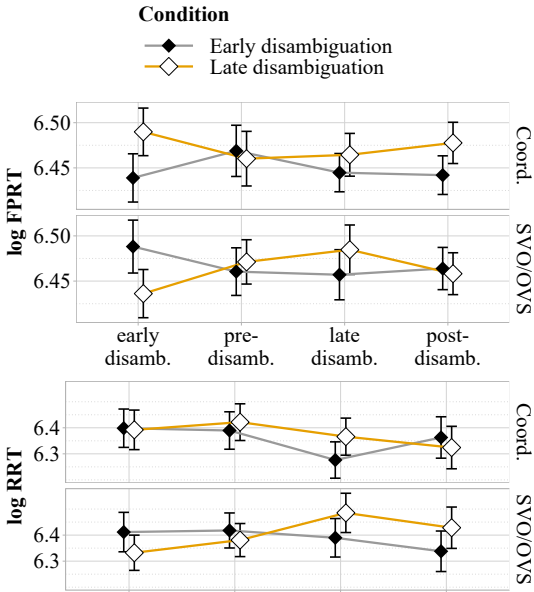


Figure 1. Left: First-pass reading times and rereading times by region of interest, residualized against region length. Error bars show 95% confidence intervals. Right: Location of clusters in scanpath space. “Regression location” refers to the region of interest in the sentence where regressions occur.

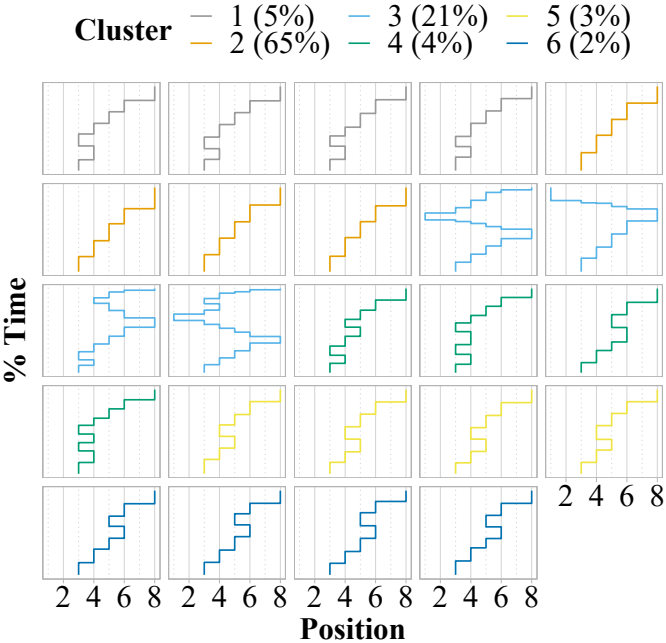


Figure 2. Typical scanpaths belonging to each of the clusters. Scanpaths were analyzed starting at the first visit to the early disambiguation region (3). Only data from regions starting with the early disambiguation region (3) and ending with the post-disambiguation region (6), as well as visits to the start (1) and end (8) regions, were included in the analysis.

References. [1] Frazier & Rayner (1982), Cognitive Psychol. [2] Frazier (1987), Nat Lang Linguist Th. [3] Hemforth (1993), Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens. [4] von der Malsburg & Vasisht (2011), J Mem Lang.

Prediction of successful reanalysis based on eye-blink rate and reading times

Lola Karsenti & Aya Meltzer-Asscher (Tel Aviv University)

lolakarsenti@mail.tau.ac.il

Introduction. Characterizing individual differences in sentence processing in general, and in recovery from processing difficulty in particular, remains a challenge [1-3]. Previous research has shown that intelligence, language experience and working memory capabilities predict overall comprehension accuracy in Garden Path (GP) sentences [2-3, see also [4] for a study of different sentence structures]; however, online syntactic effects (i.e., reading times) are unreliable as measures of individual differences [3].

In the present work, we aimed to further investigate the characteristics of participants who are (un)able to perform reanalysis of GP sentences, using a paraphrasing task [5]. Specifically, we asked whether participants' capacity to successfully reanalyze the sentence can be predicted based on (i) their reading time (RT) patterns in the critical region of the sentence, and (ii) their tonic dopamine levels, as reflected by their eye-blink rate. The performance of reanalysis in GP sentences requires efficient use of executive functions and allocation of working memory in order to remember the sentence and return to the point of difficulty, inhibit the syntactic structure built during the initial analysis, and flexibly use the new material to find alternative analyses. Working memory updating, inhibition and cognitive flexibility are all assumed to be driven by dopamine (DA) [6-8]. Tonic dopaminergic activity is correlated with resting state (tonic) eye blink rate (EBR) [6]. EBR was shown to be a predictor of individual differences in paradigms requiring inhibition or task switching, where DA followed an inverted u-shaped association with performance, such that medium DA levels corresponded to optimal performance [6-7].

We used Hebrew GP sentences with optionally transitive (OT) verbs varying in their transitivity bias, embedded in a temporal adjunct. The baseline condition included an intransitive (IN) verb (these conditions were part of a larger study with different types of sentences). (See Table 1).

Methods. Ninety-six native Hebrew speakers participated in a self-paced reading experiment with 28 sentence sets and 72 filler sentences. Each target sentence as well as some fillers were followed by an instruction to write the last sentence, without further directions. Prior to the experiment, resting-state EBR was registered with three ocular electrodes above and below the eye while participants fixated at a cross in the middle of the computer screen for 3.5 minutes. Paraphrases were coded as successfully reanalyzed (R) if the participants introduced a comma or switched the order of clauses, and as non-reanalyzed (N) in cases of a lingering misinterpretation (see Table 2). For each participant, their reanalysis performance (RP) rate was defined as their percentage of successfully reanalyzed sentences out of all GP sentences.

Results. We divided the participants to groups based on their RP for descriptive purposes. EBR results by subject group are presented in Table 3. RT results are presented in Figure 1. We fitted a GLM model for all subjects with paraphrases coded as N or R (N=71), with participants' RP rate as a dependent variable. Average log RT of the critical region across the two conditions, linear and quadratic effects of EBR, and their interaction were entered as fixed effects. We observed a main effect for RT, with longer RTs predicting better reanalysis outcome ($p < .001$). We also found a linear ($p=.019$) and polynomial inverted u-shaped ($p=.033$) effects of EBR on reanalysis performance. The interaction between EBR and average RT was not significant (see Figure 2).

Discussion. Our results show that participants with medium tonic EBR were the most successful reanalyzers, in line with previous results showing that medium dopamine levels predict high performance in tasks requiring inhibition and updating. The results also show that slow reading is associated with success at reanalysis ([9]).

Table 1: Example set, translated from Hebrew

Optionally Transitive (OT)	'After the guests drank cold water <u>flowed from the tap at the farm.</u> '
Baseline Intransitive (IN)	'After the guests woke up cold water <u>flowed from the tap at the farm.</u> '

Note: the critical region is marked in underline

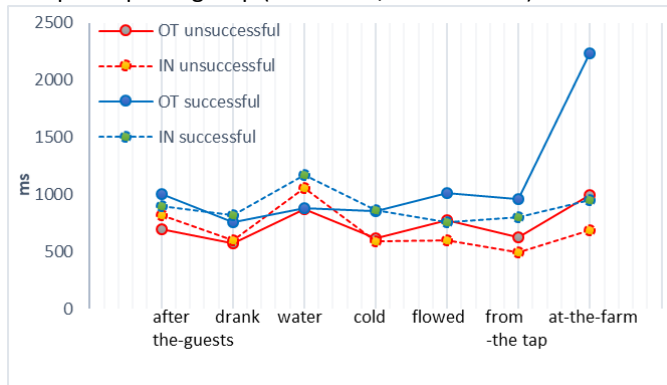
Table 2. Paraphrase coding

Category	Example paraphrase
Successful reanalysis ("R"). 98 sentences (25.5%)	"After the guests drank, cold water flowed from the tap." "Cold water flowed from the tap after the guests drank."
Lingering misinterpretation ("N"). 91 sentence (23.6%)	"After the guests drank cold water it flowed from the tap." "After the guests drank cold water, cold water flowed from the tap."
195 sentences not coded as R or N, due to obscurity	"After the guests drank cold water flowed from the tap." "Cold water flowed from the tap."

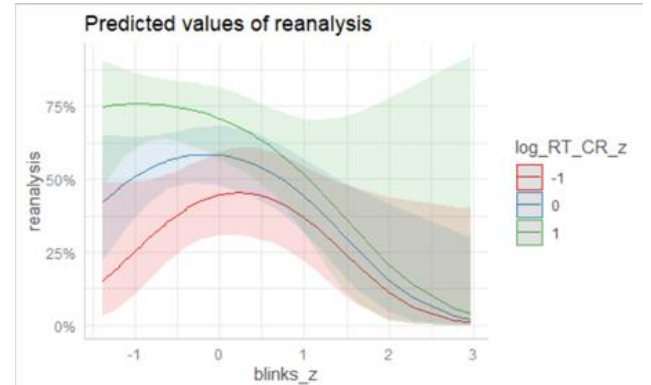
Table 3. Mean EBR by participant group

Group of participants	Average EBR (SD) (3 min)
75-100% of a participant's paraphrases exhibited N pattern, ' unsuccessful ' group (N = 17)	83 (50)
Participant's paraphrases exhibited only N pattern, but less than 75% of the time (N = 12)	70 (35)
Participant's paraphrases exhibited both R and N pattern, but less than 75% of each (N = 15)	63 (29)
Participant's paraphrases exhibited only R pattern, but less than 75% of the time (N = 11)	48 (19)
75-100% of a participant's paraphrases exhibited R pattern, ' successful ' group (N = 16)	54 (30)

Note: not all participants' paraphrases were coded as either R or N, as in Table 2 above.

Figure 1. Mean RTs of critical region by condition and participants group (successful/unsuccessful)

Note: In Hebrew, both SV and VS orders are possible, especially with unaccusative verbs. This has consequences for the GP effect, the greater processing difficulty appears at the end of the sentence.

Figure 2. Predicted values of reanalysis performance by EBR and RT

Note: log_RT_CR_z represents standardized log RT of the critical region and blinks_z represents standardized EBR

References: [1] Just & Carpenter (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149. [2] Engelhardt, P. E., Nigg, J. T., & Ferreira, F. (2017). Executive Function and Intelligence in the Resolution of Temporary Syntactic Ambiguity: An Individual Differences Investigation. *Journal of Experimental Psychology*, 70(7): 1263–1281. [3] James, A.N., Fraundorf, S. H., Lee, E-K, & Watson, D.G. (2018). *J MEM Lang.*; 102: 155-181. [4] Blott, L. M., Rodd, J. M., Ferreira, F., & Warren, J. (2020). Recovery from misinterpretations during online sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. [5] Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 280-285. [6] Jongkees, B. J., & Colzato, L. S. (2016). Spontaneous eye blink rate as predictor of dopamine-related cognitive function—A review. *Neuroscience & Biobehavioral Reviews*, 71, 58-82. [7] Akbari Chermahini, S., & Hommel, B. (2012). More creative through positive mood? Not everyone! *Frontiers in Human Neuroscience*, 6, 319. [8] Paprocki, R., & Lenskiy, A. (2017). What does eye-blink rate variability dynamics tell us about cognitive performance? *Frontiers in human neuroscience*, 11, 620. [9] Nicenboim, Logačev, Gattei and Vasishth (2016). When High-Capacity Readers Slow Down and Low-Capacity Readers Speed Up: Working Memory and Locality Effects. *Frontiers in Psychology*, 7.

Reanalysis difficulty modulates cumulative structural priming effects in sentence comprehension
Ming Xiang and Weijie Xu (The University of Chicago)

Cumulative structural priming effect refers to the observations that repeated exposure to a syntactically infrequent structure could facilitate the subsequent processing of similar structures. For example, it was found in [1] that participants who were repeatedly exposed to reduced relative clause sentences (**RR**) such as “*The experienced soldiers warned about the dangers conducted the midnight raid*” showed a reduced garden-path ambiguity effect after the exposure phase, compared to participants in a control group; conversely the same participants also showed increased difficulty for the originally preferred matrix verb parse (**MV**) “*The experienced soldiers warned about the dangers before the midnight raid*”. A number of recent replication studies [2, 3], however, failed to find any robust effects, suggesting that the original effect, if exists at all, has a very small effect size. One possible reason for the small effect size is that the garden-path ambiguity in the above RR prime sentence spans over a relatively long region, i.e. “warned about the dangers”, creating a “digging in” effect [4,5] that reduces the likelihood of successful reanalysis towards the RR parse in the first place. In three experiments, the current study investigates whether increasing the likelihood of the RR parse on the prime sentences facilitates the priming effect.

Procedure Three self-paced-reading experiments (n=240 total) were conducted on Ibex Farm, with participants answering a comprehension question after reading each sentence. Each experiment consists of three blocks. In **Block 1**, in a between-participant design, half of the participants (**Exposure Group 1A**, n=40) read 16 target sentences (8 ambiguous RR and 8 unambiguous RCs), and the other half (**Control Group 1B**, n=40) read 16 filler control sentences (examples in Table 1). Both groups of participants were then tested on the RR ambiguity in **Block 2** and MV ambiguity in **Block 3**. In Block 2, participants read 16 target sentences (8 RR ambiguity and 8 unambiguous RCs) and 16 fillers. In Block 3, participants read 16 target sentences (8 MV ambiguity and 8 unambiguous sentences) and 16 fillers. The three experiments only differ in their exposure Block 1A. The RR sentences in Experiment 1 Block 1A always contained a salient disambiguating by-phrase right after the ambiguous verb, making reanalysis relatively easy. The RR sentences in **Experiment 2** Block 1A further made the subject noun phrases inanimate, providing more cues for the correct RR parse. In these two experiments, the verbs used in Block 1 were repeated in Block 2&3. In **Experiment 3**, morphologically unambiguous verbs (e.g. *taken*) were used to signal the RR parse in Block 1A. These three experiments therefore gradually increased the likelihood of the RR parse in the exposure block. The critical sentences used in the exposure blocks were adapted from [6,7].

Analysis and results For each experiment, we performed linear mixed effects models on the log-transformed RTs on the disambiguating region and the next spill-over region, using the Bayesian statistical analysis R package brms [8] (Figure 1). The analyses reported here focus on the critical Group (exposure vs. control) x Ambiguity interaction in order to answer two questions.: (i) whether the RR ambiguity is *reduced* in Block 2 for participants from the exposure group compared to those from the control group; (ii) conversely whether there is a *larger* MV ambiguity in Block 3 for participants from the exposure group. For the RR ambiguity in Block 2, we found some evidence for a critical Group x Ambiguity interaction on the spill-over region in Experiment 1 (*got*, Figure 1, Estimate 0.016, SE 0.009, 95% CrI [-0.001, 0.048]); and the interaction effect is present on the critical disambiguating region in Experiment 2 (*by the doctor*, Figure 1, Estimate 0.024, SE 0.011, 95% CrI [0.001, 0.046]). No interaction was found in Experiment 3. For the MV ambiguity in Block 3, no interaction was found for any experiment.

Conclusion By making the prime RR sentences easier to reanalyze/parse than previous studies, we observed reduced RR ambiguity effect after repeated exposure to RR primes (Experiment 1&2). Relative to Experiment 1, the effect came online earlier in Experiment 2 when the exposure RR sentences contain an additional facilitating animacy cue on the subject, despite the fact that the same animacy cue is absent on the post-exposure target sentences, suggesting that the priming effect is not simply based on surface statistical contingencies. However, verb overlap between the exposure and the post-exposure targets is required for structural priming to take place, even when the prime sentences are unambiguously RR (Experiment 3), confirming that structural priming in comprehension is mediated through the lexical representation of the verb [9]. Finally, we found no evidence that repeated exposure to RR ambiguity increases the processing difficulty of the originally preferred MV parse, replicating the findings in [2,3].

Table 1: Example stimuli. Disambiguating regions in bold; slashes indicate SPR regions.

Block 1: Exposure Group 1A (item n=16) (unambiguous version in the parenthesis), Latin square distribution of the items

Experiment 1: The defendant\ (that was) examined\ **by the lawyer**\ turned out to be\ unreliable.

Experiment 2: The evidence\ (that was) examined\ **by the lawyer**\ turned out to be\ unreliable.

Experiment 3: The money\ (that was) taken\ **by the student**\ was\ finally\ returned.

Block 1: Control Group 1B (item n=16)

The apples on that tree are surprisingly delicious.

Block 2 (16 targets + 16 fillers), Latin square distribution of the target items

RR Ambiguous: The patient \examined\ **by the doctor**\ got\ a stomach ache\ last night.

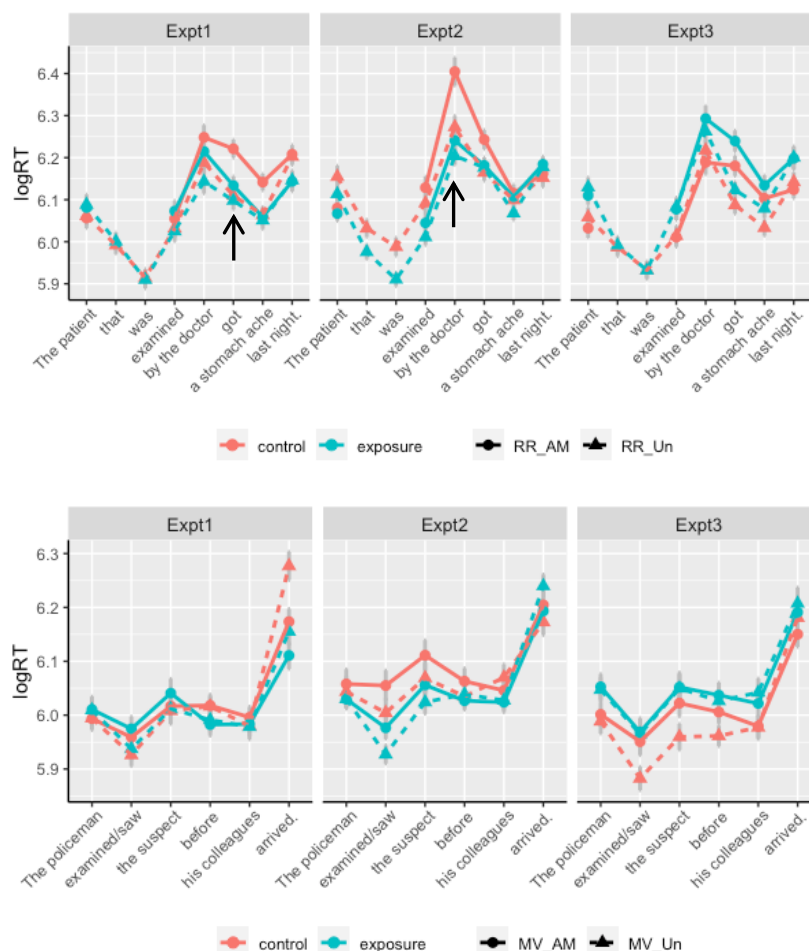
RR Unambiguous: The patient \that was examined\ **by the doctor**\ got\ a stomach ache\ last night.

Block 3 (16 targets + 16 fillers), Latin square distribution of the target items

MV Ambiguous: The policeman\ examined\ **the suspect**\ before\ his colleagues\ arrived.

MV Unambiguous: The policeman\ saw\ **the suspect**\ before\ his colleagues\ arrived.

Figure 1. Top: Block 2 RR ambiguity; **Bottom:** Block 3 MV ambiguity



Linear mixed effects model for Block 2&3 separately, under each experiment, on the disambiguating region and the spill-over region:
 $\log RT \sim \text{ExposureGroup} * \text{Ambiguity} + \text{RT. previous.region} + (1 + \text{Ambiguity} | \text{participant}) + (1 + \text{ExposureGroup} * \text{Ambiguity} | \text{Item})$

References:

- [1] Fine, Jager, Farmer & Qian. 2013. *PLoS ONE*
- [2] Stack, James & Watson, 2018. *Memory and Cognition*
- [3] Dempsey, Liu & Christianson. 2020. *Journal of Experimental Psychology*.
- [4] Ferreira & Henderson. 1991. *Journal of Memory and Language*
- [5] Tabor & Hutchins, 2004. *Journal of Experimental Psychology*
- [6] Trueswell, Tanenhaus & Garnsey. 1994. *Journal of Memory and Language*
- [7] Ferreira & Clifton. 1986. *Journal of Memory and Language*.
- [8] Bürkner, P. 2017. *Journal of Statistical Software*.
- [9] Pickering & Ferreira. 2008. *Psychological Bulletin*.

Interaction between local coherence and garden path effects supports a nonlinear dynamical model of sentence processing.

Roeland Hancock & Whitney Tabor, University of Connecticut

Predictive sentence processing models that incorporate both lexical and syntactic expectations [2,4] have treated these as additive sources of information, yet experimental data have provided support for interacting syntactic and lexical expectations [1]. Three classes of incremental parsing models—additive surprisal, noisy-channel, and self-organizing—make distinctive predictions in the case of local coherence following syntactic ambiguity. We present simulation results from a self-organizing nonlinear dynamical model (Self-Organized Sentence Processing ; SOSP [5–7]) which predicts speeded reading time when local coherence coincides with garden path disambiguation, and we present experimental data in support of this prediction.

We examined incremental lexical and syntactic effects by manipulating local coherence in a 2×2 design that fully crosses local coherence (LC) and garden pathing (GP) (see items 1a-1d).

1a (+GP +LC) The division encamped near the fierce *battle was fought* by the brigade.

1b (+GP -LC) The division encamped near the fierce *battle was pestered* by the brigade.

1c (-GP +LC) The division that was encamped near the fierce *battle was fought* by the brigade.

1d (-GP -LC) The division that was encamped near the fierce *battle was pestered* by the brigade.

In (1a), the critical disambiguating region *was fought* occurs within a locally lexically coherent fragment *battle was fought*, with low trigram surprisal. In (1b), the corresponding context has high trigram surprisal. Comparison with the unambiguous controls (1c, 1d) isolate locally coherent lexical and syntactic effects, and highlight conflicting predictions from three classes of models.

Additive surprisal models predict that both garden path and locally coherent structures can influence parsing, but that they do so independently. Noisy-channel models predict garden path effects and they sometimes predict local coherence effects if there is a locally coherent structure with high probability and a low-cost edit that can license it [3]. Here, however, the most plausible low-cost edit (“and” after “battle”) does not produce an interaction with garden pathing in our materials. In SOSP (Table 1), processor state \mathbf{x} is governed by a potential function called “Harmony” ($H(\mathbf{x})$). Each harmony peak corresponds to a stable configuration of bonds (fully grammatical structure, suboptimal structure that is a coherent tree, or an ill-formed mix of partially completed trees). SOSP uniquely predicts that the local coherence effect and the garden path effect will interact because strong bottom-up formation of fully grammatical structures in the unreduced examples (1c and 1d) dwarfs the potential of local coherence, but weaker bottom-up induction of an ill-formed mix of trees due to the garden path in 1a and 1b causes the difference between the locally coherent and nonlocally coherent structures to manifest as an effect. Such interactions are a hallmark property of nonlinear dynamical systems, where parameter changes can cause categorical change (bifurcation).

58 participants read 40 experimental items and 40 fillers in a web-based centered-window self-paced reading task. Log-transformed reading times were residualized for word length, position, and frequency. Linear mixed effects analysis at the critical region (Figure 1b) showed significantly longer reading times for garden path sentences overall ($t = 2.4, p = .02$) and an interaction with local coherence effects ($t = 2.1, p = .04, BF_{H1} = 4.9$), with shorter reading times for locally coherent (1a) than locally incoherent (1b) garden path sentences. These results confirm the distinctive prediction of the SOSP model, supporting the view that local coherence effects arise from bottom-up dynamics in the parser.

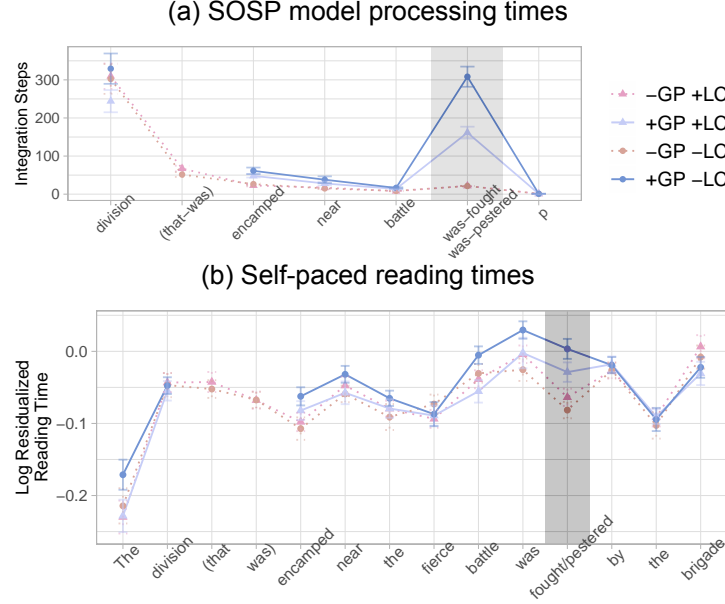


Figure 1: **(a)** SOSP predicts a garden path \times coherence interaction at the disambiguating verb (shaded). Processing time is measured in Euler integration steps to convergence at each word. **(b)** Self-paced reading data confirm this prediction.

A structure's harmony = product of bond harmonies (each reflecting degree of clash)

$$h_i = \prod_{l \in \text{bonds}} \left(1 - \frac{\text{dist}(\mathbf{f}_{l, \text{daughter}}, \mathbf{f}_{l, \text{mother}})}{n_{\text{feat}}} \right)$$

Radial basis functions define peaks at each structural locus, \mathbf{c}_i . γ is peak width.

$$\phi_i(\mathbf{x}) = \exp \left(-\frac{(\mathbf{x} - \mathbf{c}_i)^T (\mathbf{x} - \mathbf{c}_i)}{\gamma} \right)$$

In addition to peaks for individual structures, there are peaks at the averages of future possible structures for initial substrings.

$\mathbf{c}_j = \frac{1}{\sum_i w_i} \sum_i w_i \mathbf{c}_i$ for \mathbf{c}_i a destination from partial parse j , $w_i = h_i p_i$, p_i = probability in PCFG for grammatical parsing

The global harmony at any point is the height on the flank of the locally dominant peak.

$$H(\mathbf{x}) = \max_{i \in 1 \dots n} h_i \phi_i(\mathbf{x})$$

Competitive bond formation and feature specification is noisy hill-climbing on the landscape.

$$\frac{d\mathbf{x}}{dt} = \nabla_{\mathbf{x}} H(\mathbf{x}) + \eta$$

Table 1: SOSP dynamics.

References: [1] K Christianson et al., Cogn Psychol 42, 368–407 (2001). [2] MW Crocker et al., J Psycholinguist Res 29, 647–669 (2000). [3] R Levy, EMNLP, 234–243 (2008). [4] J Mitchell et al., ACL, 196–206 (2010). [5] G Smith et al., Cogn Psychol 124 (2021). [6] G Smith et al., ICCM, 138–143 (2018) [7] W Tabor et al., AMLaP (2020).

Coordination ambiguity resolution in native and non-native language comprehension

Yesi Cheng, Hiroki Fujita, Ian Cummings (University of Reading)

Garden-path sentences have been examined to investigate the similarities and differences between native (L1) and non-native (L2) comprehension [5,6,7]. Both L1ers and L2ers exhibit garden-path effects during reading, and have difficulty revising initially assigned misinterpretations during comprehension [1,7]. L1/L2 differences have also been reported. [7] reported a larger proportion of misinterpretations of garden-path sentences in L2 than L1 speakers, and [6] reported smaller garden-path effects, and lower comprehension accuracy, in L2ers, which they took to indicate that L2ers may be less likely to initiate reanalysis than L1ers [6,7]. Additionally, reanalysis is modulated by ambiguity length in both L1 and L2ers, with an increased reanalysis difficulty for a longer ambiguous region [1,4,6]. While ambiguity length effects have been examined in subject-object ambiguities, they have not been widely studied in other ambiguities. To further investigate these issues in L1 and L2 processing, we examined the co-ordination ambiguity [1,3] in an eye-tracking while reading experiment.

48 L1ers and 48 proficient L2ers (mean proficiency score = 49/60; range 40-58) read 24 sentences like (1) while their eye-movements were monitored. In (1a/c), the coordinator “and” causes temporary ambiguity, as “the cat” may be initially interpreted as the conjoined direct object of “washed”, when it is in fact the subject of “played”. (1b/d) are unambiguous controls, as the subordinating conjunction “while” renders the direct object analysis impossible. Additionally, in (1a), the temporary ambiguity is disambiguated immediately, whereas in (1c), the ambiguity is longer due to inclusion of a prepositional phrase (“in the garden”) before the disambiguated verb. We expected longer reading times at “played” in (1a/c) than in (1b/d) due to garden-path effects. If maintaining an initial interpretation for longer leads to increased reanalysis difficulty [1,6], we would expect longer reading times in (1c) than (1a). If the initial misinterpretation lingers after reanalysis [1], comprehension accuracy rates should be lower for (1a/c) than for (1b/d), and if length influences reanalysis, (1a) should have lower accuracy than (1c). If L2ers are less likely than L1ers to conduct reanalysis [6], they should show smaller garden-path effects during reading than L1ers, especially in the long conditions, and show lower comprehension accuracy rates than L1ers in ambiguous conditions only.

We pre-registered analyses (<https://osf.io/ausmx>) of first-pass, regression path and total viewing times at the disambiguating (“played”) and spillover regions (“with a ball”). There were significant effects of ambiguity in all measures (all $p < .02$). Ambiguity interacted with group only in regression path times ($p = .02$), with a larger garden-path effect in the L1 group (L1 effect = 95ms, L2 effect = 61ms). Ambiguity also interacted with length and region in regression path times ($p < .001$), with longer reading times at the spillover region in long (1c) rather than short (1a) ambiguous conditions. Comprehension accuracy rates showed a significant main effect of ambiguity ($p < .001$), with lower accuracy rates for (1a/c) than (1b/d). This main effect was modulated by length ($p = 0.044$), with lower accuracy in (1c) than (1a), and by group ($p = 0.003$). Although the L2 group showed a larger difference between ambiguous and unambiguous conditions than the L1 group, this was due to L1ers having lower accuracy in unambiguous conditions, while the groups did not differ in ambiguous conditions.

Our results conceptually replicate previously reported length effects on garden-path recovery and misinterpretation observed in the subject-object ambiguity [1,4] and extend them to the co-ordination ambiguity, in both L1 and L2 readers. Although L2ers showed smaller garden-path effects in one measure, potentially compatible with [6], we did not find evidence of increased misinterpretation in L2ers, contra [2,6,7], which would be expected if L2ers do not initiate reanalysis as successfully as L1ers. As L1ers and L2ers were affected by garden-path effects and length effects during processing and in offline comprehension, we suggest that reanalysis processes are influenced by the same factors in L1 and L2 processing.

(1a) Ambiguous, Short

Yesterday afternoon, Ken washed the dog and the cat played with a ball.

(1b) Unambiguous, Short

Yesterday afternoon, Ken washed the dog while the cat played with a ball.

(1c) Ambiguous, Long

Yesterday afternoon, Ken washed the dog and the cat in the garden played with a ball.

(1d) Unambiguous, Long

Yesterday afternoon, Ken washed the dog while the cat in the garden played with a ball.

Question: Was Ken washing the cat?

Figure 1. *Reading times.*

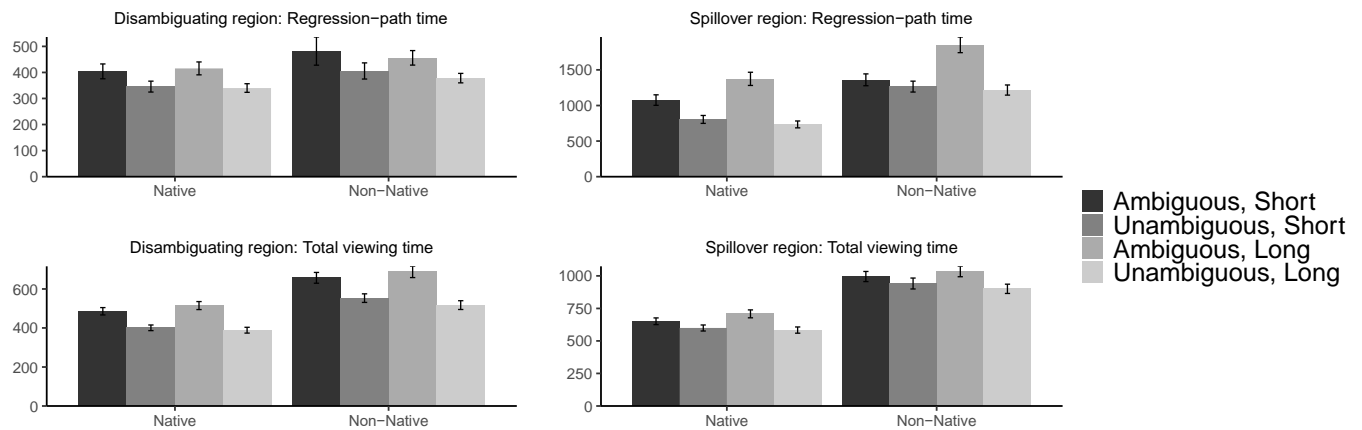
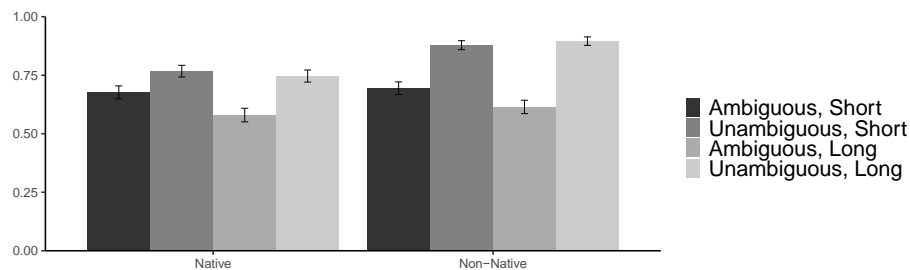


Figure 2. *Comprehension accuracy rates.*



References

[1] Christianson et al. (2001), *CP*; [2] Cunnings (2017), *BLC*; [3] Engelhardt & Ferreira (2010), *LS*; [4] Ferreira & Henderson (1991), *JML*; [5] Hopp (2006), *SLR*; [6] Jacob & Felser (2016), *TQJEP*; [7] Pozzan & Trueswell (2016), *BLC*.

Back to the Future: Do Influential Results from 1980s Psycholinguistics Replicate?

Fernanda Ferreira (fferreira@ucdavis.edu), Gwendolyn Rehrig, Madison Barker, Eleonora Beier, Suphasiree Chantavaran, Beverly Cotter, Zhuang Qiu (UC Davis), Matthew Lowder (U of Richmond), Hossein Karimi (Mississippi State University)

Background. In the 1980s, a prominent research question concerned the effects of discourse context on parsing decisions. Two highly influential and widely cited studies reported contradictory results: Ferreira and Clifton (1986; F&C86) conducted two experiments, one using eyetracking and the other using self-paced reading, in which minimal attachment (MA; syntactically easy) or nonminimal attachment (NMA; syntactically difficult) sentences were presented either in biased or neutral contexts, and they reported that helpful context affected later processing stages but not the parser's initial attachment decisions. In contrast, Altmann and Steedman (1988; A&S88) conducted a self-paced reading study in which MA and NMA sentences were embedded in appropriately or inappropriately biasing contexts, and they reported that context did drive the parser's initial structure-building operations.

Recently, experimental psychologists have been concerned with issues of replicability, with several reports of failures to replicate well-known findings (e.g. Stack et al. 2018). Replication has received less attention in psycholinguistics, which is a lost opportunity since our field is uniquely positioned to highlight the opportunities and challenges associated with conducting replication studies, particularly regarding issues of direct versus conceptual replication. Because research practices change, analysis techniques advance, and language evolves so that past stimuli may no longer appropriately instantiate key linguistic manipulations, direct replications are often difficult in psycholinguistics. It is important to ascertain whether past findings replicate given that some past studies may not conform to current best practices.

Method. The study was conducted as a single eye movement experiment and designed as a conceptual replication of F&C1986 and A&S1988. We view the replication as conceptual because, although the same design was used as in the original studies, a few essential changes were made: (a) the *N* was increased to 60; (b) the stimuli were normed; (c) sentences were updated to fit current cultural norms (e.g., sexist items were changed); and (d) analyses were conducted according to current approaches. The eyetracking measures included for analyses were those reported in F&C86: first-pass reading time, probability of a first-pass regression out of a region, and second-pass reading time. Norming data and accuracy were also analyzed.

Results. Behavioral results were as follows: First, analyses of norms suggest the contexts from both studies were less effective than assumed by the original investigators. For the F&C86 stimuli, context had no effect on offline ratings of the appropriateness of either the MA or NMA sentences; instead, overall, subjects rated MA sentences as better than NMA sentences regardless of context bias. For A&S, the NMA-biased contexts did support the NMA form, but raters given MA-biased contexts had no preference for either the MA or the NMA sentence. Question-answering accuracy did not differ across conditions either for F&C86 or A&S88 (contrary to F&C86). Eyetracking results for regressions and first-pass reading times are shown on the following page (Fig. 1). The F&C86 replication showed no clear pattern of results for first-pass reading times, and the likelihood of a first-pass regression was overall greater for NMA than for MA structures, regardless of context. For A&S88 stimuli, regression probability was higher for VP-attached (MA) than for NP-attached (NMA) forms, with no effect of context. First-pass reading times for A&S88 did not differ for either structure given NP-biased contexts and were faster for VP-attached (MA) sentences given VP-biased contexts.

Conclusions. The results of this replication study differed substantially from the findings reported in F&C86 and A&S88. The discrepancies are due to numerous factors including lack of norming data for contexts and low statistical power. Overall, replicability is an important issue in psycholinguistics, and we would suggest that psycholinguistics has much to contribute to discussions concerning how to conduct and evaluate replication studies.

- Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3), 191-238.
- Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25(3), 348-368.
- Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & cognition*, 46(6), 864-877.

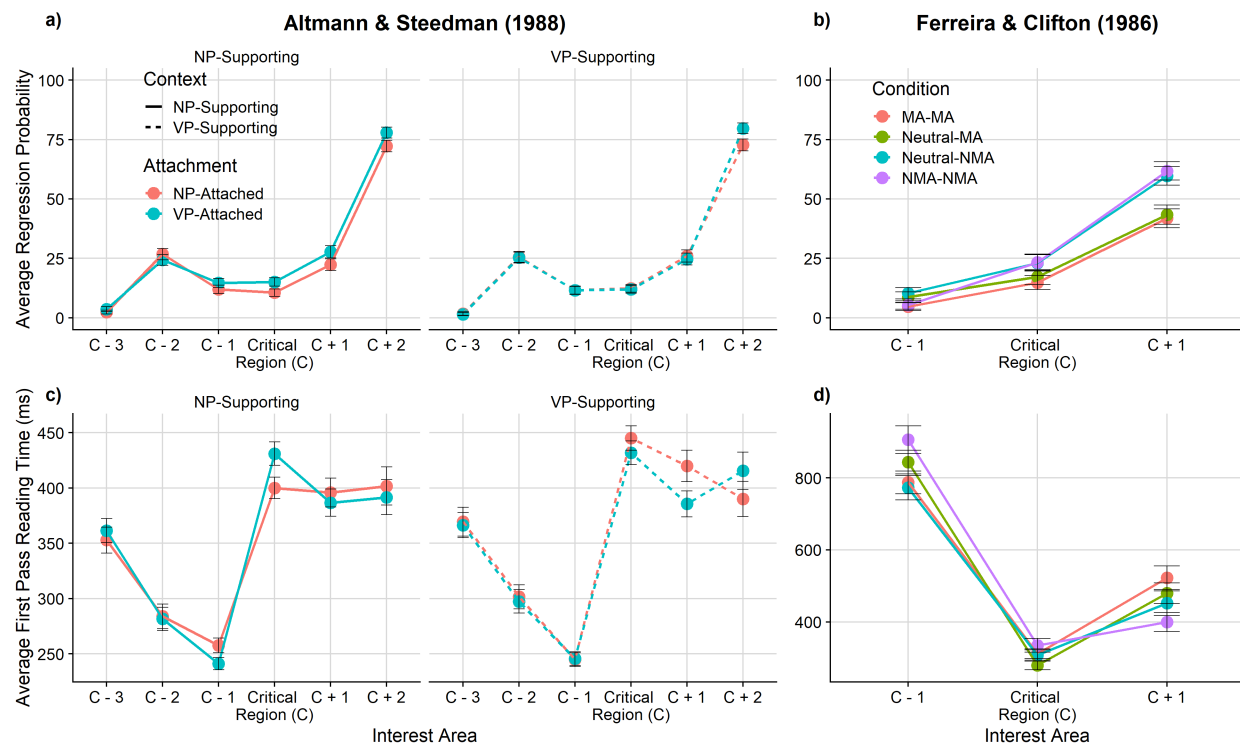


Figure 1. Average of eyetracking measures at each interest area for A&S88 (left) and F&C86 (right). The interest area relative to the critical region (C) is indicated on the x-axis. Panels a) and b) show average regression probability (y-axis) for each interest area and each condition. Panels c) and d) show average first pass reading times (y-axis) for each interest area (note that the y-axis range for average reading time differs for A&S88 and F&C86).

Table 1. Generalized Mixed-Effects Model Analysis Summary for F&C86 and A&S88

Experiment	DV	Region	Summary
F&C86	Regression probability	C - 1	Neutral-MA ($p=.04$), NMA-NMA ($p=.02$)
	First pass reading	C - 1	Neutral-MA ($p=.002$)
	First pass reading	C	Neutral-MA ($p=.03$)
	First pass reading	C + 1	Neutral-NMA ($p=.002$), NMA-NMA ($p<.001$)
A&S88	Regression probability	C + 1	Attachment ($p=.004$)
	First pass reading	C	Context ($p=.049$), Context x Attachment ($p=.003$)

Presupposition projection from disjunction is symmetric

Alexandros Kalomoiros & Florian Schwarz (*University of Pennsylvania*)

We investigate ‘Bathroom sentences’ (Partee), where the second disjunct can support a presupposition in the first, if its negation entails it, as in: ‘Either *the bathroom* is in a weird place or this house has no bathroom’. Recent accounts take projection to be fundamentally asymmetric, and account for the above by positing something extra: Schlenker (2009) posits symmetric filtering that is available at a processing cost for overriding the asymmetric default. Hirsch & Hackl (2014) propose that the presupposition is locally accommodated due to pragmatic constraints on disjunctions. Our adaptation of Mandelkern et al. (2019)’s paradigm for conjunction, which shows that asymmetry in conjunction cannot be overridden, yields support for genuinely symmetric projection for disjunction (without cost), indicating lexical encoding of projection properties.

Design: We created 6 items for different triggers (*continue, again, aware, find out, happy, stop*) in 6 conditions. Conditionals with the trigger in the antecedent in Support (S) and Explicit Ignorance (EI) contexts established a baseline for the acceptability of local accommodation (as in Mandelkern et al.). Presuppositional disjunctions (Ps) in either Order (First vs. Second) were presented in EI context to assess order effects on filtering, with non-presuppositional control variants (No-Ps) (1-4).

Predictions: Accounts positing any kind of asymmetry predict PsFirst to be less acceptable than PsSecond (in EI contexts), beyond any potential independent order effects for No-Ps, i.e., an interaction between Ps/No-Ps and Order. The local accommodation-based asymmetry view also predicts that the difference between PsFirst (Local Acc) and PsSecond (Support from 1st Disj) should parallel that between EI-Cond-Ps (Local Acc) and S-Cond-Ps (Support from Context), given that it posits a parallel contrast between local accommodation and presuppositional support; i.e., there should be NO interaction between embedding (Disj vs. Cond) and presupposition status in context.

Participants & Procedure: 255 participants from Prolific were shown 6 items, one per trigger and condition, in a latin square design. The Cond-Ps controls were shown first to establish baselines (in random order), followed by the disjunction conditions (in random order). Participants indicated on a 7-point scale how natural the sentence sounds in the given context.

Results: The overall pattern is simple (Fig. 1), and confirmed by mixed effect model analyses: S-Ps-Cond was rated higher than all other conditions, and there are no contrasts in the disjunction conditions. Contrary to asymmetric predictions, there is no interaction between Ps/No-Ps and Order here, either. And contrary to local accommodation-based asymmetry accounts, there is a significant interaction between the conditional and disjunction conditions and the contextual status of the presupposition, rather than parallel context effects as posited by such accounts.

Discussion: Our findings contrast starkly with those for conjunction in Mandelkern et al (2019), where asymmetry is reflected in a Ps/No-Ps vs. order interaction. Given the parallel paradigm, this makes a strong case that the effect of linear order on projection differs across connectives. This is incompatible with a domain-general processing account of projection asymmetries grounded in linear order alone (Schlenker 2009). Rather, it favors lexical encoding of linear order projection properties for individual connectives. Two theoretical options remain for disjunction: (i) there is no filtering mechanism at play in disjunction at all, i.e., presuppositions generally project from both disjunctions (cf. Geurts 1999). Cases of non-projection then would have to be explained in another way, e.g., by local accommodation. But note that there is no penalty for local accommodation for the Ps-conditions relative to the No-Ps disjunction conditions in our data, as one might expect on such a view. (ii) The lexical entry for disjunction allows for symmetric filtering in disjunction, in a way that does not incur any processing cost a la Schlenker (2009). While our results do not conclusively settle the choice between these options, the empirical picture clearly speaks against asymmetric treatments and in favour of connective-specific, lexical encoding of linear order

projection properties, thus constraining the space of possible theories of presupposition projection.

Example Stimuli: (font highlights for presentational purposes here only)

- (1) **Contexts:** My friend John researches 20th century literature. One day, I stopped by his house and I saw a copy of Tolkien's "The Fellowship of the Ring" lying around.
- a. I *know that John has been researching Tolkien* recently, ... **(S)**
- b. I *don't know if John has ever had research interests in Tolkien's work*,... **(EI)**
- ...so I thought:
- (2) If John continues having research interests in Tolkien, then that's why he is reading 'The Fellowship'.
- (Ps-Cond)**
- (3) Either John {has / continues having} research interests in Tolkien, or he has never had an interest in Tolkien and the book is unrelated to his research. **((No-)Ps-First)**
- (4) Either John has never had an interest in Tolkien and the the book is unrelated to his research, or he {has / continues having} research interests in Tolkien. **((No-)Ps-Second)**

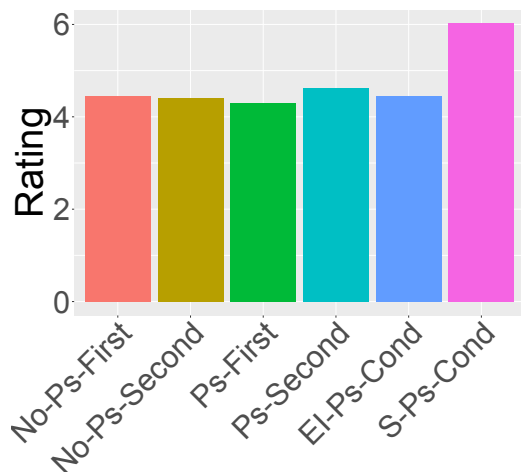


Fig. 1. Mean acceptability by condition

Table 1: Mixed-effects models summary

S-Ps-Cond vs.	Coeff.	SE	<i>p</i>
EI-Ps-Cond	-.89	0.18	<.001
No-Ps-First	-0.78	0.18	<.001
No-Ps-Second	-.54	0.18	<.01
Ps-First	-0.73	0.18	<.001
Ps-Second	-0.61	0.18	<.001
Interaction:	Coeff.	SE	<i>p</i>
Embed * Ps-Status	-.81	0.25	<.01

References: Mandelkern, M., Zehr, J., Romoli, J. et al. 2020. We've discovered that projection across conjunction is asymmetric (and it is!). *Linguistics and Philosophy* 43, 473–514. Schlenker, P. 2009. Local Contexts. *Semantics and Pragmatics* 2, 1-78. Hirsch, A. & Martin Hackl. Incremental presupposition evaluation in disjunction. *Proceedings of NELS 44*. Geurts, B. 1999: *Presuppositions and Pronouns*. Elsevier, Oxford.

Pragmatic inference facilitates word retention in school-aged children

Previous literature has shown that children can leverage social cognition to learn sound-meaning mappings through pragmatic inference. However, the focus has been on in-the-moment meaning mappings rather than meaning retention (Frank & Goodman, 2014; Gollek & Doherty, 2016; Zosh et. al., 2013). In our prior work on adults, we found novel words learned through pragmatic inferences were better retained than those learned through direct mappings and were associated with individuals' social cognition. These results suggest a specific link between social cognition and meaning retention in adults. Here, we examine how children between 4 and 8 years old, a prime stage for social cognition development, learn and retain novel words from an inferential context versus direct-mapping context.

Children ($M_{\text{age}} = 6.0$ years, $SD_{\text{age}} = 1.3$ years, $N = 61$) learned eight novel words during a toy-store tour (see Fig.1 for the overview of the design and the example sentences). During the learning phase, children learned words which could either be mapped to one unique novel object on the display – the Direct Mapping Condition (DMC) – or required pragmatic inference for referential disambiguation – the Inference Condition (IC). The two conditions were manipulated in a blocked design. The attainment of the novel words was tested in a four-alternative-forced-choice (4-AFC) task immediately after each learning block. After completing both learning blocks, children completed a Theory of Mind (ToM) task (Richardson et. al., 2018) via Zoom, lasting an average of 15 minutes, followed by a second recall task testing the retention of all eight novel words in the same 4-AFC task (Fig. 1). The experiment ended with an assessment of children's executive function (EF) skills (Flanker Task).

Learning rates were highly accurate in both conditions, with DMC having a mean of 0.96 ($CI = +0.02$) and IC a mean of 0.69 ($CI = +0.04$). Children performed above chance for DMC and IC in both the recall and retention tests as well. However, unlike adults (Fig. 2B), children showed no difference between the conditions when all children are accounted for (Fig. 2C & 2D). The advantage of IC on retention only emerged in children older than 6 years ($N = 28$, $MIC = 0.55$; $MDMC = 0.41$, estimate = -0.6094 , $z = -2.158$, $p = 0.0309$). For retention in IC over the full age range, age uniquely contributed to variance (Fig. 3), even while taking dependent variables EF, IC immediate recall accuracy, and IC learning accuracy into account (beta = 0.09, $t = 2.624$, $p = 0.0114$). Moreover, the effect of age on IC retention was partially mediated by ToM, explaining 16% of the variance ($F(2,58) = 5.66$, $p = 0.0018$), while the direct effect of age after removing the effect of ToM was no longer significant ($p = 0.1$). There were no significant predictors for retention in DMC or for immediate recall in either condition.

Our findings demonstrate that while children can successfully map and retain meanings learned via pragmatic inference, the facilitation of the pragmatic inference on meaning retention grows with development: children show better memory for pragmatically inferred words than directly mapped words, an adult-like pattern, only after they reach 6 years old. Such a developmental shift in consolidation mechanism is possibly mediated by children's developing ToM skills.

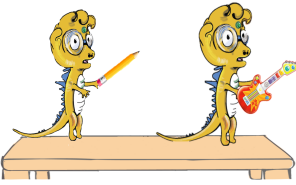
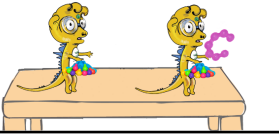


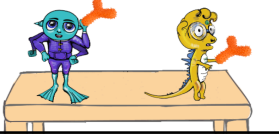

Practice (x2) Which toy does Mary like?	Condition	Learning Phase: 2 trials per word 4 word per condition	Immediate Recall (x4) 15-min break (ToM Task) Retention (x8) 4-min wrap up (Flanker)
	Inferential context		
		"Look! I like this dinosaur! It is holding a MEL !"	
	Direct mapping		
"Look! I like the dinosaur that's holding a guitar!"		"Look! I like this BINK ! It is on the dinosaur!"	"Which one is a BINK ?"

Fig. 1: Experimental procedure for word learning.

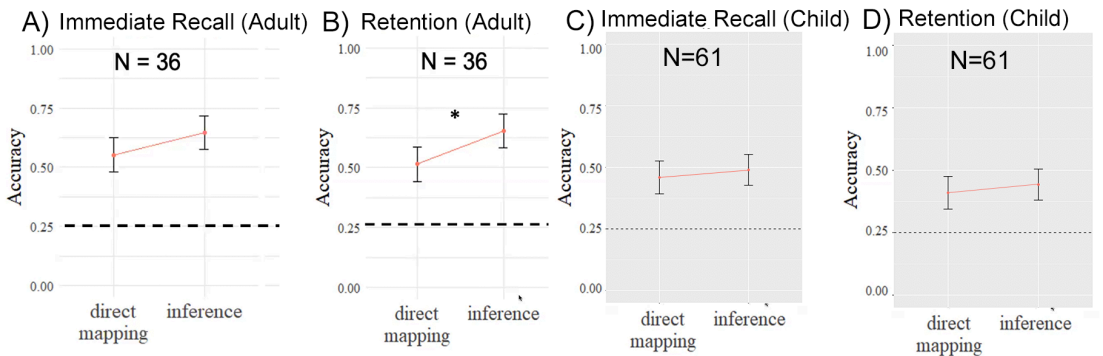


Fig. 2: A comparison of accuracy for both adults, as previously reported, and children. The dashed lines represent the chance levels

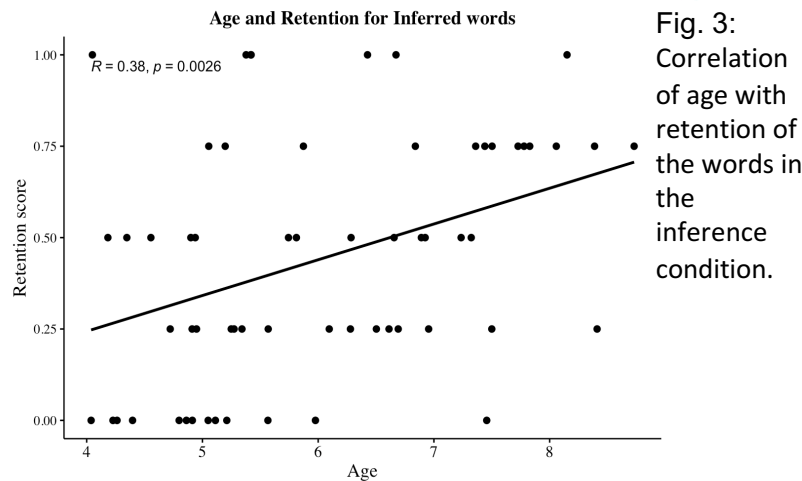


Fig. 3: Correlation of age with retention of the words in the inference condition.

References:

Frank, M.C., & Goodman, N.D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75, 80-96.

Gollek, C., & Doherty, M.J. (2016). Metacognitive developments in word learning: Mutual exclusivity and theory of mind. *Journal of Experimental Child Psychology*, 148, 51-69. <https://doi.org/10.1016/j.jecp.2016.03.007>

Richardson, H. et. al. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, 9, 1027. DOI: 10.1038/s41467-018-03399-2

Zosh, J.M., Brinster, M., & Halberda, J. (2013). Optimal Contrast: Competition Between Two Referents Improves Word Learning. *Applied Developmental Science*, 17(1), 20–28. DOI: 10.1080/10888691.2013.748420

A corpus-based study of (non-)exhaustivity in *wh*-questions

Morgan Moyer and Judith Degen (Stanford University)

mcmoyer@stanford.edu

A key issue in *wh*-question interpretation regards the distribution of exhaustive (Mention-All, MA) vs. non-exhaustive (Mention-Some, MS) question readings (see (1) and (2)):

- | | |
|------------------------------------|------------------------------------|
| (1) Who came to the party? | (2) Where can I find coffee? |
| a. Who is every person that...? MA | a. What is every place that...? MA |
| b. Who is a person that...? #MS | b. What is a place that...? MS |

Theories of question interpretation have typically assumed that a MA reading is always appropriate [1,2]. Linguistic factors that have been argued to generate variation in readings include the specific *wh*-word—e.g., *who*-questions are biased for MA, while *where/how*-questions are biased for MS [3-4]—and existential (priority) modality—e.g., *can* purportedly licenses MS, as in (2) [5-9]. Recent work [10] tested these judgements in lab-controlled experiments with artificial stimuli and found evidence for some biases, but these biases can be overridden by features of the context like speaker/discourse goals [cf. 3-4,11]. However, there is to date no systematic investigation of *naturally occurring questions* that tests the intuitions reported in the literature. We ask: (Q1) How much does question interpretation vary in natural discourse contexts? Is there indeed a bias for MA? (Q2) Is the distribution of interpretations modulated by linguistic form?

Methods. Step 1: Naturalistic Stimuli from a Corpus Database. Using Tgrep2 and the Tgrep2 Database Tools [12-14], we extracted all occurrences of *wh*-questions (10,009) from the Switchboard corpus [15] and coded the questions for syntactic structure (e.g., embedded, root), *wh*-word, and presence of modality. To curate stimuli for step 2, we constrained the database to the most frequently discussed cases: root *who*-, *where*-, and *how*-questions. We also removed degree (*How much sugar do you need?*) and identity (*Who is that?*) questions because MS and MA meanings converged, with 335 questions remaining. The distribution of *wh*-word and modality in this database is reported in Table 2. **Step 2: Paraphrase Rating Task.** The remaining cases were divided into 11 lists with occurrence of critical factors roughly proportional to the overall database. Participants (n=385) on Prolific were presented with each question and the 10 preceding lines of dialogue, and asked to rate the likely intended meanings (paraphrases), using a slider task (Fig. 1). Question paraphrases were selected to reflect MS/MA readings: *a* indicates MS, *every* MA, while in *the*-paraphrase the two readings collapse. There was a fourth option (*something else*) in case no other was appropriate. Performance on 6 catch trials functioned as exclusion criterion (n=19).

Results. Questions with highest ratings for *something else* (17%) were excluded because they were rhetorical (see Tab. 1). *The*-paraphrases, where MS=MA, had the highest mean rating (.59), suggesting that only one reading was possible for most cases. Data were analysed using linear mixed effects regression. To investigate the posited MA bias, we compared *every* vs. *a* ratings, as these represent MA and MS (Fig. 2): contrary to the literature, there was no bias for *every* (Q1). However, significant two-way interactions between paraphrase and linguistic form factors partially support reports from the literature (Q2): first, the presence of a modal resulted in higher ratings of *a* ($p < .0001$, Fig. 3) [5-9,10] but not *every*. Second, ratings for *how*-questions resulted in higher *a* than *every* ratings ($p < .04$), confirming [3-4, 10], but not for *where* or *who*-questions.

Conclusion. In contrast to theoretical predictions, we find no bias for MA question readings in naturalistic dialogue (Q1). With respect to (Q2), we find support for some, but not all, observations about the effect of linguistic form on question interpretation reported in the literature. We suggest that MS/MA readings result from reasoning about the speaker's goal in the context, consistent with a constraint-based account [16] on which hearers integrate multiple sources of information to determine meaning. These results also have methodological implications: data hand-selected during theory-building may be biased and not reflect a realistic distribution of meanings [17].

Paraphrase	Example	Mean Rating
every (MA)	<i>Where have you skied?</i> <i>Where's it all going?</i>	.66 .59
a (MS)	<i>Where do you like to eat?</i> <i>How would you achieve that?</i>	.57 .51
the (MS=MA)	<i>Where you going to school?</i> <i>Where do you work?</i>	.99 .99
something else	<i>Who knows?</i> <i>How can you watch that?</i>	.61 .53

Table 1: For each paraphrase, examples of questions that resulted in high ratings on that paraphrase.

Wh	Modal?	% of Total
who	Yes	2.4%
	No	13.7%
where	Yes	1.2%
	No	27.8%
how	Yes	8.4%
	No	46.6%

Table 2: Joint distribution of *wh*-words and modals in database of 335 root questions.

Speaker #2: pretty good.
Speaker #1: i do like to ski.
Speaker #2: pretty, pretty down there. huh?
Speaker #1: yeah, i, i said i do like to ski.
Speaker #2: *so, where, have you skied ?*

Based on the sentence in red, how likely do you think it is that the speaker wanted to know about each of the following?

What is every place...? 0

What is a place...? 0

What is the place...? 0

Something else 0

Continue

Figure 1: Paraphrase Rating Task: Participants evaluate intended question meanings by moving the slider next to paraphrases, assigning a numerical value between 0-1. Combined ratings must sum to 1 to generate a proper probability distribution.

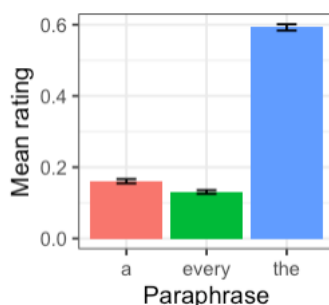


Figure 2: Surprisingly, *every* paraphrases were not preferred over *a* paraphrases.

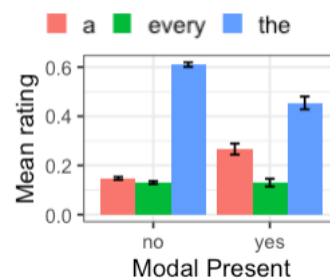


Figure 3: For modal questions, *a* received higher ratings than *every* (in line with [5-9]), but surprisingly not lower in non-modal questions (in contrast to [5-9]).

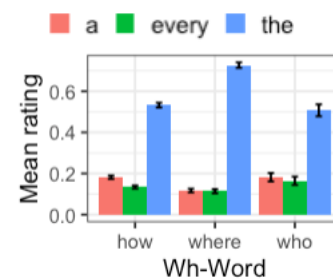


Figure 4: *A*-paraphrases received higher ratings than *every* for *how* (in line with [3-4]), but surprisingly not lower for *who* (in contrast to [3-4]).

References. [1] Karttunen (1977), [2] Groenendijk & Stokhof (1984), [3] Ginzburg (1995), [4] Asher & Lascarides (1998), [5] George (2011), [6] Nicolae (2013), [7] Fox (2014), [8] Dayal (2016), [9] Xiang (2016), [10] Moyer & Syrett (2019), [11] van Rooij (2003), [12] Rohde (2005), [13] Jaeger (2006), [14] Degen & Jaeger (2011), [15] Godfrey et al. (1992), [16] Degen & Tanenhaus (2019), [17] Degen (2015)

At least as a scalar modifier: Scalar diversity and ignorance inferences

Stavroula Alexandropoulou (University of Potsdam)

It is an established fact that utterances with *at least* convey a signal of speaker ignorance (SI). The majority of the relevant literature has focused on *at least* as a numeral modifier and the SI inference it triggers, e.g., *the speaker doesn't know the exact number n of people that were at the party* in (1). The most popular approach to these inferences derives them as (primary) Quantity implicatures. Importantly, the literature has generally overlooked uses of *at least* as an adjective modifier and their potential SI inferences, e.g., *the speaker doesn't know whether Sue is gorgeous* in (2). A few exceptions are Geurts & Nouwen (2007) and Cohen & Krifka (2014), who treat the two *at least* scalar constructions on a par.

- | |
|---|
| (1) There were at least 50 people at the party. (1') ??In fact, there were 54 people. |
| (2) Sue is at least pretty. (2') In fact, she's gorgeous. |

These analyses

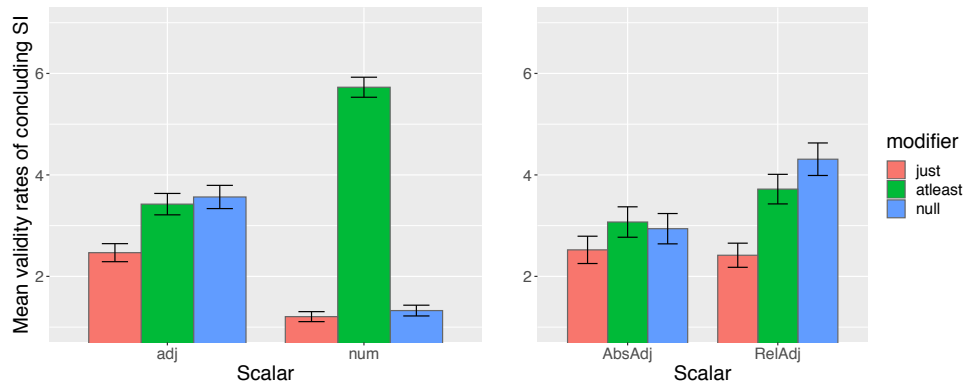
take for granted that the two *at least* constructions trigger SI uniformly, however cancellation data suggest otherwise. While (1-1') illustrate that SI inferences of *at least*+numeral are hard to cancel, (2') easily suspends the corresponding SI inference of (2). In this context, we set out to probe experimentally whether this contrast in SI robustness holds, although not captured by any theory.

Experiment—We used a web-based inference task (in Greek, $n=46$), consisting of pairs of an utterance by Maria and a conclusion. Participants had to rate how valid the conclusion was given Maria's utterance on a Likert scale from 1 (*not valid at all*) to 7 (*absolutely valid*). In the target items, Maria's utterance contained *at least*+(weak) scalar term. In **a.**, the ignorance inference of the conclusion follows from a (Quantity-based) reasoning given, e.g., a two-scale analysis of *at least*+scalar: i.e., with substitutions of *at least* from <at least, just> and of n from the number scale. Likewise for **b.**, with the difference of substituting for the adjective from the Horn scale <tipsy, drunk>. We had two manipulations

in Maria's utterance: the scalar modifier: *at least*

- | |
|---|
| a. Maria says: "There were at least thirteen actors on stage yesterday."
Conclusion: Maria doesn't know the exact number of actors that were on stage yesterday. |
| b. Maria says: "When she came back to the hotel room, Fani was at least tipsy ."
Conclusion: Maria doesn't know whether Fani was drunk when she returned to the hotel. |

just / \emptyset (*null*), and the **scalar** type: num vs. adj (3×2 Latin square). The *just* control conditions being inconsistent with the conclusion's ignorance inference are expected to obtain low rates. The same holds for the *null* conditions, if the respective scalar implicatures are computed, replicating Doran et al.'s (2009) diverse findings for bare adj and num. We had 12 items mixed with 24 fillers. The adj items were split into 6 absolute and 6 relative adj. **Results**—Mixed-effects ordinal regression analyses (baselines: *just*+adj) revealed: *At least*+num received high rates overall (see plot), confirming the robustness of SI inferences of *at least*+num. *At least*+adj was rated significantly higher than *just*+adj ($p < .01$), indicating that *at least*+adj triggers SI, though to a lesser extent than *at least*+num does (interaction: $p < .0001$). This is at odds with a uniform analysis of *at least*+num/adj. Also, the higher rates of *null*+adj vs. *just*+adj ($p < .001$) and the lack of such an effect for num are consistent with the claim that numerals are better at triggering scalar inferences than adjectives (Doran et al.'s scalar diversity). Zooming in on the two gradable adj classes, we find that the significant simple effect of *at least* in the previous analysis seems to be mainly driven by such an effect for relative adj ($p < .01$), while this was not significant for absolute adj ($p = .53$). Hence, SI inferences target a specific class of gradable adj. A potential source of this preference is the underlying scale structure of relative adj, and specifically, vagueness. **Implications**—This study provides evidence that (i) scalar diversity is relevant not only for scalar inferences but also for SI inferences, hinting at a difference in the implicature mechanism of the different scalars, (ii) the underlying scale structure of adjectives affects the availability of SI inferences, as in the case of scalar inferences of bare adjectives (Gotzner et al., 2018). The interplay of scale structure and implicature needs to be looked into.



References: Cohen, A. & Krifka, M. (2014). Superlative quantifiers and meta-speech acts. *Linguistics and Philosophy*, 37:41–90. Doran, R., Baker, R. E., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics*, 1:211–248. Geurts, B. & Nouwen, R. (2007). At least et al.: The semantics of scalar modifiers. *Language*, 83:533–559. Gotzner N, Solt S., & Benz A. (2018). Scalar Diversity, Negative Strengthening, and Adjectival Semantics. *Frontiers in Psychology*, 9:1659.

Original example items with glosses in all conditions

I Maria lei:
Det Maria says
'Maria says:'

a.' Numeral

"Ipirhan to lighotero / akrivos / ∅ dhekatris ithopii epi skinis stin parastasi pu idhame
there were at least / exactly / ∅ 13 actors on stage at the show that watched
hthes."
yesterday

'There were at least / exactly / ∅ 13 actors on stage at the show we watched yesterday.'

Simberasma: I Maria dhen kseri ton akrivi arithmo ithopion pu itan epi skinis stin
conclusion Det Maria not knows the exact number actors.gen that were on stage at the
parastasi pu idhe hthes.
show that saw yesterday

Conclusion: 'Maria doesn't know the exact number of actors that were on stage at the show she saw yesterday.'

b.' Adjective

"Otan epestrepse sto dhomatio tu ksenodhohiu apo to
when returned at the room the.gen hotel from the
nihterino maghazi dhyaskedhasis, i Fani itan to lighotero / aplos / ∅ zalizmeni."
night club Det Fani was at least / simply / ∅ tipsy
'When she came back to the hotel room from the night club, Fani was at least / just / ∅ tipsy.'

Simberasma: I Maria den kseri an i Fani itan methismeni otan epestrepse sto
conclusion Det Maria not knows whether Det Fani was drunk when returned at the
dhomatio tu ksenodhohiu apo to nihterino maghazi dhyaskedhasis.
room the.gen hotel from the night club

Conclusion: Maria doesn't know whether Fani was drunk when she came back to the hotel room from the night club.'

Priming pragmatic reasoning in the verification and evaluation of comparisons

Vishakha Shukla, Madeleine Long, Vrinda Bhatia & Paula Rubio-Fernandez (University of Oslo)
paula.rubio-fernandez@ifikk.uio.no

Most studies on scalar implicature focus on the lexical scale ‘some’ vs ‘all’, which tends to elicit high rates of pragmatic responses [1-4]. Here we examined an understudied scale formed by two syntactic constructions: categorizations and comparisons (e.g., ‘A robin is a bird’ vs ‘A robin is like a bird’). Unlike ‘some’ statements, superordinate comparisons have been found to elicit high rates of logical responses [5], even though they are under-informative when interpreted pragmatically (SI: *A robin is not a bird*). Following recent work on enrichment priming [6-9], we predicted that ‘some’ and ‘all’ statements would introduce an *informativity bias* in sentence verification and evaluation, increasing pragmatic responses to under-informative comparisons.

EXP 1 aimed to replicate previous findings by testing whether under-informative comparisons would elicit high rates of logical (vs pragmatic) responses in a sample of 22 UCL students. Replicating prior work [5], high rates of True responses (83%) were observed, in stark contrast to the high rates of True responses previously reported for ‘some’ and ‘all’ [1-4].

EXP 2 employed a rating task to test whether ‘some’ and ‘all’ statements are more readily perceived as scalemates and elicit scalar implicatures, than categorizations and comparisons. 68 adults from the UK were recruited via Prolific to rate statements on a scale (1=Very bad, 7=Very good). In line with previous work [5], we predicted higher ratings for stronger statements (‘all’ and categorizations) than weaker ones (‘some’ and comparisons). Critically, we also predicted comparisons would be rated higher than ‘some’ sentences (despite both being under-informative). An LMER model of Rating with Sentence Form (Weak vs Strong) and Group (Some & All vs Categorization & Comparison) as FE and maximum RE structure revealed a main effect of Sentence Form ($p<.001$), with lower ratings for weak forms, and a main effect of Group ($p<.001$) with higher ratings for categorizations and comparisons. The Sentence Form x Group interaction was also significant ($p<.001$), driven by a main effect of Group for weak forms ($p<.001$) (Fig. 1).

EXP 3 tested our main prediction that ‘some’ and ‘all’ would prime pragmatic reasoning. 156 adults from the UK were recruited via Prolific and were administered one of two online tasks: sentence verification or evaluation. In both tasks, participants read comparisons and categorizations alone, or randomized with ‘some’ and ‘all’ sentences, and had to judge whether the statements were True or False (verification) or Good or Bad (evaluation). An LMER model of Response (True/Good=1, False/Bad=0) with Sentence Type (Categorization, Comparison), Condition (Without Some/All, With Some/All) and Instruction (Verification, Evaluation) as FE and maximum RE structure revealed a marginal Sentence Type x Condition interaction ($p=.056$), driven by a difference in comparisons ($p=.007$), but not categorizations ($p=.135$) across conditions (Fig. 2). Specifically, the rate of True/Good responses was lower for comparisons With Some/All, supporting our hypothesis that participants engaged in pragmatic reasoning when processing ‘some’ and ‘all’ statements and as a result responded pragmatically to comparison statements (for full model output, see Table 1). Further support comes from an RT LMER analysis of True/Good responses using the same variables. We found a main effect of Condition ($p=.013$), with faster RTs in the Without Some/All condition than the With Some/All condition (Fig. 3, Table 2), likely because pragmatic reasoning slows down processing. Along these lines, the B&N effect [3,10] posits that participants will take longer to respond False than True for ‘some’ statements precisely because deriving scalar implicatures is cognitively costly. We tested this with our own data and replicated these findings with slower RTs for False/Bad than True/Good for ‘some’ items ($p=.040$) (Fig. 4). These results extended to comparison statements where RTs were slower for False/Bad than True/Good ($p<.001$), suggesting the inference *An X is not a Y* is also costly [10].

Our study is the first to show that ‘some’ and ‘all’ statements prime pragmatic reasoning in both sentence verification and evaluation tasks. This finding suggests that different scalar terms not only give rise to different rates of scalar implicatures [10-13], but can also affect the interpretation and processing of other types of scalar expressions.

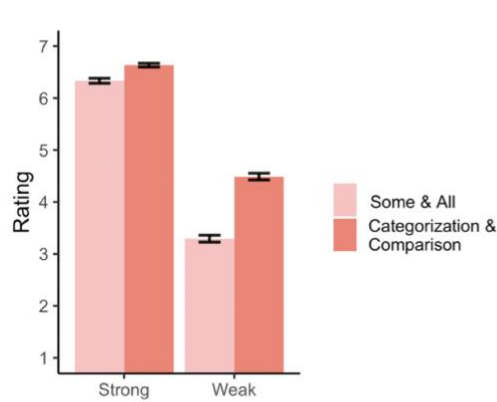


Fig 1. Average sentence ratings in Experiment 2, showing an interaction between Sentence Form and Group.

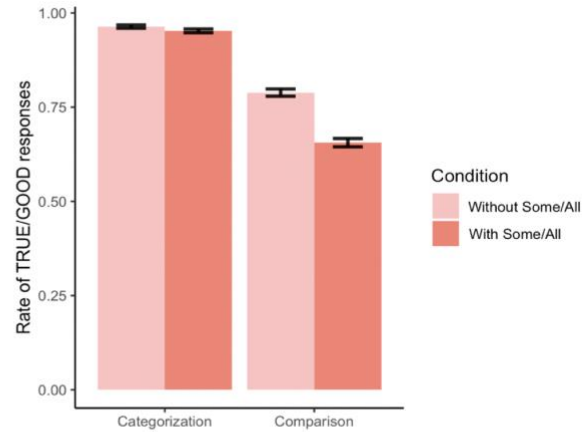


Fig. 2. Average rates of TRUE/GOOD (logical) responses to categorizations and comparisons in Experiment 3, showing the Sentence Type by Condition interaction.

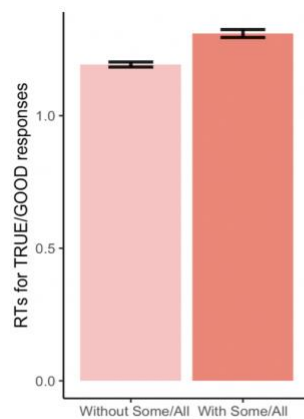


Fig. 3. Average RTs to categorizations and comparisons in Experiment 3, showing the main effect of Condition.

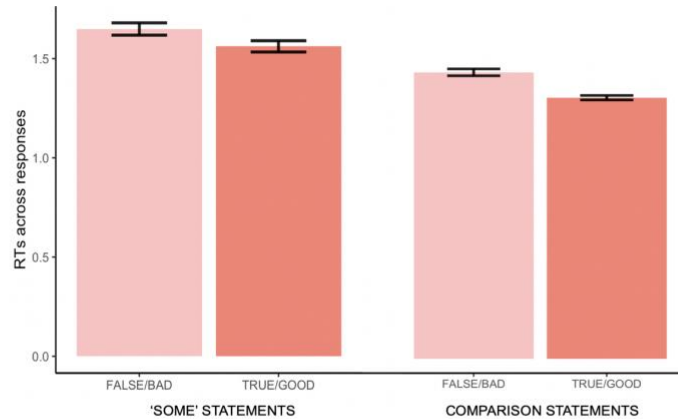


Fig. 4. Average RTs for FALSE/BAD (pragmatic) responses and TRUE/GOOD (logical) responses for 'some' items and comparison items in Experiment 3. Pragmatic responses to both types of under-informative statement were significantly slower than logical responses, confirming that the derivation of the corresponding scalar implicatures is cognitively costly.

Table 1. Full model output for accuracy rates.

Fixed effect	Coefficient	SE	p value
Sentence Type	-1.435	0.308	<.001
Condition	-0.922	0.316	0.004
Instruction	-1.330	0.338	<.001
Sentence Type x Condition	-1.108	0.580	0.056
Sentence Type x Instruction	2.034	0.566	<.001
Condition x Instruction	-0.078	0.620	0.900
Sentence Type x Condition x Instruction	1.057	1.143	0.355

Note: Significant and marginal main effects and interactions are shaded.

Table 2. Full model output for RTs to True/Good responses.

Fixed effect	Coefficient	SE	p value
Sentence Type	0.128	0.019	<.001
Condition	0.110	0.044	0.013
Instruction	-0.022	0.043	0.620
Sentence Type x Condition	-0.008	0.038	0.843
Sentence Type x Instruction	-0.042	0.041	0.313
Condition x Instruction	0.108	0.087	0.216
Sentence Type x Condition x Instruction	-0.013	0.079	0.870

Note: Significant main effects and interactions are shaded.

References

- [1] Noveck, 2001. *Cognition*. [2] Noveck & Posada, 2003. *Brain and Language* [3] Bott & Noveck, 2004. *Journal of Memory and Language* [4] Feeney, Scafton, Duckworth & Handley, 2004. *Canadian Journal of Experimental Psychology* [5] Rubio-Fernandez, Geurts & Cummins, 2017. *Review of Philosophy and Psychology* [6] Bott & Chemla, 2016. *Journal of Memory and Language* [7] de Carvalho, Reboul, der Henst, Cheylus & Nazir, 2016. *Frontiers in Psychology* [8] Rees & Bott, 2018. *Cognition* [9] Rees, Bott & Schumacher, 2019. *Neuroscience Letters* [10] van Tiel, Pankratz & Sun, 2019. *Journal of Memory and Language* [11] Doran, Baker, McNabb, Larson & Ward, 2009. *International Review of Pragmatics* [12] Doran, Ward, Larson, McNabb & Baker, 2012. *Language* [13] van Tiel, van Miltenburg, Zevakhina & Geurts, 2016. *Journal of Semantics*.

Self-reported inner speech salience moderates implicit prosody effects
Mara Breen (Mount Holyoke College) and Evelina Fedorenko (MIT)

According to the Implicit Prosody Hypothesis (Fodor, 1998), readers generate imagined sound representations during silent reading which can influence comprehension. These representations are imagined, so cannot be measured directly. Inspired by prior work demonstrating that individuals' self-reported auditory imagery salience predicts memory for pitch contours (Hishitani, 2009), in the current pre-registered study, we investigated whether self-reported inner speech salience predicts the correlation between silent reading processes, as measured by eye-tracking, and overt reading behavior, as measured by spoken duration. The Varieties of Inner Speech Questionnaire (VISQ) predicts activation in brain areas associated with inner speech tasks (Alderson-Day et al., 2016) and self-reported imagery during silent reading (Alderson-Day, et al., 2017). If implicit prosodic representations are similar to explicit ones, participants with higher VISQ scores should exhibit stronger correlations between silent and aloud reading durations. We also assessed standardized measures shown to predict spoken durations, including the Peabody Picture Vocabulary Test (PPVT) (Spear-Swerling, 2006), Author Recognition Test (ART) (Moore & Gordon, 2015), Digit Span test (Naveh-Benjamin & Ayres, 1986), and Rapid Automatized Naming test (RAN) (Vukovic, et al., 2004), to determine whether they also predict silent reading durations. Finally, the Communication subscale of the Autism Quotient test (AQ-C) modulates implicit prosodic effects (Jun & Bishop, 2015). Therefore, participants with higher scores on the AQ-C should also exhibit stronger correlations between silent and aloud reading word durations.

Participants (N = 62) read 176 syntactically and semantically diverse English sentences twice – once silently and once aloud: 128 were 12-word naturalistic sentences with variable syntactic structure; 48 were syntactically controlled sentences. Twenty-four sentences were read silently and aloud twice, in order to assess the reliability of the reading measures. During silent reading, participants' eyes were tracked with an EyeLink 1000+. During overt reading, participants' voices were recorded with a head-mounted microphone. Participants read both silently and aloud on two separate days, at least a week apart, and completed standardized assessments on the second day. Order of list presentation and modality was counter-balanced.

Using linear mixed-effects regression, we predicted first pass reading time on each word in each sentence from spoken duration, with participant and sentence as random effects. We tested whether the standardized measures individually, and the interaction of the AQ-C and the VISQ with spoken duration, improved model fit by comparing models with and without each term using a likelihood ratio test. Effects which lead to significantly, or marginally, better fit were retained. The final model (Table 1) includes main effects of spoken duration, and the ART and RAN, demonstrating that shorter first pass times are predicted by faster spoken word duration, higher ART scores, and faster RAN times. In addition, the interactions of duration and AQ-C and duration and VISQ were significant, indicating that spoken durations are more predictive of silent reading times for speakers who report a) more salient inner speech, and b) less autistic-like communication skills (contrary to the prediction). In summary, these results demonstrate that the correspondence between silent and over reading processes is modulated by individual differences, providing support for the role of implicit prosody in sentence processing.

<i>Fixed Effects</i>	First Pass Duration			
	<i>Estimates</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	208.81	27.13	7.70	<0.001
Spoken Duration	84.08	16.18	5.20	<0.001
ART	-0.86	0.53	-1.62	0.11
RAN	3.28	1.53	2.14	0.04
Spoken Duration x VISQ	0.58	0.21	2.74	0.006
Spoken Duration x AQ_C	-3.05	1.25	-2.44	0.01

Table 1: Fixed effects in the model predicting first pass duration during silent reading. ART = Author Recognition Test; RAN = Rapid Automatized Naming; VISQ = Varieties of Inner Speech Questionnaire; AQ-C = Autism Quotient, Communication subscale

References

1. Alderson-Day, B., Bernini, M., & Fernyhough, C. (2017). Uncharted features and dynamics of reading: Voices, characters, and crossing of experiences. *Consciousness and Cognition*, 49, 98–109.
2. Alderson-Day, B., Weis, S., McCarthy-Jones, S., Moseley, P., Smailes, D., & Fernyhough, C. (2016). The brain's conversation with itself: Neural substrates of dialogic inner speech. *Social Cognitive and Affective Neuroscience*, 11(1), 110–120.
3. Fodor, J. D. (1998). Learning To Parse? *Journal of Psycholinguistic Research*, 27(2), 285–319.
4. Hishitani, S. (2009). Auditory Imagery Questionnaire: Its factorial structure, reliability, and validity. *Journal of Mental Imagery*, 33, 63-80.
5. Jun, S.-A., & Bishop, J. (2015). Priming Implicit Prosody: Prosodic Boundaries and Individual Differences. *Language and Speech*, 58(4), 459–473.
6. Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods*, 47(4), 1095–1109.
7. Naveh-Benjamin, M., & Ayres, T. J. (1986). Digit Span, Reading Rate, and Linguistic Relativity. *The Quarterly Journal of Experimental Psychology Section A*, 38(4), 739–751.
8. Spear-Swerling, L. (2006). Children's Reading Comprehension and Oral Reading Fluency in Easy Text. *Reading and Writing*, 19(2), 199–220.
9. Vukovic, R. K., Wilson, A. M., & Nash, K. K. (2004). Naming Speed Deficits in Adults with Reading Disabilities: A Test of the Double-Deficit Hypothesis. *Journal of Learning Disabilities*, 37(5), 440–450.

Guiding Implicit Prosody with Delexicalized Melodies: Evidence from a Mismatch Task

Nicholas Van Handel, Matthew Wagers, Amanda Rysling (UC Santa Cruz)

In [1]’s “reading with delexicalized melodies” task, subjects heard low-pass filtered sentences, which lack segmental content but retain prosody, then replicated these melodies during silent reading of a target sentence. This method seems to hold promise for addressing when/how implicit prosody manifests in reading and how implicit prosody interacts with syntactic parsing [2, 3]. There is growing interest in extending this task [4, 5] and in using overt speech to guide reading [6]. Conclusions from this method depend on the extent to which subjects accurately replicate melodies in reading. Holding a sentence melody in memory is potentially difficult, but previous work has not explicitly assessed subjects’ ability to project full melodies onto read sentences. Here, we report 4 match/mismatch tasks using more complex stimuli than [1], contrasting simultaneous and sequential presentation of the melody and written sentence.

Method. 36 items manipulated STRUCTURE (NP vs. Z) and MELODY (MATCH vs. MISMATCH); see Table 1. The NP/Z garden path [7, 8] was chosen because NP and Z have clear prosodic differences [9]. A native American English-speaking phonetician recorded all sentences. MISMATCH melodies cross-spliced NP and Z recordings, such that the boundary occurred in the wrong location. Accurate performance required subjects to remember the relative position of the boundary, providing a strong test of subjects’ ability to replicate the melody; cf. [1], which only varied the presence of a boundary after the second word in a sentence. In Expt 1, melody and sentence were presented simultaneously, while in Expt 2, the sentence appeared after the melody. Subjects judged the melody as “Match” or “Mismatch” and rated their confidence on a 3-point scale. Responses were converted to a 6-point scale (1=confident “Mismatch”; 6=confident “Match”) [10]. Bayesian cumulative link mixed models [11] were fit to ratings (fixed effects: STRUCTURE, MELODY, interaction; maximal random effects).

Expt 1. Simultaneous (n=65). Ratings are summarized in Figure 1. There were main effects of STRUCTURE, such that Z sentences were rated lower (-.69, [-1.02, -.36]), and MELODY, such that MISMATCH were rated lower (-3.03, [-3.55, -2.52]), with no interaction. The MISMATCH penalty confirms that subjects were sensitive to mismatches. The Z penalty suggests that it is harder to compare a melody to a written sentence when the latter contains a garden path.

Expt 2. Sequential (n=38). There was a main effect of MELODY, such that MISMATCH were again rated lower, (-.83, [-1.08, -.59]), but no effect of STRUCTURE, nor an interaction. The MISMATCH penalty shows that subjects distinguish MATCH from MISMATCH, but the effect size was small relative to Expt 1, with worse performance in the MATCH conditions in particular. Subjects also reported difficulty with the task; data from an additional 18 (32%) were excluded for giving higher ratings to MISMATCH melodies, indicating poor performance. Lack of a Z penalty may be the result of compressed ratings and poor performance.

Discussion. Expt 1 showed that subjects compared melodies and written sentences when presented simultaneously, but the Z penalty suggests that the melody did not override default parsing to prevent garden paths. This method may be appropriate to make certain phrasings available, but not to direct implicit prosody in first-pass reading. Expt 2, which required more memory load with sequential presentation, was less effective: while the effect of MELODY suggests a (limited) ability to replicate melodies, we are skeptical that subjects do so reliably enough for the melody to direct their implicit prosody. Poor performance with NP/Z raises doubts about generalizing the task to longer melodies or subtler cues. Preliminary results from replications with slower melodies, to make the task easier, show qualitatively the same effects (Expts 3 and 4, below). We note that [1]’s study included extensive production training, shorter sentences, and easier boundary placement conditions; the present study did not. We thus advise caution with any sequential method, closer to [1]’s original, as the assumption that subjects accurately read with a melody may not hold without [1]’s training and conditions.

STRUCTURE	MELODY	Item
NP	MATCH	[After Anne visited the British relatives %] _{NP} [the cousins moved to the countryside.] _{NP}
	MISMATCH	[After Anne visited % the British relatives] _Z [the cousins moved to the countryside.] _{NP}
Z	MATCH	[After Anne visited % the British relatives] _Z [moved to the countryside.] _Z
	MISMATCH	[After Anne visited the British relatives %] _{NP} [moved to the countryside.] _Z

Table 1. Sample NP/Z item.

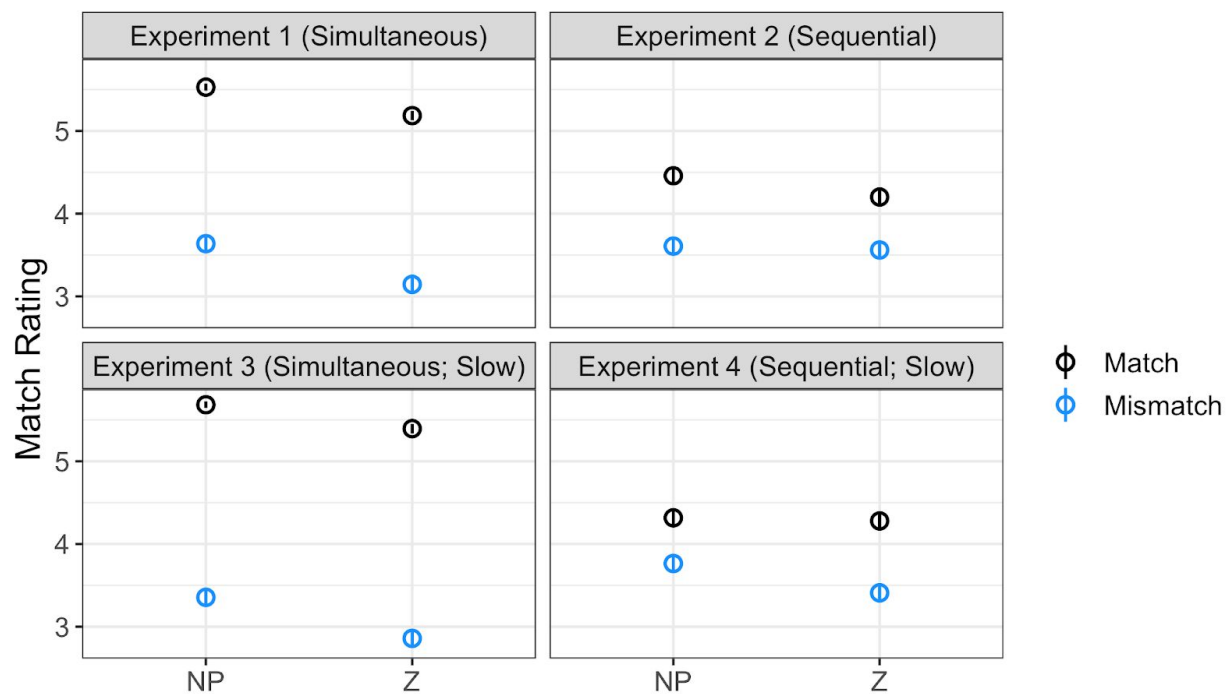


Figure 1. Mean ratings by experiment. Error bars indicate standard error of the mean.

References.

- [1] Steinhauer, K., & Friederici, A.D. (2001). *Journal of Psycholinguistic Research*
- [2] Fodor, J. (2002). *Speech Prosody 2002*
- [3] Breen, M. (2014). *Language and Linguistics Compass*
- [4] Luo, Y., Yan, M., & Zhou, X. *Journal of Experimental Psychology: Learning, Memory, and Cognition*
- [5] Mills, J. (2020). *Proceedings of Speech Prosody 2020*
- [6] Zhang, G., & Husband, E.M. (2019). *Poster at the 32nd CUNY Sentence Processing Conference*
- [7] Frazier, L., & Rayner, K. (1982). *Cognitive Psychology*
- [8] Frazier, L., Carminati, M.N., Cook, A.E., Majewski, H., & Rayner, K. (2006). *Cognition*
- [9] Kjelgaard, M.M., & Speer, S.R. (1999). *Journal of Memory and Language*
- [10] Macmillan, N.A., & Creelman, C.D. (2005).
- [11] Bürkner (2017). *Journal of Statistical Software*

Using eye movements to predict performance on reading comprehension tests

Diane Mézière (IDEALAB, Macquarie University, Rijksuniversiteit Groningen, University of Potsdam, Newcastle University), Lili Yu (Macquarie University), Erik Reichle (Macquarie University), Titus von der Malsburg (University of Potsdam, Massachusetts Institute of Technology), Genevieve McArthur (Macquarie University)

Reading comprehension is one of the most complex cognitive tasks that we engage in on a daily basis. Although many theories of reading comprehension exist, the essential cognitive skills that are predictive of reading comprehension remain unclear, making the design of valid measurements of reading comprehension difficult. In this study, we use eye-movements to examine the extent to which three different reading comprehension tests measure various cognitive skills.

We gave three widely-used standardised reading comprehension tests to 79 adults with no history of reading difficulties: the *York Assessment for Reading Comprehension* (YARC; Snowling et al., 2009), the *Gray Oral Reading Test* (GORT-5; Wiederholt & Bryant, 2012), and the sentence comprehension subtest of *Wide Range Achievement Test* (WRAT-4; Wilkinson & Robertson, 2006). In the YARC, participants read two long passages silently, followed by comprehension questions. In the GORT, participants read eleven short passages aloud, also followed by comprehension questions. In the WRAT, participants were asked to read thirty-one sentences with a missing word, and were asked to provide the missing word (cloze procedure). Participants' eye movements were monitored while the tests were administered.

The correlations between the three comprehension scores were moderate and statistically significant (0.59-0.63). Correlations between the comprehension scores and the eye-movement measures yielded a different pattern for each test. Scores from the YARC tended to be more highly correlated to early eye-movement measures, indicative of early reading processes such as lexical processing. Scores from the GORT showed similar correlation coefficients for both early and late eye-movement measures - typically associated with higher-level integration processes. Scores from the WRAT were more highly correlated to late eye-movement measures.

To further investigate the relationship between eye movements and comprehension scores, we ran a second set of analyses to test if eye movements could predict comprehension scores. Bayesian linear models were used to evaluate the efficacy of all combinations of our eye movement measures. Leave-one-out cross-validation (Vehtari, Gelman & Gabry, 2017) was then used to compare these models and identify the 'best' model to predict comprehension. Results from these analyses also yielded test variance. For the YARC, the best model included both early and late eye-movement measures. For the GORT, early measures appeared as the best predictors, closely followed by total reading time. For the WRAT, the best set of predictors did not include any fixation time measures but rather skipping and regression rates. Models run with the average comprehension score across the three tests indicated reading speed (number of words read per minute) and late measures as the best predictors of comprehension. In all cases, eye movements explained substantial amounts of variance over and above reading speed alone. Full models for the comprehension tests explained an average of 39% of the variance in comprehension scores (YARC: 29%; GORT: 42%; WRAT: 46%).

The results from these analyses are in line with previous studies showing that reading comprehension tests do not measure the same cognitive skills to the same extent (Keenan, Betjemann & Olson, 2008; Keenan & Meenan, 2014). Results from both sets of analyses shed light on the complexity of the relationship between eye movements and reading comprehension – eye movements can predict comprehending scores, however, the best predictors and their predictive ability are modulated by the task demands. These results have important practical implications for the use of reading comprehension tests in research and clinical settings, as well as theoretical implications about the relationship between eye movements and reading comprehension.

Table 1: Correlations between comprehension scores and eye movements

Measure	YARC	GORT	WRAT	Average
Global				
Speed	0.23*	0.30*	0.57*	0.48*
Av. Fix. Dur.	-0.17	-0.11	-0.28*	-0.22 ^{p=0.058}
Saccade Length	0.15	0.41*	0.26*	0.32*
First-Pass				
Skipping	-0.03	0.09	0.16	0.07
First-Fix. Dur.	-0.19	-0.07	-0.26*	-0.21 ^{p=0.06}
Gaze Dur.	-0.26*	-0.35*	-0.29*	-0.33*
Late				
Regression	-0.02	0.12	-0.09	-0.03
Go-Past	-0.09	-0.30	-0.43*	-0.34*
Total Time	-0.17	-0.36*	-0.50*	-0.41*

Note: This table shows the correlation coefficients between eye-movement measures and comprehension scores for each test. * = $p < 0.05$

Table 2: Outputs of the ‘Best’ and Full Models

Predictors	YARC		GORT		WRAT		Average	
	Best Model	Full Model	Best Model	Full Model	Best Model	Full Model	Best Model	Full Model
Intercept	90.61	90.58	90.96	90.93	105.75	105.74	95.60	95.61
Speed (wpm)	5.95	9.29	-5.49	-5.32	11.84	10.32	7.84	7.47
Av. Fix. Dur.		8.55	-12.94	-11.47		-3.17		-6.86
Saccade Length		-4.29	4.64	5.31		0.87		1.98
Skipping	-4.82	-3.30		-1.64	-4.66	-4.89	-3.94	-5.68
First-Fix. Dur.	7.30	1.55	15.21	14.85	1.82	4.95		8.64
Gaze Dur.	-13.15	-19.09		-2.73		0.67		-3.30
Regression		-0.05		-0.97	4.13	4.20		-0.60
Go-Past	9.34	7.66		0.14		-0.81	9.81	11.09
Total Time		6.10	-8.28	-6.24		-1.63	-6.90	-6.61

Note: This table shows the estimated coefficients of the Bayesian linear models for the three comprehension tests and the average of the three test scores. For each, the output of the “best” model according for the leave-one-out cross-validation and the output of the full model are presented. Green cells indicate the 95% credibility interval does not include zero, yellow cells indicate the 90% credibility interval does not include zero, blank cells indicate the 90% credibility interval includes zero.

References: Snowling, M.J, et al. (2009). YARC. GL Publishers • Wiederholt, J. L., & Bryant, B. R. (2012) GORT5. Pro-Ed. • Wilkinson, G. S., & Robertson, G. J., (2006). WRAT4. Pearson. • Vehtari, A., Gelman, A., & Gabry, J. (2017). Statistics and computing, 27(5), 1413-1432. • Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281-300. • Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of learning disabilities*, 47(2), 125-135.

Reading Minds, Reading Stories: Social-Cognitive Abilities are Related to Linguistic Processing of Narrative Viewpoint

Lynn S. Eekhof, Kobie van Krieken, José Sanders, Roel M. Willems (Centre for Language Studies, Radboud University, The Netherlands)

Introduction. Narratives have a unique ability to disclose the inner worlds of others, leading various scholars to argue for a relationship between exposure to narratives and social-cognitive abilities such as empathy and theory of mind (e.g., Hakemulder, 2000; Mar & Oatley, 2008; Zunshine, 2006). One of the ways in which narratives can represent the inner worlds of characters is through the use of viewpoint markers, i.e., lexical elements that signal that a part of the narrative has to be constructed from the subjective viewpoint of a character (van Krieken et al., 2017). In this study, we investigated the link between narratives and social cognition by studying how the linguistic processing of viewpoint is related to social-cognitive abilities.

Method. Eye-tracking data was collected from 90 participants reading a Dutch 5000-word narrative that was scored for the presence of lexical markers of perceptual (PVP; e.g., *to listen*, *unrecognizable*; 93 words), cognitive (CVP; e.g., *to want*, *sceptic*; 148 words), and emotional viewpoint (EVP; e.g., *to feel*, *desperation*; 59 words) using a validated identification procedure (Eekhof et al., 2020; $\kappa = .82$). In addition, various social-cognitive measurements were collected, including the Interpersonal Reactivity Index (IRI; Davis, 1983; $\alpha = .83$), a Visual Perspective-Taking Task (VPT; Samson et al., 2010), and the Spontaneous Theory of Mind Protocol (STOMP; Rice & Redcay, 2015), a measure of the spontaneous tendency to mentalize.

Results. We used (generalized) linear mixed models to study the effect of PVP, CVP, and EVP markers on gaze duration, skip rate, and rereading rate using non-viewpoint marking content words as a baseline, and controlling for word length, frequency, and print exposure (Author Recognition Test; Stanovich & West, 1989). We found that PVP markers were read faster ($\beta = -3.54$ ms, $p = 0.01$), whereas markers of EVP ($\beta = 4.83$ ms, $p = 0.002$) and CVP ($\beta = 4.87$ ms, $p < .001$) were read slower compared to non-viewpoint markers. Furthermore, the odds of skipping were decreased by both CVP (by 0.71 times, $p < .001$) and EVP markers (by 0.88 times, $p < .001$). Finally, EVP markers increased the odds of rereading by 1.16 times ($p < .001$). Crucially, the effect of viewpoint markers on skip rate and rereading rate was found to interact with individual differences in social-cognitive abilities. IRI scores increased the odds of skipping PVP markers by 1.13 times ($p < .001$). Altercentric intrusion scores (i.e., altercentric interference during egocentric perspective-taking) on the VPT decreased the odds of skipping CVP markers by 0.95 times ($p = .02$). Finally, egocentric intrusion (i.e., egocentric interference during altercentric perspective-taking) increased the odds of rereading CVP markers by 1.08 times ($p = .01$).

Conclusion. We found diverging patterns of reading behavior for perceptual viewpoint markers on the one hand, and emotional and cognitive viewpoint markers on the other, suggesting that the processing of emotional and cognitive viewpoint is possibly more effortful (see also Mak & Willems, 2018), whereas the processing of perceptual viewpoint is rather fast. Interestingly, these findings align with developmental literature showing that perception verbs are generally acquired before cognitive verbs (e.g., E. E. Davis & Landau, 2020). Moreover, our findings reveal an interesting interplay between linguistic and social-cognitive processing and suggest that readers with relatively poor social-cognitive abilities are also slower to process linguistic elements related to the emotional, cognitive, and perceptual viewpoint of fictional others (as evidenced by decreased skipping and increased rereading); perhaps because these readers need to rely more on these explicit markers to make sense of the inner world of story characters. Although more research is needed to shed light on the causal mechanisms behind this relationship, this study underlines the promising role of narrative viewpoint techniques in the study of the social-cognitive potential of narratives.

References

- Davis, E. E., & Landau, B. (2020). Seeing and Believing: The Relationship between Perception and Mental Verbs in Acquisition. *Language Learning and Development*, 1–21. <https://doi.org/10.1080/15475441.2020.1862660>
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126. <https://doi.org/10.1037/0022-3514.44.1.113>
- Eekhof, L. S., Van Krieken, K., & Sanders, J. (2020). VIP: A Lexical Identification Procedure for Perceptual, Cognitive, and Emotional Viewpoint in Narrative Discourse. *Open Library of Humanities*, 6(1), 18. <https://doi.org/10.16995/olh.483>
- Hakemulder, J. (2000). *The Moral Laboratory: Experiments Examining the Effects of Reading Literature on Social Perception and Moral Self-concept*. John Benjamins Publishing.
- Juhász, B. J., & Rayner, K. (2003). Investigating the Effects of a Set of Intercorrelated Variables on Eye Fixation Durations in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1312–1318. <https://doi.org/10.1037/0278-7393.29.6.1312>
- Mak, M., & Willems, R. M. (2018). Mental simulation during literary reading: Individual differences revealed with eye-tracking. *Language, Cognition and Neuroscience*, 34(4), 511–535. <https://doi.org/10.1080/23273798.2018.1552007>
- Mar, R. A., & Oatley, K. (2008). The Function of Fiction is the Abstraction and Simulation of Social Experience. *Perspectives on Psychological Science*, 3(3), 173–192. <https://doi.org/10.1111/j.1745-6924.2008.00073.x>
- Rice, K., & Redcay, E. (2015). Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social Cognitive and Affective Neuroscience*, 10(3), 327–334. <https://doi.org/10.1093/scan/nsu081>
- Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5), 1255–1266. <https://doi.org/10.1037/a0018729>
- Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly*, 24(4), 402–433. <https://doi.org/10.2307/747605>
- van Krieken, K., Hoeken, H., & Sanders, J. (2017). Evoking and Measuring Identification with Narrative Characters – A Linguistic Cues Framework. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01190>
- Zunshine, L. (2006). *Why We Read Fiction: Theory of Mind and the Novel*. Ohio State University Press.

The interaction between subject-verb agreement and register-situation formality congruence in German sentence processing: an eye-tracking-reading pilot study

Camilo R. Ronderos^{1,2}, Aine Ito¹, Katja Maquate¹, Pia Knoeferle^{1,3,4}

¹Humboldt-Universität zu Berlin, ²University of Oslo, ³Berlin School of Mind and Brain, ⁴Einstein Center for Neurosciences Berlin

c.r.ronderos@ifikk.uio.no

The present research assesses the (representational and procedural) similarities between comprehenders' processing of standard-language grammar and their processing of register (situation-dependent language variation). Can we parsimoniously assume a single mechanism and closely-linked mental representations or must we model standard-language and register processing via distinct mental representations and / or mechanisms? Incongruence between language input and our knowledge of morphosyntax elicits rapid brain and eye-gaze responses during sentence comprehension (Pearlmutter et al., 1999; Tannen et al., 2013). Likewise, we know that incongruence in social aspects of meaning elicits rapid effects on language processing (e.g., van Berkum et al., 2009). Investigating these two kinds of incongruence can help refine extant models of sentence processing (e.g., Altmann & Kamide, 2009; Münster & Knoeferle, 2018; Venhuizen et al., 2018) also considering research on modelling social meaning (e.g. Burnett, 2019).

The present eye-tracking reading pilot study compares the processing cost of encountering morpho-syntactic (in)congruence and situation-(in)appropriate register. Participants (N=16, 40 critical items, 56 fillers) read two-sentence stories in German (context and target sentence, see example 1). They read each sentence one by one and answered comprehension questions on 1/3 of trials (fillers only). We crossed two independent factors: register-situation formality congruence (congruous vs. incongruous) with subject-verb agreement congruence (congruous vs. incongruous with linguistic knowledge). To establish register-situation-formality congruence, we paired verbs with a formal (*bereden*) and an informal (*belabern*) variant of the verb 'talk' either with a register- / formality-matching or mismatching context sentence (see 1). The register congruence factor was counterbalanced for formality. The subject-verb agreement congruence factor was established by varying grammaticality of subject-verb agreement (grammatical: 3rd person singular: ... *beredet* 1.a. vs. ungrammatical: the infinitive: ... **bereden*, 1.b).

We expected to replicate longer reading times during the verb or subsequent noun region (spillover) for morpho-syntactic incongruence (vs. congruence, Pearlmutter et al., 1999). Observing rapid interaction of morphosyntactic congruence with register-congruence would support accounts of one conceptual store and set of mechanisms. Delayed or no interaction of these two stimuli aspects would by contrast suggest the implicated mechanisms are distinct (eliciting delays and / or more subtle processing effects). We fitted linear mixed-effects models (sum contrast coded) to the log-transformed first-pass, regression path, and total reading times of the verb and spill-over regions, as well as to the total target sentence reading times. The results replicated longer reading times for sentences with subject-verb agreement mismatches than matches (all measures on verb region, regression path duration of the spill-over region: $t=2.7$, $p < .01$; total sentence reading times: $t=2.5$, $p < .05$; Bonferroni-corrected, von der Malsburg & Angele, 2017). No significant main effect or interaction involving register congruence emerged (see Figures 1, 2 and 3).

The results show clear subject-verb agreement effects, an absence of any register congruence effects and no interaction of these two factors. It is possible that overt subject-verb agreement incongruence overshadowed any subtle situation-dependent register incongruence effects that might have otherwise been observed. It is also possible that the implementation of the register incongruence was not strong enough. Follow-up research will omit incongruence in subject-verb agreement and strengthen the implementation of register-context congruence, giving us a more sensitive paradigm for investigating the processing of social meaning.

1. Example critical item

Formal context

Die Empfangsdame verkündigte der vornehmen Gesellschaft:

Informal context

Die Groupies posteten auf der Fanpage:

Target sentence versions:

formal, grammatical

a. Der Rocksänger /beredete VERB/ den Schlagzeuger SPILLOVER/.

formal, ungrammatical

b. Der Rocksänger /bereden VERB/ den Schlagzeuger. SPILLOVER /.

informal, grammatical

c. Der Rocksänger/ belaberte VERB/den Schlagzeuger. SPILLOVER/.

informal, ungrammatical

d. Der Rocksänger/ belabern VERB/den Schlagzeuger. SPILLOVER/.

Figure legend: 'grammatical' refers to sentences with subject-verb agreement match; 'ungrammatical' refers to sentences with subject-verb agreement mismatch

Figure 1

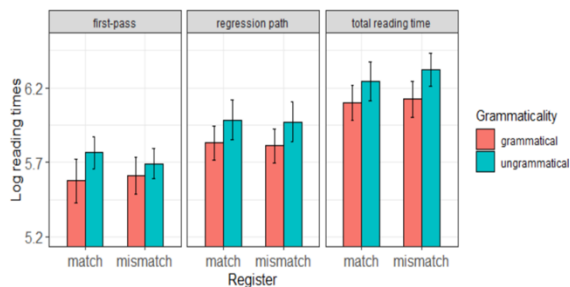


Figure 2

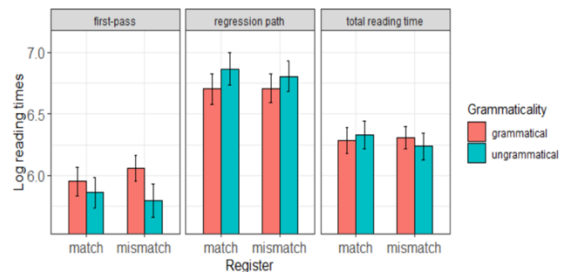
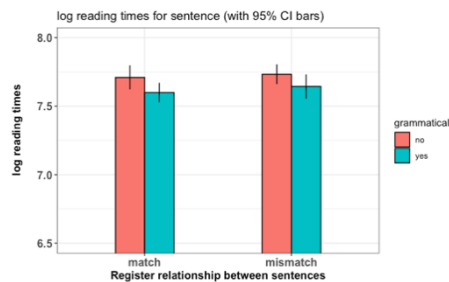


Figure 3



References

- Burnett, H. (2019). Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 42(5), 419-450.
- Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and language*, 41(3), 427-456.
- Tannen, D. & van Hell, J. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia*, 56, 289-301.
- Van Berkum, J., Van den Brink, D., Tesink, C., Kos, M., and Hagoort, P. (2008). The neural integration of speaker and message. *J. Cogn. Neurosci.* 20, 580-591.
- Venhuizen, N., Crocker, M.W., Brouwer, H. (2018). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, 56, 229-255.
- Von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of memory and language*, 94, 119-133.

Event Completion, Not Ongoingness, Is Language Dependent: Crosslinguistic Evidence from ERPs in English and Russian

Anna Katikhina^{1,2} & Vicky Tzuyin Lai^{1,3} (¹Cognitive Science, ²Second Language Acquisition & Teaching, ³ Department of Psychology, University of Arizona)
akatikhina@email.arizona.edu, tzuyinlai@email.arizona.edu

Verb aspect is a lexico-grammatical feature that defines the temporal distribution of an event. According to English and Russian linguistic theories, English past simple (perfective: *washed*) is associated with completion, but is not morphologically marked for aspect and can refer to both completed and in-progress events. Aspectually marked past progressive (imperfective: *was washing*) is restricted to unfolding events. In contrast, Russian marks aspect obligatorily. Perfective carries a completed connotation, while imperfective, although associated with ongoingness, can be used as a general past reference. Extant literature in English suggests that aspect serves to build a mental model of an event [1]. Perfective emphasizes event completion within a temporal boundary [2, 3], while imperfective presents an event in progress, providing richer details [4], but no specific event stage. To date, little examined effects of aspect on the mental representations of events in non-English languages.

In this ERP study we examined (1) whether differences in aspect usage influence the mental representations of event stage (completed, in-progress); (2) whether aspect processing is semantic or morphosyntactic in languages with different degrees of aspect marking obligatoriness. Our hypotheses and predictions are that (1) Russian perfective and English imperfective will result in specific mental representations of event stage (completed for Russian; in-progress for English); (2) Obligatoriness of aspectual marking determines whether semantic (N400) or morphosyntactic (P600) mechanism is engaged.

Participants were native speakers of English (N=19) and Russian (N=19). The design was 2 Event (In-progress, Completed) x 2 Aspect (Perfective, Imperfective) (Table 1). The stimuli were 256 pictures and descriptions, presented in 4 blocks. In the two experimental blocks, the events in the pictures and the verb stems in the descriptions matched semantically. In the perfective block, all verbs were perfective. Half were preceded by completed events (congruent), and half, in-progress events (incongruent). Likewise, in the imperfective block, all verbs were imperfective, and were preceded by completed (incongruent) and in-progress events (congruent). In the two control blocks, the events and the verb stems did not match semantically in incongruent trials, leading to an outright semantic violation. The order of blocks was counterbalanced with subjects. In each trial, a picture was presented for 500 ms, followed by a description, word-by-word. Comprehension questions appeared after each trial.

In English, only perfective verbs preceded by semantically-matched in-progress events elicited a sustained negativity starting at 300 ms, which reached statistical significance 500–900 ms, compared to perfective verbs preceded by semantically matched completed events ($p=0.002$), anteriorly (Fig.1a,c). This suggests recomputation of a mental model to integrate information about a previously held assumption regarding event completion [5]. **In Russian**, only perfective violations resulted in a wide-spread enhanced positivity that reached significance in the 600–900 ms time window ($p=0.015$) (Fig.1b,d), suggesting morphosyntactic mechanisms and consistent with previously reported morphosyntactic P600 effect for perfective violations in Slavic languages [6]. Being more semantically specific and less flexible in aspectual meaning interpretations, Russian perfective likely elicited greater attention to its grammatical features. In control blocks, semantic violation at lexical verbs in both perfective and imperfective blocks elicited N400 effects in both groups (Fig.2).

In conclusion, imperfective in both English and Russian was not associated with a specific event stage, consistent with previous literature [1,3]. Obligatory aspect marking engages morphosyntactic processing, i.e. specific verb morphology is associated with event stage. Less obligatory marking likely engages semantic processing, with the match between verb form and event stage processed more holistically, as a function of verb semantics. We found crosslinguistic similarities in the association between aspect and mental representations of event stage, but the processes supporting this association differed based on language-specific aspectual system.

Table 1. Design and Examples of Stimuli. Asterisk (*) indicates violation.



Picture	Condition	Sentence
 (in progress)	Exp (aspect) Ctrl (semantic)	She *cleaned / was cleaning the glasses. She was *licking / cleaning the glasses.
 (completed)	Exp (aspect) Ctrl (semantic)	She *was shredding / shredded the cabbage. She *ate / shredded the cabbage.

Figure 1. Aspect Violations. ERPs for the aspect violation conditions (red) and the aspect match conditions (black). 1a, 1b: Averaged waveforms of 6 anterior electrodes (AF3, AF4, AFz, F1, F2, Fz). 1c, 1d: Averaged waveforms of 9 central electrodes (FC1, FC2, FCz, C1, C2, Cz, CP1, CP2, CPz). The topographies are based on the difference waves between the two conditions.

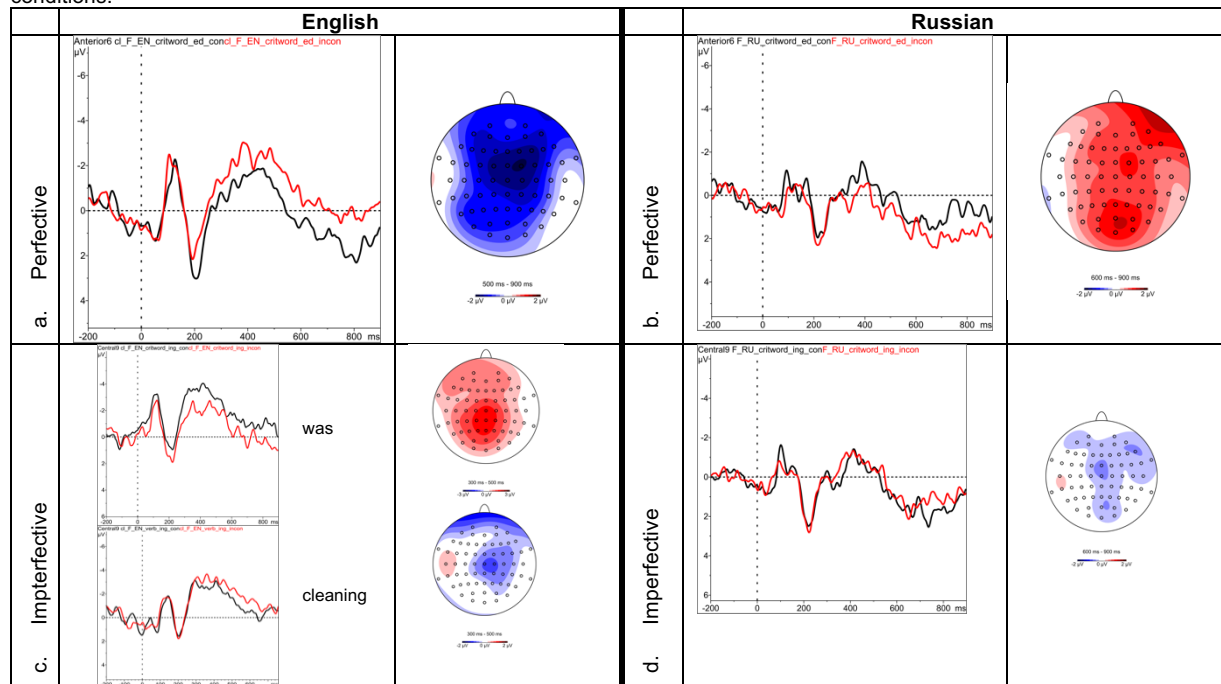
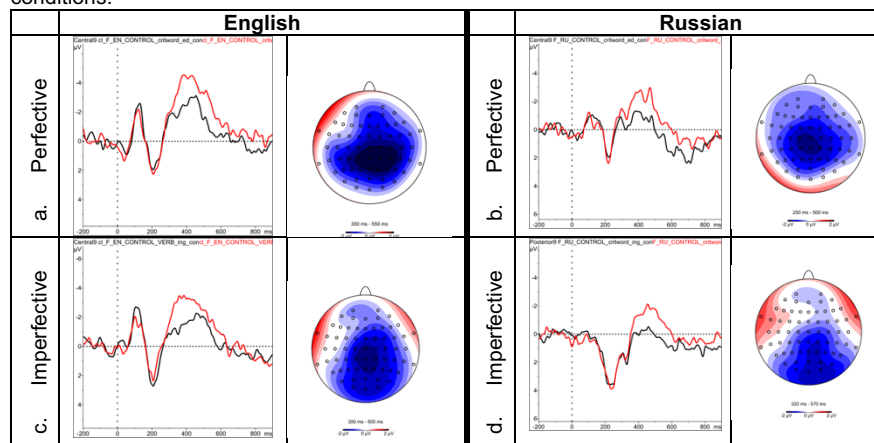


Figure 2. Semantic Violations. ERPs for semantic violation conditions (red) and semantic match conditions (black). 2a, 2b, 2c: Averaged waveforms of 9 central electrodes (FC1, FC2, FCz, C1, C2, Cz, CP1, CP2, CPz). 2d: averaged waveforms of 9 posterior electrodes (P1, P2, Pz, PO3, PO4, POz, O1, O2, Oz). The topographies are based on the difference waves between the two conditions.



References

1. Madden, C. J., & Zwaan, R. A. *Memory & cognition*, 2003.
2. Ferretti, T. R., Rohde, H., Kehler, A., & Crutchley, M. *Journal of memory and language*, 2009
3. Misersky, J., Slivac, K., Hagoort, P., & Flecken, M. *Cognition, in press*.
4. Ferretti, T. R., Kutas, M., & McRae, K. *Journal of Experimental Psychology: Learning, memory, and cognition*, 2007.
5. Baggio, G., Van Lambalgen, M., & Hagoort, P. *Journal of Memory and Language*, 2008.
6. Błaszczak, J., Jabłońska, P., & Klimek-Jankowska, D. *The Processing of Lexicon and Morphosyntax*, 2014.

Case marking influences the apprehension of briefly exposed events

Arrate Isasi-Isasmendi (U. of Zurich), Caroline Andrews (U. of Zurich), Sebastian Sauppe (U. of Zurich), Monique Flecken (U. of Amsterdam), Moritz Daum (U. of Zurich), Itziar Laka (U. of the Basque Country), Martin Meyer (U. of Zurich) and Balthasar Bickel (U. of Zurich)

When preparing to describe an event depicted in a picture, speakers need to apprehend its gist, including event roles (the “who does what to whom”), rapidly—sometimes in as little as 100–300 ms [1]. Event apprehension has been argued to be a prelinguistic process [2], i.e., grammar should play no role (yet) in speakers’ gist extraction, but only later should impact the message and linguistic encoding. Here we present two experiments using the brief exposure paradigm [3,4,5] that test whether case differences in Basque and Spanish affect not only linguistic encoding but can impact event apprehension. The two languages differ in their case marking systems: In Basque, agentive subjects are marked by ergative case (-k), while patients (subjects of unaccusative intransitive verbs and objects of transitive verbs) receive absolutive case. In Spanish, by contrast, subjects always carry the same unmarked nominative case regardless of their thematic role (cf. Figures 1–3). This may require Basque speakers to commit to a level of subject agentivity (ergative or absolutive) early during planning [6], which may in turn afford heightened attention to agents in event apprehension, as compared to Spanish.

In our experiments, participants saw photographs of events with four different actors (e.g., a man watering a plant) for 300 ms in a randomly assigned corner of the screen. As planning and executing saccades requires up to 200 ms [7], this left only approximately 100 ms to take up visual information foveally after a gaze shift from a central fixation cross into the picture. Following this brief exposure, participants either produced a sentence description (Event Description task) or determined whether a subsequent picture matched a participant from the primary picture (Probe Recognition task). In Exp. 1 (online, without eye tracking), native speakers of Basque ($N=90$) and Spanish ($N=88$) completed a block of 58 trials per task and typed sentences after each picture in the event description task. In Exp. 2 (in-lab, with eye tracking), native Basque and Spanish speakers ($N=32$ each) received two blocks per task and described the pictures orally. In Exp. 2 we tracked the location of fixations on the briefly presented pictures. We analyzed first and second fixations to event pictures with Bayesian hierarchical binomial regression [8] to test whether the marked vs. unmarked status of agents in Basque and Spanish was reflected in increased looks towards agent areas of interest in Basque speakers as compared to Spanish speakers. Accuracy in Exp. 1 and Exp. 2 was analyzed with Bayesian hierarchical ordinal models.

Results: In probe recognition, Basque speakers were more accurate in recognizing agents than Spanish speakers in both experiments (mean log odds: Exp. 1, $\hat{\beta} = 0.06$, $P(\hat{\beta} > 0) = 0.81$; Exp. 2, $\hat{\beta} = 0.13$, $P(\hat{\beta} > 0) = 0.93$). In event description, Basque speakers were more accurate describing agents than Spanish speakers ($\hat{\beta} = 0.15$, $P(\hat{\beta} > 0) = 0.98$) in Exp. 2. A different pattern was found in Exp. 1: Basque speakers were more accurate describing patients ($\hat{\beta} = -0.06$, $P(\hat{\beta} > 0) = 0.87$), possibly due to the modality difference in this task across experiments (i.e., written vs spoken descriptions). Analyses of first and second fixations revealed that Basque speakers fixated more often on agents than Spanish speakers ($\hat{\beta} = 0.08$, $P(\hat{\beta} > 0) = 0.95$), while Spanish speakers fixated more often on patients than Basque speakers ($\hat{\beta} = -0.09$, $P(\hat{\beta} > 0) = 0.98$). In addition, in event description these effects were stronger than in probe recognition (Agent, $\hat{\beta} = 0.04$, $P(\hat{\beta} > 0) = 0.91$; Patient, $\hat{\beta} = -0.05$, $P(\hat{\beta} > 0) = 0.93$).

Our results suggest that the grammatical features of a language shape not only structural and linguistic encoding [9] but can also affect event apprehension [2,10]. In particular, these results show that language-related task requirements can influence attention to agents during event apprehension. These findings suggest the possibility of an interaction between language and event cognition [11].

Example event pictures and sentences in Basque (B) and Spanish (S)



Fig. 1. Transitive human-human event:
(1) B: Lisa-k Emma orraztu du
Lisa-ERG Emma brushed has
(2) S: Lisa-ø ha peinado a Emma
Lisa-NOM has brushed DOM Emma
"Lisa has brushed Emma"



Fig. 2. Transitive human-inanim. event:
(1) B: Tim-ek bizikleta konpondu du
Tim-ERG bike fixed has
(2) S: Tim-ø ha arreglado la bici
Tim-NOM has fixed the bike
"Tim has fixed the bike"



Fig. 3. Intransitive event:
(1) B: Paul-ø makurtu da
Paul-ABS crouched has
(2) S: Paul-ø se ha agachado
Paul-NOM REFL has crouched
"Paul has crouched"

Fixations to event roles by language across tasks

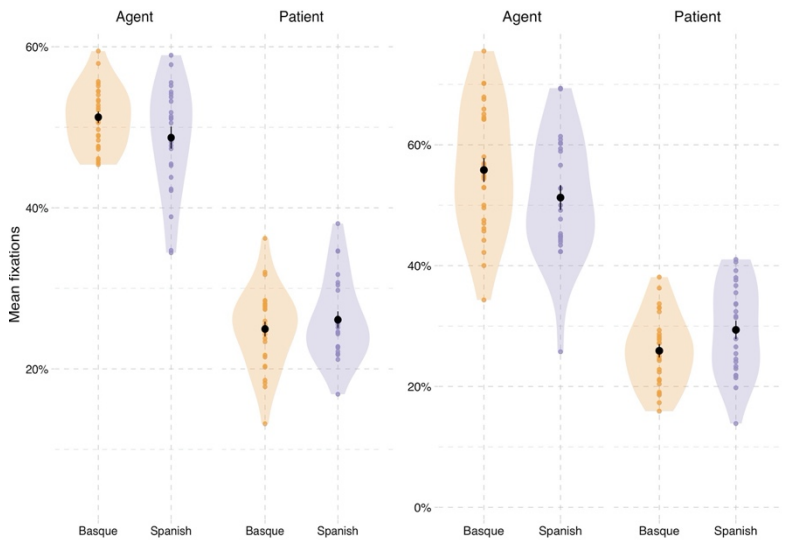


Fig 4. Proportion of first and second fixations to agents and patients in the briefly exposed event pictures, by language. Black dots represent participant means.

Language	Event Role	Fix Mean	SE
Basque	Agent	51.2	0.7
Spanish	Agent	48.7	1.3
Basque	Patient	24.9	0.9
Spanish	Patient	26.0	1

Table 1. Mean proportions of *first* fixations by language and event role.

Language	Event Role	Fix Mean	SE
Basque	Agent	55.8	1.9
Spanish	Agent	51.2	2
Basque	Patient	25.9	1.1
Spanish	Patient	29.3	1.5

Table 2. Mean proportions of *second* fixations by language and event role.

References: [1] Dobel et al., 2007. *Acta Psychol.* [2] Griffin & Bock, 2000. *Psychol Sci.* [3] Gerwien & Flecken, 2016. *Proceed Cog Sci.* [4] Sauppe & Flecken, 2021. *Cognition.* [5] Hafri et al., 2018. *Cognition.* [6] Sauppe et al., 2021, *PLOS Biology.* [7] Pierce et al., 2019. *Eye movement research.* [8] Burkner, 2017. *J Stats Soft.* [9] Norcliffe et al., 2015. *Lang Cog Neurosci.* [10] Gleitman et al., 2007. *JML.* [11] Lupyan et al., 2020. *Trends Cogn Sci.*

Conceptualization and formulation of motion event sentences in L2.

Matias Morales, Martin Pickering, Holly Branigan (University of Edinburgh).

Speakers encode message-level representations based on the preferences of the language they speak (Levelt, 1996; Slobin, 1996). This implies that bilinguals may use different encoding strategies while planning their utterances depending on whether they use their L1 or L2. We address this issue by looking at the way bilinguals linguistically encode conceptual information that is strongly preferred in their L2, but not in their L1. Thus, we (1) investigate how bilinguals select and distribute this information in L2 (exp.1), and 2) examine the level of representation for this information in bilinguals and whether this affects their L2 sentence formulation (exp.2-4). We use the well-studied cross-linguistic variation for motion events that indicates speakers of different languages show different preferences in the lexicalization of manner and path information (Talmy, 2000). For example, for the event (A) 'a penguin skiing into an igloo', English speakers typically encode manner in the main verb (e.g., *A penguin is skiing into an igloo*), while Spanish speakers usually encode path (e.g., *Un pingüino está entrando en un iglú*, 'A penguin is entering an igloo'). Therefore, we focus on two analyses: (1) the probability participants use a manner verb in their motion event descriptions, and (2) the probability they use manner-dominant descriptions (for examples, see Table 1).

In Exp. 1, monolingual L1 English speakers (N=24, *L1 English*) and late proficient L1 Spanish-L2 English bilinguals tested in their L2 (N=24, *L2 English*) freely described animations depicting boundary-crossing motion events like (A) (see Fig.1), and their utterances were compared to Spanish descriptions by late proficient L1 Spanish-L2 English bilinguals tested in their L1 (N=24, *L1 Spanish*). L1 English and L2 English speakers were more likely to use a manner verb compared to L1 Spanish speakers (92% vs. 33%; $p < .001$ and 57% vs. 33%; $p < .05$) (see Fig. 2A). Additionally, L1 English speakers were more likely to produce manner-dominant sentences than L1 Spanish speakers (95% vs. 70%, $p < .001$), but L2 English participants did not differ from L1 Spanish participants in this respect (see Fig. 2B). These results indicate that L2 speakers used the lexical preferences, but not the structural choices of their L2 for motion events.

In Exp. 2, L1 English (N=48) and L2 English (N=48) speakers described the same set of animations after reading aloud a prime sentence that described an unrelated event either with a manner or a path interpretation (e.g., *The man is skiing skilfully* vs. *The nurse is entering quietly*, see Table 2). Crucially, prime sentences contained a lexical overlap with the target (i.e. the verb was repeated across prime/target). L2 speakers were more likely to use manner verbs and manner-dominant descriptions after manner primes vs. path primes (70% vs. 50%; $p < .001$ and 79% vs. 62%, $p < .001$). L1 speakers did not show either of these differences.

Exp. 3 was a version of exp. 2 with the critical difference that prime sentences contained a conceptual overlap with the target event (i.e. the verb was not repeated across prime and target) (e.g., *The girl is crawling happily* vs. *The boy is circling senselessly*). Results indicate that neither L1 nor L2 speakers were more likely to use manner verbs after manner vs. path primes.

Exp. 4 was another version of exp.2 with only L2 speakers (N=72). Critically, we added a baseline condition with sentences that described non-motion events (e.g. *The pirate is whispering loudly*) to test whether the effect found in exp.2 was due to a lexical effect (i.e. participants just repeated the prime verb) or a conceptual priming effect. Results indicate that L2 speakers were more likely to use manner verbs after manner primes vs. baseline (70% vs. 55%, $p < .001$), but not after a path prime vs. baseline (56% vs. 55%, $p > .05$). Additionally, they were more likely to produce more manner-dominant responses after manner primes vs. baseline (85% vs. 77%, $p < .05$), but not after path primes vs. baseline (79% vs. 77%, $p > .05$).

Overall, results show that L2 speakers were primed by the manner information contained in the verb of prime sentences, and that the locus of these representations was lexical and not conceptual. In addition, this lexical priming affected the formulation of L2 sentences in bilinguals in ways that do not reflect their L1 preferences, suggesting that the encoding strategies in L2 were language-specific.



Figure 1. Example of a target motion animation representing the event *A penguin skiing into an igloo* used in all experiments: (1) start, (2) middle , and (3) end of the video.

Target utterances	Analyses
<i>A penguin is skiing.</i>	manner verb and manner-dominant.
<i>A penguin is skiing <u>into</u> an igloo.</i>	manner verb and manner-dominant.
<i>A penguin <u>on skis/skiing</u> is entering an igloo.</i>	manner-dominant.

Table 1. Examples of target responses that entered in the analyses of manner verb use and manner-dominant utterances (i.e. responses that included manner content only or manner preceding path information). Manner content is in bold, while path information is underlined in target utterances.

Target Event: <i>A penguin skiing into an igloo.</i>	Manner Prime	Path Prime
Experiment 2: Lexical overlap	The man is skiing skillfully.	The nurse is entering quietly.
Experiment 3: Conceptual overlap	The girl is crawling happily.	The boy is circling senselessly.

Table 2. Examples of prime sentences used in Experiments 2 and 3.

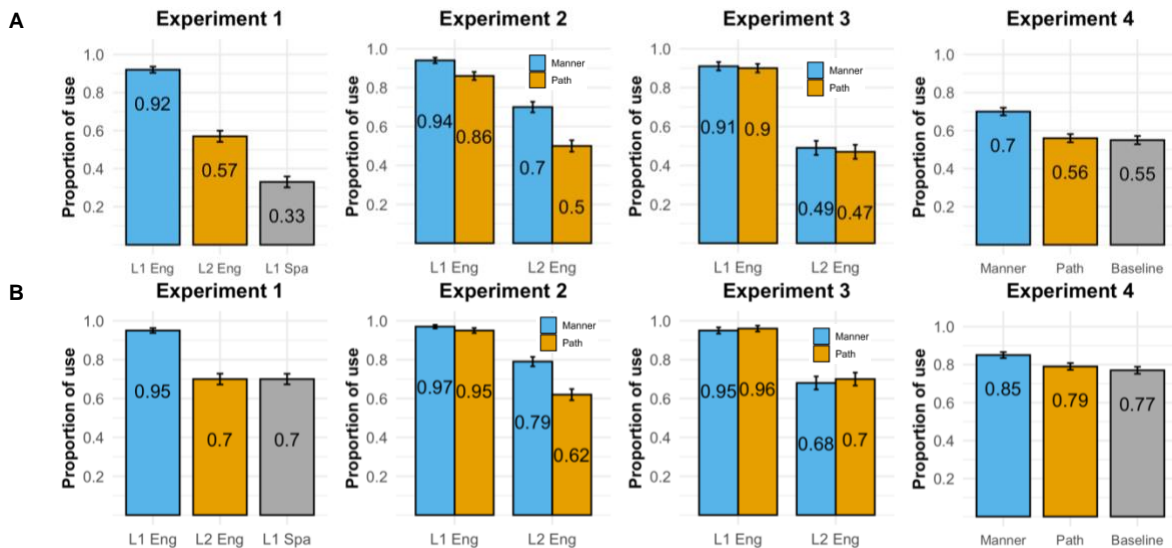


Figure 2. Mean proportion of manner verbs (panel A) and manner-dominant responses (panel B) across all experiments.

References.

- Levelt, W. J. (1996). Perspective taking and ellipsis in spatial descriptions. In P. Bloom, Peterson, M.A., Nadel, L., Garrett, M.F. (Ed.), *Language and space* (pp. 77-107): MIT Press.
- Slobin, D. I. (1996). From “thought and language” to “thinkingfor speaking”. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking Linguistic Relativity* (pp. 70-96): Cambridge University Press.
- Talmy, L. (2000). *Toward a cognitive semantics: Typology and process in concept structuring* (Vol. 2): MIT Press.

Patterns of motion expression in children with or without a language disorder

Samantha N. Emerson (Boys Town National Research Hospital), Karla K. McGregor (Boys Town National Research Hospital), & Şeyda Özçalışkan (Georgia State University)

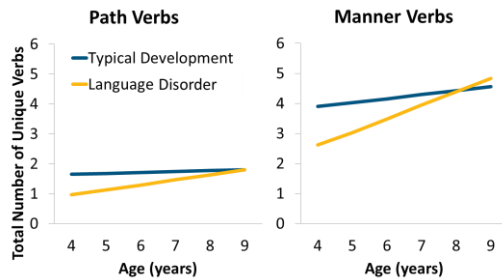
Children with language disorders (LD) have smaller vocabularies with shallower semantic knowledge [1] and more difficulty mastering some grammatical markings [2] than their typically developing (TD) peers. Recent evidence suggests that these deficits may stem from an impaired ability to extract distributional statistics from sequenced stimuli [3]. We asked whether children with LD are sensitive to the distributional patterns of expression in their native language. For example, while it is grammatical to say “He is walking” or “He enters the room,” English speakers tend to conflate the two motion components in a single utterance by saying “He walks into the room,” using one among a diverse array of manner verbs (“run”, “fly”, “crawl”) tightly packaged with a path particle/preposition (“up”, “across”, “to” [4]). Despite having more semantic elements than separated packaging—in which only manner or path is encoded—by the age of 3, English-speaking TD children produce a greater number of conflated motion utterances than children who are speakers of languages that typically use separated packaging with a smaller variety of manner verbs to describe similar motion events (French, Turkish; [5-6]). Such cross-linguistic findings suggest that the use of the conflated motion packaging does not simply reflect a developmental trajectory toward using more complex expressions but, instead, reflects a sensitivity to the distribution of semantic information (rate of production for conflated vs. separated constructions) in adult language. In this preregistered study, we asked whether children with LD attune their descriptions of motion events to language-specific patterns akin to TD children. We predicted that if children with LD were sensitive to distributions of motion information in English, they would show similar rates of expression as TD children or would show lower rates of use if they were not.

We examined narratives in the Edmonton Narrative Norms Instrument Database [7] produced by 4- to 9-year-old English-speaking children with LD ($n=77$; enrolled in services) and age- and gender-matched TD peers ($n=77$; teacher report). Children described six scenes in a picture book and the two groups produced narratives that were comparable in length. Each expression of a motion event was coded for verb vocabulary type (manner verbs, path verbs) and packaging type as conflated (manner and path in a single utterance) or separated (manner or path in separate utterances) following earlier work [6]; utterance grammaticality was not considered. Data were analyzed with mixed effects models. Results showed age \times diagnosis interactions for both verb and packaging: Children with LD produced a smaller variety of manner and path verbs and fewer conflated and separated descriptions than their TD peers, but only at the younger ages. Furthermore, our results showed a sex \times diagnosis interaction suggesting that, boys—but not girls—with LD were 2.5 times less likely to use conflated packaging for motion descriptions than TD children.

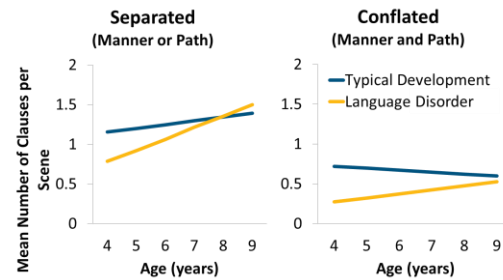
Our results showed both weaknesses and strengths in children with LD in attuning to language-specific patterns in their expression of motion. In line with past research [1], children with LD had smaller motion vocabularies than TD children but only at the earlier ages; however, as the vocabulary of the children with LD caught up to their TD peers, so did their rates of use for each of the motion packaging types. This indicates that children with LD were in fact sensitive to the distribution of motion expression types in English, a result that mirrors earlier findings suggesting that the use of certain verb types can drive the use of associated syntactic constructions [8]. At the same time, boys—but not girls—with LD were less likely to conflate motion when describing a motion event, consistent with previous findings showing a female advantage for language abilities in both individuals with LD or TD [1,9]. Overall, our findings demonstrate that children with LD attune their patterns of expression to the distributional properties of motion expression in their language—once they have acquired the prerequisite vocabulary. However, matching such distributions may be more challenging for boys than girls.

Figures

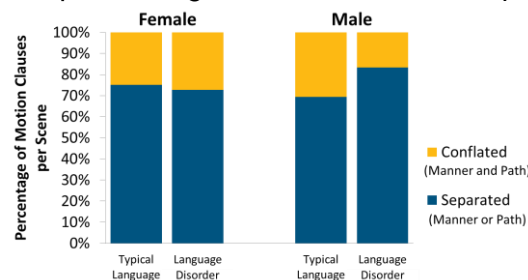
(1) Are there group differences in children's production of motion vocabulary?



(2) Are there group differences in children's production of motion packaging?



(3) Are there group and sex differences in the likelihood of a motion description being conflated in children's production?



References

- McGregor, K. K., Oleson, J., Bahnsen, A., & Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language & Communication Disorders*, 48(3), 307–319.
- Leonard, L. B., & Kueser, J. B. (2019). Five overarching factors central to grammatical learning and treatment in children with developmental language disorder. *International Journal of Language & Communication Disorders*, 54(3), 347–361.
- Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, 60(12), 3474–3486.
- Talmy, L. (2000). *Toward a cognitive semantics: Vol. II: Typology and process in concept structuring*. Cambridge, MA: MIT Press.
- Hickmann, M., Taranne, P., & Bonnet, P. (2009). Motion in first language acquisition: Manner and path in French and English child language. *Journal of Child Language*, 36(4), 705–741.
- Özçalışkan, Ş., & Goldin-Meadow, S. (2018). How early does speaking shape the native language of gesture? Paper presented at the 43rd Boston University Conference on Language Development. Boston, MA
- Schneider, P., Hayward, D., & Dubé, R. V. (2006). Storytelling from pictures using the Edmonton Narrative Norms Instrument. *Journal of Speech-Language Pathology and Audiology*, 30, 224–238.
- Owen Van Horne, A. J., & Lin, S. (2011). Cognitive state verbs and complement clauses in children with SLI and their typically developing peers. *Clinical Linguistics & Phonetics*, 25(10), 881–898.
- McGregor, K.K., Arbisi-Kelm, T., Eden, N., & Oleson (2020). The word learning profile of adults with Developmental Language Disorder. *Autism & Developmental Language Impairments*, 5, 1–19.

The role of prior discourse in the context of action: Insights from pronoun resolution

Tiana V. Simovic, Craig G. Chambers (University of Toronto)

The comprehension of a pronoun (*she*, *they*...) involves using linguistic and non-linguistic cues to select an intended candidate from entities in a comprehender's mental model of the discourse or situational context. These entities have often been previously mentioned, giving rise to the notion of a "linguistic antecedent". But what kind of information in a mental model is needed for resolving coreference? Given their status as deep anaphors [1], pronouns do not need to "match" linguistic antecedents with the same surface form (i.e., agreement or constituency: "I need a knife, where do you keep them?", "Jo ran into Sue while shopping. They..."), yet the notion of *retrieval processes* is evoked in many theoretical accounts [2, 3, 4, 5, 6]. Here, we explore the role of the antecedent term's *semantics* by using novel situations where the content of this expression is no longer viable when pronoun interpretation occurs. Fig. 1 shows a visual environment where objects are located within a grid with numbered squares. Critically, in this context, the outcome of an instruction like "Move the house on the left to area 12" entails that the unmoved/unmentioned house is now the leftmost one. If a subsequent instruction contains a pronoun (e.g., "Now move it..."), the key point is that the antecedent expression in memory no longer accurately describes the intended referent. Thus, if retrieving the antecedent term's semantics is a fundamental part of the process, some measurable processing cost should be observed relative to when the semantics are still valid, despite the intuition that the previously-mentioned object is ultimately the intended referent. **Expt 1** (production, $N=56$) was conducted to confirm certain background assumptions. After encountering the first instruction and viewing its outcome (Fig. 1, version *a/b*), speakers were prompted to describe various objects in the display. When prompted to describe the previously-moved object, results showed that, when speakers used a spatial description, the content reflected the updated visual scene (i.e., speakers did not treat the NP in the initial sentence as a "linguistic precedent" [7]). This tendency was stable regardless of whether the past action required a switch (e.g., from "on the left" to "on the right": 96% of descriptions reflecting updated scene) or not (97%). This behavior was largely expected but the findings validate the idea that the original description is no longer adequate following the action, and thus should cause difficulty if relied upon in some subsequent process. Results also showed modifiers like "on the left" are readily produced alongside other modifier types (10.25% overall), suggesting expressions of this type would be perfectly adequate as antecedents in a pronoun interpretation task. Our key evidence comes from **Expt 2** (Visual World, $N=24$), where participants also heard a second instruction (S2), and the earlier semantic viability manipulation was retained. In control conditions, S2 contained a full NP ("Now move the same/other house to area 4"). For fixations to the previously-moved house, the control conditions showed the expected unambiguously distinct patterns (Fig. 2). Critically, when S2 contained a pronoun ("Now move it to area 4"), mouse clicks on the intended referent showed no differences in reaction times, regardless of whether the antecedent term's semantics were still relevant or not (Fig. 3). Further, fixation patterns were strikingly similar for the two pronoun conditions (Fig. 2). Notably, there was no momentary consideration of the referent that now matched the antecedent term's semantics. The similarity across pronoun conditions was corroborated by analyses using bootstrapped group mean curves (Fig. 4), where strong overlap was still found. Together the data suggest a pronoun is effortlessly linked to an intended referent regardless of whether the semantics of its linguistic antecedent are still relevant. We then ask, if neither antecedent form nor semantics are relevant, what is "retrieved" on a retrieval account? Instead, real-world referents seem to be linked to mental variables via attentional bindings [8] that are indifferent to information in the linguistic record that can change or become irrelevant downstream [9]. Among other things, this helps explain cases where there is a shift in precisely what is being referred to in antecedent-pronoun sequences (A: "Speaking of pets, Ty got a capybara", B: "Huh? How do you spell it?", where the antecedent denotes a conceptual kind, yet the pronoun denotes an orthographic pattern).







1			4
5			8
9			12

Figure 1: Display before first sentence is heard.
 "Move the house on the left to...
 a. ...area 9" (original desc. remains viable)
 b. ...area 12" (original desc. no longer viable)
 (Display is updated accordingly)

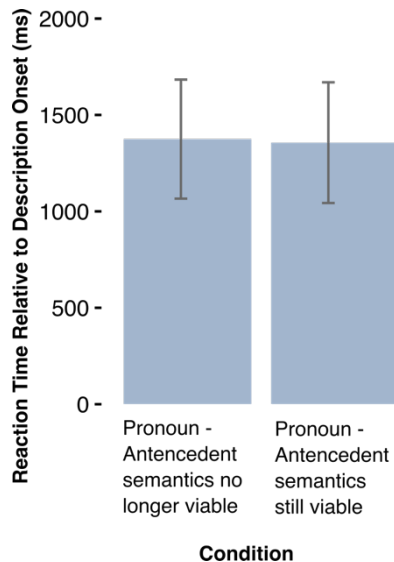


Figure 3: Mean reaction times for pronoun conditions in Expt 2.

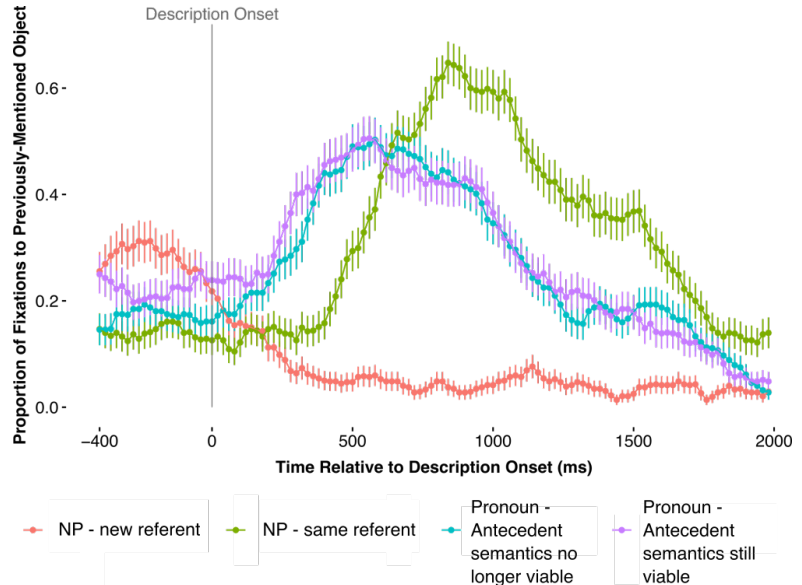


Figure 2: Proportion of fixations over time relative to pronoun onset (experiment conditions) or ADJ onset (controls) as indicated by grey line.

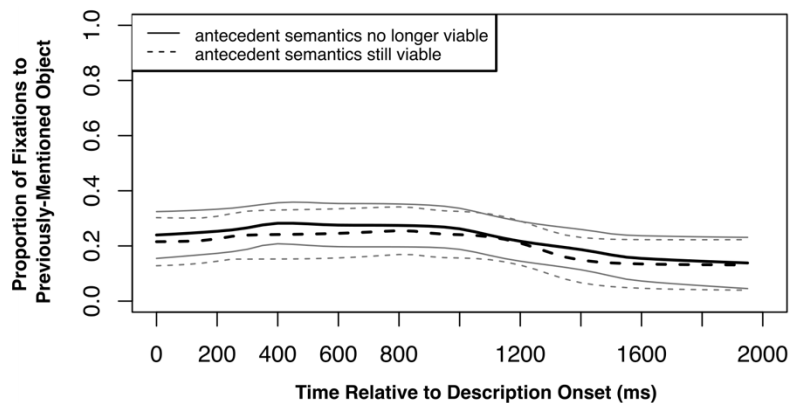


Figure 4: Difference between bootstrapped group mean fixations over time for pronoun conditions.

References

- Hankamer, J., & Sag, I. (1976). Deep and surface anaphora. *Linguistic Inquiry*, 7(3), 391-428.
- Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10, 447-454.
- Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2), 85-103.
- Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, 5(1252), 1-12.
- Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, 82, 18-40.
- Kush, D., Johns, C. L., & Van Dyke, J. A. (2019). Prominence-sensitive pronoun resolution: New evidence from the speed-accuracy tradeoff procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(7), 1234.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482.
- Landman, F. (1986). Pegs and Alecs. In *Theoretical Aspects of Reasoning About Knowledge* (pp. 45-61). Morgan Kaufmann.
- Webber, B., & Baldwin, B. (1992). Accommodating context change. In *30th Annual Meeting of the Association for Computational Linguistics* (pp. 96-103).

The social cost of maxims violation: Pragmatic behavior informs speaker evaluation

Andrea Beltrama and Anna Papafragou
University of Pennsylvania

Classic pragmatic theories treat communication as a cooperative enterprise ([1]), showing how listeners draw *pragmatic* inferences to compute a speaker's intended message. At the same time, work in sociolinguistics ([2-3]) and social psychology ([4]) has shown that interlocutors systematically draw *social* inferences from speech — i.e., they form impressions about the interlocutor's social or personal qualities: such inferences are usually independent of what the speaker intended to convey, and have thus mostly escaped the domain of pragmatics. Bridging pragmatic and social approaches to communication, we show that a speaker's choice to obey or violate the pragmatic maxims of Relevance and Informativeness — as well as the reasons behind these choices (Inability vs. Unwillingness) — affect how the speaker is perceived, revealing a connection between pragmatic cooperativeness and social evaluation.

EXP1. A 2x2 design was implemented in a conversation between two co-workers, Kim and John, in which John talked about a recent skiing vacation (see Table 1). In the Relevance manipulation, John either addressed Kim's dilemma, when she expressed interest in a skiing vacation (+Relevance); or failed to address it, when she expressed interest in a beach vacation (-Relevance). In the Informativeness manipulation, John either provided a detailed description of his vacation (+Informativeness), or simply disclosed its location (-Informativeness). Before his description, John claimed familiarity with all places mentioned by Kim; this ensured that his uncooperative responses would be attributed to *unwillingness* to provide the needed information. Participants evaluated John with a 1(min)-7(max) rating targeting two dimensions central to person perception: Warmth — reflecting someone's intentions towards others — and Competence — reflecting their individual skills and intellectual standing ([4]; see Table 1). We predicted that irrelevant utterances, by completely ignoring Kim's request, should be seen as especially uncooperative, and thus elicit a high social penalty for the speaker in both Competence and Warmth. Under-informative ones, by still retaining some value for the listener, might instead incur a lesser cost. The study consisted of a single trial: 400 subjects were recruited on MTurk (100 per 2x2 cell). Results are shown in **Fig 1**. Two-way ANOVAs performed separately for Competence and Warmth showed that both Competence and Warmth were influenced by Relevance, with John rated as both more competent and warmer when his contribution was relevant (all $ps < .001$). Competence only was affected by Informativeness ($p < .05$) with more informative utterances eliciting higher ratings than less informative ones.

EXP2. Exp2 consisted of a partial replication of Exp1: the Informativeness manipulation was retained, but only irrelevant utterances were included. Contrary to Exp1, these were introduced by the phrase "I've never been to these places", indicating that the maxim violation was due to *inability*, and not *unwillingness*. As they are compatible with the speaker being well-intentioned towards the interlocutor, we expect *inability*-driven violations to be less socially costly than *unwillingness*-driven ones in terms of Warmth. 200 subject were recruited on MTurk. The average ratings for Exp2 and the —Relevance condition in Exp1 are displayed in **Fig 2**. Separate two-way ANOVAs were performed for Warmth and Competence on pooled data from Exp2 and the —Relevance data from Exp1 (factors: Informativeness and Experiment). A main effect of Experiment was found for Warmth ($p < .001$), with irrelevant responses yielding higher warmth ratings when driven by inability. No effect was found on Competence.

DISCUSSION. These results suggest that listeners draw social inferences based on their interlocutor's conversational behavior, with the most disruptive pragmatic violations — i.e., Relevance — emerging as the most socially costly. Moreover, listeners reason about the cause that might have driven a violation, as shown by the mitigated Warmth-related penalty of inability-driven Relevance violations. A lingering puzzle concerns why the social effects of Informativeness are only observed for Competence: a possibility is that the choice of disclosing more information enhanced John's perceived individual ability as a speaker, but not his

perceived propensity to help out Kim. We predict that, by making the under-informative condition more disruptive to the interlocutor's goals, violations of Informativeness should also affect Warmth. In sum, these findings suggest that, even after a brief exposure to someone's conversational behavior, people draw social inferences about the speaker — and do so by reasoning along the same principles that inform pragmatic inferences in the Gricean framework.

Table 1: Manipulations and dialogue for Exp 1

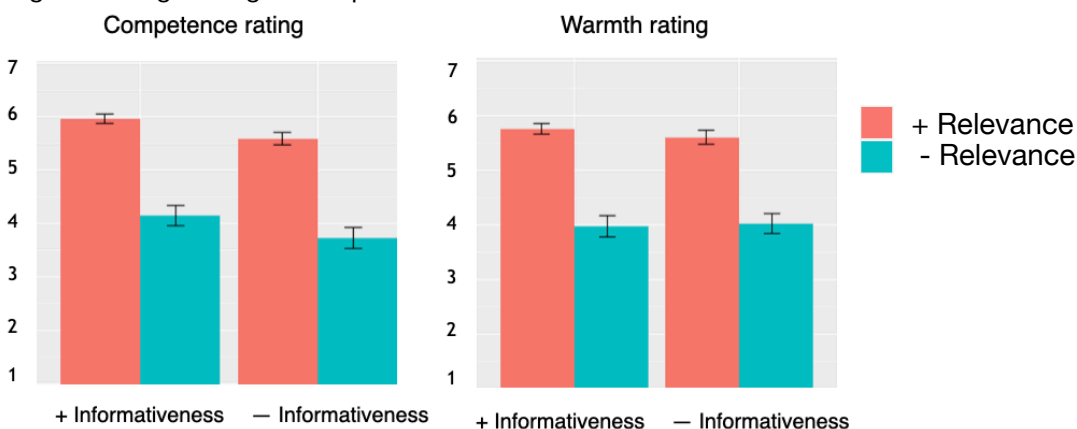
Speaker	Utterance	Manipulation
Kim	Either: I'd like to go on a skiing vacation. I'm thinking Austria, Switzerland or Italy.	Sets up + Relevance of John's description
	Or: I'd like to go on a Caribbean vacation. I'm thinking Antigua, Barbados or Bahamas.	Sets up — Relevance of John's description
John	I've been to all these places.*	
John	Either: I recently went to Zermatt, Switzerland. Best slopes of all places I've been to.	+ Informativeness
	Or: I recently went to Zermatt, Switzerland.	— Informativeness

*Exp2: "I haven't been to any of these places"

Table 2: Questions for Competence vs. Warmth,

Question	Dimension
How knowledgeable do you think John is in this conversation?	Competence
How competent do you think John is as a person?	
How considerate towards Kim do you think John is in this conversation?	Warmth
How likable do you think John is as a person?	

Fig. 1: Average ratings for Exp 1



References

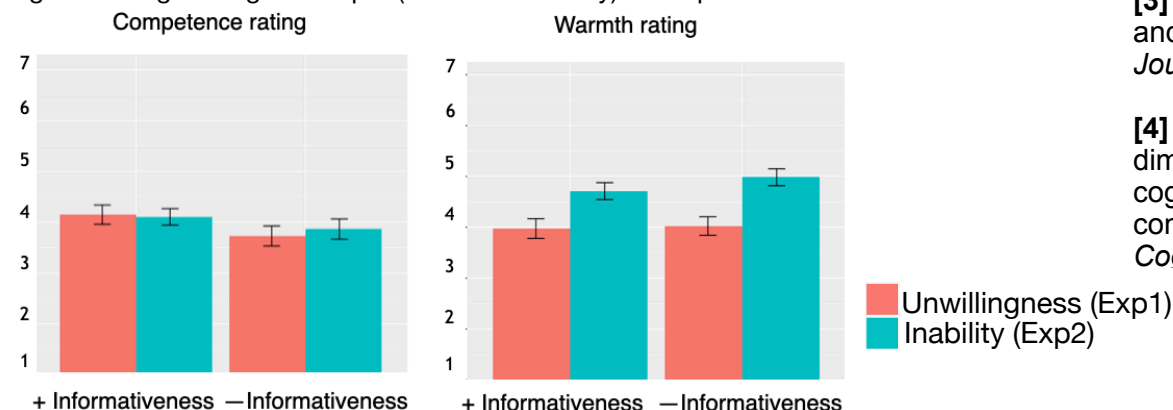
[1] Grice, 1975. Logic and Conversation. *Syntax and Semantics*, vol.3.

[2] Campbell-Kibler, 2007. Accent, (ING), and the Social Logic of Listener Perceptions. *American Speech*.

[3] Eckert 2008. Variation and the indexical field. *Journal of Sociolinguistics*.

[4] Fiske, 2007. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*.

Fig. 2: Average ratings for Exp 1 (—Relevance only) vs. Exp 2



Perceptual contrast as a visual heuristic in the formulation of referential expressions

Madeleine Long (U Oslo), Isabelle Moore (U Virginia), Francis Mollica (U Edinburgh) & Paula Rubio-Fernandez (U Oslo) paula.rubio-fernandez@ifikk.uio.no

We propose that speakers rely on perceptual contrast as a visual heuristic to produce efficient referential expressions *efficiently*. That is, to produce referential expressions that may facilitate the listener's visual search, while requiring limited effort on the speaker's part. Under a contrast perception heuristic, significant perceptual contrast will trigger modification, even when it may be redundant. We understand this visual heuristic as a form of 'low-cost pragmatics' in line with Victor Ferreira's *feedforward audience design* [1]: according to this mechanistic framework, speakers need not engage in reflective processes to be sensitive to their listeners' needs; instead, they can make use of contextual cues prior to utterance onset and rely on previously learned strategies that facilitate communication (see also [2,3]). A number of psycholinguistic studies have shown that redundant modification can facilitate the listener's visual search for a referent [4-9], confirming that over-specification can be efficient [10,11]. **Here we report two language production experiments testing whether perceptual contrast triggers efficient over-specification.**

Experiment 1: Koolen et al. [12] (see also [13]) have argued for an alternative account in which color over-specification is triggered by 'scene variation' (i.e. the number of dimensions along which the objects in a scene vary). Their results support their predictions, but they tested high scene variation in polychrome displays and low scene variation in monochrome displays, so their results could have been driven by color contrast rather than by scene variation. Here we pitched scene variation against color contrast (see Fig. 1). UCL students ($n=31$) requested a target in two blocks of monochrome and polychrome displays (lab task). An LMER model of Over-specification with Scene Variation level (high vs low) as FE and maximal RE structure revealed more over-specification in low scene variation (polychrome) than high scene variation (monochrome) ($\beta=8.7$, $95\%CI=[4.8-13.8]$), contra to [12,13]. The perceptual contrast hypothesis was tested in another LMER model with Modifier Type (Color vs Other: size, border type and border weight), Display Type (Monochrome vs Polychrome), and Block as FE and maximal RE structure. Supporting our hypothesis, we observed more color over-specification in polychrome than monochrome displays ($\beta=7.1$, $95\%CI=[4.1-10.5]$), and more over-specification of size, border type and border weight in monochrome than polychrome displays ($\beta=-17.8$, $95\%CI=[-31.0 - -10.6]$) (see Fig. 2).

Experiment 2: Previous studies have shown that speakers over-specify atypical colors (e.g., 'pink banana') more than typical colors (e.g., 'yellow lemon') [13-15], which some have interpreted as a cooperative strategy to aid the listener's visual search [10]. We predicted that atypical colors would be over-specified in polychrome displays, but not in monochrome displays. According to the alternative view that atypical colors are salient because they violate world knowledge, color contrast should not make a difference. MTurk participants ($n=38$) had to instruct a virtual partner to click on a target object in a series of displays (see Fig. 3). We ran an LMER model of Over-specification with Display Type and Target Typicality (Atypical, Typical, Variable) as FE and maximal RE structure. Replicating [10], we found higher over-specification in atypical polychrome compared to typical polychrome ($\beta=-8.7$, $95\%CI=[-16.7 - -4.6]$) or variable polychrome ($\beta=-3.9$, $95\%CI=[-6.4 - -1.9]$). As predicted, we found a decrease in over-specification in atypical monochrome compared to atypical polychrome ($\beta=-19.9$, $95\%CI=[-42.0 - -7.7]$) and no effect of target typicality across monochrome displays (see Fig. 4). These results suggest that over-specifying atypical colors is an efficient, cooperative strategy [10].

Our findings support the view that speakers use perceptual contrast as a visual heuristic for efficient referential communication [1,11]. In this view, deciding whether to use modification in referential communication need not be costly (e.g., speakers need not identify competitors in the visual context prior to producing an efficient referential expression; see [16,17]). **Relying on perceptual contrast as a visual heuristic would allow speakers to adapt their referential expressions to their listener's needs with minimal expenditure of cognitive resources, in line with Ferreira's feedforward audience design.**

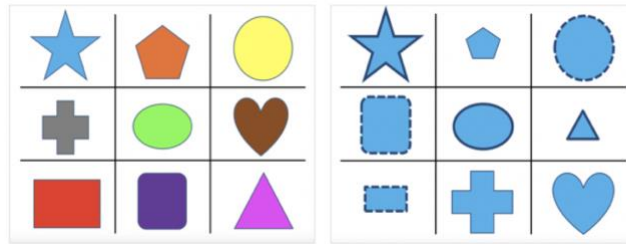


Figure 1. Sample polychrome display with low scene variation (shape and color vary) and monochrome display with high scene variation (shape, size, border type and border weight vary) from Exp. 1.

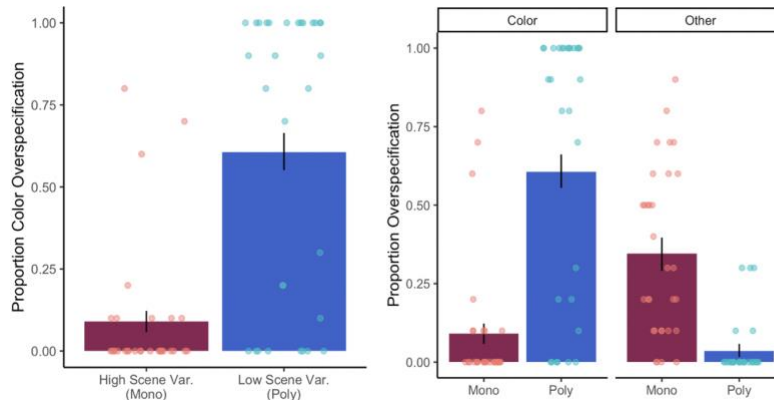


Figure 2. Mean proportion of over-specification from Exp. 1 testing the scene variation hypothesis (left panel) and the perceptual contrast hypothesis (right panel), aggregated by display type (both panels) and response type (right panel). Line ranges reflect 95% bootstrapped CIs and points reflect participant means.

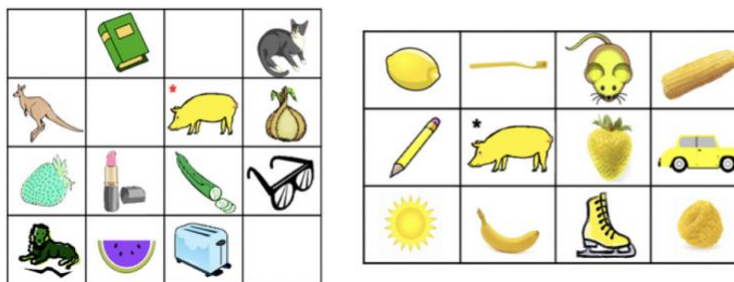


Figure 3. Sample polychrome and monochrome displays from Exp. 2. The polychrome displays were taken from Rubio-Fernandez (2016).

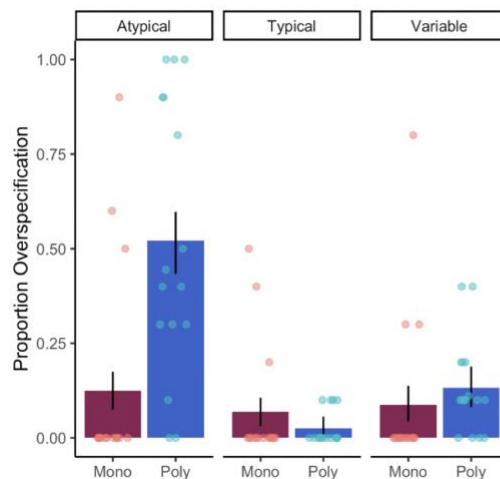


Figure 4. Mean proportion over-specification from Exp. 2 by display type and target typicality. Line ranges reflect 95% bootstrapped CIs and points reflect participant means.

References: [1] Ferreria, V., 2019. *Ann Rev Psych* [2] Jaeger & Ferreira, V., 2013. *Behav & Brain Sci* [3] Kurumada & Jaeger, 2015. *JML* [4] Sonnenschein & Whitehurst, 1982. *J Psycholing Res* [5] Mangold & Pobel, 1988. *Lang & Soc Psych* [6] Paraboni, Van Deemter & Masthoff, 2007. *Comp Lings* [7] Arts, Maes, Noordman & Jansen, 2011. *J Prags* [8] Paraboni & Van Deemter, 2014. *Lang, Cog & Neurosci* [9] Tourtouri, Delogu, Sikos & Crocker, 2019. *J Cul Cog Sci* [10] Rubio-Fernandez, 2016. *Front Psych* [11] Rubio-Fernandez, 2019. *Cog Sci* [12] Koolen, Goudbeek & Krahmer, 2013. *Cog Sci* [13] Degen, Hawkins, Graf, Kreiss & Goodman, 2020. *Psych Rev* [14] Sedivy, 2003. *J Psycholing Res* [15] Westerbeek, Koolen & Maes, 2015. *Front Psych* [16] Brown-Schmidt & Tanenhaus, 2006. *JML* [17] Davies & Kreysa, 2017. *Acta Psych.*

But what can I do with it?: Speakers name interactable objects earlier in scene descriptions

Madison Barker (msbarker@ucdavis.edu), Gwen Rehrig, Fernanda Ferreira (UC Davis)

Introduction: Spoken language requires speakers to decide what to say and when; deciding on a linear order is the linearization problem of language production (Levelt, 1981). Previous research has suggested that image salience influences word order (Gleitman et al., 2007). More recent work found that image salience and meaning are correlated (Henderson & Hayes, 2017; Henderson et al., 2018), but neither image saliency nor scene meaning predicted the order in which objects are mentioned (Rehrig et al., 2020). Perhaps linearization decisions are based on another type of information that is more relevant to a human agent, such as object affordances. One type of object affordance, graspability, has been shown to predict visual attention (operationalized as fixation density) as well as meaning (Rehrig et al., 2020a). This study investigates whether object affordances more generally, which we term “interactability”, predicts the order in which objects are mentioned in speakers’ verbal descriptions. We hypothesized that objects that received higher ratings of interactability would be more task-relevant and would occur earlier in speakers’ descriptions of the scenes.

Methods: Thirty native English speakers verbally described 30 real-world scenes, each for 30s, while eye-movements and speech were recorded (Henderson et al., 2018; Rehrig et al., 2020; see Fig.1a). To measure interactability, a separate group of participants was shown a black and white version of the scene with a single object shown in color (Figure 1b). Participants were asked to indicate on a scale from 1 (Very Unlikely) to 7 (Very Likely) the degree to which a human would interact with the highlighted object (Figure 1c). To obtain meaning and saliency values, the same objects that were rated for interactability were parsed into polygons using CVAT and LabelMe (Figure 1c). Object name referents were identified using a window of time and fixation data based on eye-voice span estimates (see Rehrig et al., 2020b).

Results: To assess word order, object mentions were identified with respect to their temporal onset in the verbal description. Meaning map ($M = 0.43$, $SD = 0.13$), saliency map ($M = 0.37$, $SD = 0.12$), and object interactability values ($M = 4.52$, $SD = 0.95$) were used as predictors of word onset ($M = 13623.61$ ms, $SD = 8253.49$ ms; Figure 2). Object map values were averaged over the entire polygon (parsed in CVAT/LabelMe). The correlations revealed that neither meaning map values ($r = -0.021$, $p = 0.54$, Fig.1a) nor saliency map values ($r = -0.034$, $p = 0.34$, Fig.1b) were correlated with the order in which objects were mentioned. Consistent with our hypothesis, whole object interactability values did predict the order in which objects were mentioned ($r = -0.14$, $p < 0.001$, Fig.1c).

Discussion: Consistent with previous results, we observed that neither meaning nor saliency values predicted the order in which objects were mentioned. In contrast, object interactability did predict sequencing: Objects rated as more interactable were mentioned earlier in participants’ verbal descriptions. These results add to our growing understanding of how complex verbal descriptions are planned and sequenced, suggesting that the specific aspect of meaning that influences utterance sequencing decisions is object interactability. When speakers plan multi-utterance sequences such as scene descriptions, they begin by identifying objects with which they would be inclined to interact. Overall, this work provides compelling evidence for the role of object affordance information in language processing.

- Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, 57(4), 544-569.
- Levelt, W. J. (1981). The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 295(1077), 305-315.
- Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1(10), 743.
- Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports*, 8, 13504.
- Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020a). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Rehrig, G., Hayes, T. R., Henderson, J. M. & Ferreira, F. (2020b). Setting the scene: Saliency and meaning in linearization during scene description. Poster presented on March 20th, 2020 at the 33rd Annual CUNY Human Sentence Processing Conference, University of Massachusetts, Amherst.



Figure 1. A) Real-world scene presented to subjects in the description task. B) Object and scene context presented in the interactability rating task. C) Parsed object polygon overlaid on the scene. The average of the map values for pixels within the polygon served as meaning and saliency measures in the correlations.

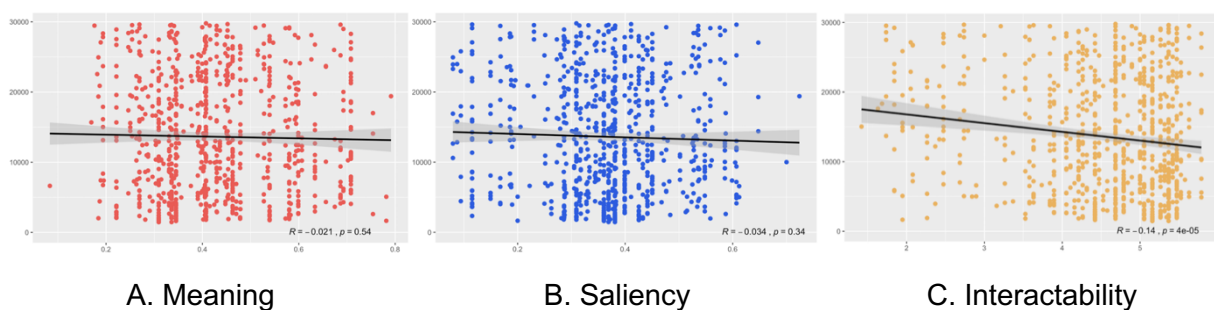


Figure 2. Scatterplots showing object name onset in the description on the x-axis (in ms) plotted against A) object meaning map values, B) object saliency map values, or C) whole object interactability ratings on the y-axis. Black regression lines indicate correlations.

Culture, collectivism, and second language use affect perspective taking in language production

¹Dunn, M.S., ¹Cai, Z.G., ¹Xu, Z., ²Branigan, H.P., & ²Pickering, M.J.

¹Chinese University of Hong Kong & ²University of Edinburgh

1155131488@link.cuhk.edu.hk

Language allows interlocutors to depict spatial positions from a range of perspectives. For example, an interlocutor can use an egocentric self-perspective (e.g., on my right), or an allocentric non-self-perspective¹ (e.g., to her left). Tosi et al.² conducted a study whereby native English speakers produced spatial descriptions of objects. These participants were shown pictures with two objects on the left or right of the screen. One condition had no person in the picture (no agent condition). In the other conditions, the person faced either away from the participant (same orientation) or towards (opposite orientation); and could see/act on the objects (can-act action potential) or could not do so (cannot-act action potential). Tosi et al. found that Orientation affected the use of allocentric perspective taking, especially in the can-act condition.

Tosi et al. along with other papers focus on how environmental factors and audience design affect perspective taking. However, there is a lack of research on how factors internal to an interlocutor affect this phenomenon. We therefore conducted two experiments, the first comparing perspective taking by Chinese and English speakers, who grew up in more collectivistic and individualistic cultures respectively. Collectivism entails a self perception grounded in relationships, with Asian cultures being generally more collectivist than Western cultures³. We hypothesize that higher collectivism may lead to greater allocentrism, due to more relational emphasis that could evoke an increase in simulating the perspectives of others⁴.

Experiment 1 replicated Tosi et al.'s Experiment 3 (described above) but with 93 native Mandarin speakers. We built a logistic mixed model (binary DV of egocentric/allocentric response) on our and Tosi et al.'s data, with Language (Mandarin vs. English), Orientation (same vs. opposite perspective) and Action Potential (can act/see vs. cannot act/see items) along with their interactions as fixed effects, and with data justified maximal random effects. More allocentric responses were produced by the Mandarin speakers than the English speakers ($z = 5.01$), and for opposite than same orientations ($z = 11.74$), but the effect of Orientation was greater for the Mandarin speakers ($z = 6.06$). In addition, the effect of Orientation was greater when the person could than could not act ($z = 2.93$). Therefore, consistent with their collectivist culture, Mandarin speakers used more allocentric perspectives compared to English speakers, and especially when viewing people with opposite orientations.

Experiment 2 tasked 109 native Mandarin speakers with the same task, except in both their native/second languages (Mandarin/English respectively; within subjects) with only the opposite-orientation can-act condition and the no-agent condition, with participants answering scales on collectivism and English proficiency. Participants produced more allocentric responses in their second language ($z = -0.914$), and the effect of language was modulated by both presence of an agent ($z = 1.181$) and Collectivism ($z = -0.831$). These results suggest that second language use increases allocentrism, especially when no agent is present, and that individual differences in collectivism modulated perspective taking more in second language use. Overall these two experiments replicate and extend Tosi et al.'s findings, showing that culture and collectivism affect perspective taking, and that the increased allocentrism of the Mandarin speakers was not due to idiosyncratic properties of Mandarin. On the contrary, use of English as a second language increased allocentrism, possibly due to an increase in deliberative thinking associated with second language use⁵.

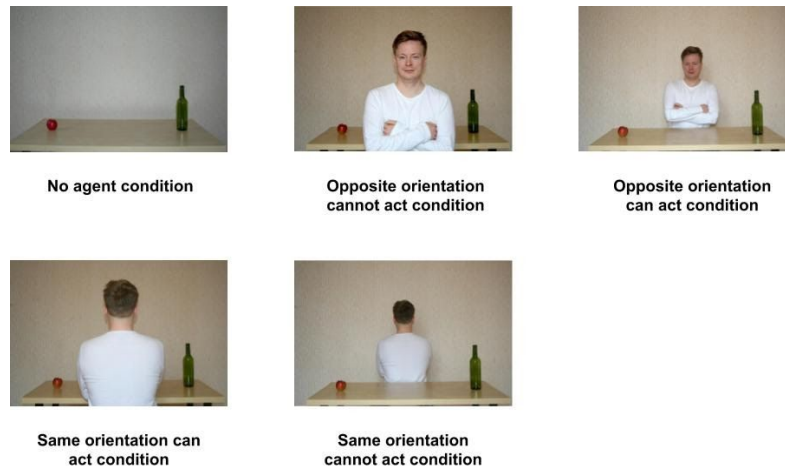


Figure 1: The Conditions of Experiment One and Two

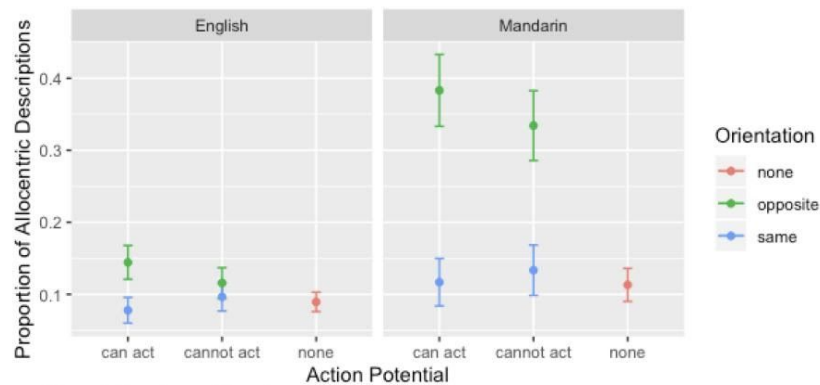


Figure 2: Experiment 1 Results

English data source: Tosi et al. (2020)

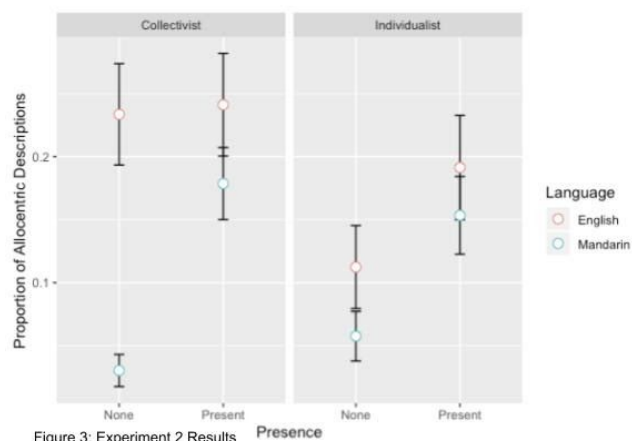


Figure 3: Experiment 2 Results

¹Tversky, B. (1996). Spatial perspective in descriptions. *Language and space*, 3, 463-491.

²Tosi, A., Pickering, M. J., & Branigan, H. P. (2020). Speakers' use of agency and visual context in spatial descriptions. *Cognition*, 194, 104070.

³Singelis, T. M., Triandis, H. C., Bhawuk, D. P., & Gelfand, M. J. (1995). Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-cultural research*, 29(3), 240-275.

⁴Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological science*, 18(7), 600-606.

⁵Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4), 569.

The Role of Relatedness on Sentence Production

Jacqueline Erens & Jessica Montag (The University of Illinois – Urbana-Champaign)

In psychology, interference is observed in many domains. Specifically, semantic interference is observed in cyclical naming (Oppenheim et. al, 2010, Howard et al. 2006) and picture word interference tasks (Rosinski, 1977). Likewise, sentences containing semantically similar entities take longer to produce (Smith & Wheeldon, 2004) and speakers make structure choices to alleviate interference (Gennari et al., 2012). In other tasks, semantic relatedness is facilitatory. In semantic priming tasks, related words lead to speeded responses (Neely, 1976). Likewise, speakers make fewer agreement errors when similar items are closer in a sentence (Gillespie & Pearlmutter, 2011). Semantic relatedness can lead to either interference or facilitation.

One possible reason for inconsistent findings is differences in the types of relatedness examined. Studies finding interference often investigate semantically *replaceable* entities: category members with a high feature overlap (a baseball player and basketball player). Studies finding facilitation often use entities that *co-occur* (a baseball player and a coach). *Co-occurring* entities may not interfere as they are not *replaceable* and would not be activated as competitors during lexical access (Levelt et al., 1991). *Co-occurring* entities may be easier plan and produce in a sentence, whereas *replaceable* entities should be more difficult. We test predictions about co-occurring vs. replaceable entities in a picture-description sentence production task.

Speakers often make planning and production easier via implicit structure choices (Bock, 1982; MacDonald, 2013). We investigate the English dative alternation, which allows flexibility in speaking about transfer events. We investigate whether speakers choose sentence structures that allow them to separate semantically replaceable (interfering) entities (e.g., Prepositional Dative: The farmer is giving the bell to the fisherman vs. Double Object: The farmer is giving the fisherman the bell) or group co-occurring (facilitatory) entities (PD: The farmer is giving the corn to the ninja vs. DO: The farmer is giving the ninja the corn). This task allows us to investigate structure choices and speaking duration as a consequence of relatedness between entities.

Method: Stimuli were sets of images of two people transferring an item. We created 21 item quads (Table 1). All items in a quad had the same agent and included a related and unrelated recipient and item. Related people were chosen to be *replaceable*, defined by cosine similarity calculated using Spacy (Honnibal & Montani, 2017). Related items were chosen to co-occur using Wikipedia (Davies, 2015). Participants (N=23) saw one item per quad and items from each condition, and were given the verb to use on each trial but not the labels for entities in the pictures. To ensure name agreement, only items with 80% or higher name agreement on a norming task were used. The study was run using Psychopy3 (Pierce et al., 2019) on Zoom.

Results: We found no difference in the use of PD versus DO constructions across conditions (~70% PD/30% DO). However, effects of our manipulation are seen in speaking durations (Figures 1 & 2). When participants used the DO construction, they produced related people more quickly (Table 2). The results did not change when the Pointwise Mutual Information (PMI) values, a measure of co-occurrence that controls for word frequency (Bouma, 2009), between the agent/recipient and agent/item were added. In a model with only the two PMI values and cosine similarity between the agent and recipient, we saw an effect of cosine similarity (Table 3), suggesting that the relatedness between people, as defined by cosine similarity (not co-occurrence) accounted for the facilitation for related people. For the PD constructions, higher PMI values between the agent and item were associated with faster speaking times, but when producing the recipient (Table 4). Perhaps when easier-to-produce items precede recipients, speakers have extra planning time during the easier item phrase to plan the recipient.

Discussion: We saw no difference in PD or DO use across conditions. Effects of relatedness in timing measures were in the opposite direction as expected: relatedness between people appeared to speed, not slow, speaking times in DO constructions. In PD constructions, as predicted, related items did speed speaking times, but for the following (recipient) phrase. Potential reasons for these unexpected results and planned follow-up studies will be discussed.

Table 1: Experimental Design

Agent (Farmer)	Agent (Farmer)
Related Person (Fisherman)	Related Person (Fisherman)
Related Item (Corn)	Unrelated Item (Bell)
Agent (Farmer)	Agent (Farmer)
Unrelated Person (Ninja)	Unrelated Person (Ninja)
Related Item (Corn)	Unrelated Item (Bell)

Table 2: Mixed-effects model for DO Duration 1 (Recipient Duration)

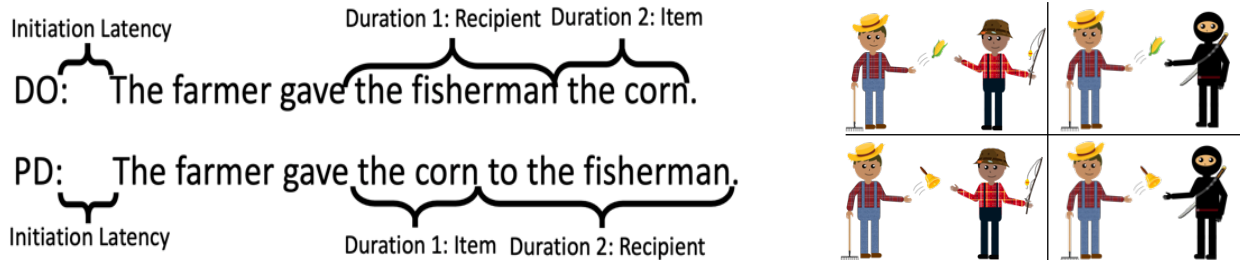
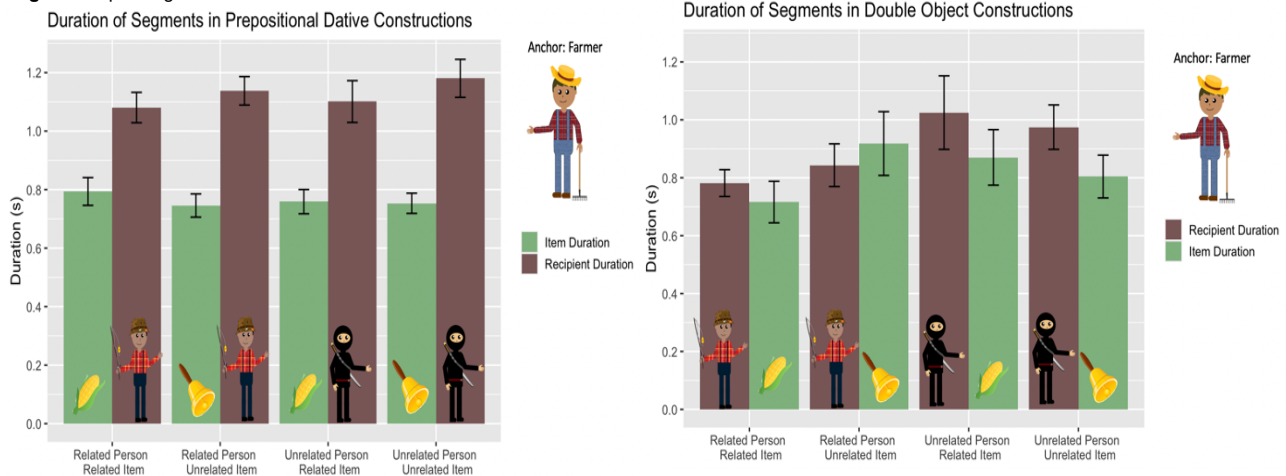
DO Construction: Duration 1	β	SE	t	p
Intercept	0.927	0.076	12.169	< .001
Person Related	-0.169	0.082	-2.062	0.042
Item Related	0.004	0.081	0.047	0.963
Person Related x Item Related	-0.087	0.162	-0.538	0.592

Table 3: Mixed-effects model for DO Duration 1 with PMI & Cosine

DO Construction: Duration 1	β	SE	t	p
Intercept	0.937	0.075	12.512	< .001
PMI of Anchor and Person 2	0.076	0.067	1.145	0.255
PMI of Anchor and Item	0.039	0.04	-0.973	0.333
Cosine similarity of Anchor & Person2	-0.136	0.066	-2.061	0.042

Table 4: Mixed-effects model for PD Duration 2 (Recipient Duration)

PD Construction: Duration 2	β	SE	t	p
Intercept	1.127	0.063	17.989	< .001
PMI of Anchor and Person 2	0.022	0.059	0.376	0.709
PMI of Anchor and Item	-0.085	0.041	-2.054	0.041
Cosine similarity of Anchor & Person2	-0.046	0.055	-0.834	0.407

Figure 1: Boundaries and words included in each speaking duration for DO and PD sentences in an example item**Figure 2:** Speaking durations for DO and PD sentences

References

- Bock, J.K. (1982). Towards a cognitive psychology of syntax: Information processing contributes to sentence formulation. *Psychol Review*, 89, 1-47.
- Bouma, G.. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31-40.
- Davies, M. (2015). The Wikipedia Corpus: 4.6 million articles, 1.9 billion words. Adapted from Wikipedia. Available online at <https://www.english-corpora.org/wiki/>
- Gennari, S.P., Mirkovic, J., & MacDonald, M.C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, 65, 141-176.
- Gillespie, M. and Pearlmutter, N.J. (2010). Hierarchy and scope of planning in subject-verb agreement production. *Cognition*, 118, 377-397.
- Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, 100, 464-482.
- Levitt, W.J.M., Schriefers, H., Vorberg, D., Meyer, A.S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, 98, 122-142.
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, 226.
- Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cog.*, 4, 648-654.
- Oppenheim, G.M., Dell, G.S., & Schwartz, M.F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, 114, 227-252.
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51, 195-203.
- Rosinski, R. R. (1977). Picture-word interference is semantically based. *Child Development*, 48, 643-647.
- Smith, M. and Wheeldon, L. (2004). Horizontal information flow in spoken sentence production. *JEP: Learning, Memory, and Cognition*, 30, 675-686.

Speech Rate Convergence in Spontaneous Conversation

Maya Ricketts (Vanderbilt University), Benjamin Schultz (University of Melbourne), Duane Watson (Vanderbilt University)

Features of interlocutors' speech become more similar over the course of a conversation (Giles & Ogay, 2007). This convergence exists at the lexical (Garrod & Anderson, 1987), syntactic (Branigan et al., 2000), acoustic (Natale, 1975), and rhythmic (Tarr et al., 2014) level. Previous work has found speech rate convergence in scripted conversations and conversations with a confederate (e.g., Tierney, Patel, & Breen, 2018). We investigated speech rate convergence in spontaneous, non-scripted conversations to test two theories of linguistic convergence. The Communication Accommodation Theory (CAT) posits that individuals adjust communicative behaviors to increase or decrease social distance in the context of an interaction (Giles & Ogay, 2007). In this framework, convergence of rhythmic features serves as an indicator of greater social affiliation. The Interactive Alignment Model (IAM) suggests that speech convergence occurs to facilitate comprehension (Garrod & Pickering, 2015). As speech features become more aligned, interlocutors better understand the semantic content of utterances. Both CAT and IAM predict that speech rates converge within real-world interactions when speakers are positively affiliated. However, CAT specifically predicts that divergence occurs when speakers have competing sociocultural affiliations. We tested whether convergence is modulated by social factors, as predicted by CAT or whether it occurs universally, independent of affiliation, as predicated by IAM. Dyads ($N=56$) engaged in conversations in which we manipulated awareness of interlocutors' beliefs on politically loaded statements.

Each participant first performed a three-minute monologue describing their favorite trip. Then, interlocutors engaged in six conversations. Before each conversation, each participant was asked to respond to a statement which was either political (e.g. "Abortion is morally wrong in most cases") or neutral (e.g. "Pineapple belongs on pizza"). Using cards marked "Agree" or "Disagree", participants were asked to indicate, in full view of their partner, how their opinions or beliefs aligned with the statement. After this selection process, dyads were told to discuss a case study describing an apolitical dilemma to arrive at a solution. They completed this process six times, and the number of times their opinions differed on the statements was counted as the polarization score. Participants then completed a questionnaire that probed their honesty on statement responses. Speech was recorded throughout the task, and speech rates were measured using a beat-tracking algorithm in MATLAB (Schultz et al., 2016).

All speakers changed their speech rate in the conversations compared to the monologue (see Figure 1a). The speech rates of interlocutors also converged over the course of dialogues and significantly differed from baseline speech rate differences in monologues in the final half of the conversation (see Figure 1b). Cross-correlational analyses were used to assess how speech rates of conversational partners covaried over the course of the entire conversation. These revealed moderate positive correlations between patterns of speech rate within dyads (Mean $r = 0.35$, $SD = 0.08$).

Exploratory comparisons revealed greater convergence in dyads with high agreement (i.e., agreement in 5 or more conversations), independent of whether statements were political or apolitical (see Figures 2a and 2b). Exploratory analyses also revealed that dyads containing a speaker who was dishonest converged less than honest dyads (see Figure 2c), and same-sex dyads converged less than opposite-pair dyads (see Figure 2d). These findings suggest that social factors can modulate the degree of convergence, as predicted by CAT. However, speech rate convergence occurred across all conditions, a result more in line with the IAM than the CAT, which predicts divergence when speakers have competing opinions.

Overall, these findings suggest that speech rate convergence manifests regardless of conversational topic but may vary as a function of social factors, lending support to both CAT and IAM. To the knowledge of the authors, this is the first study to demonstrate speech rate convergence in spontaneous, non-scripted speech.

Figures

Deviation from baseline and speech rate convergence

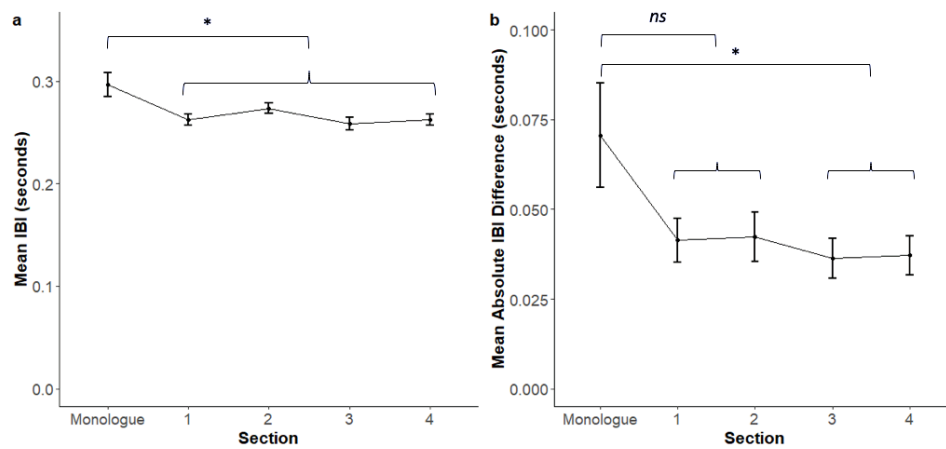


Figure 1. a) Mean inter-beat intervals (IBIs) in the monologue and across dialogue sections 1 to 4, and b) Mean IBI difference within dyads for the monologue and across dialogue sections 1 to 4. Error bars represent standard error of the mean.

Sociopolitical influences of speech rate convergence

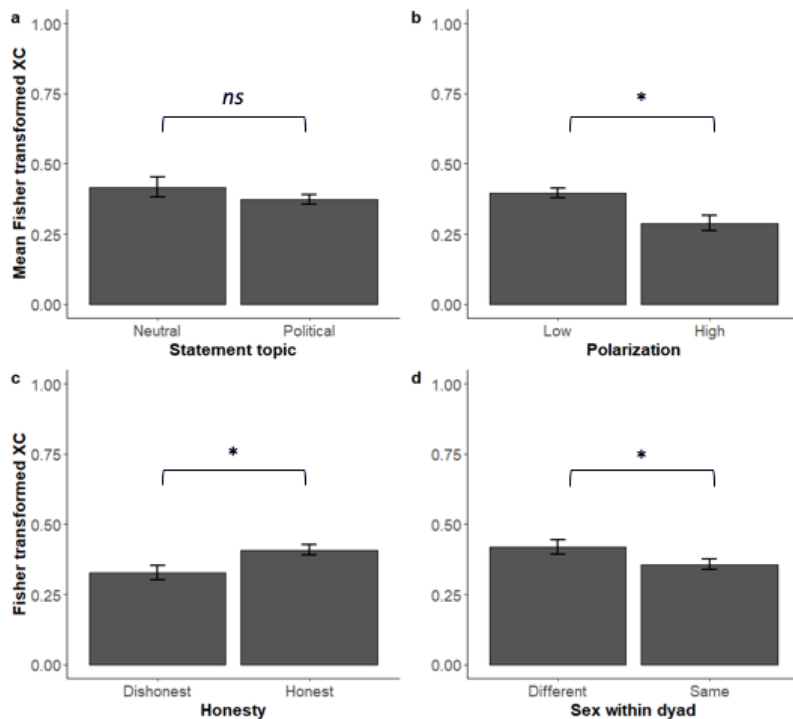


Figure 2. Fisher transformed cross-correlation coefficients between a) statement topics, b) polarization, c) dyads containing at least one dishonest individual and honest dyads, and d) dyads containing members of opposite sexes (different) or the same sex (same). Error bars represent standard error of the mean.

A cross-cultural study of the use and comprehension of color words: English vs Mandinka

Paula Rubio-Fernandez (University of Oslo) and Julian Jara-Ettinger (Yale University)

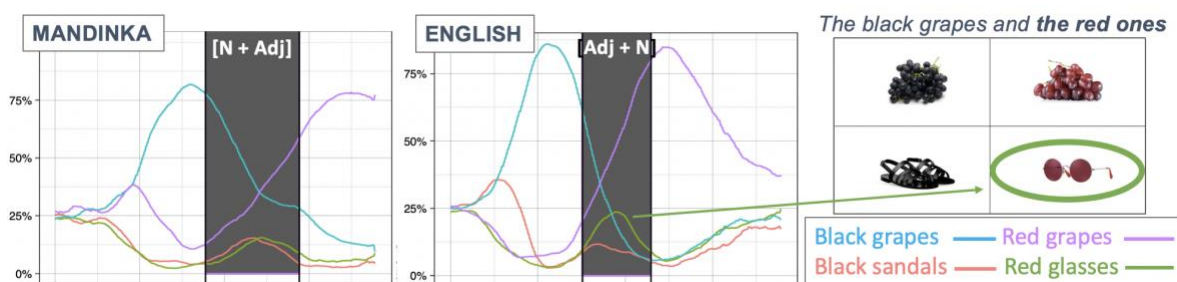
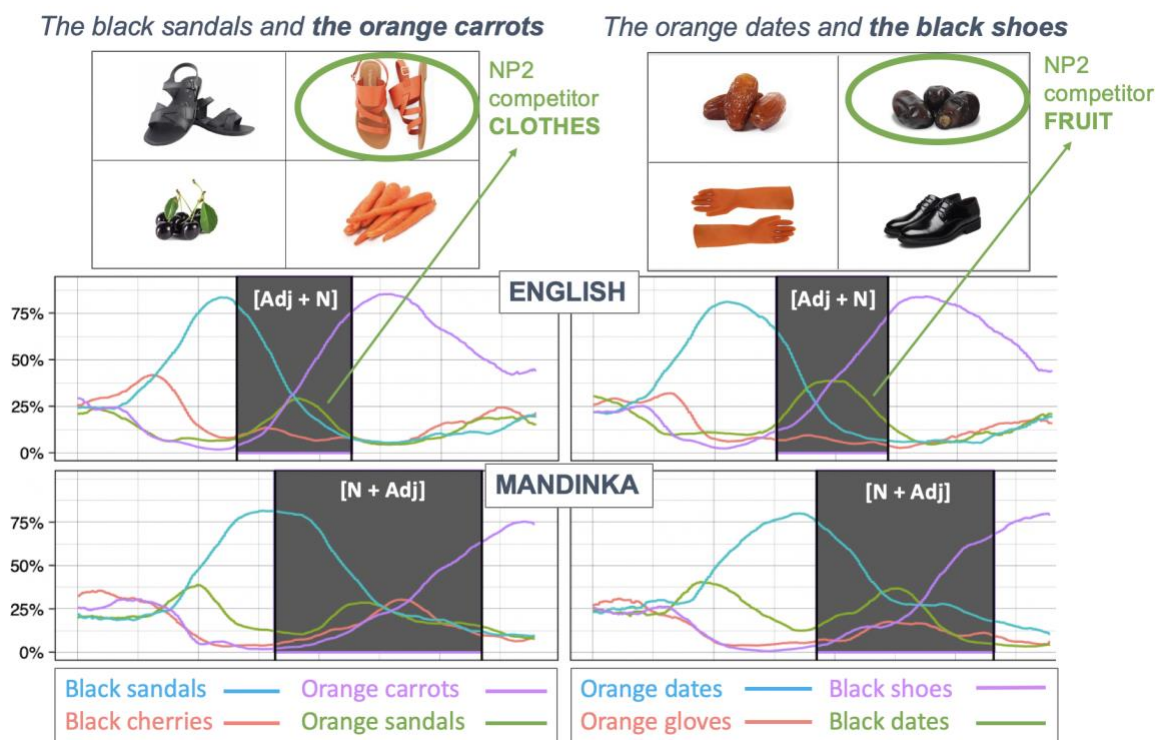
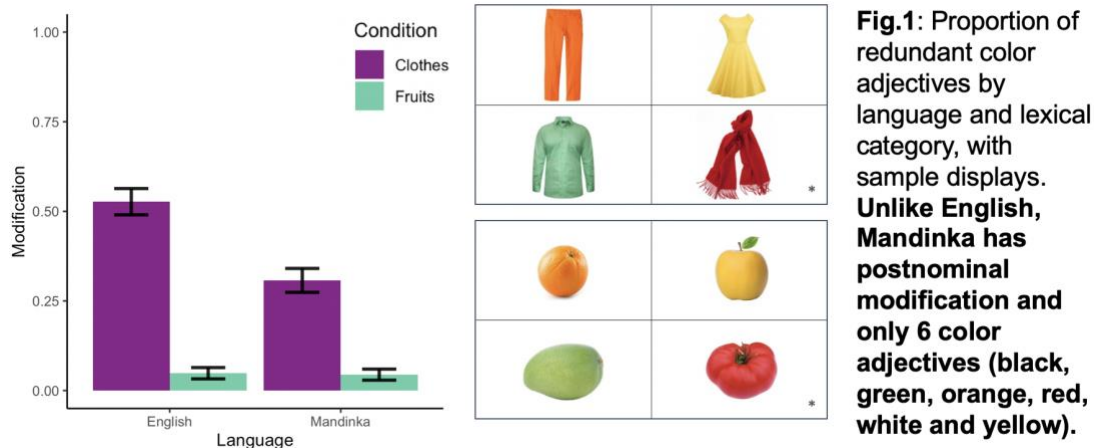
paula.rubio-fernandez@ifikk.uio.no

It has been extensively documented that people often use color adjectives redundantly. However, this tendency is modulated by a number of factors, including the lexical category of the noun (e.g., people may refer to a single dress as ‘the yellow dress’, but would not refer to a single banana as ‘the yellow banana’ [1-4]). Adjective position also affects color over-specification, with prenominal adjectives being used redundantly more often than postnominal adjectives [4-7]. On the comprehension side, contrastive inferences may be derived pragmatically for prenominal adjectives [1,2], but not for postnominal ones [8]. **Here we extended this line of research in a cross-cultural direction to address three questions: (Q1) Would speakers from a non-industrialized society with a reduced color vocabulary also use color words redundantly to refer to clothes, but not to fruits with predictable colors? (Q2) Does the tendency to use color adjectives redundantly with some lexical categories affect the processing of color words accordingly? (Q3) Can contrastive inferences be derived on the basis of adjective position for postnominal adjectives?** (i.e. on a syntactic, rather than a pragmatic basis).

Positive responses to Q1-Q3 were obtained from a reference-production task and an eye-tracking task with native speakers of English (MIT) and Mandinka (a Mande language spoken in The Gambia, West Africa). **EXP1/Q1:** Participants ($n=31+31$) requested a target from a series of displays of clothes or fruits, in a block design (see Fig.1). An LMER model of Over-specification with Language (English, Mandinka) and Lexical Category (Clothes, Fruits) as FE and maximal RE structure revealed a main effect of Language ($\beta=-8.467$, $SE=2.736$, $p<.002$), with more redundant modification observed in English (prenominal) than in Mandinka (postnominal) (replicating [4-6]). There was also a main effect of Lexical Category ($\beta=-9.350$, $SE=2.781$, $p<.001$), with color being used to refer to clothes but not to fruits (also replicating [1,3,4]).

EXP2/Q2: Participants ($n=30+30$) were presented with displays containing a pair of clothes or fruits and another two objects of the other category, in two colors (see Fig.2) and had to click on two of the objects following twice color-modified instructions such as ‘The orange dates and the black shoes.’ When the first NP referred to a member of a pair, processing of the second NP revealed the relative expectation that color would be used contrastively again (e.g., that ‘black’ referred to the other dates, rather than the shoes). An LME model of Percentages of Fixations on the Competitor (e.g., the black dates) during the NP2 window with Language and Lexical Category as FE and maximal RE structure revealed a main effect of Lexical Category ($\beta=8.344$, $SE=2.925$, $p<.017$), with more fixations on the Fruit competitor than the Clothes competitor. These results confirm that speakers of both languages expected color to be used contrastively for fruits more than for clothes. It is remarkable that Language did not have a significant effect or interaction, since Mandinka speakers processed the noun *before* the adjective (“The dates orange and the shoes black”) and were therefore fixating on the contrast object (the black dates) while processing the second noun (“the shoes”; Fig.2). This suggests a **strong expectation that color be used contrastively, possibly because that is how color is used most frequently in Mandinka.**

EXP2/Q3: A prenominal adjective may distinguish two objects of the same kind, revealing a contrastive inference (e.g., in hearing ‘The black...’, participants would fixate on the grapes, not the sandals; see Fig.3) [1,2]. However, this form of pragmatic reasoning is not possible in languages with postnominal modification [8]. Interestingly, some of those languages (including Mandinka, and several Romance languages) use nominalized adjectives to refer to another object of the same kind (“The grapes black and the red”). Thus, **the second adjective in ‘The black grapes and the red ones’ is temporarily ambiguous in English, yet the same construction in Mandinka should elicit a contrastive inference triggered by the syntactic position of the adjective.** An LME model of Percentage of Fixations on the Competitor with Language as FE and maximal RE structure showed a significant main effect ($\beta=-6.685$, $SE=1.958$, $p<.002$), **revealing, for the first time, a contrastive inference that is syntactic – rather than pragmatic.**



References [1] Sedivy, 2003. *J Psycholinguistic Research* [2] Sedivy, 2005. MIT Press (Ch.17) [3] Tarenskeen, Broersma & Geurts, 2015. *Frontiers in Psychology* [4] Rubio-Fernandez, 2016. *Frontiers in Psychology* [5] Rubio-Fernandez, 2019. *Cognitive Science* [6] Rubio-Fernandez, Mollica & Jara-Ettinger, 2020. *JEP:G* [7] Wu & Gibson, 2021. *Cognitive Science* [8] Rubio-Fernandez & Jara-Ettinger, 2020. *PNAS*.

Recall and production of singular *they/them* pronouns

Bethany Gardner & Sarah Brown-Schmidt (Vanderbilt University)

The use of singular *they/them* pronouns is becoming increasingly common as nonbinary identities gain more visibility, with a third of Gen Z and a quarter of Millennials knowing someone who uses *they/them*¹. An exciting opportunity surrounding this cultural and linguistic change is to examine how people learn to associate pronouns with a person. The learning process may require a change from automatically accessing pronoun gender based on semantic/conceptual features of a person², or based on syntactic gender associated with a person's name³, and instead recalling episodic information about a person's stated pronouns. People can learn to interpret *they/them* as singular instead of plural, especially when given explicit instructions to do so⁴. However, speakers often fail to consistently use the correct pronouns when referring to individuals who use *they/them*⁵. Here we ask: When a person is introduced with their pronouns, how accurately are their pronouns remembered and produced, and what is the relationship between memory and production?

Methods: Participants (Ps) [N=102] were introduced to 12 characters, each associated with 4 facts: name (6 masculine, 6 feminine), pronouns (he/him [H], she/her [S], singular *they/them* [T]), job (one of 12), and pet (one of 3). Four characters were associated with masculine names and H, 4 with feminine names and S, and 4 with T (2 masculine, 2 feminine names), such that the use of T could not be predicted from the name. Characters were introduced one-by-one in the frame "[Name] uses [pronouns]. Name works as a [job] and has a [pet]." After a brief delay, we tested memory and production accuracy: For each name, Ps completed a multiple-choice memory test for that character's pronouns, job, and pet. Next, Ps saw each character referenced in the prompt "After [Name] got home from [job]..." and were asked to finish the sentence. Prompts were designed to easily continue using subject pronouns.

Predictions: As T forms are lower frequency than H/S, we expect more accurate memory and production for H/S over T. If learning to produce singular *they/them* requires a shift to a new type of thinking-for-speaking based on episodic memory for a person's stated pronouns, Ps may correctly recall T but fail to accurately use T in production. If episodic retrieval is a necessary first step in production, memory accuracy should predict production accuracy, but more so for T, which is less frequently produced and not always fully incorporated into participants' dialects.

Results: Analysis using mixed-effects models revealed that Memory for pronouns (Fig1) was significantly more accurate for H/S vs. T ($z=11.36$), with no H vs. S difference ($z=0.43$). For characters whose pronouns are T, Ps correctly remembered their pronouns above 33% chance ($t(101)=3.42$, $p<.001$) and at a similar rate as the control item (pet) ($t(101)=0.70$, $p=0.49$). When incorrect, Ps responded with H and S at similar rates (Fig2). Production (Fig3) was more accurate for H/S vs. T ($z=8.80$), with no H/S difference ($z=-0.33$). When referencing characters whose pronouns are T, accuracy was *not* significantly different than 33% chance ($t(101)=-1.11$, $p=.27$), with Ps producing H/S/T at roughly equal rates (Fig4). As predicted, memory accuracy predicted production accuracy ($z=7.40$). Further, this relationship was modulated by pronoun type ($z=-2.44$): When Ps correctly recalled a character's pronouns, the relative difficulty in producing T was somewhat alleviated (Fig5), and further, they produced T at above chance levels ($t(80)=2.69$, $p<.01$).

Conclusion: While memory and use of H/S was more accurate than T, memory for T was above chance, suggesting speakers can learn a person's pronouns when pronouns cannot be automatically inferred. While successful retrieval of T facilitated accurate production of T, speakers were not always successful even when they correctly identified a person's pronouns when explicitly asked. Our findings demonstrate that learning to use *they/them* pronouns may require targeting multiple aspects of learning: remembering that a person uses *they/them*, but also updating the processes by which personal pronouns are produced.

- [1] Parker, K., Graf, N., & Igielnik, R. (2019). Generation Z looks a lot like millennials on key social and political Issues. Pew Research Center.
- [2] Antón-Méndez, I. (2010). Gender bender: Gender errors in L2 pronoun production. *Journal of Psycholinguistic Research*, 39(2), 119-139.
- [3] Schmitt, B. M. (1997). Lexical access in the production of ellipsis and pronouns. *MPI Series in Psycholinguistics*, Nijmegen.
- [4] Arnold, J., Mayo, H., & Dong, L. (2020). Personal pronouns matter: Singular they understood better after explicit introduction. 33rd CUNY Human Sentence Processing Conference.
- [5] The Trevor Project. (2019). National survey on LGBTQ youth mental health.

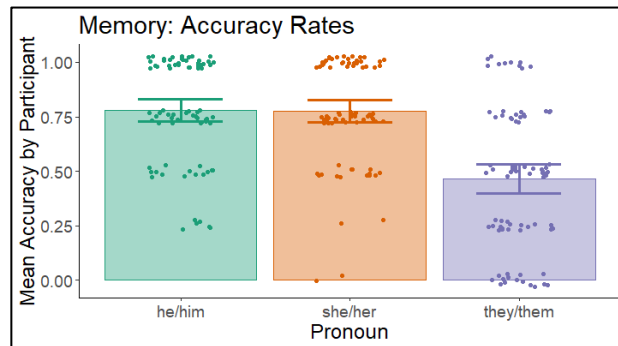


Figure 1. Multiple choice accuracy rates by pronoun condition, with participant means and by-participant standard errors.

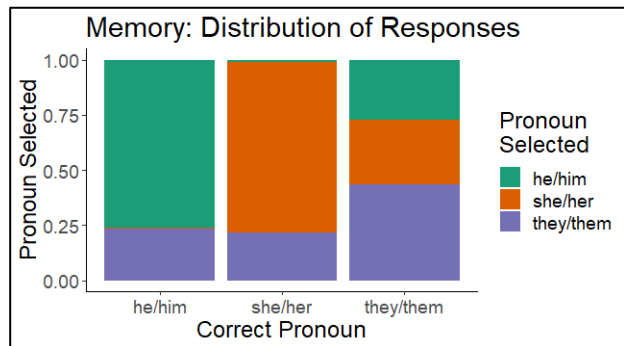


Figure 2. Distribution of multiple choice responses, with the correct pronoun on the x axis and the selected pronoun as the color.

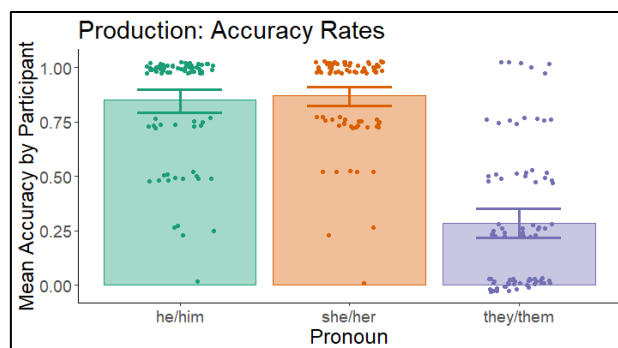


Figure 3. Sentence completion accuracy rates by pronoun condition, with participant means and by-participant standard errors.

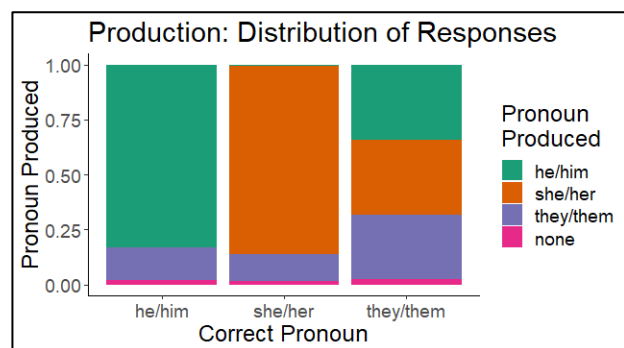


Figure 4. Distribution of sentence completion responses, with the correct pronoun on the x axis and the recalled pronoun as the color.

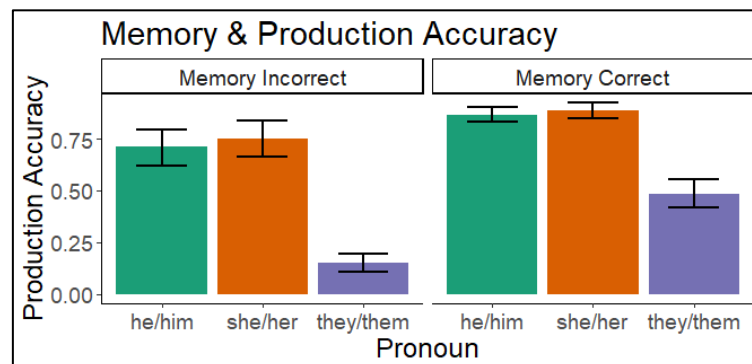


Figure 5. Accuracy on the production task, split (on a trial-by-trial basis) based on whether the P correctly remembered the pronoun in the memory test.

Gender-inclusivity in English pronoun selection by L1 English and Spanish speakers

Cara Walker & Lauren Ackerman (Newcastle University)

Building on the rapidly expanding body of literature on learning and processing of ‘singular *they*’, this study investigates how the interaction of L1 and gender identity influence its use and uptake. Recent work suggests that singular *they* is increasingly acceptable in reference to specific individuals, whether used to indicate the referent’s nonbinary gender or the speaker’s own uncertainty (Bjorkman 2017, Conrod 2018, 2019, Konnelly & Cowper 2020, a.o.). However, these studies investigate L1 users of English. Since instruction in English often occurs in classrooms and in formal contexts where singular *they* might not have been adopted yet, we predict L2 users of English will be less familiar with it, thus be less likely to use it. If so, how do L2 English users reference specific individuals of unknown, ambiguous, or nonbinary gender? We anticipate that, if L2 English users are native users of a grammatically gendered L1 like Spanish, language transfer will lead to L2 English users to produce more gendered pronouns than L1 users. Additionally, there is evidence that certain communities of practice are more likely to produce and accept singular *they* for specific individuals (Ackerman 2020, Conrod 2018), and this is visible in coarse-grained measures like user gender (women, men, nonbinary, and ‘other’), with ‘nonbinary’ and ‘other’ genders leading in use, and ‘male’ trailing. In parallel, innovations in Spanish (while slow to be adopted generally but gaining ground in transgender and nonbinary communities) include pronouns such as *elle*, a gender-neutral alternative to *él/ella* (López 2019). Therefore, we also anticipate both L1 and L2 usage of singular *they* to vary with participant gender. If so, this supports the hypothesis that extant gendered social structures directly influence adult language acquisition and use of singular *they*.

A survey was conducted to identify the patterns in use of singular *they* in adult L1 English and L1 Spanish users (N=100, Table 1). Participants were presented with a drawing of a person doing an activity (Ribb 2020) and asked to assemble a sentence describing the image using a pool of subjects (pronouns) and predicates (past tense verb phrases). Only one predicate accurately described the image (e.g., “read a book.”), while the subjects consisted of the words “She”, “He”, “They” (Figure 1). Subjects were also asked for demographic information, including gender, age of English acquisition, and location of exposure to English.

Figure 2 shows that, contrary to our predictions, L1 Spanish users are more likely to use singular *they* than L1 English users ($z(1)=3.5$, $p<0.001$). When considering participant gender (Figure 3), nonbinary individuals lead in use of singular *they* ($z(1)=4.08$, $p<0.0001$). This contrasts sharply with English L1 men, who use of singular *they* least, whereas Spanish L1 users who aren’t nonbinary show remarkable consistency. This is consistent with the hypothesis that community of practice, specifically for trans and nonbinary individuals, is an important influence on adoption of singular *they*. We also examined response as a function of stimulus gender (previously normed) (Figure 4). All participants used *they* more frequently with nonbinary/unknown images ($z(1)=2.7$, $p=0.008$). Interestingly, nonbinary L1 Spanish users were highly consistent across stimuli, indicating that this community of practice applies singular *they* more radically than the other genders across both L1s. The higher frequency use of *they* by the nonbinary and ‘other’ individuals supports the hypothesis that community of practice is a major influence on adoption of singular *they*.

Curiously, the degree that gender categories differ from each other appears larger than the degree that L1 categories differ. This suggests that gendered social structure is a stronger influence on adoption of singular *they* than language of origin or formal language instruction. We therefore posit that L2 English users might have an *easier* time learning a novel pronominal paradigm which includes singular *they*, as compared to L1 English users, for whom the task requires reanalysis and reassembly of long-established syntactic features on a single element of the pronominal paradigm (Konnelly & Cowper 2020, Lardiere 2008).

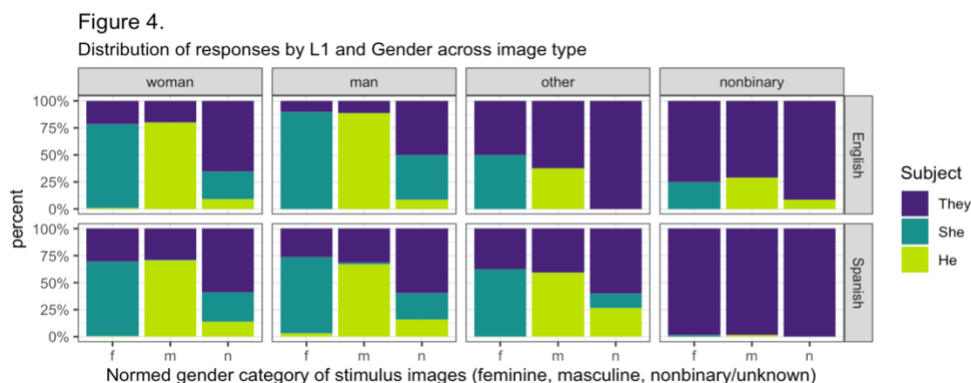
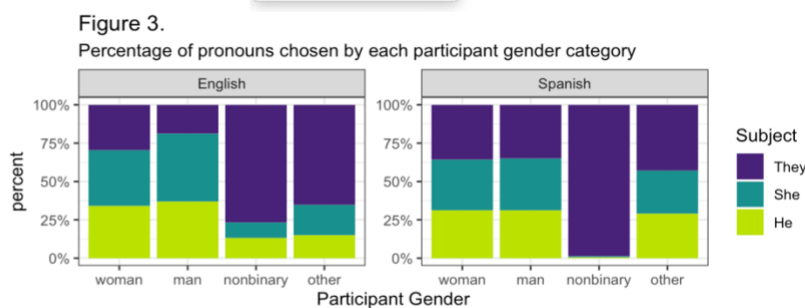
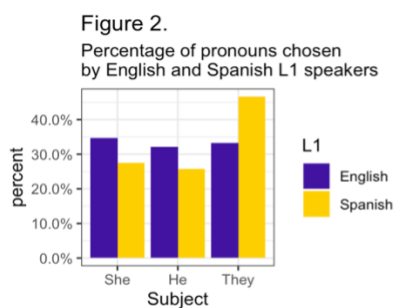
References

Ackerman. 2020. Social exposure to gender-variance influences the real-time processing of pronouns. *CUNY2020*.; Bjorkman. 2017. Singular *they* and the syntactic representation of gender in English. *Glossa*.; Conrod. 2018. Changes in Singular *They*. *Cascadia Workshop in Sociolinguistics (CWSL2018)*.; Conrod. 2019. *Pronouns Raising and Emerging*. PhD dissertation, University of Washington.; Conrod. 2020. Pronouns and Gender in Language. In: *The Oxford Handbook of Language and Sexuality*.; Konnelly & Cowper. 2020. Gender diversity and morphosyntax: An account of singular *they*. *Glossa*.; Lardiere. 2008. Feature-Assembly in Second Language Acquisition. In: *The role of formal features in Second Language Acquisition*.; López. 2019. Tú, yo, elle y el lenguaje no binario. *La Linterna del Traductor*.; Ribu. 2020. *Language and cognition in healthy aging and dementia*. PhD dissertation, University of Oslo.

Table 1: Participant demographic categories

	L1 English	L1 Spanish
Woman	33	28
Man	12	8
Nonbinary	6	8
Other	1	4

Figure 1. Example stimulus with nonbinary image.



Bias against “she” pronouns can be rapidly overcome by changing event expectations

Till Poppels (University of Paris); Veronica Boyce (Stanford University), Chelsea Ajunwa (MIT), Titus von der Malsburg (University of Potsdam), Roger Levy (MIT)

Changing expectations about a future event can manifest rapidly in language use. During the 2016 US presidential election, von der Malsburg et al. (2020) elicited Americans’ production and comprehension preferences for pronoun references to the then-future next president, potentially a woman (Hillary Clinton) or a man (Donald Trump). Participants’ pronoun production rates changed in close lockstep with expectations regarding the likely election winner, whereas reading times in comprehension were less labile. The study’s main result, however, was a persistent disadvantage for “she” relative to “he” in both production and comprehension, even when the female candidate was expected to win. Since the male candidate won the 2016 election, this study could not address whether and how quickly this disadvantage for “she” pronouns might be overcome in case the female candidate won. Here we address this open question in the context of the 2020 U.S. Presidential election by examining pronoun references to the future Vice President (VP), either a woman (Kamala Harris) or a man (Michael Pence). Additionally, we widen the scope of inquiry with references to the future VP’s race.

We collected data from 1611 US-based Mechanical Turk participants in two rounds: pre-election (10/30-11/2); and post-election (11/7-11/10, starting immediately after major news media projected a Biden/Harris victory). Each participant completed an **event expectation** task (“How likely do you think each candidate is to win?”) paired in random order with either a Cloze **production** task or a **comprehension** task using the A-Maze paradigm (Forster et al., 2009; Boyce et al., 2020). Following von der Malsburg et al. (2020), participants in the production component read a context sentence, shown in (1), and completed a partial version of one of 12 target sentences, exemplified in (2). Pre-election, “she” references were much rarer than “he” references (Fig 2) even though the female candidate was expected to win (Fig 1), but “she” references were numerically more frequent post-election (effect of round: $p < 0.05$). Also following von der Malsburg et al. (2020), half the participants in the comprehension component read (1) followed by two target sentences on the pattern of (3–4), each with a “he”, “she”, or “they” pronoun reference. At the first pronoun, “she” references elicited much slower RTs than “he” or “they” (pre-election); but post-election, “she” was read faster than “he” (Fig 3; all $p < 0.001$ except pairwise she/he post-election $p < 0.1$). Pronoun 2 results: “she” references have faster RTs post-election than pre-election, and *he*-references have slower RTs post-election than pre-election (interaction $p < 0.05$). In order to widen the scope of inquiry to mentions of the future VP’s race, half of the participants in the comprehension task were presented with either (5) or (6) after (1). We see an interaction between experimental round and mentioned race ($p < 0.01$), with faster RTs post-election to the word “black” than to the word “white” ($p < 0.05$), but no differences pre-election. Finally, following all comprehension components, participants indicated who they thought the writer would expect to become the next Vice President. “He” references yielded more “writer is unsure” responses than “she” references (Fig 6; $p < 0.05$), suggesting that comprehenders may be taking into account the production biases against “she” relative to the event expectations observed in Fig 1. **In conclusion**, this study reconfirms the large, persistent dispreference for using “she” pronouns in references to future office-holders even when explicit event expectations favor the female candidate. However, this dispreference can be rapidly reversed by sufficient changes in event expectations (here, the election outcome).

- (1) January 20, 2021, is Inauguration Day for the next term of the vice president of the United States.
- (2) Because the vice president breaks ties in the US Senate, if there is a 50–50 party split in 2021 then...
- (3) Because the vice president breaks ties in the US Senate, if there is a 50–50 party split in 2021 then **she|he|they** may cast many tie-breaking votes.
- (4) The vice president holds nuclear launch codes, which will be a great responsibility for **her|him|them** to carry as the second in command for the country.
- (5) The vice president will be **black|white|Black|White** and this is likely to be mentioned in discussions of US race relations.
- (6) The vice president will be a **black|white|Black|White** person and this is likely to be mentioned in discussions of US race relations.

[All error bars are standard errors of the mean.]

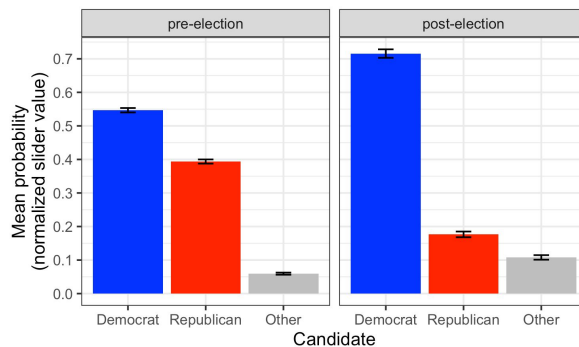


Fig 1: Event expectations

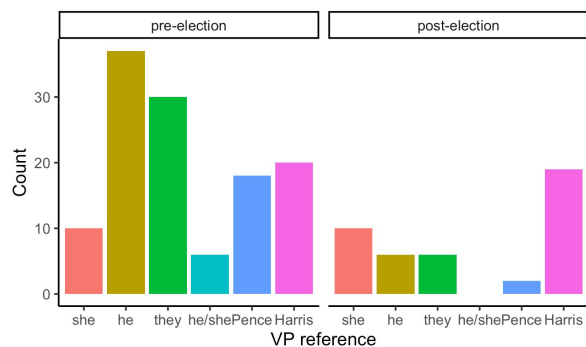


Fig 2: Cloze continuation VP references

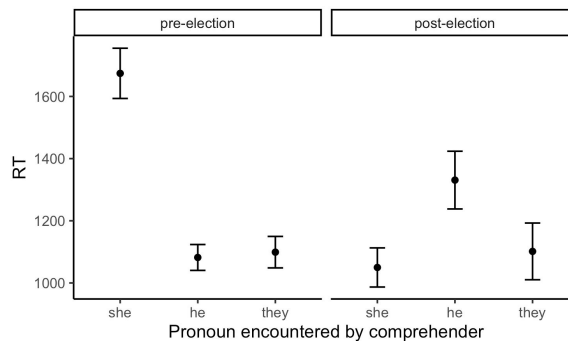


Fig 3: A-Maze RTs at pronoun 1

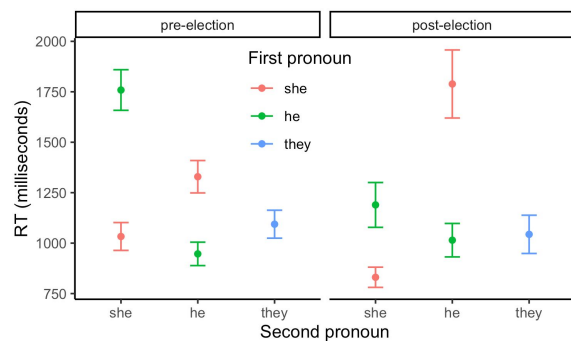


Fig 4: A-Maze RTs at pronoun 2

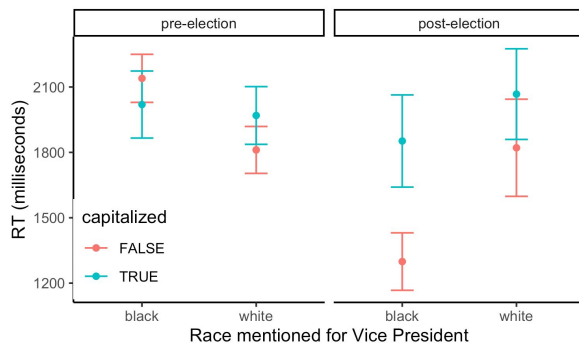


Fig 5: A-Maze RTs at mention of VP race

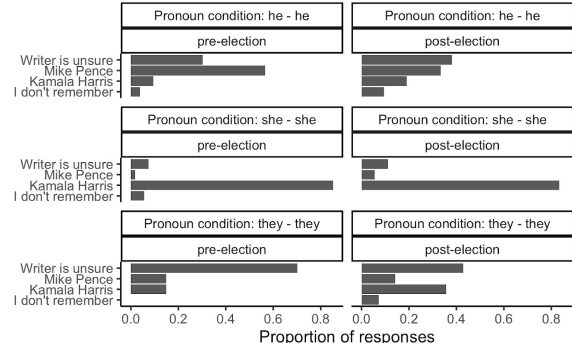


Fig 6: Inferred writer's expectations of next VP

The online application of structural and semantic biases during pronoun resolution

Markus Bader, Yvonne Portele (Goethe University Frankfurt)

Pronoun interpretation is known to be subject to semantic biases (e.g., implicit causality) and structural biases (e.g., subject bias). For p(ersonal)-pronouns, Koornneef and Van Berkum (2006) showed that implicit causality is used in a top-down fashion during on-line comprehension for predicting up-coming referents. In sentences as *Linda apologized to David because he according to the witnesses was not the one to blame.*, the bias-inconsistent pronoun *he* contradicts the gender of the predicted referent *Linda*, slowing down reading times. Anaphoric d(emonstrative)-pronouns in German are subject to the same semantic biases as p-pronouns, but have different structural preferences. P-pronouns have a moderate bias toward the subject, d-pronouns a strong bias toward the object (Authors, 2019; Patil et al., 2020). The main question of our experiments is whether the strong structural object-orientation of d-pronouns prevents the gender inconsistency effect (due to semantic biases) that Koornneef and Van Berkum (2006) found for p-pronouns.

We addressed this question in two self-paced reading experiments (word-by-word moving-window presentation). 20 sentences composed of a main and an embedded clause were created. The main clause contained an object-experiencer (OE) or a subject-experiencer (SE) verb (factor *Verb Type*). A pretest with a no-pronoun prompt confirmed a strong semantic bias toward the stimulus for both verb types. The embedded clause contained a p- or a d-pronoun (factor *Pronoun*).

Both experiments were presented on Ibx farm. In Experiment 1, sentences were presented in advance with characters replaced by understrikes. 97 participants recruited via Prolific read 20 sentences as well as 66 fillers. Results are shown in Figure 1. Accuracy on comprehension questions was higher for expected (SE verbs) than for unexpected continuations (OE verbs), especially when the question probed the embedded clause, with no difference between pronouns. Thus, the final interpretation was not affected by pronoun type. Reading times, however, showed a difference. Reading times on the complementizer *weil* were significantly faster for p-pronouns following a SE verb than for the remaining three conditions for which no further differences were significant. This replicates the gender inconsistency effect of Koornneef and Van Berkum (2006) for p-pronouns but at a position immediately preceding the pronoun. We hypothesize that this surprisingly early effect was caused by participants anticipating the upcoming pronoun from the word-length information given in the sentence preview. The male p-pronoun (*er*) is expected for SE verbs but the female p-pronoun (*sie*) for object-experiencer verbs. Thus, there is a match between expectation and preview information in the condition SE verb/p-pronoun, leading to fast reading times, but a mismatch in the remaining three conditions, resulting in increased reading times.

To corroborate the preview hypothesis, Experiment 2 (61 participants) was identical to Experiment 1, but sentences were no longer presented in advance by means of understrikes. Thus, preview information was not available while reading. Results for Experiment 2 are shown in Figure 2. Question accuracy was similar to Experiment 1, but reading times differed. On the complementizer, there was only a main effect of Verb Type. A significant interaction between Verb Type and Pronoun, however, was now visible on the pronoun and its spill-over region. A semantically unexpected p-pronoun caused longer reading times than an expected p-pronoun, whereas no significant difference was observed for d-pronouns. A comparison of Figures 1 and 2 reveals a similar overall pattern, except for the complementizer, which showed an interaction in Experiment 1 but not in Experiment 2. On the pronoun and its spill-over region, the interaction between Verb Type and Pronoun was significant in Experiment 2 but only numerically visible in Experiment 1.

In sum, our results replicate the top-down gender inconsistency effect for p-pronouns found by Koornneef and Van Berkum (2006). For d-pronouns, in contrast, no inconsistency effect showed up. We hypothesize d-pronouns to gain direct access to the object referent independently of the verb's semantic bias due to their strong structural preference.

Table 1: A complete stimulus for Experiment 1 and Experiment 2

Condition
<i>Object-experiencer verb: semantic bias toward the subject</i> Sabine beeindruckt den Fischer, weil er/dieser niemand anderen mit derart viel Erfolg kennt. “Sabine impresses the fisher because he/lit. this does not know anybody else with as much success.”
<i>Subject-experiencer verb: semantic bias toward the object</i> Sabine achtet den Fischer, weil er/dieser immer die bei weitem höchsten Fangzahlen aufweist. “Sabine respects the fisher because he/lit. this has by far the highest catch counts.”

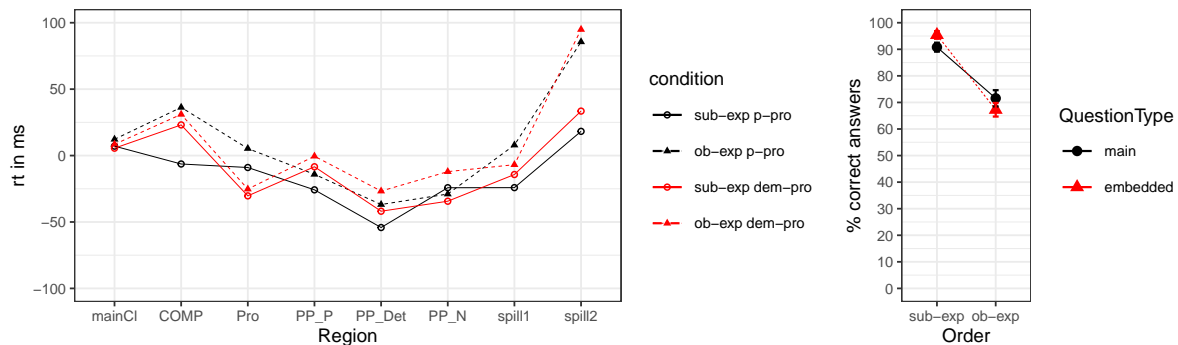


Figure 1: Residual reading times (left) & percentages of correct answers (right) in Experiment 1.

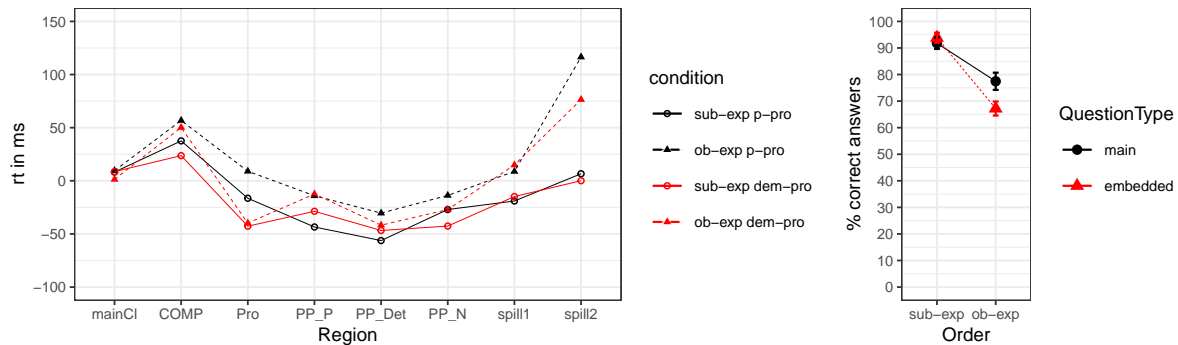


Figure 2: Residual reading times (left) & percentages of correct answers (right) in Experiment 2.

References

- Koornneef, A. W. and Van Berkum, J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54(4):445–465.
- Patil, U., Bosch, P., and Hinterwimmer, S. (2020). Constraints on German *diese* demonstratives: Language formality and subject-avoidance. *Glossa: a journal of general linguistics*, 5(1).

Singular vs. Plural *Themselves*: Evidence from the Ambiguity Advantage

Nicholas Van Handel, Lalitha Balachandran, Stephanie Rich, Amanda Rysling (UC Santa Cruz)

Background. Recent work has documented changes in the distribution of *they* and the antecedents to which *they* refers [1, 2]. Other work has investigated the processing of *they* and *themselves* with both singular and plural antecedents [3, 4]. In an eyetracking while reading study, [3] showed that *they* incurs a processing cost when its antecedent is singular (*someone*) rather than plural (*some people*). [3] proposed that *they* first initiates a search for a plural antecedent, and only accommodates a singular antecedent when no plural is found. [4] found that *themselves* elicits a P600 with singular antecedents that are gendered (*John*), but not with singular antecedents with ambiguous gender (*the participant*). [4] suggest that *they* is unspecified for gender, and the processing cost of singular *they* is due to a gender rather than number mismatch. However, studies have not examined the processing of *themselves* when both singular and plural antecedents are available in the same sentence. This configuration is necessary to test if *themselves* preferentially refers to plural antecedents, as proposed by [3].

Experiment. $n=57$; 12 observations/participant/condition. We extend previous work on the ambiguity advantage [5-7] to test if *themselves* first triggers a search for plural antecedents, as proposed by [3]. In sentences like those in Table 1, we disambiguate relative clause (RC) attachment height with the reflexive *themselves*. In AMBIG, both N1 and N2 are plural. In LOW, only N2 is plural, and in HIGH, only N1 is plural. Thus, if *themselves* first searches only for a plural antecedent, then the RC must attach to N2 in LOW and to N1 in HIGH. Previous work [5-7] has demonstrated an ambiguity advantage when RC attachment height is disambiguated by reflexive gender and semantic plausibility, i.e. reading times at the point of disambiguation were faster in AMBIG compared to when the RC must attach LOW or HIGH. We expect this same ambiguity advantage if *themselves* prioritizes plural antecedents. If, instead, singular and plural antecedents were treated equally by *themselves*, all three conditions would be ambiguous because the number of the antecedents would not force low or high attachment, and there should be no differences in reading times across conditions.

Method. Participants read sentences in the Maze task [8], in which participants are presented with two words at a time and must pick the word that forms a grammatical continuation with the preceding material in order to advance through the sentence. This task is thought to encourage incremental processing and more localized effects than self-paced reading.

Results. Reaction times are plotted in Figure 1. We fit a Bayesian linear mixed effects model [9] to RTs at the disambiguating reflexive and spillover prepositions. Attachment was coded into two contrasts: High Attachment (HIGH vs. AMBIG) and Low Attachment (LOW vs. AMBIG). No effect of Low Attachment was found at either reflexive or spillover, but English has a low attachment bias, so any cost of disambiguating to low attachment in the LOW condition would be small; this is not evidence against an ambiguity advantage. We found a main effect of High Attachment at the reflexive (66.29 ms, [39.00, 92.81]) and preposition (24.27 ms, [1.02, 46.43]). This is a clear replication of the ambiguity advantage: there was a processing cost when only N1 was plural. This cost indicates that the reflexive *themselves* does preferentially refer to plurals, forcing disambiguation to the dispreferred high attachment parse.

Discussion. We found evidence of an ambiguity advantage: participants spent more time reading *themselves* in HIGH compared to AMBIG conditions. This is only expected if *themselves* preferentially refers to plural antecedents, forcing high attachment in HIGH. This constitutes novel evidence for [3]'s proposal that *they* accommodates singular antecedents only when no plural is available. However, many nouns in our experiment were gendered, and [4] found that singular *themselves* is costly only when a singular antecedent is also gendered. It is thus possible that *themselves* does not prioritize plurals over non-gendered singulars. Follow-up work testing different antecedents in a retrieval interference paradigm is underway.

ATTACHMENT	[...] received a lot of media attention.
AMBIG(UOUS)	The partners _{N1} of the attorneys _{N2} who paid themselves from the settlement
LOW	The partner _{N1} of the attorneys _{N2} who paid themselves from the settlement
HIGH	The partners _{N1} of the attorney _{N2} who paid themselves from the settlement

Table 1. Sample item.

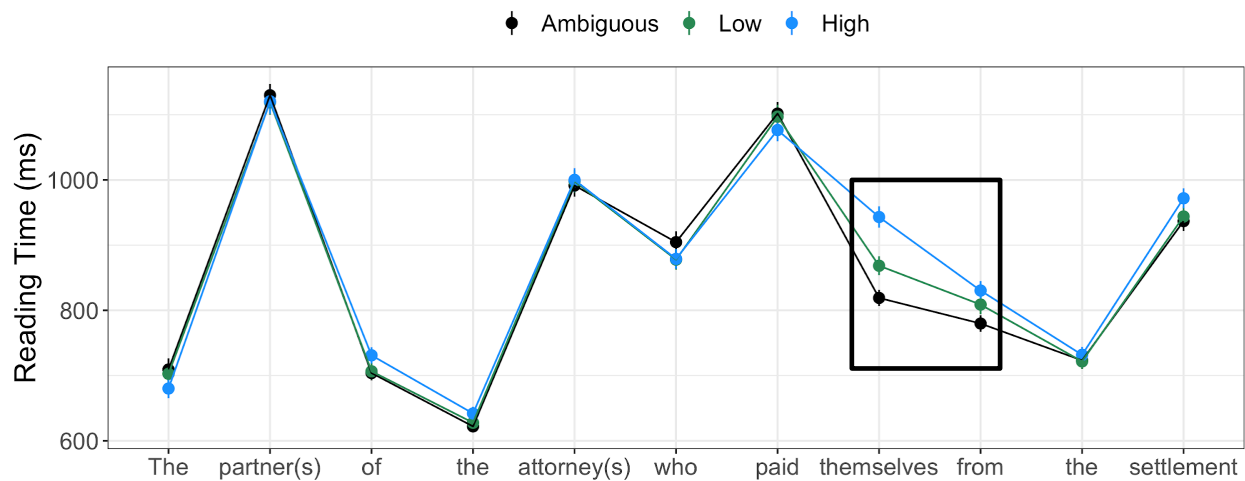


Figure 1. Mean reading times by word. Error bars indicate standard error of the mean.

References.

- [1] Bjorkman, B. (2017). *Glossa: A Journal of General Linguistics*
- [2] Conrod, K. (2019). Doctoral dissertation, University of Washington.
- [3] Sanford, A., & Filik, R. (2007). *Quarterly Journal of Experimental Psychology*
- [4] Prasad, G., Feinstein, M., & Morris, J. (2018). *Poster at the 31st CUNY Sentence Processing Conference*
- [5] Traxler, M., Pickering, M., & Clifton, C. (1998). *Journal of Memory and Language*
- [6] Van Gompel, R., Pickering, M., & Pearson, J., & Liversedge, S. (2005). *Journal of Memory and Language*
- [7] Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). *Memory & Cognition*
- [8] Forster, K. I., Guerra, C., & Elliot, L. (2009). *Behavior Research Methods*
- [9] Bürkner (2017). *Journal of Statistical Software*

Singular *they* in transition: ERP evidence and individual differences

Peiyao Chen (Swarthmore College), Olivia Leventhal (UCSD), Sadie Camilliere (University of Chicago), Amanda Izes (Hofstra University), & Daniel Grodner (Swarthmore College)

The English use of singular *they* to refer to a non-specific antecedent or an individual of unknown gender dates back to the 1300s [1]. Recently, *they* has emerged as a the most common personal pronoun for individuals who identify as gender nonbinary, and a coherent subset of English speakers will accept *they* when referring to a specific, antecedent of known gender (e.g., *Sarah*, slept because *they*, were tired.) [2,3]. Most research in this area adopts explicit offline judgments rather than online processing. The present work employs ERPs to examine the processing of nonbinary *they*, in comparison with binary gender pronouns. Gender mismatches such as *The boy thought that she would win the race* typically evoke a larger P600 than gender matches [4-7]. This P600 is thought to reflect the processes involved in diagnosing and attempting to repair a structural mismatch. Another component that can be elicited in this situation is an Nref, which has been argued to reflect extra work involved in either positing an unheralded referent outside the sentence [6] or linking the pronoun with a counter stereotypically gendered antecedent within the sentence [5].

The present work compared the processing of singular (*he/she*) and plural (*they*) pronouns that matched or mismatched the subject in the sentence. 120 items like (1) were constructed and pseudorandomly presented with 30 matching pronoun filler items using a Latin square design. Participants were 78 undergraduates attending a school where every student is introduced to preferred pronouns, taught about nonbinary gender identities, and encouraged to provide their preferred pronouns as part of orientation. They were told they were going to read sentences about named individuals who would be referenced with their preferred pronouns. The names were strongly associated with either male or female identities, which was established via a web-based survey on a separate group of participants. As an attention check, after each trial, participants were asked to identify the gender they would associate with each name. After the ERP study, participants completed a survey querying their attitudes towards and familiarity with transgenderism and nonbinary gender, as well as an acceptability survey of *they* with various antecedents. All analyses and the study design were preregistered.

Both mismatched singular pronouns and mismatched plural pronouns elicited a larger posterior positivity compared to their matched controls during the 450-1150 ms time window after the pronoun was presented (i.e., P600 effects). The mismatched singular pronouns also elicited a larger frontal negativity compared to matched controls in this window, consistent with an Nref effect. In contrast, the mismatched plural pronouns showed little or no reliable enhanced frontal negativity, which was confirmed by a cluster-based permutation analysis. These results replicate our previous finding with a smaller sample size ($n=21$) from the same population. This finding suggests that both types of mismatch triggered processing difficulty, but the mismatching singular pronouns also initiated additional referential work. Though robust for all groups, the P600 effect between the mismatched and matched plural pronouns decreased as participants' age increased. This could be because processing singular *they* becomes easier with increased exposure to it in a college environment. Intriguingly, offline acceptability judgments did not affect online ERPs. We compared 26 participants who were accepting of *they* with various singular named antecedents with 44 participants who rejected *they* in these contexts. These two groups did not show reliable differences in terms of their P600 and Nref effects. Thus, even individuals who are familiar with and robustly accepting of singular *they* exhibit difficulty processing it in comprehension. Importantly, this difficulty does not result in referential failure as it does for mismatched *he/she*. This work sheds light on the way in which the grammar of *they* is in transition. We see clear evidence for a coherent group of speakers who explicitly accept judgments of singular *they*. This group still exhibits implicit processing difficulty in online ERP measures. At the same time, this processing difficulty may be reduced for individuals with increased exposure to a non-binary accepting environment.

References

- [1] Balhorn, M. (2004) The rise of epicene they. *Journal of English Linguistics* 32(2), 79-104.
- [2] Camilliere, Izes, Leventhal & Grodner (2019). Multiple grammars for singular they. CUNY
- [3] Konnelly, L. & Cowper, E. (2019) The future is they: the morphosyntax of the epicene pronoun
- [4] Osterhout & Mobley. (1995). Event-Related potentials elicited by failure to agree. *JML*
- [5] Canal, Garnham & Oakhill (2015). Beyond gender stereotypes in language. *Frontiers*
- [6] Nieuwland, M. (2014). "Who's he?" Event-related potentials and unbound pronouns. *JML*
- [7] Prasad, Morris, Feinstein (2018) The P600 for singular 'they'. CUNY

(1) Sample item with critical pronoun in bold (actual stimuli were not bolded)

Matched Singular (MA_SI): *Lillian had just gotten back from vacation, so **she** felt exhausted.*

Mismatched Singular (MM_SI): *Lillian had just gotten back from vacation, so **he** felt exhausted.*

Matched Plural (MA_PL): *Lillian and Paul had just gotten back from vacation, so **they** felt exhausted.*

Mismatched Plural (MM_PL): *Lillian had just gotten back from vacation, so **they** felt exhausted.*

Figure 1. Scalp topographies in the 450-1150 ms time window of the comparisons between singular matched (MA_SI) and mismatched (MM_SI), plural matched (MA_PL) and mismatched (MM_PL), as well as singular and plural mismatched (MM_SI and MM_PL).

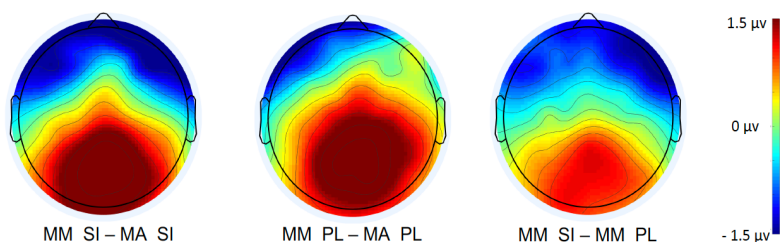


Table 1. By-subject and by-item analyses for the P600 effect. ($df_{F1} = 1,77$, $df_{F2} = 1,119$).

Comparison	450-750 ms	750-950 ms	950-1150 ms
MM_SI-MA_SI	$F_1 = 53.43^{***}$, $F_2 = 66.08^{***}$	$F_1 = 42.63^{***}$, $F_2 = 48.87^{***}$	$F_1 = 29.14^{***}$, $F_2 = 26.07^{***}$
MM_PL-MA_PL	$F_1 = 75.18^{***}$, $F_2 = 111.6^{***}$	$F_1 = 55.54^{***}$, $F_2 = 79.66^{***}$	$F_1 = 36.57^{***}$, $F_2 = 45.73^{***}$
MM_SI-MM_PL	$F_1 = 11.69^{**}$, $F_2 = 12.54^{***}$	$F_1 = 20.53^{***}$, $F_2 = 24.19^{***}$	$F_1 = 8.52^{**}$, $F_2 = 9.03^{**}$

*** $p < .001$, ** $p < .01$

Table 2. By-subject and by-item analyses for the Nref effect. ($df_{F1} = 1,77$, $df_{F2} = 1,119$).

Comparison	450-750 ms	750-950 ms	950-1150 ms
MM_SI-MA_SI	$F_1 = 14.86^{***}$, $F_2 = 25.32^{***}$	$F_1 = 3.54^{\dagger}$, $F_2 = 4.93^*$	$F_1 = 3.85^{\dagger}$, $F_2 = 4.67^*$
MM_PL-MA_PL	$F_1 < 1$, $F_2 < 1$	$F_1 < 1$, $F_2 < 1$	$F_1 = 1.39$, $F_2 = 1.74$
MM_SI-MM_PL	$F_1 = 12.44^{***}$, $F_2 = 20.47^{***}$	$F_1 = 3.71^{\dagger}$, $F_2 = 5.93^*$	$F_1 = 3.21^{\dagger}$, $F_2 = 4.07^*$

*** $p < .001$, ** $p < .01$, * $p < .05$, $^{\dagger}p < .08$

Mismatches in Subject-Verb Agreement: The Processing of Numeral Quantifiers in Turkish

Ayşe Gül Özyay-Demircioğlu (TED University)

Background. In Turkish, numerally quantified phrases in a subject position generally agree with 3SG verbs, but sometimes it is possible to see them agree with 3PL verbs, as in (1) (Göksel & Kerslake, 2004; Kornfilt, 1997).

- (1) Üç kişi gel-di-Ø / gel-di-ler.
three person come-PAST-3SG / come-PAST-3PL
'Three persons came.'

Match Condition

In my experiment, sentences as in (1) where the numerally quantified plural subject agrees with a third singular (3SG) or third plural verb (3PL) represent the so-called match condition because the number marking on the verb matches the features of the subject. Besides, Turkish allows a mismatch in person agreement when the verb agrees with a quantified subject. Therefore, the verb may show the first plural (1PL) agreement, and second plural (2PL) agreement, as in (2), which is called a mismatch condition in this study.

- (2) Üç kişi gel-di-niz/ gel-di-k.
three person come-PAST-2PL / come-PAST-1PL
'Three persons came.'

Mismatch Condition

The possible explanation of this variation in agreement is that a numeral phrase can agree with a 1PL and 2PL verb is the existence of a silent subject *biz* 'we'/*siz* 'you.PL', which controls the PRO subject of the adverbial clause headed by the converb *olarak* 'being/as', which are not present in the surface structure, as in (3) (Göksel & Kerslake, 2004; Özyıldız, 2017).

- (3) Buraya biz_i [PRO_i üç kişi ol.arak] gel-di-k.
here we PRO three person be.GER come-PAST-1PL
'We were three people to come here.'

Especially in Turkish, the possibility of various agreement patterns with numerally quantified subjects creates a necessity to test what is said in theoretical and empirical research perspectives. This study examines whether agreement mismatches with numerally quantified subjects (for example, the subject=3SG and the verb=1PL) harder to process than the absence of mismatches (subject=3SG and the verb=3SG). In predictive processing, speakers integrate what is seen and make predictions about what kind of structure will come next (Altmann & Kamide, 1999; Kaan, 2014; Levy, 2008). On seeing the numerally quantified subject, speakers expect 3SG or 3PL. If this expectation is not met and when they see 2PL or 1PL, speakers use the retrieval mechanism and reanalyze the whole structure. This reasoning underlies the design of the experiment with which I examined agreement mismatches in Turkish.

Methods. In present study, data were collected from 134 Turkish Learners of English via a self-paced reading task. To eliminate any effect of English on Turkish, English level was chosen as A1. The experimental items consisted of 24 items distributed across four lists with four conditions as in (4) and mixed with eight fillers. Every sentence consisted of seven regions as in (5). Although the agreement is on the verb, and so it is the critical region, but the verb differed in length. Therefore, Region 5 and 6 are taken as critical regions as in (5). Data was collected through Ixet Farm, an online platform used for online tasks. The experiment started with five practice items. Regarding analysis, 4 (Agreement) x 2 (Regions) Repeated Measures ANOVA and following post-hoc comparisons were conducted. The agreement variable had four levels (3SG, 3PL, 2PL, 1PL) and region one had two levels (Region 5 and Region 6). The purpose of this analysis was to discover which agreement interpretation of numerally quantified subjects is preferred most, as revealed by the speed with which participants read sentences with different agreement morphology on the verbs.

Results and Discussion. The Repeated-Measures ANOVA Analysis showed that there was a significant main effect of the agreement type [$F(1, 132) = 2.905, p = .008$], the region [$F(1, 132) = 129.32, p < .001, F(1, 20) = 112.20, p < .001$], as well as a two-way interaction between agreement type and region [$F(2, 264) = 7.62, p < .001, F(2, 20) = 4.5, p = .054$]. Bonferroni correction showed a significant difference between 3SG and 1PL agreement type ($p < .05$): 3SG verbs were slower to process than 1PL verbs as in (6). Also, it revealed that participants were significantly slower in 2PL condition than in 1PL condition ($p < .05$). Results indicated that the agreement mismatch, which does not cause ungrammaticality, does not lead to any extra processing load or any increase in the reading time. By contrast, mismatch one is actually preferred to match condition, contrary to the findings of previous literature, which indicated speakers' sensitivity towards agreement mismatches (Bock & Miller, 1991; Bock et al., 1999). Moreover, my findings indicate that 1PL is the most preferred agreement morphology with numerally quantified subjects contrary to the fact that syntactically simple structures are easier to process than syntactically complex structures (Kemper, 1987). As this is the case, I propose that the possible explanation for the absence of contrast between mismatch and match in the processing of agreement pattern may be that 3SG and other options are all equally complex because the underlying structure is the same across all agreement types as in (3).

- (4) **Sample item.** 4x2 design context (a: Third person singular, b: Third Person Plural, c: Second Person Plural, d: First Person Plural; a: Region 5, b: Region 6 (24 test items across 4 lists and 8 filler items).

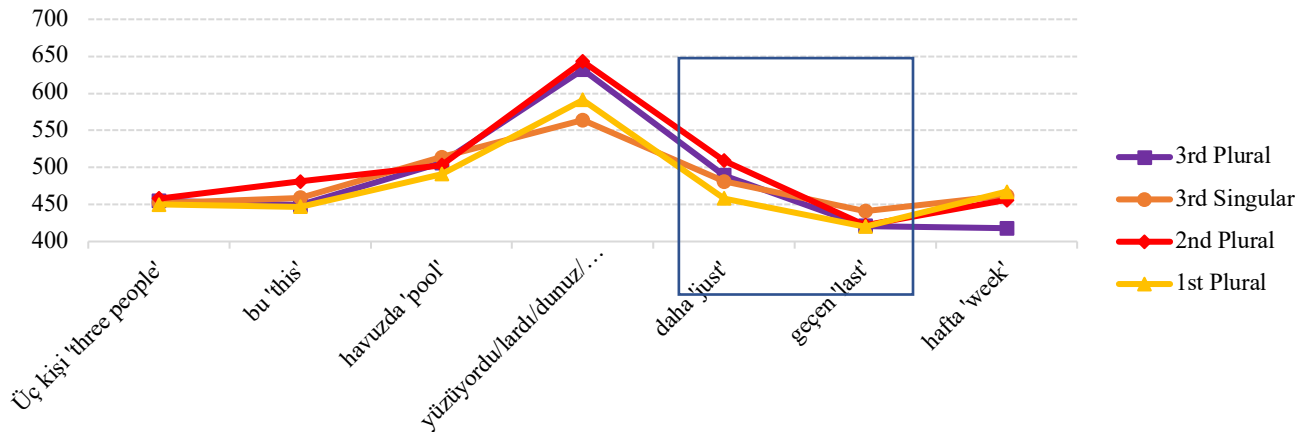
- a) Üç kişi bu havuz-da yüz-üyor-du daha geçen hafta. Third Person Singular
three person this pool-LOC swim-PROG-PAST-3SG just last week.
*‘Three person was swimming in this pool just last week.’
- b) Üç kişi bu havuz-da yüz-üyor-lar-dı daha geçen hafta. Third Person Plural
three person this pool-LOC swim-PROG-3PL-PAST just last week.
*‘Three person were swimming in this pool just last week.’
- c) Üç kişi bu havuz-da yüz-üyor-du-nuz daha geçen hafta. Second Person Plural
three person this pool-LOC swim-PROG-PAST-2PL just last week.
*‘Three person (you) were swimming in this pool just last week.’
- d) Üç kişi bu havuz-da yüz-üyor-du-k daha geçen hafta. First Person Plural
three person this pool-LOC swim-PROG-PAST-1PL just last week.
*‘Three person (we) were swimming in this pool just last week.’

Region 5 and Region 6
are spillover regions

(5) *Regions of the Items with Numerally Quantified Subjects*

R1	R2	R3	R4	R5	R6	R7
Üç kişi	bu	havuz-da	yüz-üyor-du	daha	geçen	hafta
Three person	this	pool- loc	swim-prog-past-3sg	just	last	week
‘Three persons were swimming in this pool just last week.’						

(6) Figure 1. Mean Reading Times for Every Region



References

- Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264. [https://doi.org/10.1016/s0010-0277\(99\)00059-1](https://doi.org/10.1016/s0010-0277(99)00059-1)
- Bock, J. K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23(1), 45-93. [https://doi.org/10.1016/0010-0285\(91\)90003-7](https://doi.org/10.1016/0010-0285(91)90003-7)
- Bock, K., Nicol, J., & Cutting, J. (1999). The ties that bind: Creating number agreement in speech. *Journal of Memory and Language*, 40(3), 330–346. <https://doi.org/10.1006/jmla.1998.2616>
- Göksel, A., & Kerslake, C. (2004). *Turkish: A comprehensive grammar*. Routledge.
- Kaan, E. (2014). Predictive sentence processing in L2 and L1. *Parsing to Learn*, 4(2), 257–282. <https://doi.org/10.1075/lab.4.2.05kaa>
- Kemper, S. (1987). Syntactic complexity and elderly adults' prose recall. *Experimental Aging Research*, 13(1), 47-62. <https://doi.org/10.1080/03610738708259299>
- Kornfilt, J. (1997). *Turkish*. Routledge.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177. <https://doi.org/10.1016/j.cognition.2007.05.006>
- Özyıldız, D. (2017). Quantification in Turkish. In D. Paperno & E. Keenan (Eds.), *Handbook of Quantifiers in Natural Language*. Vol. 2, (pp. 857-937). Springer.

Regional constructions still need learned after adaptation

Emily Atkinson & Julie Boland (University of Michigan)

Comprehenders unfamiliar with a syntactic structure that is acceptable in some regional dialects of American English, the *needs* + past participle construction (*The car needs washed*), adapt to this structure following exposure [1-3]. After exposure, comprehenders also generalize this adaptation to the construction in a new sentential context (e.g., *John thinks that what the meal needs is cooked*) [2], which some researchers claim entails having learned the *needs* construction. It is not clear, however, that these faster reading times reflect specific knowledge about the construction. Rather, comprehenders could be adapting to the presence of a set of related structures (i.e., the *needs* construction and *be*-dropping) [3] or perhaps 'odd' structures in general. Experiment 1 (*needs* construction) addresses the question of learning by directly probing the comprehenders' knowledge through acceptability judgments. Experiment 2 is a parallel experiment that uses the double modal regional construction (*You might could go there*), which heretofore has not been examined in this adaptation literature.

In each experiment, half of the participants were randomly assigned to an exposure group (control vs. regional dialect). The experiments proceeded in 2 phases: adaptation via self-paced reading then acceptability judgment. **Adaptation (self-paced reading):** Participants read 2 narratives (35 sentences) one word at a time. Interspersed were 15 target sentences (example target sentences in Table 1). Participants assigned to a regional exposure group read either *needs* constructions (Exp 1) or double modals (Exp 2). If they were assigned to a control group, they read versions of these sentences that are acceptable in standard American English (SAE). **Acceptability judgment:** As a test of their learning after the adaptation phase, participants rated 18 target sentences from each regional grammar on a 7-point Likert scale. Sentences were either acceptable in the regional grammar (*needs*: "These bills need paid"; double modal: "You might should eat"), acceptable in SAE (*needs*: "These bills need to be paid"; double modal: "You should eat"), or unacceptable in either (*needs*: "These bills need pay"; double modal: "You should might eat"). Filler sentences ($n=36$) were constructions accepted in other regional dialects. If participants in the exposure groups have learned, they should find sentences in the regional grammar that they were exposed to more acceptable than ungrammatical sentences; those in the control groups should treat both as equally ungrammatical. After the experiment, participants were asked to rate the variety of contexts in which they had experienced the relevant regional construction (6-point scale) as a measure of familiarity. All groups' average familiarity scores were less than 3 (*needs*: control=2.57, exposure=2.77; double modal: control=1.96, exposure=1.86).

Exp1: Needs ($N=48$) The construction is disambiguated at the verb, but the adaptation effects first appear in the first spillover region (verb+1) (Fig1). In this region, there were main effects of exposure (i.e., the *needs* exposure group was slower, $\beta=54.57, p<0.01$) and order (i.e., reading times decreased across the experiment, $\beta=-7.03, p<0.001$). Crucially, exposure and order interacted ($\beta=-3.53, p<0.05$), indicating that the *needs* exposure group adapted more dramatically than the control group. In the acceptability judgments (Fig1), participants rated the regional *needs* sentences higher than ungrammatical sentences, and much lower than SAE sentences, regardless of their exposure group ($\beta=0.18, p<0.01$).

Exp2: Double Modal ($N=48$) Again, the critical results for adaptation appeared at the region following the potential second modal (*look*), see Fig2. There is a main effect of order (i.e., reading times decrease across the experiment, $\beta=-6.73, p<0.001$) and an interaction of exposure and order ($\beta=-2.15, p<0.05$), which indicates that the double modal exposure group adapts more dramatically than the control group. Regardless of exposure, regionally grammatical constructions were not rated higher than ungrammatical sentences (Fig2, $\beta=0.04, p>0.1$).

In both experiments, participants exposed to a regional construction adapted to it, but did not demonstrate knowledge of that construction compared to the control groups. Implications for the interpretation of adaptation effects as learning will be discussed.

References [1] Kaschak & Glenberg 2004. *JEP: General*. [2] Kaschak 2006. *Memory & Cognition*. [3] Franundorf & Jaeger 2016. *JML*.

Table 1. Sentence examples from the self-paced reading portion of the experiments.

	Exposure Group	Control Group
Needs Construction	The dog will <u>need walked</u> in the morning.	The dog will <u>need to be walked</u> in the morning.
Double Modal Construction	I was thinkin' you <u>might could</u> look at it quick.	I was thinkin' you <u>might just</u> look at it quick.

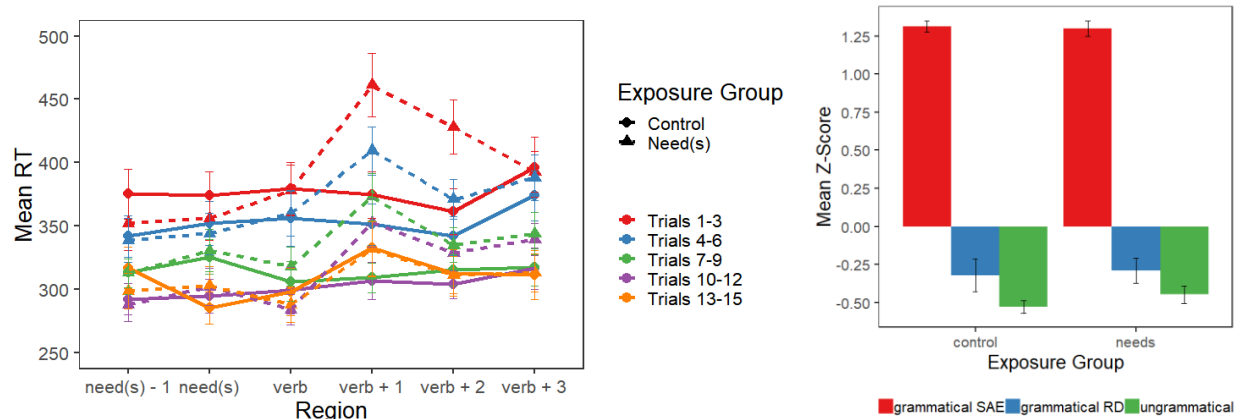


Figure 1. Results from the **needs construction** experiment. The self-paced reading results (left) include 3 critical regions: the verb and the following 2 regions. The acceptability judgment results (right) present mean z-scores.

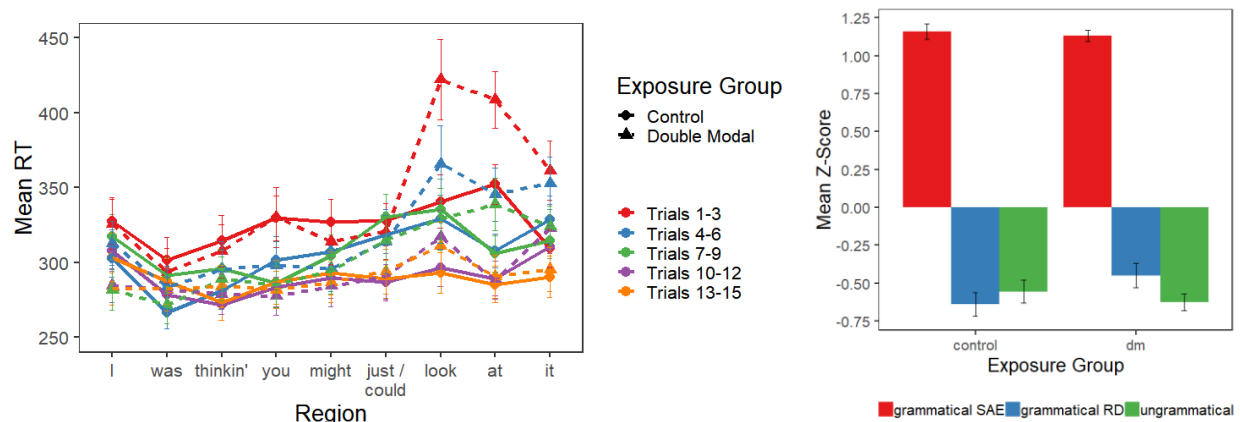


Figure 2. Results from the **double modal** experiment. The self-paced reading results (left) include 3 critical regions: the second modal (*just/could*) and the following 2 regions (spillover 1 = *look*; spillover 2 = *at*). The acceptability judgment results (right) present mean z-scores.

Understanding center embedding sentences: Can agreement and resumption help?

Hila Davidovitch, Maayan Keshev, Aya Meltzer-Asscher (Tel Aviv University)

Introduction. Center Embedding (CE) sentences, which consist of two nested object-relative filler-gap dependencies (e.g. 'The reporter who the senator who the professor met attacked resigned') are notoriously difficult to process (Chomsky & Miller, 1963; Baltin & Collins, 2008). Two main explanations have been offered for this difficulty. Gibson (1998) argues for prohibitively high integration costs at the second verb, exceeding the working memory capacity of most comprehenders. Lewis, Vasishth, & Van Dyke (2006) claim that the difficulty arises at retrieval: in the absence of sufficient cues, retrieval of the filler at the verb site fails due to the similarity between the different NPs, leading to interference.

The present study focuses on Hebrew CE sentences and examines whether their comprehension can benefit from the presence of (i) agreement features differentially marking the different NPs and identifying every verb's subject, and (ii) Resumptive pronouns (RPs, grammatical in Hebrew), which can aid retrieval by allowing more processing time and/or exhibiting the filler's agreement features, thus unambiguously identifying the verb's object.

Experiment 1 (160 participants; 8 sets + 24 grammatical filler sentences) used a comprehensibility rating task. It included four conditions crossing the factors DISTINCT AGREEMENT (agreement features on the three subject NPs and verbs are all identical vs. all different) and RESUMPTION (embedded verbs are followed by an RP or not). See Table 1 for sample materials. Sentences were presented in full. Participants read the sentences at their own pace and rated their comprehensibility on a 1-7 scale.

Results revealed that neither DISTINCT AGREEMENT nor RESUMPTION significantly affected comprehensibility. The interaction between the factors was significant ($p = .03$), signaling an advantage of distinct agreement only in the absence of resumption (Figure 1).

Experiment 2 (192 participants; 8 sets + 24 grammatical filler sentences) used end-of-sentence comprehension questions. Experimental sentences were of the same four conditions as in Experiment 1. The comprehension questions manipulated VERB POSITION, targeting either the first or second verbs' objects (see Table 1). Sentences were presented word by word at a rate of 400ms per word + 200ms inter-stimulus interval.

Results showed that DISTINCT AGREEMENT significantly improved comprehension ($p = .004$), while RESUMPTION did not. The interaction between the two factors was non-significant, i.e. in contrast to Experiment 1, here RPs did not reliably cancel out the advantage of distinct agreement. The results also revealed an effect of VERB POSITION ($p = .001$), such that the resolution of the dependency at the first, most embedded verb presented the most difficulty. The interaction between VERB POSITION and DISTINCT AGREEMENT was significant ($p = .001$), showing that while resolution of the dependency at the most embedded verb was not aided by distinct agreement, distinct agreement did aid the comprehension of the second verb (Figure 2).

Discussion. The results of Experiment 2 suggest that CE sentences are comprehensible to some extent, especially given aid by distinct agreement. It could be that by aiding to identify each verb's subject, distinct agreement also indirectly helps to identify the verbs' correct objects (targeted by the comprehension questions in Experiment 2). In contrast, resumption, though potentially identifying each verb's object unambiguously, did not help comprehension. These results suggest either that RPs are not used by comprehenders for retrieval, or that interference had arisen already during the encoding of the three similar NPs (Gordon, Hendrick, & Johnson, 2004; Villata, Tabor, & Franck, 2018), rendering the fillers not sufficiently distinct for successful retrieval at the verb.

Not only was resumption unhelpful, but it cancelled out the advantage offered by distinct agreement in Experiment 1. This finding can be explained similarly to the 'missing V2' illusion, the observation that center embedding is better accepted when only two of the verbs appear (Frazier, 1985; Gibson & Thomas, 1999). Gibson & Thomas suggest that in such cases one of the dependencies is compromised, thus concealing the processing difficulty. Adopting this idea, it can be assumed that resumption blocks the option to neglect one of the dependencies, leading to decreased ratings.

Different agreement features	<i>The child.</i> SG-M <i>that the neighbors.</i> PL-M <i>that the guest.</i> SG-F <i>frightened.</i> SG-F {Ø/them} <i>liked.</i> PL-M {Ø/him} <i>fell.</i> SG-M
Same agreement features	<i>The child.</i> SG-M <i>that the neighbor.</i> SG-M <i>that the guest.</i> SG-M <i>frightened.</i> SG-M {Ø/him} <i>liked.</i> SG-M {Ø/him} <i>fell.</i> SG-M
Comprehension questions: First verb's object: Who did the guest(SG-F/SG-M) frighten? The child / The neighbor (PL-M/SG-M) Second verb's object: Who did the neighbor(PL-M/SG-M) like? The child / The guest (SG-F/SG-M)	

Table 1. Sample sentences (Experiments 1 and 2) and comprehension questions and answer options (correct in bold) (Experiment 2), translated from Hebrew

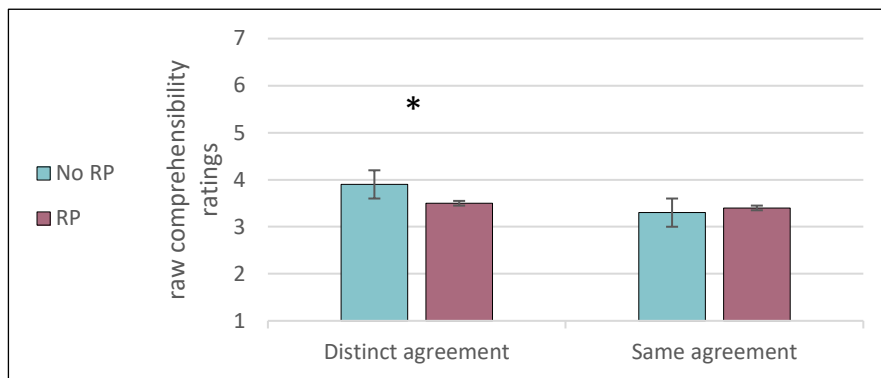


Figure 1. Results of Experiment 1. Error bars mark ± 1 SE; * represents $p < .05$; Analysis was conducted with a linear mixed-model regression.

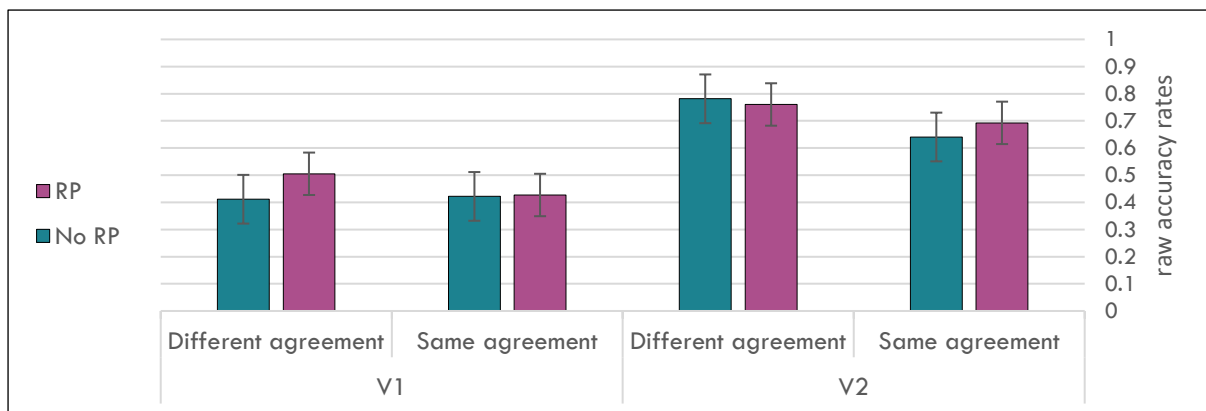


Figure 2. Results of Experiment 2. Error bars mark ± 1 SE; Chance level is 50%; Analysis was conducted with a linear mixed-model regression.

References

- Baltin, M., & Collins, C. (Eds.). 2008. *The handbook of contemporary syntactic theory*, vol. 23.
- Chomsky, N., & Miller, G. 1963. Introduction to the formal analysis of natural languages. In Luce, R.D., Bush, R.R., Galanter, E. (Eds.), *Handbook of Mathematical Psychology*, vol. 2, pp. 269–321.
- Frazier, L. 1985. Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, 129-189.
- Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1-76.
- Gibson, E., & Thomas, J. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3): 225-248.
- Gordon, P. C., Hendrick, R., & Johnson, M. 2004. Effects of noun phrase type on sentence complexity. *Journal of memory and Language*, 51(1): 97-114.
- Lewis, R., Vasishth, S., & Van Dyke, J. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10: 447-54.
- Villata, S., Tabor, W., & Franck, J. 2018. Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in psychology*, 9, 2.

When singular morphology meets notional plurality: another puzzle for agreement

Martina Abbondanza, Francesca Foppolo (University of Milano-Bicocca)

Subject-verb agreement reveals interesting phenomena of interference or attraction in both production and comprehension, as documented by several studies in many languages (since Bock & Miller, 1991). Agreement variability has been documented also for coordinated phrases (Keung & Staub, 2018; Foppolo & Staub, 2020). Different explanations have been advocated to explain speakers' errors or listeners' preferences in subject/verb agreement. One processing explanation is the Marking and Morphing model (Bock et al, 2001), according to which semantic features are assumed to impact the agreement process in production prior to morphosyntax. Extending this account, the *self-organized sentence processing* model (Smith et al. 2018) explains the variability in agreement in production and comprehension as the result of a dynamic interplay between semantics and syntax. **Our study.** To explore the dynamic interplay of semantics and morphosyntax, we tested conjunctive subjects containing notionally plural, but morphologically singular/plural quantifiers in Italian followed by either a singular or a plural verb. We present two experiments. **Experiment 1** (N=42) was an acceptability judgement task (on a 7-point Likert scale) on sentences containing a conjunction of quantified nouns in a latin-square design consisting of 2 (quantifier) x 2 (verb number) conditions, 24 items each (Table 1): the quantifiers were notionally plural in all conditions but they were morphologically singular (*ogni/qualche*) in conditions A-B and morphologically plural (*tutti/alcuni*) in conditions C-D; they were followed by either singular (A-C) or plural (B-D) verbs. Condition C was expected to be the most natural and acceptable, while D was expected to be unacceptable. The critical conditions were A and B, in which the quantifiers' morphology was singular. **Hypotheses.** If notional plurality takes precedence over morphological agreement, we predict higher judgments for A and C sentences, in which the verb is plural, compared to B and D sentences, in which the verb is singular. If morphosyntax overrides notional plurality, we predict an asymmetry between A-B sentences (in which the quantifiers are morphologically singular) compared to C-D sentences (in which the quantifiers are morphologically plural) when these are followed by a singular or plural verb. **Results.** Results showed that C and D received the highest and lowest ratings, respectively (Figure 1). We set contrasts to compare the conditions in a CLMM with the package "ordinal" in R (Christensen, 2019): while in A-B sentences: (i) the mean ratings of B, in which notionally plural/morphologically singular quantifiers were followed by a singular verb, were significantly higher than the ratings of D; (ii) the mean ratings of A, in which notionally plural/morphologically singular quantifiers were followed by a plural verb, were significantly lower than the mean ratings of C (Table 2). **Experiment 2** tested the same sentences in a self-paced reading task in a different group of participants (N=82). Singular/plural agreement always appeared on the auxiliary of the verb followed by a past participle. **Results.** Longer RTs were recorded in D (Figure 2). The interaction between subject morphology and verb agreement significantly predicted RTs ($t=-3.1$, $p=0.002$). We then ran a linear mixed-effect model on log-transformed RTs on the auxiliary and the past participle that immediately followed, including Condition type as the dv and subject and items as random intercepts. Results confirmed the findings of Experiment 1, showing that RTs on condition D were significantly longer than those in condition B ($t=4.3$, $p<.0001$). RTs in condition C were faster than RTs in condition A ($t=-2.2$, $p=0.03$) and, remarkably, RTs in Condition A and in Condition B were not significantly different ($t=1.8$, $p=0.08$). **Conclusions.** (i) neither singular nor plural verbs are considered optimal in the case of conjoined morphologically singular quantifiers; (ii) no disruption is revealed when a singular verb follows notionally plural subjects if this is morphologically singular. These findings show that notional plurality does not take precedence over morphosyntax in subject-verb agreement, suggesting a more dynamic interplay between semantics and morphosyntax in agreement phenomena.

Table 1. Conditions involved in the study. The English translation of the sentences is: For security reasons, all mechanic(s) and some engineer(s) has/have inspected the airplane prior departure.”

Condition	Example	Quantifiers' morphology	Verb number
A	<i>Per sicurezza, ogni meccanico e qualche ingegnere hanno ispezionato l'aereo prima della partenza.</i>	sing	plur
B	<i>Per sicurezza, ogni meccanico e qualche ingegnere ha ispezionato l'aereo prima della partenza.</i>	sing	sing
C	<i>Per sicurezza, tutti i meccanici e alcuni ingegneri hanno ispezionato l'aereo prima della partenza.</i>	plur	plur
D	<i>Per sicurezza, tutti i meccanici e alcuni ingegneri ha ispezionato l'aereo prima della partenza.</i>	plur	sing

Table 2. Output of the Cumulative Link Mixed Model (CLMM) of experiment 1 with the acceptability ratings as dependent variable, sentence type as predictor and subjects and sentences as random intercepts. Contrasts were set as follows: contrast $<- \text{cbind}(c(-0.5, 0, +0.5, 0), c(0, -0.5, 0, +0.5), c(-0.5, +0.5, 0, 0))$. We checked for a possible influence of the word-length of the auxiliary (ha/hanno) adding word-length as covariate in the model and it did not affect the results.

	Estimate	Std. Error	z value	P value
A compared to C	4.7	0.1	45.9	<.0001
B compared to D	-4.8	0.1	-45.8	<.0001
A compared to B	-5.8	0.1	-48.8	<.0001

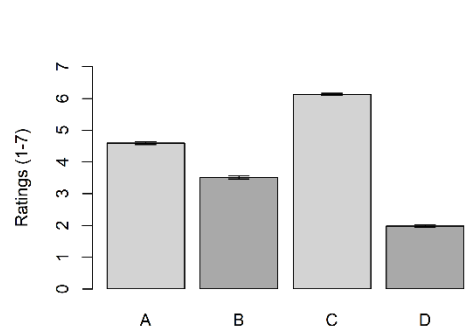


Figure 1. Bar plot showing the mean values of the acceptability rating (Experiment 1)

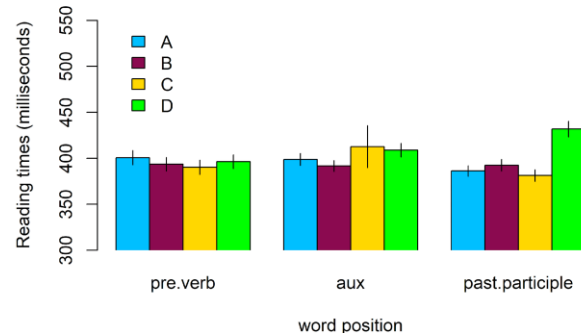


Figure 2. Reading times on pre-verb, auxiliary, and past participle in Experiment 2.

Selected References

- Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, 23, 45-93.
- Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, 43(2), 83-128.
- Christensen, R.H.B (2019). Ordinal—Regression Models for Ordinal Data. R package version 2019.12-10.
- Foppolo, F., & Staub, A. (2020). The puzzle of number agreement with disjunction. *Cognition*, 198, 104161.
- Keung, L., & Staub, A. (2018). Variable agreement with coordinate subjects is not a form of agreement attraction. *Journal of Memory and Language*, 103, 1-18.
- Smith, G., Franck, J., & Tabor, W. (2018). A Self-Organizing Approach to Subject-Verb Number Agreement. *Cognitive Science*, 42, 1043-1074.

Distribution matters: change in relative frequency affects syntactic processing

Valerie J. Langlois & Jennifer E. Arnold (University of North Carolina – Chapel Hill)

valeriel@live.unc.edu

Comprehenders encounter a variety of syntactic structures in everyday life, whether through reading or spoken conversation. Some theoretical models of syntactic processing claim that comprehenders can acquire the frequency statistics of syntactic structures from exposure, which in turn leads to syntactic expectations (Levy, 2008; MacDonald et al., 1994; MacDonald & Thornton, 2009). These models imply that not only do comprehenders have implicit statistical knowledge of the relative frequencies of syntactic structures given a verb, but also that they can adapt to distributional changes. Infrequent structures (e.g. reduced relative clauses, such as *The soldiers warned about the dangers conducted the raid*) impose more difficulty as measured by reading time (MacDonald et al., 1994), and previous work has shown that this difficulty decreases with repeated exposure (Fine et al., 2013; Wells et al., 2009). Yet theoretically, processing is specifically impacted by distribution – i.e., the relative frequency of target and competing structures. But the role of distribution has only been investigated by correlating data from corpora and reading times (Gennari & MacDonald, 2009). In the current study we provide the first experimental test of whether comprehenders keep track of the distribution of syntactic structures.

We investigate whether comprehenders acquire syntactic distributional information by directly manipulating the relative frequency of two competing syntactic structures: the dialectal *needs* structure and the modifier structure (Table 1). The dialectal *needs* structure is unfamiliar to most people (apart from those in Western Pennsylvania; Murray et al. 1996). Despite this unfamiliarity, comprehenders can rapidly adapt to the dialectal structure with enough exposure (Fraundorf & Jaeger, 2016; Kaschak & Glenberg, 2004). Critically, both structures are syntactically ambiguous until two words after *needs*. If comprehenders implicitly keep track of the distribution of structures that co-occur with *needs*, then a distribution with a higher proportion of dialectal *needs* structures should result in less processing difficulty during disambiguation, independent of overall exposure.

Methods: We used a 2x2 between-subjects design with two distributions and an ambiguous and unambiguous condition. 233 participants were assigned to one of two distributions (80-20 vs. 40-60) with either the dialectal structure or the standard structure. The numbers in each distribution represent the relative percentages of the two syntactic structures (dialectal/standard and modifier structure respectively). In the 80-20 distribution, participants completed a self-paced reading task in which they read 20 target *needs* (80%), 5 modifier structures (20%), and 55 unrelated fillers. Likewise, in the 40-60 distribution, participants read 20 target *needs* (40%), 30 modifier structures (60%), and 30 unrelated fillers. Modifier structures were presented at specific timepoints in the experiment, so that at any given target structure, the distribution of target to modifier sentences would be as close to the target distribution as possible. At the end of the sentence, participants answered one comprehension question to ensure they read the sentence. Notably, participants in both distributions read precisely the same number of target structures in the same order. Thus, if mere exposure drives facilitation, no difference is expected across the ambiguous conditions. In contrast, if comprehenders track the distribution of the dialectal and modifier structures, then there should be a difference even when controlling for overall exposure.

Results: Reading times were corrected for word length, baseline reading speed, and task adaptation. The target *needs* structures were analyzed at the same word (e.g. *before*). There was a significant three-way interaction between distribution, ambiguity, and order ($p < .05$, Fig.1); Reading times for the disambiguating word decreased faster for the ambiguous 80-20 condition than the 40-60 condition ($p < .05$), but not for the unambiguous conditions ($p = .93$).

Conclusion: A higher proportion of dialectal *needs* sentences led to a faster rate of syntactic adaptation, independent of overall exposure. The difference in reading rate across the two distributions in the ambiguous condition suggests that comprehenders are sensitive to the change in distribution. This shows that comprehenders can acquire syntactic distributional information, consistent with experience-based models of syntactic processing (e.g. MacDonald et al., 1994).

Table 1: Example sentence for each structure.

Dialectal structure:	The fire needs stoked <u>to</u> keep it from burning out.
Standard structure:	The fire needs <i>to be</i> stoked <u>to</u> keep it from burning out.
Modifier structure:	The meal needs cooked <u>vegetables</u> so the guests will be happy.

Table 2: Summary of model results at the critical word (e.g. *to*) for target structures.

Model Parameters	Estimate	Std. Error	df	t-value	p-value
Intercept	10.546	4.402	27.329	2.396	0.02369
Distribution (80-20 vs. 40-60)	5.515	5.227	220.71	1.055	0.29256
Order	-17.117	5.547	18.022	-3.086	0.00637
Ambiguity (1 vs. 0)	31.078	5.227	220.7	5.945	< .001
Distribution*Order	-7.027	4.914	4155.581	-1.43	0.15285
Distribution*Ambiguity	-4.379	10.455	220.714	-0.419	0.67571
Ambiguity*Order	-10.738	4.914	4155.633	-2.185	0.02894
Distribution*Order*Ambiguity	-22.237	9.829	4155.704	-2.262	0.02373

Order of presentation (log-transformed) was regressed out in the length-corrected reading time model and centered in the final model. Distribution and ambiguity were contrast-coded, with 40-60 and ambiguity=0 as the reference level.

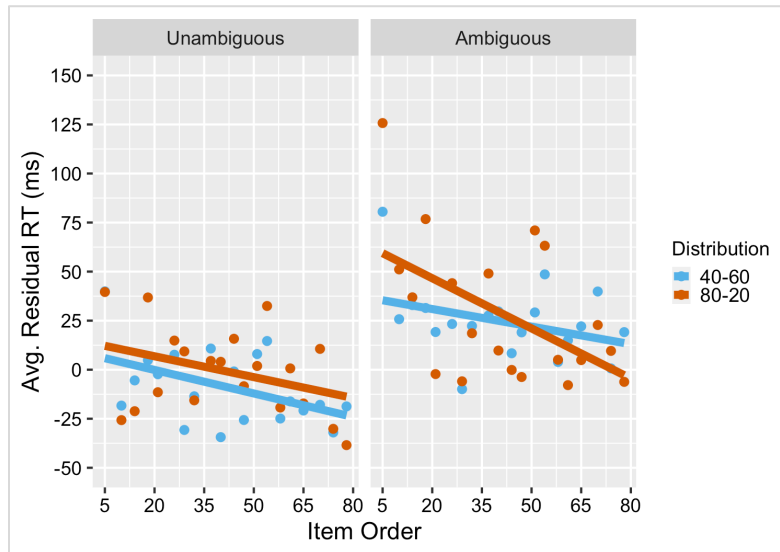


Figure 1: Average residual RT during the disambiguating region over the course of the experiment, broken down by distribution and ambiguity condition.

References: Fine et al. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, 8(10). • Fraundorf & Jaeger (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *Journal of Memory and Language*, 91, 28–58. • Gennari & MacDonald (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, 111(1), 1–23. • Kaschak & Glenberg (2004). This construction needs learned. *Journal of Experimental Psychology*, 133(3), 450–467. • Levy (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177. • MacDonald et al. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4), 676–703. • MacDonald & Thornton (2009). When language comprehension reflects production constraints: Resolving ambiguities with the help of past experience. *Memory and Cognition*, 37(8), 1177–1186. • Wells et al. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58(2), 250–271.

Cognitive Control and Ambiguity Resolution: Beyond Conflict Resolution

Varvara Kuz, Keyue Chen, Clement Veall, Andrea Santi (UCL)

Cognitive control is a multi-layered function involved in highly demanding, goal-oriented behaviours, including the processing of garden path sentences. In the ambiguous version of (1), cognitive control is argued to facilitate the switch from the preferred main clause interpretation of **'fed the hot dogs'** to the less preferred relative clause interpretation at the disambiguating *'got a stomach ache'*.

(1) The sunburned boys (that were) **fed the hot dogs** *got a stomach ache*.

Causal evidence for this has been provided with the visual world paradigm [1]. Participants completed a Stroop task (congruent/incongruent) before hearing a sentence like (2) (ambiguous/unambiguous) that they acted out with objects in the visual display.

(2) Put the frog (that is) on the napkin into the box.

In the ambiguous condition, there were fewer incorrect goal actions and less looks to the incorrect goal when the sentence was preceded by an incongruent Stroop condition (compared to congruent). For unambiguous sentences, there was no effect of the preceding Stroop condition. Hsu and Novick [1] argue that the incongruent Stroop condition activates a conflict resolution mechanism that is sustained and facilitates syntactic reanalysis.

To generalise this finding to syntactic processing that is independent of a visual context we used a similar interleaved Stroop-Sentence design, but with self-paced reading and sentences like (1). Fine and Jaeger [2] found an ambiguity effect (ambiguous > unambiguous) in self-paced reading times at both ambiguous **'fed the hot dogs'** and disambiguating *'got a stomach'* regions. Based on [1], we predicted the incongruent Stroop condition would reduce the ambiguity effect at the disambiguating region, but not the ambiguous region, where all information is compatible (i.e., not conflicting) with the preferred interpretation of the verb.

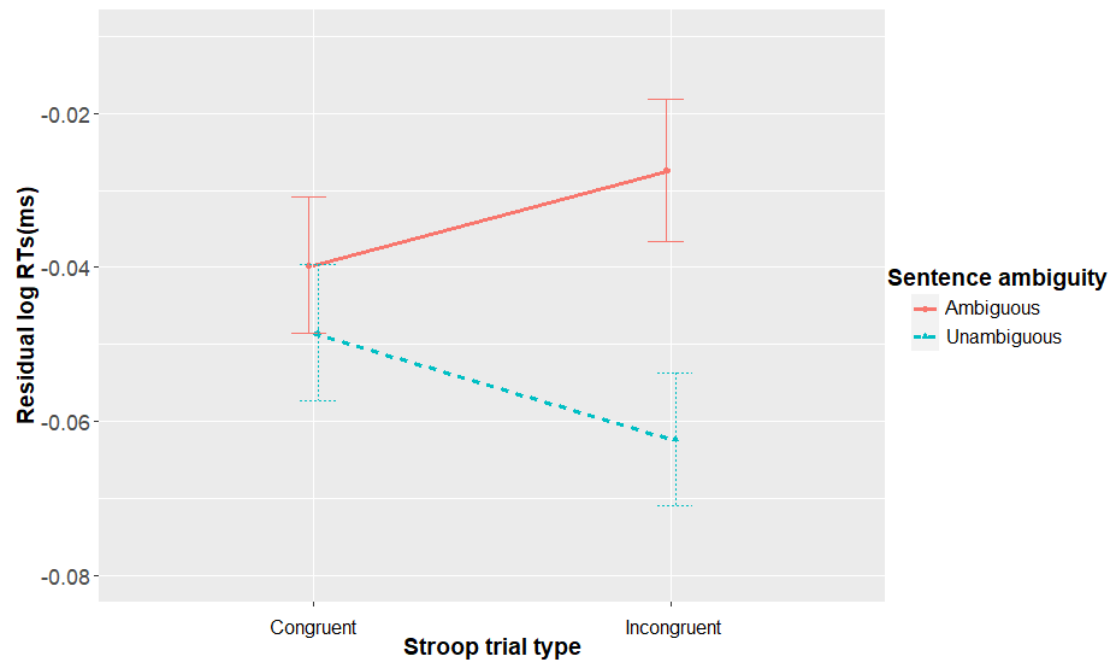
Method & Results: Stroop (congruent, incongruent) and Ambiguity (ambiguous, unambiguous) were crossed in 36 items (N= 96 native English speakers recruited via prolific.co). A Stroop task was followed by self-paced sentence reading and a yes/no comprehension question (see Figure 2). These were presented with filler items (54 sentence-Stroop and 18 Stroop-sentence) in a pseudorandomised order. Data was analysed using linear mixed effects models.

We replicated the ambiguity effect at ambiguous and disambiguating regions [2] (see Table 1). Contrary to expectation, we did not observe a Stroop x Ambiguity interaction at the disambiguating region ($p=.64$). Potentially the lag between Stroop completion and disambiguation was too long for sustained cognitive control [3]. Critically, however, we observed an interaction at the ambiguous region ($t=-2.15$, $p<.05$; see Figure 1, Table 1), where the standard ambiguity effect was present when the preceding Stroop was incongruent ($t=4.29$, $p<.001$), but eliminated when congruent ($p=.275$).

Conclusions:

This is the first study to demonstrate that congruent, or low conflict, trials can eliminate the ambiguity effect standardly observed at the ambiguous region. Contrary to previous findings [1], Stroop did not affect processing of directly conflicting parses at disambiguation, but the consideration of potential parses at the ambiguous region. While a conflict resolution mechanism seems necessary, it is insufficient to explain this transfer effect at ambiguous region. Like work outside language processing, that has also found processing adaptation from a congruent Stroop condition [4], we suggest attentional mechanisms to underlie our transfer effect. This gives rise to interesting new avenues to explore the interaction between attentional mechanisms and sentence processing in future work.

Figure 1. Residual log reading times at ambiguous region.



Note: The figure illustrates mean residual log reading times with 95% confidence intervals.

Figure 2. Trial dynamics.

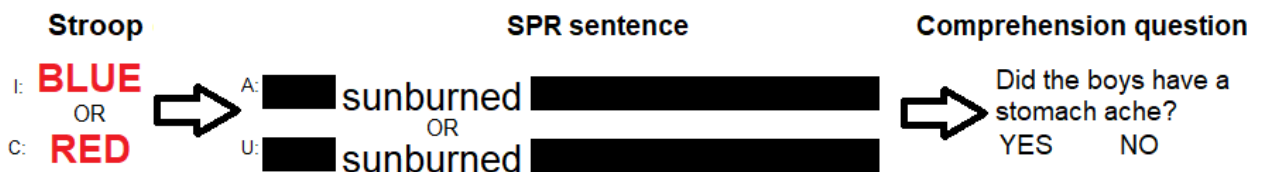


Table 1. Results of fixed effects at the three regions of analysis.

	Ambiguous region RT		Ambiguous verb only RT		Disambiguating region RT	
Fixed effects	β	t	β	t	β	t
Ambiguity	0.014	2.95	0.034	6.69	0.018	4.21
Stroop	0.001	0.31	-0.003	-0.75	-0.002	0.62
Stroop x Ambiguity	-0.006	-2.15	-0.012	-2.69	-0.001	-0.47

Note: significant effects ($p < .05$) are marked in bold.

References

- [1] Hsu, N. S., & Novick, J. M. (2016). *Psychological science*, 27(4), 572-582.
- [2] Fine, A. B., & Jaeger, T. F. (2016). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1362.
- [3] Egner, T., Ely, S., & Grinband, J. (2010). *Frontiers in psychology*, 1, 154.
- [4] Lamers, M. J., & Roelofs, A. (2011). *Quarterly Journal of Experimental Psychology*, 64(6), 1056-1081.

Six-month-old infants' abilities to represent regularities in speech

Irene de la Cruz-Pavía (University of the Basque Country, Basque Foundation for Science Ikerbasque) and Judit Gervain (University of Padova, CNRS-Université de Paris)

Introduction. In order to acquire grammar, infants need to extract regularities from the linguistic input. From birth, infants can detect certain regularities from speech, notably repetitions. Thus, newborns show strong neural activation (as compared with a silent baseline) for syllable sequences that contain adjacent repetitions (ABB: *mubaba*). Meanwhile, their activation in response to random syllable sequences (e.g. ABC: *mubage*) is very weak (Gervain et al. 2008, *PNAS*), and does not differ from their response to the silent baseline. Here, we seek to uncover when in development infants begin to also represent sequences containing a diversity-based rule — such as the random sequences — as strongly as sequences containing a repetition-based rule. We examine thus 6-month-old-infants' abilities to represent the two types of structures. As infants begin to learn their first word forms at this age, we hypothesize that the ability to represent sequences of different syllables might become important for them.

Methods. We used NIRS to examine whether 6-month-old French learning infants' (n = 24) representation of repeated and random sequences in speech. We presented infants with Gervain and colleague's (2008) original materials (ABB vs. ABC: *mubaba* vs. *mubage*), and measured, using a NIRx NIRScout system, infants' brain responses in the bilateral temporal, parietal and frontal areas, that is, in the brain network known to be involved in language processing in adults and infants (10 channels/hemisphere). Procedure consisted of an alternating/non-alternating design (see Figure), a paradigm used extensively in developmental NIRS to test discrimination. In this design, infants listen to two types of blocks. Alternating blocks contained tokens of the two types of structures presented in strict alternation (6 blocks: half ABB-ABC, e.g. ABB-ABC: *talulu*_{ABB1} *zimuta*_{ABC1} *toffi*_{ABB2} *dufeto*_{ABC2}..., the remaining half ABC-ABB). In turn, non-alternating blocks contained tokens of a single structure (6 blocks: half only ABB, e.g. ABB: *dufefe*_{ABB1}, *fibaba*_{ABB2}, *zepipi*_{ABB3}, *lokuki*_{ABB4}..., the remaining half ABC). If infants discriminate both types of structures, they are expected to exhibit different neural activation in response to the alternating and non-alternating blocks. Blocks with artifacts in the signal were discarded, and we averaged responses across the remaining blocks of each condition.

Results & discussion. Using cluster-based permutation tests we examined infants' brain activation in response to the alternating and non-alternating blocks, and found an advantage for non-alternating blocks in right frontal regions. This result shows that the 6-month-old infants discriminated the two sequence types. Crucially, analysis of only non-alternating blocks revealed equally strong neural activation to the blocks containing only ABB or only ABC tokens, higher than during the silent baseline. That is, while newborns show high activation only in response to repetition-based structures (i.e. ABB), 6-month-old infants show high activation in response to repetition- and diversity-based structures (i.e. ABC).

This finding contrasts with infants' failure to detect diversity-based rules even at 12 months of age in behavioral studies (Kovács, 2014). Our results provide thus the earliest evidence that young infants encode diversity-based patterns, i.e. represent difference, in speech. This research has important implications for language development, furthering our knowledge of infants' processing of rules in linguistic stimuli.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *PNAS*, 105(37), 14222-14227.

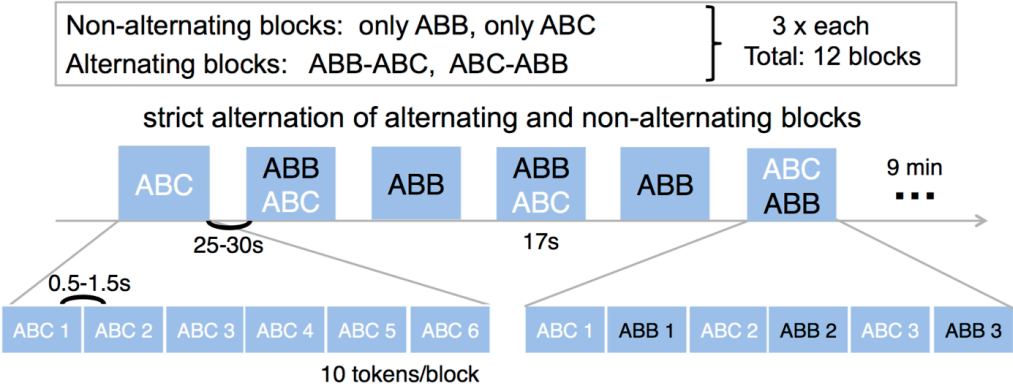
Kovács, Á. M. (2014). Extracting Regularities From Noise: Do Infants Encode Patterns Based on Same and Different Relations? *Language Learning*, 64, 65–85.

Figure. Stimuli (A), procedure (B), and layout of the regions measured and channels showing significant differences between alternating and non-alternating blocks (C)

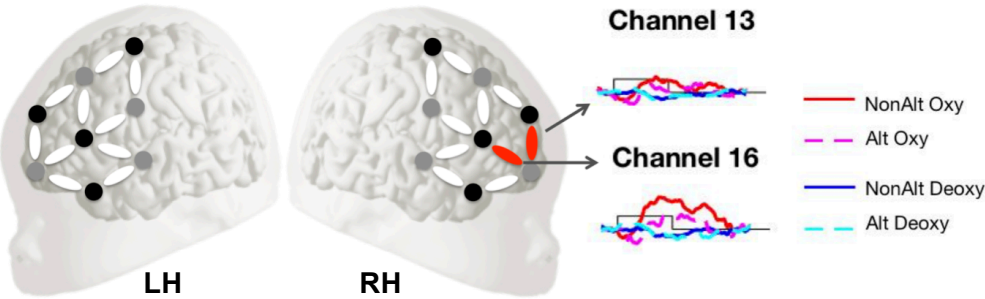
A. Stimuli

ABB	ABC
tobibi	tobisha
lukoko	lukobi
bazeze	bazeko
kushasha	kushape
fetata	fetamu
... (x60)	...(x60)

B. Alternating/non-alternating design



C. Layout of the regions measured and results



The newborns' brain detects utterance-level prosodic contours

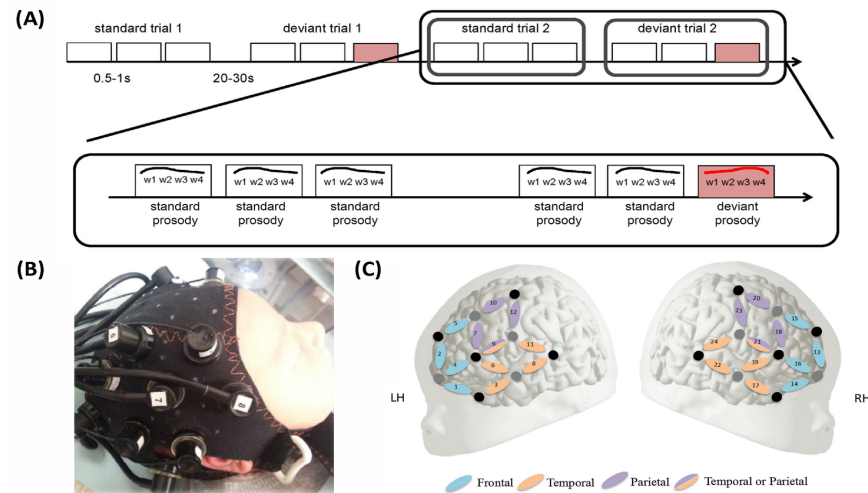
Martinez-Alvarez, Anna¹, Benavides-Varela, Silvia¹ & Gervain, Judit^{1,2}

¹ University of Padova

² CNRS- Université Paris Descartes

Introduction. Infants' prenatal experience with speech mostly entails prosody due to the filtering effect of the mother's womb (Gerhardt et al. 1990). How prosody perceived in utero influences early speech perception remains to be understood. In this study, we asked whether it allows infants to recognize and discriminate utterance-level prosodic contours at birth. *Methods.* The current study investigated this question using near-infrared spectroscopy (NIRS) in 1-5-day-old French-exposed newborns ($n=25$). We used a paradigm (Figure 1) similar to the newborn NIRS study of Benavides-Varela & Gervain (2017) testing newborns' ability to detect word order violations in the absence or presence of prosody. We used the same 4-word-long ungrammatical sequences as utterances (e.g. *et appelle de aller*) as in Benavides-Varela & Gervain (2017). Ungrammatical sequences were chosen to avoid potentially familiar word combinations biasing infants' performance. The sequences were recorded with well-formed declarative utterance prosody by a professional actress. (Trained speakers are able to achieve natural-sounding prosody with nonsense sequences.) Each such sequence was presented three times identically in a Standard Block. Each Standard Block was followed by a Deviant block, in which the same sequence was repeated twice with the same prosody as before, and a third time carrying a prosodic violation (Figure 1A). This deviant prosodic contour was obtained by time-reversing the original one, and super-imposing it on the intact segmental information with word order, and all other properties preserved. The resulting prosodic contour was thus time-reversed, unfamiliar to the infants and ill-formed in French (and universally, as energy increased in it). We compared newborns' ($n=25$) hemodynamic responses to the Standard and Deviant Blocks using a 24-channel NIRS probe (Figure 1B), which queried the frontal, temporal and parietal areas bilaterally (Figure 1C), i.e. the areas known to respond to speech and language (e.g. Peña et al. 2003, Gervain et al. 2008, Benavides-Varela & Gervain 2017). *Results and Conclusion.* The obtained grand average results are shown in Figure 2. A cluster-based permutation tests revealed a difference between Standard and Deviant Blocks with oxyHb concentrations for the Deviant condition being greater than for the Standard one in a spatial cluster including channels 17, 19, 21, 22, and 24, i.e. the parietal-temporal areas in the RH (Figure 2). These results suggest that newborns are already capable of detecting utterance-level prosodic violations at birth. The localization in right parieto-temporal areas of the differential response confirms previous results regarding the right lateralization of speech prosody since birth. This is a key ability for newborns to start breaking into their native language. Future investigations will allow us to disentangle whether discrimination in the current study was based on familiarity, i.e. experience with speech prosody prenatally, the ill-formedness of the time-reversed contours or simply a detection of change.

Figure 1.



(A) Experimental design (adapted from Benavides-Varela & Gervain, 2017). **(B)** Picture of a neonate with the cap located upon the head (right view). **(C)** Probe configuration overlaid on a schematic neonate brain.

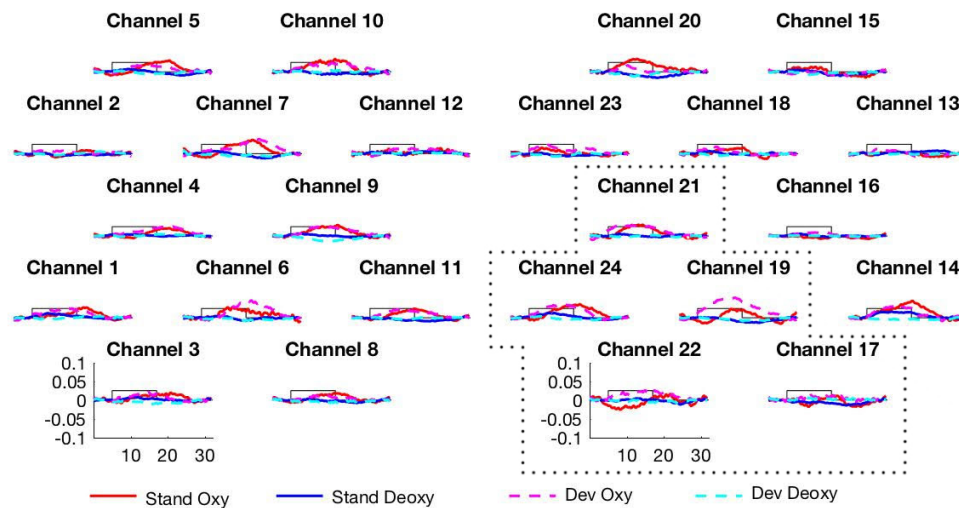


Figure 2. Grand average results. The x axis represents time in sec. The y axis shows concentration change in mmol × mm. The curves indicate grand average responses for standard (oxyHb: continuous red line, deoxyHb: continuous blue line) and deviant blocks (oxyHb: dashed pink line, deoxyHb: dashed turquoise line). The rectangle along the x axis indicates time of stimulation. The ROI obtained through the permutation test is encircled using dotted lines.

References

- Benavides-Varela, S. & Gervain, J. (2017) Learning word order at birth: A NIRS study. *Developmental Cognitive Neuroscience* 25, 198–208.
- Gerhardt, K. J., Abrams, R. M., & Oliver, C. C. (1990). Sound environment of the fetal sheep. *American journal of obstetrics and gynecology*, 162(1), 282-287.
- Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, 105(37), 14222- 14227.
- Peña, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: an optical topography study of language recognition at birth. *PNAS*, 100(20), 11702-11705.

Distributional learning as a driver of robust speech processing

Xin Xie, Chigusa Kurumada (U. of Rochester) & Andrés Buxó-Lugo (U. of Maryland)

Many influential theories of language processing assume that listeners **learn and store previously experienced distributional statistics of the input** (Dell & Chang, 2014; Frank & Goodman, 2012; Futrell et al., 2020; Johnson, 2006; Levy, 2008; MacDonald, 2013; Maye et al., 2008; Pierrehumbert, 2001; Tanenhaus & Trueswell, 1995). This knowledge is considered critical for guiding listeners' expectations to achieve efficient language processing. Further, recent work suggests that learning distributions *specific to a talker* can be a key to accommodating the cross-talker variability ubiquitous in spoken language (Kleinschmidt & Jaeger, 2015; Theodore & Monto, 2019). However, approximations of the relevant long-term or talker-specific experiences of distributions often remain unattainable or unreliable because large-scale data of sufficient resolution (e.g., estimates of means and variances of cues to a particular linguistic category) are lacking. So far, most evidence that is taken to support distributional learning as a mechanism underlying speech processing has been based on a short-/mid-term exposure to researcher-curated distributional statistics (Clayards et al., 2008; but see McMurray & Jongman, 2011).

The current study addresses this critical gap in the domain of speech prosody. We, for the first time, combine production, modeling, and comprehension experiments to examine **whether listeners indeed store distributional statistics in productions and draw on them in comprehension**. We built a corpus of 65 talkers, each producing 24 questions vs. 24 statements in the form of "*It's X-ing*" (e.g., "It's changing?" vs. "It's changing") resulting in a total of 2974 tokens (after excluding speech errors). Recorded utterances were segmented into three sections 1) *it's*, 2) *X* (the stressed syllable), and 3) *-ing*. F0 and duration of each syllable were extracted ([Fig.1A, B](#)) and examined with respect to the structure of variability in the cue distributions ([Fig.1C](#)).

Experiment 1) Do long-term statistics predict listeners' categorization of a novel talker's speech?
We trained two 65 classifiers (multivariate ideal-observers, extending Kleinschmidt, 2019), one for each talker, based on means and variances of the question vs. statement categories directly estimated from the corpus ([Fig.1D](#)). We then bundled these talker-specific models by the talkers' gender to create two "gender-specific" models, each simulating a prototypical female and a prototypical male talker. Additionally, we created a model without the knowledge of talker gender (the "gender-independent" model). We tested these models against human judgments (N = 240) on categorization of items from a 11-step continuum constructed based on recordings of two new talkers (1 male and 1 female). The *gender-specific* models significantly outperformed the gender-independent one ([Fig.2](#)), suggesting that **the long-term statistics estimated from male vs. female talkers' productions directly predict listeners' categorization of the prosodic input** ($R^2 = .95$). Listeners *do* seem to store gender-specific distributions and apply the knowledge in comprehension when first encountering a *novel* talker of a particular gender.

Experiment 2) Do listeners accommodate unexpected distributional statistics from a novel talker?
The same human listeners from Experiment 1 were randomly assigned to three conditions: Q(uestion)-biasing, No-bias, S(tatement)-biasing. Those in the Q-biasing condition heard prototypical statements (step 1) and the ambiguous item (step 7 for the female and step 8 for the male talker) disambiguated as questions via feedback. Those in the S-biasing condition instead heard the prototypical Questions (step 11) and the ambiguous items as statements. In the No-bias condition, listeners received only prototypical questions and statements. Results show that **the listeners incrementally adjusted their responses to the ambiguous items throughout the 30 trials** ([Fig.3](#), green lines), rapidly learning the underlying, talker-specific, distributions.

In sum, the current study is among the first to empirically demonstrate that speech processing does indeed leverage the implicit knowledge derived from long- and short-term learning of distributional statistics. Listeners process the variable linguistic input by applying distinct sets of

expectations derived through prior experiences, which continue to be fine-tuned in response to new exposure.

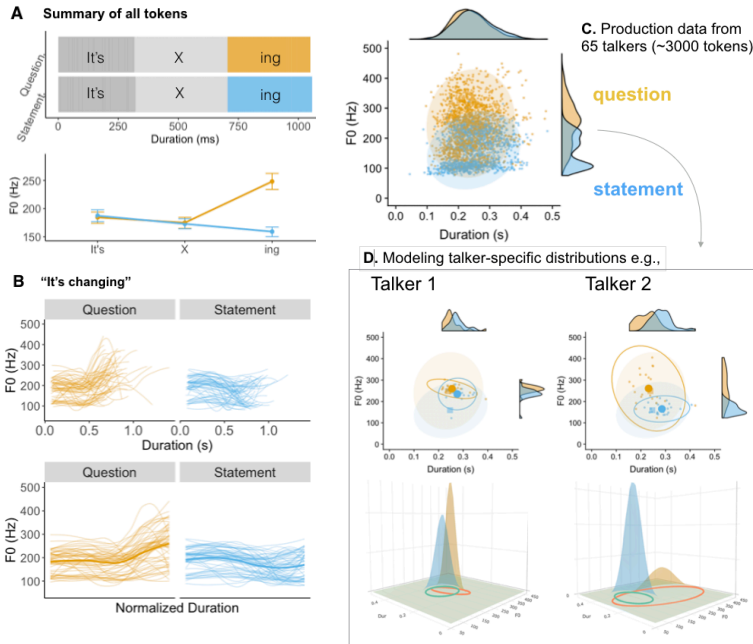


Figure 1.

A. Summary statistics of duration (top) and fundamental frequency (F0, bottom) in the intonation contours for "It's X-ing" utterances produced by 65 native English speakers.

B. F0 values of individual tokens of "It's changing" to illustrate the magnitude of talker variability seen for each of the 24 item types.

C. Group-level variations of syllable mean F0 (y-axis) and duration (x-axis) in the ~3000 tokens collected;

D. Talker-specific ideal observer models of productions for two example talkers (Talker 1 and Talker 2).

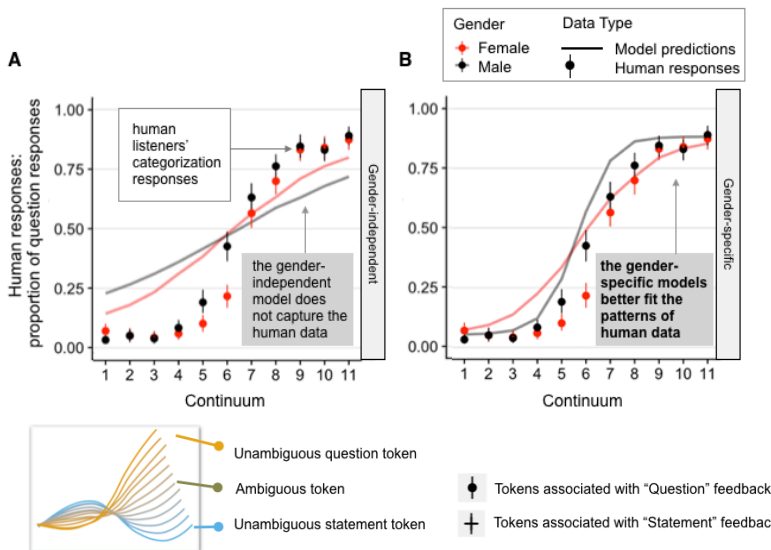


Figure 2.

Categorization functions predicted by ideal observers (lines) and actual categorization by human listeners (point ranges). (The points indicate the by-item means averaged across listeners. Error bars indicate bootstrapped 95% confidence intervals. The human data plotted are identical between the two panels.) **A**: gender-independent model, wherein the two lines represent predictions of one model for the female vs. male talker data. **B**: gender-specific models.

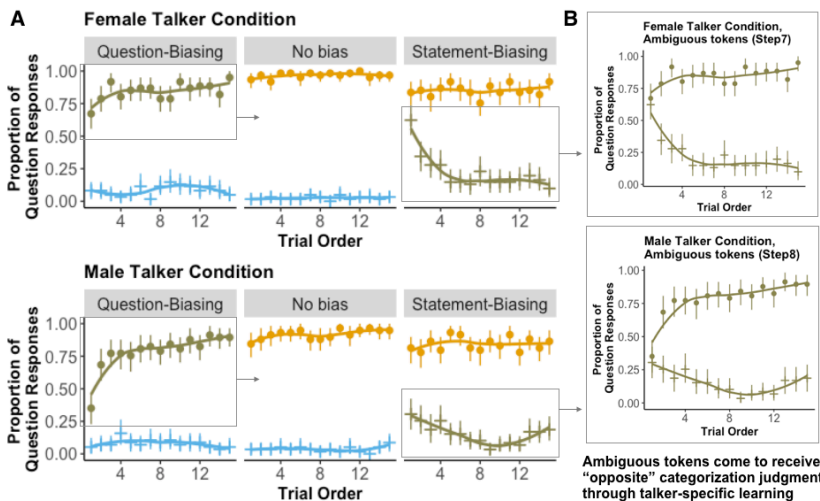


Figure 3.

A. Overall response patterns across the between-subject conditions. X-axis: The relative ordering of the 15 exposure tokens associated with the question vs. the statement feedback. Blue and yellow indicate unambiguous tokens (Step 1 and Step 11, respectively) and green represents the ambiguous items. Error bars indicate bootstrapped 95% confidence intervals.

B. Responses given to the prosodically ambiguous tokens in the female vs. the male talker conditions; the top line (circles) and the bottom line (crosses) represent the Question-Biasing and the Statement-Biasing conditions. Ambiguous tokens come to receive "opposite" categorization judgment through talker-specific learning

The Identifiability of Consonants and of Syllable Boundaries in Infant-Directed English

Corpus-based models of infant phonetic category learning usually assume that infants induce categories from experienced instances, clustering segments into categories defined by statistical distributions. Corpus-based models of word segmentation, in turn, usually assume that infants can categorize each phonetic segment, and sometimes assume that syllable boundaries are given in the signal. Both of these starting assumptions apparently conflict with the well-known result that even whole words extracted from conversation are frequently unidentifiable by adult native speakers (e.g. Pollack & Pickett, 1963) even in child-directed speech (e.g. Bard & Anderson, 1983). Existing learning models are only tenable if segments and boundaries are sometimes locally identifiable. Here, we asked: when are consonants of infant-directed speech recognizable, and to what extent are onset and coda consonants identifiable as such? Setting quantitative bounds on identifiability helps evaluate the plausibility of learning models.

An hourlong session from each of two American English mothers speaking to their 10-month-olds (Brent & Siskind, 2001) was orthographically transcribed, hand-aligned at the word and phone levels, and phonetically transcribed (e.g., Adriaans & Swingley, 2017). The words in virtually all sentences were readily interpretable in context. All vowel-consonant-vowel sequences where the consonant was at a word boundary (either as coda or onset) were extracted into v.cv (n=1008) or vc.v (n=407) 3-segment audiofiles consisting of the consonant and the entirety of the surrounding vowel segments. These files were divided into sets, and presented online to 51 trained native-English adults who judged, for each vcv clip, the identity of the consonant, and the consonant's word position {coda, onset}. Each token was judged by at least 6 listeners.

Analysis of these responses showed that many instances of maternal-speech consonants from these sessions were unintelligible, and many were impossible to assign to a syllable. Considering all consonant categories, the median identification proportion for nominal word onsets was 55.6% (25th %ile, 52.2%; 75th, 59.4%); and for codas, 26.7% (25th, 9.2%; 75th, 35.2%). Figure 1 shows the confusion matrices for onsets and codas, organized by manner of articulation. In most cases, particularly for onsets, the modal response was the correct one. However, many errors remained, and for some sounds at onset (and nearly all of the codas), most sounds were not correctly identified. Perhaps surprisingly, in most cases this was not because sounds competed with phonologically similar competitors, like 1-feature mismatches. There were some such cases, such as voicing errors in fricatives and stops, but for the most part, it seems that sounds were either correctly identifiable, or unintelligible, leading to guessing.

Participants were also quite poor at telling whether a consonant was an onset or coda. Over items, the median success proportion was 60% (25th %ile, 44; 75th %ile, 75). Though significantly above 50%, these proportions also reflected a bias toward responding "onset", which matched the stimuli (71% onsets, following expected distributions from English). Only a third of participants showed a significant contingency (by chi-square test) between their responses and the true syllable position. Could it be that syllable positions were more discernable for the consonants that were easier to identify? Some, but not much. As Figure 2 shows, for only some sounds (mainly codas), consonants' positions tracked identifiability (numbers are *rs*). Put another way, even the easiest-to-identify consonants' syllable affiliations were often a mystery to adult listeners.

Although these results only concern vcv sequences from two mothers speaking to their 10-month-olds, they suggest several conclusions relevant to modeling of early word processing in infants. First, models should not assume that all tokens are good instances for training phonetic categories. More likely, some instances are superior training tokens; the question is whether infants can identify them as such. Second, models that take syllables as inputs to "statistical learning" should not presuppose that syllable boundaries are given in the signal (see Jusczyk et al., 1999). Third, models assuming the emergence of protolexical islands of familiarity seem more plausible than full segmentation models (Goodsitt et al., 1993). Ongoing work assesses the generality of these effects and seeks correlates of identifiability.

References

- Adriaans, F., & Swingle, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *JASA*, 141, 3070-3078.
- Bard, E.G., & Anderson, A.H. (1983). The unintelligibility of speech to children. *JCL*, 10, 265-292.
- Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, B33-B44.
- Goodsitt, J.V., Morgan, J.L., & Kuhl, P.K. (1993). Perceptual strategies in prelingual speech segmentation. *JCL*, 20, 229-252.
- Pollack, I., & Pickett, J.M. (1963). The intelligibility of excerpts from conversation. *Language and Speech*, 6, 165-171.
- Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, 61, 1465-1476.

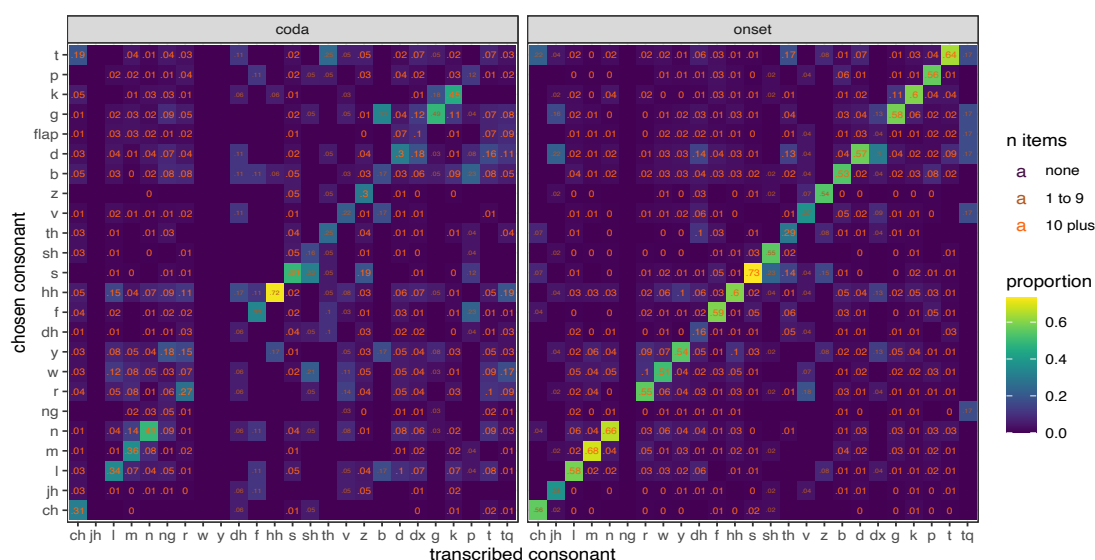


Figure 1. Confusion matrices for codas (left) and onsets (right). Proportion is shown in each cell. Numbers based on <10 items are shown in a smaller, darker font. Warmer colors show greater convergence. Onsets were ID'd correctly about 56% of the time; codas, about 27% of the time.

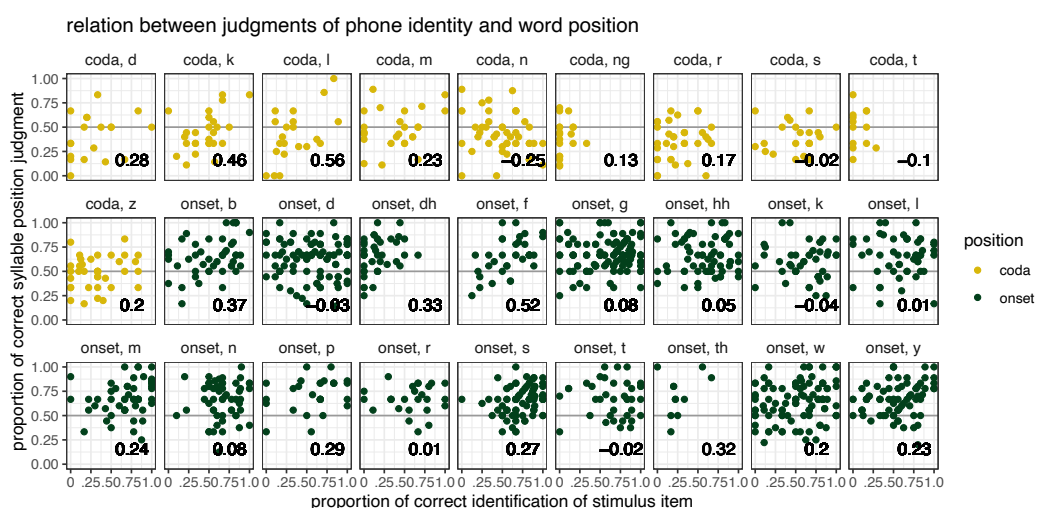


Figure 2. For each item, how often its syllable position was correctly judged (y axis) plotted against how often its identity (which consonant) was correctly judged (x axis), considering consonants tested with more than 10 stimulus items. For most sounds, identifiability of the consonant did not make judgments of syllable position more accurate, even for stop consonants.

The presence of background noise reduces interlingual phonological competition during non-native speech recognition

Florian Hintz (Max Planck Institute for Psycholinguistics), Cesko C. Voeten (Leiden University), Odette Scharenborg (TU Delft)

Language users experience interlingual competition when listening to non-native speech. Using the visual world paradigm, listeners have been shown to fixate objects whose word name overlapped phonologically in participants' native language with a simultaneously unfolding non-native target word (e.g., Spivey & Marian, 1999). This finding has been replicated numerous times and contributed to the notion of 'non-selective lexical access' during non-native language processing (Dijkstra et al., 2019). To the best of our knowledge, all previous experiments studied interlingual phonological competition under 'ideal' circumstances, involving carefully produced speech and high-quality audio recordings. In the real world, speech comprehension rarely takes place under ideal circumstances. Moreover, previous research has shown that noise has more dramatic effects on non-native than on native speech recognition (Scharenborg & van Os, 2019). The reasons for this asymmetry are not well understood.

In the present study, we tested the effects of background noise on interlingual competition, i.e. co-activation of listeners' native language when listening to non-native speech. We conducted a visual world experiment and recorded the eye movements of 35 native Dutch participants (all proficient users of English) as they listened to English sentences while looking at displays featuring four objects. Each sentence contained a target word. On filler trials ($n = 22$), the visual referent depicting the target word was present, along with three unrelated distractors. On experimental trials ($n = 22$), the picture of the spoken target (e.g., 'wizard') was absent. Instead, the display featured an English competitor, overlapping with the spoken English target in phonological onset (e.g., 'window'), a Dutch competitor, whose Dutch (but not English) word name overlapped with the English target in phonological onset (e.g., Dutch 'wimpel', English: 'pennant'), and two unrelated distractors (e.g., 'bike', 'jeans'). Half of the sentences was masked by speech-shaped noise at a signal-to-noise ratio (SNR) of +3 dB. This SNR was chosen based on an earlier Dutch study (Scharenborg et al., 2018) such that intelligibility was reduced but floor effects were avoided. The other half of the sentences were presented in the clear. Participants previewed the displays for three seconds before target word onset. Eye movements were analyzed using logistic GAMMs (generalized additive mixed models).

Our analyses showed that participants fixated the target objects on filler trials shortly after they were mentioned. Target fixations occurred later when the signal was masked by background noise. On experimental trials, we observed fixation biases for English onset competitors (relative to the distractors) in the clear and in noise demonstrating that participants engaged in non-native phonological onset competition. In contrast, the likelihood of increased looks to the Dutch onset competitors varied across listening conditions: Replicating earlier research (Spivey & Marian, 1999), participants looked at the Dutch competitors in the clear condition when hearing the English target word, reflecting the (partial) activation of their native lexicon (i.e., interlingual competition). However, the likelihood of looks to the same objects was substantially reduced when speech was masked by background noise (Panel D in Figure 1).

Our data thus demonstrate that the presence of background noise reduces the likelihood of interlingual competition during non-native listening, casting new light on the situational influences on non-selective lexical access. Interestingly, while earlier research showed that noise enhances *intralingual* phonological competition (in both native and non-native listeners, e.g., Scharenborg et al., 2018), the present data suggest the opposite for the involvement of one's native language during non-native speech recognition. We believe that our results are most compatible with the notion that hearing non-native speech in noise enforces a re-allocation of cognitive resources in the service of achieving the present task goal. This happens at the expense of the task-irrelevant co-activation of one's native language.

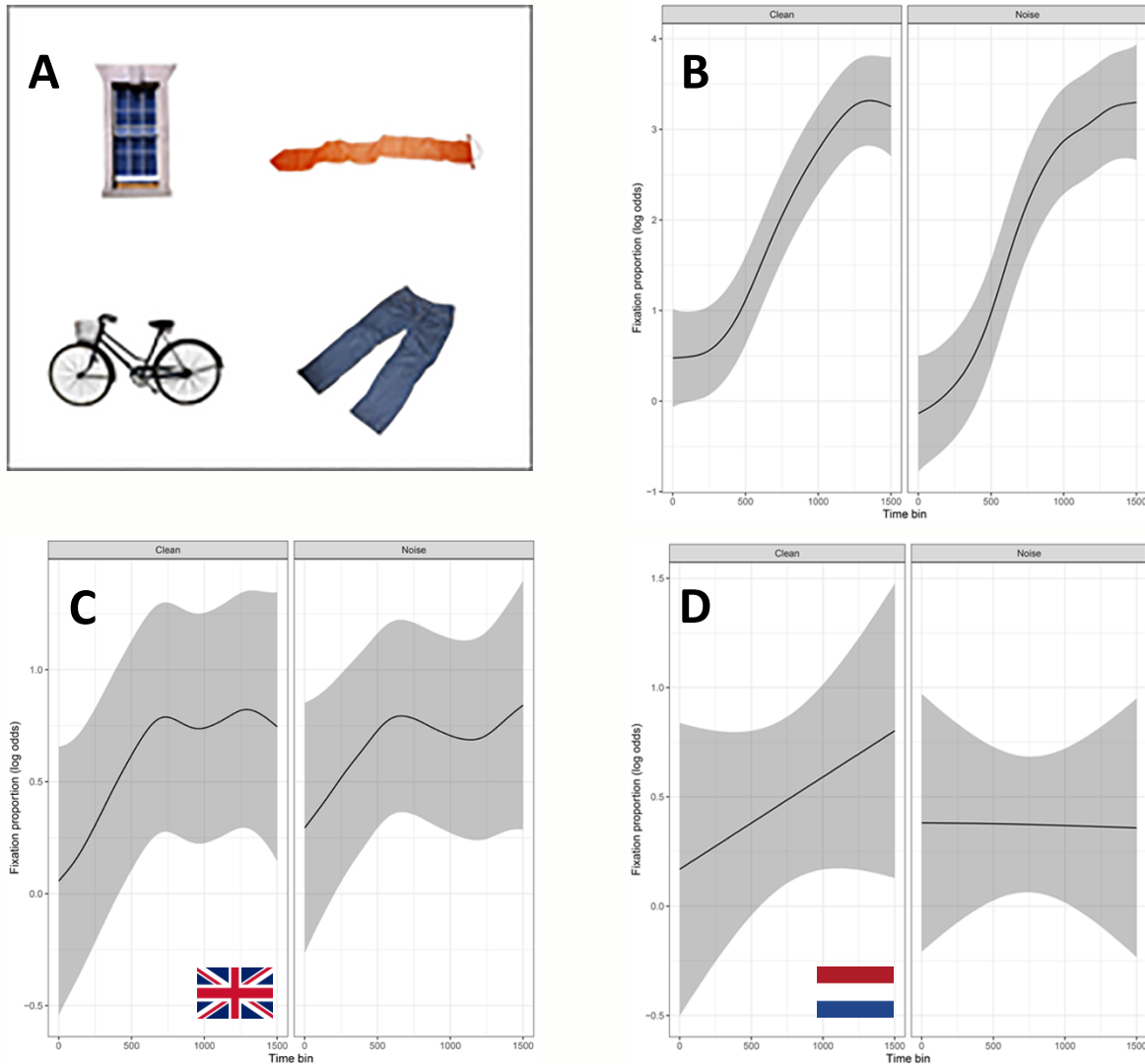


Figure 1. *Panel A:* Example of a visual stimulus used on experimental trials. English target was ‘wizard’; ‘window’ was English phonological competitor; ‘pennant’ (Dutch: ‘wimpel’) was Dutch phonological competitor; ‘bike’ and ‘jeans’ were unrelated distractors. *Panel B:* Results of logistic additive mixed-model for filler items (left: clear trials, right: noise trials). *Panels C and D:* Results of logistic additive mixed-models for English and Dutch phonological competitors (experimental items; left: clear trials, right: noise trials). As a shorthand, fixation biases can be considered meaningful when confidence intervals (gray ribbons) do not cross zero.

References

- Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(4), 657-679.
- Scharenborg, O., & Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Communication*, 108, 53-64.
- Scharenborg, O., Coumans, J. M., & van Hout, R. (2018). The effect of background noise on the word activation process in nonnative spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(2), 233.
- Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological science*, 10(3), 281-284.

An investigation of the time-course of syntactic and semantic interference in online sentence comprehension

Daniela Mertzen¹, Brian W. Dillon², Dario Paape¹, Ralf Engbert¹ and Shravan Vasishth¹

¹University of Potsdam, ²University of Massachusetts Amherst
mertzen@uni-potsdam.de

Introduction. One central question in sentence comprehension research is when syntactic and semantic information are used during the formation of non-adjacent dependencies (e.g., [1:6]). In the cue-based parsing literature, this question has been addressed by studying the time-course of similarity-based interference effects (e.g., [9:14]). Cue-based parsing theories assume that items are encoded and later retrieved from memory using retrieval cues [7:10]. These cues can be syntactic or semantic, and both sources of information can be used in parallel during retrieval. Interference occurs when the retrieval cues cannot uniquely identify a target item because other syntactically and/or semantically similar (distractor) items are encoded in memory. In subject-verb dependencies, interference from syntactically similar distractors was observed at the retrieval point (a verb), while semantic interference was reported at a later, sentence-final region [10]. This finding may suggest that syntactic information is used to reactivate dependents in memory before semantic information. A similar proposal was made for antecedent-reflexive dependencies in [14] [see also 12]. However, the time-course for semantic interference remains unclear: [11] reports a different time-course than [10] for semantic interference in subject-verb dependencies. Cue-based theories predict that syntactic and semantic interference occur simultaneously during retrieval. We reinvestigated this prediction in English. Furthermore, to study the generality of these effects, we conducted a second, large-sample experiment in German.

Design and materials. Our two eye-tracking (reading) experiments (English, N=61; German, N=121) used a 2 x 2 design with the factors distractor subjecthood (–subject, +subject) and distractor animacy (–animate, +animate) [10]. Table 1 shows an English example item. In all conditions, the manipulated distractor (the meeting/visitor) intervenes between the critical verb (complained) and the target subject (the attorney).

Predictions. Cue-based theories predict a reading time slowdown for +subject compared to –subject conditions, indicating syntactic interference. Similarly, a reading time slowdown is expected for +animate compared to –animate conditions (semantic interference). Crucially, both effects should be observable at the critical verb.

Results. Figure 1 shows the results from our Bayesian analysis. For both languages, +subject conditions showed reading time slowdowns in regression-path durations and total reading times at the critical verb, consistent with a syntactic interference effect. Only English exhibited semantic interference (a slowdown for +animate conditions) at the critical verb; in German there was an indication of this slowdown post-critically. Surprisingly, both languages exhibited slower reading times at the pre-critical adverb for +subject and +animate distractors.

Discussion. In English, the observed reading time slowdowns indicate that both syntactically and semantically similar distractors can cause interference during retrieval. These results are compatible with cue-based theories' predictions. The pattern in our German data is consistent with the observation that semantic effects can continue to slow down processing in later sentence regions [10]. In both languages, the unexpected pre-critical effects are consistent with spillover from prior regions. Further analyses are underway to investigate this possibility.

Conclusions. We tentatively conclude that both syntactic and semantic interference can arise simultaneously, i.e., both types of information can be used in parallel during real-time dependency formation. However, in line with previous research, the German data show that semantically similar distractors may continue to interfere further downstream in the sentence.

Table 1. English example item. The critical target subject and the critical verb (the retrieval point) are shown in bold. The manipulated distractor is underlined. +/-subject: distractor is (not) a subject; +/-animate: distractor is (not) animate.

a. *–subject, –animate*

It turned out that **the attorney** whose secretary had forgotten about the important meeting frequently **complained** about the salary at the firm.

b. *–subject, +animate*

It turned out that **the attorney** whose secretary had forgotten about the important visitor frequently **complained** about the salary at the firm.

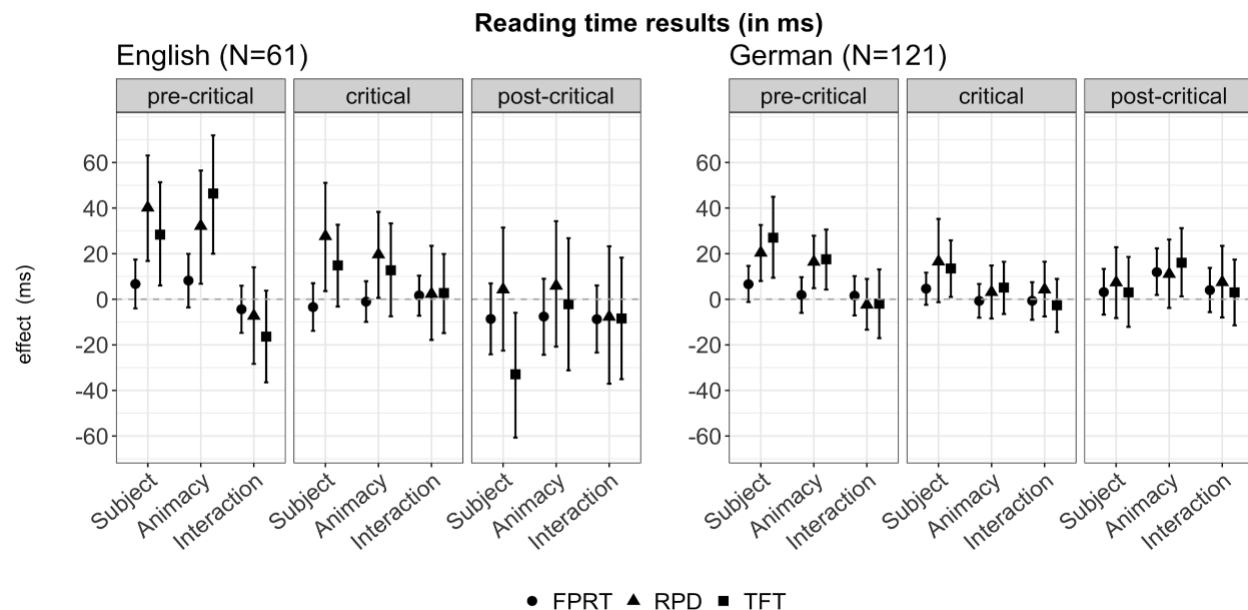
c. *+subject, –animate*

It turned out that **the attorney** whose secretary had forgotten that the meeting was important frequently **complained** about the salary at the firm.

d. *+subject, +animate*

It turned out that **the attorney** whose secretary had forgotten that the visitor was important frequently **complained** about the salary at the firm.

Figure 1. Reading measure results for the English and the German experiment. We fit maximal Bayesian hierarchical models [15]. Shown are the means of the posterior distributions with their 95% Bayesian credible intervals. These give the range in which the true parameter lies with 95% probability, given the data and model. A positive sign means that a slowdown is observed for +subject or +animate conditions. FPRT = first-pass reading times, RPD = regression-path duration, TFT = total fixation times. Pre-critical: adverb, critical: verb, post-critical: prepositional phrase



References. [1] Frazier & Rayner (1982). *Cogn Psychol.* [2] Clifton et al. (2003). *J Mem Lang.* [3] MacDonald et al. (1994). *Psychol Rev.* [4] McRae et al. (1998). *J Mem Lang.* [5] Tabor et al. (2004). *J Mem Lang.* [6] Van Gompel et al. (2005). *J Mem Lang.* [7] Lewis & Vasishth (2005). *Cogn Sci.* [8] McElree (2000). *J Psycholinguist Res.* [9] Van Dyke & Lewis (2003). *J Mem Lang.* [10] Van Dyke (2007). *J Exp Psychol. Learn Mem Cogn.* [11] Van Dyke & McElree (2011). *J Mem Lang.* [12] Dillon et al. (2013). *J Mem Lang.* [13] Cunnings & Sturt (2018). *J Mem Lang.* [14] Sturt (2003). *J Mem Lang.* [15] Gelman et al. (2014).

A CUE-BASED APPROACH TO PROCESSING ADJUNCTS

Ethan Myers Masaya Yoshida (Northwestern University)

Introduction: Cue-based retrieval models [1-2] lack consensus about the types of features available as cues during sentence-processing [3-7]. Active debate in the field circles the question whether retrieval cues should be “lexically specific” [8] or “semantically general” [9]. We show that lexically specific semantic features are active and may interfere with wh-dependency resolution in online sentence processing. Specifically, we show preliminary data from an eye-tracking experiment (n=30) that locative and temporal PPs (e.g., *in the park*, *in the morning*) may cause a similarity-based interference effect [1] with the resolution of wh-gap dependencies involving locative (i.e., *where*) or temporal (i.e., *when*) wh-phrases. Our basic observation is that when a locative PP intervenes in a locative wh-verb dependency, the verb is read slower in early eye-tracking measures, consistent with other studies of interference phenomena (cf. [6,8]). Similarly, when a temporal PP intervenes a temporal wh-verb dependency, the verb is likewise read slower. These slowdown effects, we argue, are caused by the semantic feature of the PPs that is similar to that of wh-phrases and thus, they created a similarity-based interference effect. From this, we argue cue-based models must be sensitive to semantic features specific to particular lexical items.

Experiment: An eye-tracking experiment was conducted with 30 English speaking undergraduates at Northwestern University. Experimenters manipulated (i) the type of PP (Temporal/ Locative) and (ii) the degree of semantic overlap (Match/Mismatch/No Match), using a 1x3 factorial design. To avoid the PP being interpreted as the modifier of the embedded verb (*ate*), the PP is embedded inside the relative clause attached to the subject NP. The critical region, the main verb ‘ate’ in (1) where retrieval is expected to take place [9-10], as the temporal/locative adjunct is interpreted modifying event represented by the main verb.

Weak, but significant main effects of semantic overlap were observed using linear mixed effects regression (lme4) in the first-pass ($\beta = 305.00$, $se = 15.12$, $t = 20.17$, $p < .01$) and first-fixation ($\beta = 262.45$, $se = 12.18$, $t = 21.55$, $p < .01$) reading times of Matched conditions (2) suggesting an inhibitory effect of interveners. This is consistent with the belief that semantic features of wh-adjuncts remain active in memory during wh-resolution, and that structurally unavailable PPs interfere with the processing of the matrix wh-dependency. Furthermore, these effects being limited only to Matched conditions, despite all interveners being PPs, indicates that the interference effect is not from morphological or structural cues.

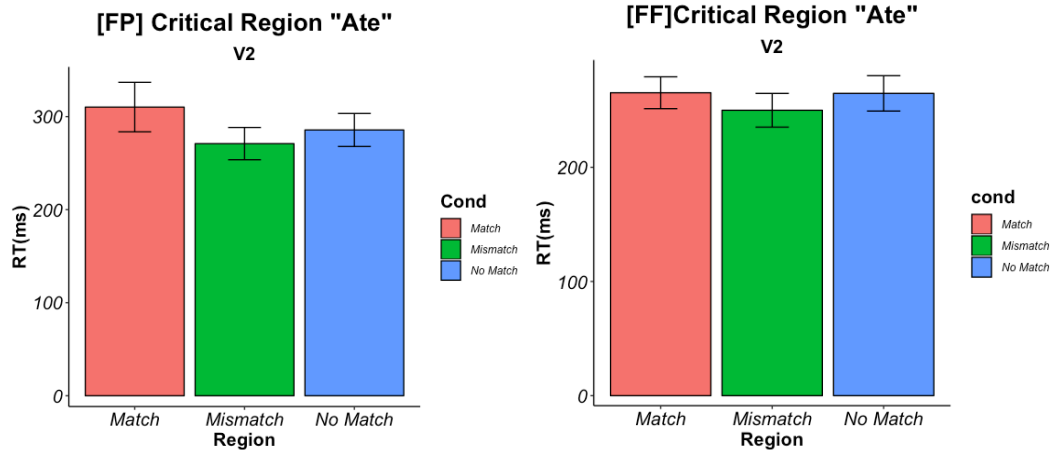
Discussion: The similarity-based interference effect we observed in the wh-verb dependency formation supports the position that semantic features like *+locative* or *+temporal* may be accessible to either retrieval or encoding mechanisms [2,6,9] in online dependency resolution of adjunct wh-phrases like *when* or *where*. Thus, this means that on top of the overt morphological features, or structural features, lexically specific semantic features may also be relevant for cue-based parsing models.

Examples/Charts:

(1) John inquired **when/where** the girl that danced ... ate sushi and donuts.

- a.where in the park (Match)
- b.where in the morning (Mismatch)
- c.if in the park (No Match)

(2)



(3) **Model used:**

`lmer(RT ~ condition+(subj|item), data = data)`

References

- [1] Gordon, P.C., Hendrick, R. & Johnson, M. Memory interference during language processing. (2001) [2] Lewis & Vasishth. (2005). An activation-based model of sentence processing as skilled memory retrieval. [3] Dillon, B. (2011). Structured access in sentence comprehension [4] Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. (2013). [5] Dillon, B. (2014). Syntactic memory in the comprehension of reflexive dependencies: an overview. [6] Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., & Vasishth, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. [7] Parker, D., Shvartsman, M., & Van Dyke, J. A. The cue-based retrieval theory of sentence comprehension: New findings and new challenges. (2017). [8] Cummings, I., & Sturt, P. Retrieval interference and semantic interpretation. (2015) [9] Smith, G., & Vasishth, S. A principled approach to feature selection in models of sentence processing. (2020). [10] Biondo, N., & Vespignani, F., Dillon, B. Structural constraints strongly determine the attachment of temporal adverbs. (2016)

Retrieval interference in the processing of RCs: Evidence from the visual-world paradigm

Gwynna Ryan & Matthew W. Lowder (University of Richmond)

Although a large literature demonstrates that object-relative clauses (ORCs) are harder to process than subject-relative clauses (SRCs) (see Table 1, Example 1), there is less agreement regarding where during processing this difficulty emerges, as well as how best to account for these effects. Explanatory frameworks that focus on the role of memory retrieval conceptualize the ORC-SRC asymmetry as resulting from the memory demands associated with processing ORCs as compared to SRCs. In contrast, experience-based accounts argue that the asymmetry reflects the fact that ORCs are less frequent than SRCs. Although both accounts may to some extent explain the mechanisms underlying RC processing, the two make very different predictions regarding the matrix verb: whereas memory-based accounts tend to predict processing differences to emerge at the matrix verb, experience-based accounts do not.

Several studies have found robust ORC-SRC effects at the matrix verb (e.g., Gordon et al., 2006; King & Just, 1991; Lowder & Gordon, 2012), but others have not (e.g., Staub, 2010), or have failed to find these effects when a prepositional phrase (PP) intervenes between the RC and the matrix verb (Staub et al., 2017). Notably, the vast majority of previous experiments on the processing of RCs have been conducted in the written domain relying on self-paced reading or eyetracking during reading. In contrast, there is very little work on RC processing in the spoken domain (cf. Kowalski & Huang, 2017), and we are not aware of any previous research that has carefully examined ORC-SRC differences at the matrix verb when sentences are presented aurally. Accordingly, the current visual-world eyetracking experiment was designed to test whether ORC-SRC differences would emerge at the matrix verb during spoken sentence processing. Memory-based accounts posit that the matrix verb cues the comprehender to retrieve the matrix subject (NP1) from memory; crucially, this process is predicted to be easier for SRCs than for ORCs because the embedded noun (NP2) in ORCs creates interference, as it must serve as the subject of the embedded verb. In contrast, experience-based accounts predict processing differences early in the RC and predict no differences at the matrix verb once the word orders of the two sentences are again identical.

Participants ($n = 40$) listened to sentences containing ORCs and SRCs in which the order of the two noun phrases (NPs) was counterbalanced across lists (see Table 1, Example 2). The visual display consisted of four pictures representing the two NPs (e.g., a cat and a dog) and two unrelated distractors (e.g., a plant and a towel). A PP always intervened between the RC and the matrix verb. This was important to ensure that any processing differences observed at the matrix verb could not be attributed to spillover from the RC. There were 40 sets of critical items, counterbalanced across four lists and mixed with 64 filler trials that did not contain RCs. A written true-or-false comprehension question followed each trial.

Accuracy on the comprehension questions was significantly worse for sentences containing ORCs ($M = 82\%$) than for sentences containing SRCs ($M = 91\%$), $p < .001$, replicating a pattern that has been obtained in many previous reading studies. Fixation plots for the two sentence types are presented in Figure 1. Participants tended to look at NP1 followed by NP2 while listening to the RC (analysis of this region is complicated by the different word orders), with fixations to these two images returning to equal levels during the PP. Crucially, at the matrix verb, the preference to fixate NP1 versus NP2 was larger in the SRC condition than the ORC condition. This observation was confirmed by statistical analyses that tested the magnitude of this preference over 200-ms time bins. The difference was significant beginning at 1400 ms after onset of the matrix verb and lasted until 2200 ms after onset of the matrix verb.

These results are most readily explained under a memory-retrieval account of RC processing; that is, retrieval of the matrix subject (i.e., NP1) was easier with less interference from NP2 in the SRC than the ORC sentences. The findings also highlight the visual-world paradigm as a useful approach for studying the processing of complex syntactic structures.

Table 1. Example sentences.

Example 1
The reporter that attacked the senator admitted the error. (SRC)
The reporter that the senator attacked admitted the error. (ORC)
Example 2
The cat that watched the dog in the living room jumped onto the couch. (SRC, Order1)
The cat that the dog watched in the living room jumped onto the couch. (ORC, Order1)
The dog that watched the cat in the living room jumped onto the couch. (SRC, Order2)
The dog that the cat watched in the living room jumped onto the couch. (ORC, Order2)

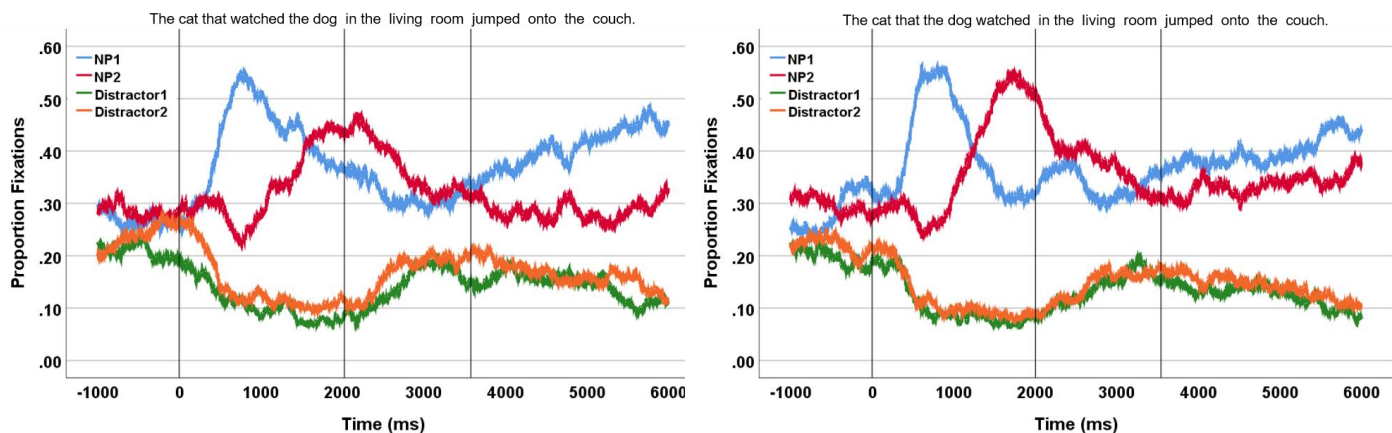


Figure 1. Fixation plots for sentences containing SRCs (left) and ORCs (right). The first vertical line (at time 0) marks the onset of the first noun (e.g., “cat”). The second vertical line represents the mean onset of the prepositional phrase. The third vertical line represents the mean onset of the matrix verb.

References

- Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 1304-1321.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580-602.
- Kowalski, A., & Huang, Y. (2017). Predicting and priming thematic roles: Flexible use of verbal and nonverbal cues during relative clause comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1341-1351.
- Lowder, M. W., & Gordon, P. C. (2012). The pistol that injured the cowboy: Difficulty with inanimate subject-verb integration is reduced by structural separation. *Journal of Memory and Language*, 66, 819-832.
- Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116, 71-86.
- Staub, A., Dillon, B., & Clifton, C., Jr. (2017). The matrix verb as a source of comprehension difficulty in object relative sentences. *Cognitive Science*, 41, 1353-1376.

Longer encoding times facilitate subsequent retrieval during sentence processing

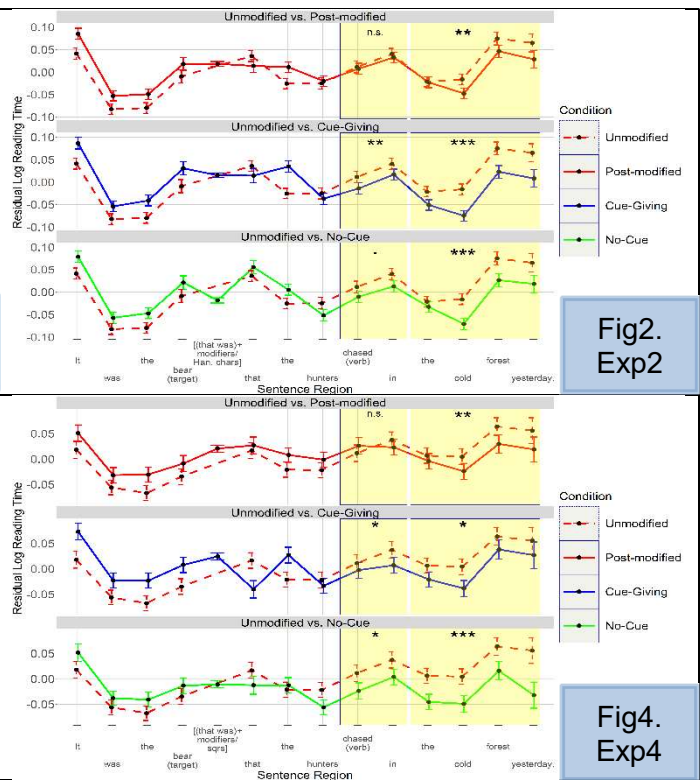
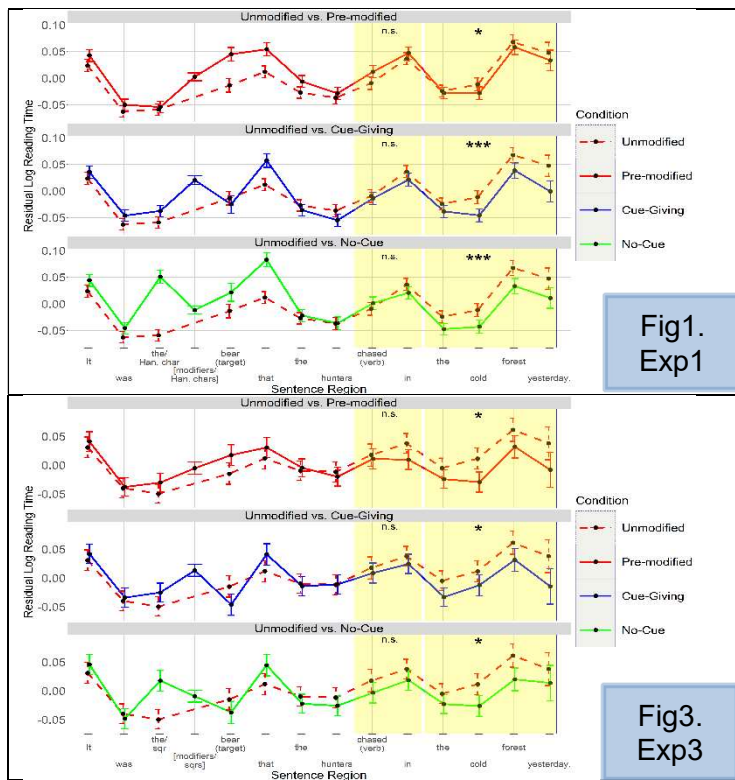
Hossein Karimi (Mississippi State), Michele Diaz (Penn State), Eva Wittenberg (UCSD)

Numerous studies have shown that modified words (i.e., *the injured and dangerous bear*) result in faster reading times compared to unmodified words (i.e., *the bear*) at a subsequent point where the retrieval of the head noun (*bear*) is triggered [e.g., 1-4]. This “modification effect” has been shown for both pre-modified (i.e., *the injured and dangerous bear*) and post-modified words (i.e., *the bear that was injured and dangerous*, [5]). Two main memory mechanisms have been proposed to explain the modification effect: (1) the *distinctiveness* account, according to which added semantic information result in representations that are more distinct from other representations in memory, rendering them less susceptible to interference. And (2) the *head-reactivation* account, which states that processing modifying words (e.g., *injured and dangerous*) causes the target word (*bear* as syntactic head of the noun phrase) to be re-activated in memory, leading to higher ultimate activation levels [e.g., 1,2]. This project challenges these accounts and provides evidence for a “time-induced attention” hypothesis: Modifying information provides more encoding time, which in turn heightens attention to the head noun, rendering encoding more robust and subsequent retrieval easier [6]. In the case of post-modified words, the processor necessarily spends more time *maintaining* the representation of the head noun when it is post-modified than when it is unmodified. In the case of pre-modified words, because the determiner (*the*) predicts an upcoming head noun, the processor spends more time *expecting* the head noun relative to unmodified words. Longer maintenance and expectation of the head noun’s representation in memory may heighten attention to it, facilitating subsequent retrieval.

Design. In addition to using UNMODIFIED (1a & 2a, see below), and PRE-, or POST-MODIFIED words (1b & 2b), we also included a CUE-GIVING (1c & 2c) condition in which modifying words were replaced with masking characters (Exps 1&2) or symbols (Exps 3&4), but the determiner *the* (in the case of pre-modifiers) or the relative pronoun and the auxiliary verb *that was* (in the case of post-modifiers) were kept in English; as well as a NO-CUE condition in which these syntactic cues were replaced with masking characters/symbols as well (1d & 2d). Masking characters/symbols were used to ensure that readers spent a comparable amount of time on the head noun as in the PRE- or POST-MODIFIED conditions, without additional content. Note that the syntactic complexity of the whole noun phrase is maintained in the CUE-GIVING condition, but in the NO-CUE condition, head noun merely enjoys more encoding time. **Analysis.** Following [1], the residuals of an initial model (predicting log-transformed RTs by sentence type (filler vs. experimental), trial number, word length, word position, and RT on the preceding word) were used as the Dependent Variable to test the effects of the predictors of interest in maximal mixed-effects models, with UNMODIFIED condition as baseline. To minimize multiple comparisons, analyses were limited to an “early” region including the verb and the immediately following word, and a “late” region including the next four words. **Results.** In all experiments (self-paced reading, N=413, n=57), we replicated the standard modification effect on the late region. Critically, we also observed faster reading times on both the early and late regions for both CUE-GIVING and NO-CUE conditions relative to the UNMODIFIED condition (regardless of modifier position; see Figures 1-4). There were no significant accuracy differences between conditions in any of the experiments, eliminating shallow processing as a function of masking characters [10]. **Discussion.** These results call into question both accounts previously developed to explain the modification effect: The *distinctiveness* account states that ease of retrieval is predicated on additional semantic information; however, we found easier retrieval despite no information added by the masking symbols/characters. Similarly, the *reactivation* account cannot explain our data either, because the character/symbol masks necessarily could not trigger the integration needed to initiate head-reactivation [1,7-9]. Instead, our results suggest that sheer time spent expecting or maintaining a representation in memory, and the concomitant heightened attention, facilitates its subsequent retrieval, carrying important implications for the current memory-based theories of language processing by highlighting the role of encoding time and attention.

Example sentences with critical words highlighted. Symbols (■) were displayed in chunks, corresponding to word-by-word presentation. The experimental sentence for Experiments 1 and 2 can be constructed by replacing each symbol chunk with a random Korean character (participants were screened to be unfamiliar with Korean).

Exps 1 & 3	(1a) UNMODIFIED	It was the bear that the hunters chased in the cold forest yesterday.
	(1b) PRE-MODIFIED	It was the injured and dangerous bear that the hunters chased in the cold forest yesterday.
	(1c) CUE-GIVING	It was the ■■■ ■■■ ■■■■ bear that the hunters chased in the cold forest yesterday.
	(1d) NO-CUE	It was ■■ ■■■ ■■■ ■■■■ bear that the hunters chased in the cold forest yesterday.
Exps 2 & 4	(2a) UNMODIFIED	It was the bear that the hunters chased in the cold forest yesterday.
	(2b) POST-MODIFIED	It was the bear that was injured and dangerous that the hunters chased in the cold forest yesterday.
	(2c) CUE-GIVING	It was the bear that was ■■■ ■■■■ that the hunters chased in the cold forest yesterday.
	(2d) NO-CUE	It was the bear ■■■ ■■■■ that the hunters chased in the cold forest yesterday.



Figures 1-4. The left and right highlighted areas correspond to “early” and “late” regions, respectively. Tables 1-4. Results for all experiments.

Table1. Exp1 Results. N=112					Table2. Exp2 Results. N=113				
Region	Condition	t	p		Region	Condition	t	p	
Early “chased in”	Pre-modified	1.72	.08		Early “chased in”	Pre-modified	-1.01	.31	
	Cue-Giving	-1.24	.21			Cue-Giving	-2.78	.005	
	No-Cue	-1.09	.27			No-Cue	-1.84	.06	
Late “the cold forest yesterday”	Pre-modified	-2.15	.03		Late “the cold forest yesterday”	Pre-modified	-3.29	.001	
	Cue-Giving	-3.96	<.001			Cue-Giving	-6.03	<.001	
	No-Cue	-4.51	<.001			No-Cue	-4.99	<.001	

Table3. Exp3 Results. N=89					Table 4. Exp4 Results. N=99				
Region	Condition	t	p		Region	Condition	t	p	
Early “chased in”	Pre-modified	-1.37	.17		Early “chased in”	Pre-modified	-.88	.42	
	Cue-Giving	-.90	.36			Cue-Giving	-1.98	.04	
	No-Cue	-1.21	.22			No-Cue	-2.42	.01	
Late “the cold forest yesterday”	Pre-modified	-2.46	.01		Late “the cold forest yesterday”	Pre-modified	-2.81	.005	
	Cue-Giving	-2.33	.02			Cue-Giving	-2.11	.03	
	No-Cue	-2.45	.01			No-Cue	-5.03	<.001	

References.

- Hofmeister, P. (2011). *Language and Cognitive Processes*.
- Hofmeister, P., & Vasishth, S. (2014). *Frontiers in Psychology*
- Karimi, H., & Ferreira, F. (2016). *Psychonomic Bulletin & Review*.
- Karimi, H., Swaab, T. Y., & Ferreira, F. (2018). *JML*.
- Karimi, H., Diaz, M., & Ferreira, F. (2019). *JEP:LMC*.
- Corley, M., & Hartsuiker, R. J. (2011). *PLoS one*.
- Lewis, R. L., & Vasishth, S. (2005). *Cognitive science*.
- Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). *Trends in cognitive sciences*
- Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). *JML*.
- Ferreira, F. Bailey, K.G.D & Ferraro. V. (2002). *Current Directions in Psychological Science*.

Cue-based retrieval of parsing rules

Jakub Dotlačil

Cue-based retrieval models have been successful in simulating behavioral measures in the resolution of syntactic dependencies, e.g., subject-verb or relative pronoun-verb dependencies ([1], [2], [3], a.o.). In this work, it will be shown that cue-based retrieval can go beyond modelling the resolution of dependencies. It is straightforwardly compatible with a class of parsers studied in computational linguistics (transition-based parsers, see [4]). Combining cue-based retrieval with transition-based parsing leads to novel psycholinguistic parsers, which (i) can be embedded in the cognitive architectures ACT-R, yet are data-driven, not manually coded (unlike previous ACT-R parsers) (ii) make predictions for psycholinguistic data like on-line behavioral measures without any extra stipulated linking function (unlike previous transition-based parsers in computational linguistics) (iii) are conceptually appealing since they provide a single mechanism for the retrieval of syntactic dependencies and for parsing and a single explanation of either cognitive difficulty (in contrast to previous data-driven psycholinguistic parsers, e.g., [5], [6]). The parser is tested on two data sets: Natural Stories corpus [7] and a self-paced reading study of relative clauses [8].

Cue-based retrieval assumes that memory items are content-addressable and that the memory system uses retrieval cues (e.g., *subject*, *plural* for plural subject retrieval) to find relevant items in memory. The activation of an item increases when it is matched by more retrieval cues and when the retrieval cues are more discriminating. An increase in activation increases a chance of retrieval success and decreases retrieval times ([1], among many others). **Transition-based parsing** is a parsing system that predicts transitions from one parsing state to another by finding the correct parsing step, see (1) for a shift-reduce parser for the sentence *John dances*. The parser has information about its context, represented here as S and \mathcal{W} , and chooses an action which leads to a new context (*shift* shifts the leftmost word in \mathcal{W} to the list of trees and assigns a label to it, *reduce* reduces the rightmost tree structure(s) into a novel tree). Parsing is finished when no upcoming word is present and no reduction can be done among trees. Assuming that finding the right parsing step is a case of memory retrieval and the parsing context (S and \mathcal{W}) serves as the list of retrieval cues, we can conceptualize parsing as just a special case of cue-based retrieval. In parallel with other cases of cue-based retrieval, the model predicts cognitive difficulties (increased latencies, decreased accuracies) if only few retrieval cues find match in memory and/or when retrieval cues are not discriminating because they are shared by many items in memory (cue overload).

Testing cue-based parsing: We construct and collect all parsing steps (assuming a shift-reduce parser) with their context in Penn Treebank, up to section 21 ([9]). We assume that these steps+contexts constitute the memory of the parser. When the parser parses a new sentence, it uses the cues from the current context to find the parsing step with the highest activation in its memory. The model predicts that the activation of the retrieved parsing step should negatively correlate with reading times (RTs). We test this on [7]. Using a mixed-effect model, we see that Activation is indeed a significant negative predictor of RT even after accounting for frequency, position, word length and bigram and trigram frequency, see the left table in (2). The negative Activation effect is moreover driven by the number of matching cues between the currently parsed context and the retrieved parsing step, just as cue-based retrieval predicts, see the right table in (2). To show that the approach allows us to provide a single account of parsing and the resolution of dependencies, we consider self-paced reading data from [8], which has been used to model cue-based retrieval for relative pronoun-verb dependencies. We model reading times by connecting activations (from retrieved lexical items, dependents and parsing steps) to latencies using the standard ACT-R formula (see (3) and [1]). After estimating parameters F and f *once for all types of retrieval*, we get a good fit to the data, Fig. 1. The fit is decreased when the parsing component is switched off, which shows that the good fit is (also) driven by the cue-based model of parsing.

(1) 1.Starting position:

$$\mathcal{S} = [], \mathcal{W} = [\langle \text{John}, \text{PN} \rangle, \langle \text{dances}, \text{V} \rangle]$$

2.shift

$$\mathcal{S} = [\langle \text{John} \text{PN} \rangle], \mathcal{W} = [\langle \text{dances}, \text{V} \rangle]$$

3.reduce (unary)

$$\mathcal{S} = [\langle \text{John} \text{PN} \rangle], \mathcal{W} = [\langle \text{dances}, \text{V} \rangle]$$

4.shift

$$\mathcal{S} = [\langle \text{John} \text{PN} \rangle, \langle \text{dances} \text{V} \rangle]$$

5.reduce (unary)

$$\mathcal{S} = [\langle \text{John} \text{PN} \rangle, \langle \text{dances} \text{VP} \rangle]$$

6.reduce (binary)

$$\mathcal{S} = [\langle \text{John} \text{NP} \text{VP} \rangle]$$

(2)

	Estimate	t-value
Position	0.034	1.87
Word length	10.75	16.06
Log(freq)	-0.26	-1.95
Length:Log(freq)	-0.53	-13.56
Log(bigram)	-0.004	-0.02
Log(trigram)	-0.56	-2.64
Activation	-0.14	-2.04

	Estimate	t-value
Position	0.02	1.09
Word length	11.24	18.60
Log(freq)	-0.40	-2.98
Length:Log(freq)	-0.56	-15.96
Log(bigram)	-0.20	-1.21
Log(trigram)	-1.00	-7.04
Number of matching cues	-0.29	-5.04

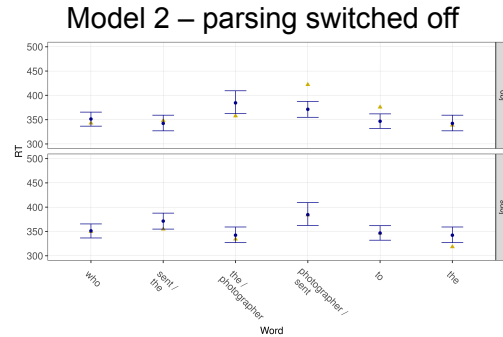
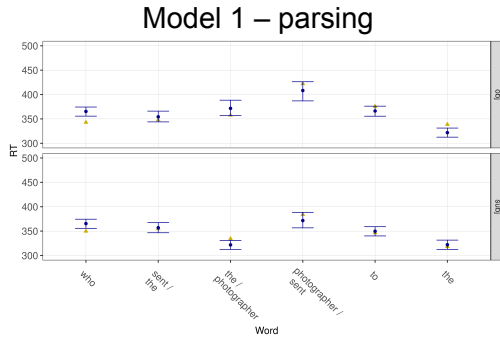


Figure 1: Modelling object-relative and subject-relative self-paced reading data from [8]. The left graphs show the predictions of the model with parsing. The right graphs show the predictions of the model without parsing. The blue dots are predicted mean RTs. The bars provide the 95% credible intervals. The yellow triangles are observed mean RTs.

(3) $T_i = Fe^{-f \cdot A_i}$ (A - activation of item i ; F, f – free parameters)

[1] Lewis et al. 2005. An activation-based model of sentence processing as skilled memory retrieval. *CogSci* 29:1–45. [2] Dillon et al. 2013. Contrasting intrusion profiles for agreement and anaphora. *JML* 69:85–103. [3] Jäger et al. 2017. Similarity-based interference in sentence comprehension. *JML* 94:316–339. [4] Nivre. 2004. Incrementality in deterministic dependency parsing. *Workshop on Incremental Parsing*, 50–57. [5] Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. The 2nd Meeting of the NAACL, 159–166. [6] Boston et al. 2011. Parallel processing and sentence comprehension difficulty. *LCP* 26:301–349. [7] Futrell et al. 2018. The natural stories corpus. *LREC 2018*, 76–82. [8] Grodner et al. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *CogSci* 29:261–291. [9] Marcus et al. 1993. Building a large annotated corpus of English. *CompLing* 19:313–330.

Competing Effects of Syntax and Animacy in Priming of Relative Clause Attachment

Melodie Yen (UCLA), Idan A. Blank* (UCLA), Kyle Mahowald* (UCSB)

Background. Structural Priming [1,2]—the increased tendency to produce a certain syntactic structure following comprehension of a sentence with the same syntactic structure—has traditionally been interpreted as evidence for the activation of abstract syntactic representations. Whereas this phenomenon is robust, its effects are often small [3], and it has been argued that priming partially (and, sometimes, completely) depends on cues that are not purely syntactic (e.g., particular words, constructional features) [4].

Here, we test priming of prepositional-phrase attachment height in relative clauses [5,6]:

1. They increased the *salary* of the **landscapers** that... **did a wonderful job** (low attachment)
 2. They increased the *salary* of the **landscapers** that... *was originally too low* (high attachment)
- In (1) the relative clause “did a wonderful job” attaches locally to “landscapers,” whereas in (2) attachment occurs higher up in the syntactic tree, to “salary.” Past studies report a priming effect for these structures: sentence stems ending before the relative clause overall elicit low-attachment (LA) completions, but high-attachment (HA) completions increase after participants read a prime sentence with a HA structure.

However, past studies of PP-attachment priming did not control for the animacy of nouns (here, “salary” and “landscapers”), which has been shown to modulate priming of other structures [7-10]. If primes and targets match in noun animacy order (e.g., inanimate before animate), then structural priming conflates a syntactic effect (activation of HA vs. LA structures) with a semantic one (activating modification of animate vs. inanimate nouns). Moreover, it remains unclear how the animacy of the target nouns themselves affects priming: for instance, given the “privileged” status of animate nouns in production [11], they might also be the preferred targets for relative clause attachment. Thus, could low attachment to an animate noun (“landscapers”) be overridden by priming of high attachment to an inanimate noun (“salary”)?

Methods. We constructed 24 pairs of prime-target stems (**Table 1**). Each stem contained a prepositional phrase with a singular, inanimate (IN) noun and a plural, animate (AN) noun. In primes, IN occurred before AN (IN/AN, see example above). We used a 2x2 factorial design, crossing prime attachment with target animacy order. Specifically, half of the primes ended with “who were” (mandating LA completions), and the other half with “which was” (encouraging HA). Target stems ended in “that”; half were IN/AN (matching the prime), and half AN/IN. 60 participants completed 24 item pairs and 30 completed 12 critical items ($n=86$, after exclusions). There were two fillers between critical items. Target completions (HA vs. LA, with HA as the dependent variable) were modeled in a Bayesian mixed-effects logistic regression with a semi-informative prior. Fixed effects were coded for 3 contrasts: whether the animate noun in the target was in high or low position, the overall structural priming effect (whether the prime is HA), and the “animacy priming” effect (whether animacy priming would predict high attachment).

Results and discussion. As shown in Fig. 1., we found a baseline animacy preference (i.e., attaching to the animate noun in the target: $\beta=1.61$, 95% credible interval [.51, 2.68]) and a robust structural priming effect ($\beta=2.53$, 95% credible interval [1.60, 3.61]). When prime attachment and target animacy conflicted (see cyan data), animacy prevented priming: participants made descriptively *more* HA completions when the primes were LA but the AN target was high, and *fewer* HA completions when the primes were HA but the AN target was low. We did not find a significant effect of “animacy priming.” We conclude that structural priming in RC-attachment cannot be fully explained in terms of animacy effects. However, the structural priming effect is modulated by animacy attachment preference in a way not addressed by prior work on relative clause attachment priming. This work points towards the need for a broader effort to integrate syntactic accounts of structural priming with semantic and cue-based factors.

		Target Animacy Order	
		IN/AN (inanimate before animate)	AN/IN (animate before inanimate)
Prime Attachment	LA (low)	Prime: "We passed the property of the landowners who were..." Target: "They increased the salary of the landscapers that..."	Prime: "The assistant announced the score of the contestants who were..." Target: "The police arrested the inhabitants of the building that..."
	HA (high)	Prime: "The reporter visited the district of the voters which was..." Target: "The spy described the hideout of the rebels that..."	Prime: "The landlord reviewed the lease of the tenants which was..." Target: "The secretary contacted the signers of the petition that..."

Table 1. Experimental design and examples of stimuli

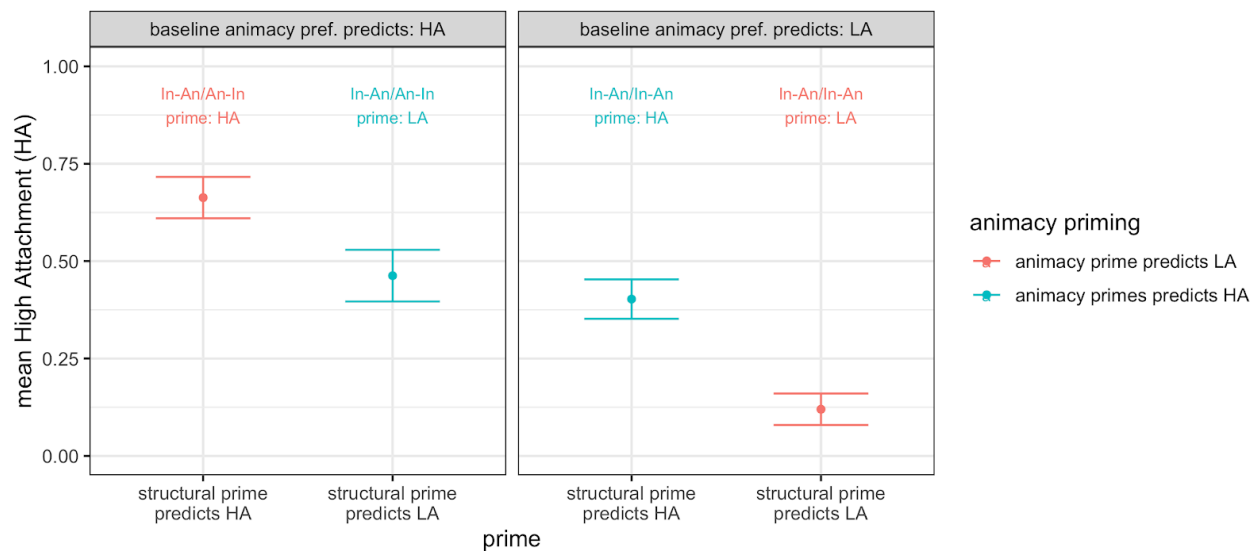


Figure 1. Mean high attachment (HA) completions of the target stem, as a function of whether (1) the animate item in the target is high or low (left vs. right panel); (2) the structural prime predicts high or low attachment (x-axis: structure primes HA vs. structure primes LA); and (3) whether “animacy priming” predicts high or low attachment (red vs. cyan). Error bars are 95% confidence intervals over participant means.

References

- [1] Bock (1986) *Cog. Psych.*; [2] Pickering & Ferreira (2008), *Psych. Bull.*; [3] Mahowald *et al.* (2016), *J. Mem. Lang.*; [4] Ziegler *et al.* (2019), *Cognition*; [5] Scheepers (2003), *Cognition*; [6] Desmet & Declerq (2006), *J. Mem. Lang.*; [7] Carminati, *et al.* (2008), *J. Exp. Psych: LMC*. [8] Bucklet *et al.* (2017), *Front. Psych.* [9] Bock *et al.* (1992), *Psych Rev.* [10] Huang, *et al.* (2016), *J. Mem. Lang.*; [11] Desmet, Brysbaert, & De Baecke (2002), *Quart. J. Exp. Psych. A*.

Language modeling using a neural network shows effects on N400 beyond just surprisal

Don Bell-Souder, Shannon McKnight, Vladimir Zhdanov, Sean Mullen, Akira Miyake, Phillip Gilley, and Albert Kim

(University of Colorado Boulder, Institute of Cognitive Science)

Electroencephalography (EEG) has provided evidence that the brain makes word-level predictions (Kuperberg & Jaeger, 2016; Van Berkum et al., 2005). However, such evidence comes from target words appearing in highly constraining sentence contexts, raising important questions about the generalizability of the effects. Here, we examined brain activity elicited by sentence-embedded words that varied substantially in their contextual support. We used a Long Short-Term Memory (LSTM) neural network to generate context-driven predictions about each word, modelling the predictions that human comprehenders might make (Sundermeyer et al., 2015). By comparing the model-generated predictions to the brain activity at each word, we evaluated the degree to which comprehenders were actually predicting words during sentence comprehension (inspired by earlier work by Frank et al., 2015; Frank & Hoeks, 2019).

The LSTM predictions were tested against EEG data collected from 190 young adults (ages 18-30) reading 400 experimental sentences (RSVP format), of which 240 were well-formed. For 5440 words, we quantified the amplitude of the N400 ERP component as mean voltage 300-500 ms post-stimulus-onset averaged across seven central-parietal channels. Substantial past research indicates that the N400 amplitude reflects the ease with which a word is accessed (Kutas and Federmeier, 2011). N400's for each word were averaged across participants.

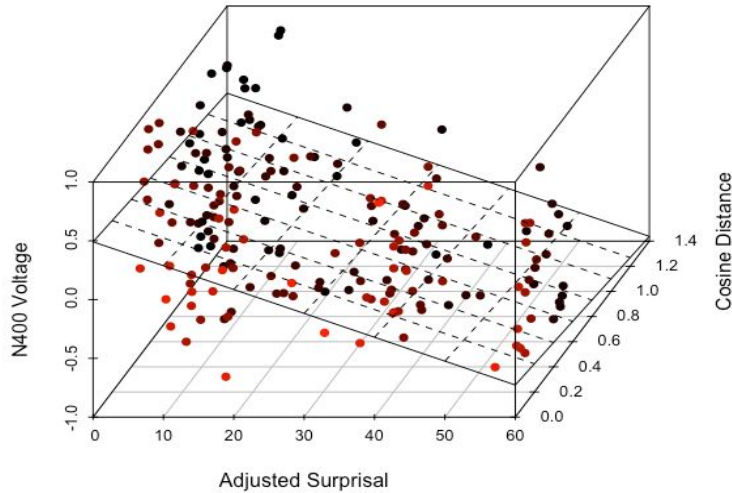
We initialised our LSTM on pretrained 300-dimensional Word2Vec embeddings, which were pretrained on the English Wikipedia corpus, and fine-tuned the LSTM using 1043 well-formed sentences from prior studies that were not included in the current study.

We used the LSTM to generate four predictors of brain activity. First, we used the LSTM to generate a distribution of conditional probabilities for words, given the previous context, and from this calculated the surprisal of each presented word. Second, we developed a novel measure of adjusted surprisal, by subtracting the surprisal of the word most predicted by the LSTM from that of the presented word. This quantifies the surprisal of a word after accounting for the level of surprisal that might have been anticipated given the context. Third, we calculated each perplexity at each word with it's left context. Finally, we calculated the cosine distance between Word2Vec embeddings for the presented word and the LSTM predicted word, reflecting the semantic distance between the most likely and actually presented words.

Regressing the N400 measures on our four candidate predictors, we found that cosine distance, surprisal, and adjusted surprisal were significant predictors. Greater cosine distance predicted more negative N400s ($F(1,5437) = 11.9, p < 0.001$) and increased adjusted surprisal also predicted more negative N400s ($F(1,5437) = 1159.4, p < 0.001$) when controlling for the other (Figure 1). Surprisal and adjusted surprisal were highly correlated and therefore could not be evaluated in the same multiple regression model, but a model with cosine distance and adjusted surprisal outperformed one with cosine distance and surprisal.

These analyses show that the LSTM framework is a useful tool to examine EEG responses. More importantly, it shows that adjusted surprisal is a valuable way to quantify how the N400 is reflecting the difference in what the brain was already prepared to process and what it actually received. We plan to continue this analysis to examine if the same measures also explain other EEG components like the P600 or if they are differentially explained by different measures.

Figure 1



The regression plane of N400 voltage on adjusted surprisal and cosine distance. Also plotted are a random selection of 200 points representing individual words in the analysis.

Definitions of computed measures

Surprisal

$$S(w_i) = \log\left(\frac{1}{P(w_i|w_1 \dots w_{i-1})}\right)$$

Perplexity

$$PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i|w_1 \dots w_{i-1})}}$$

Cosine Distance

$$CD(u, v) = 1 - \frac{u \cdot v}{\|u\|_2 \cdot \|v\|_2}$$

Table 1 - example sentences

Type	Example
Simple	My brother came into the room and looked around.
Complex	The quarterback that the fat bully ran from yelled for help.
Well-formed Control	The webs were spun by a spider this morning.
Semantic Anomaly	The webs were <i>spun</i> by a clown this morning.
Syntactic Anomaly	The webs were spun by from spider this morning.

References

- Frank, S. L., & Hoeks, J. C. (2019). The interaction between structure and meaning in sentence comprehension. *Recurrent neural networks and reading times*.
- Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, 140, 1-11.
- Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience*, 31(1), 32-59.
- Sundermeyer, M., Ney, H., & Schlüter, R. (2015). From feedforward to recurrent LSTM neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3), 517-529.
- Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 443.

The Posterior P600 reflects Reanalysis but not Repair

Edward Alexander¹, Trevor Brothers¹, Gina Kuperberg¹

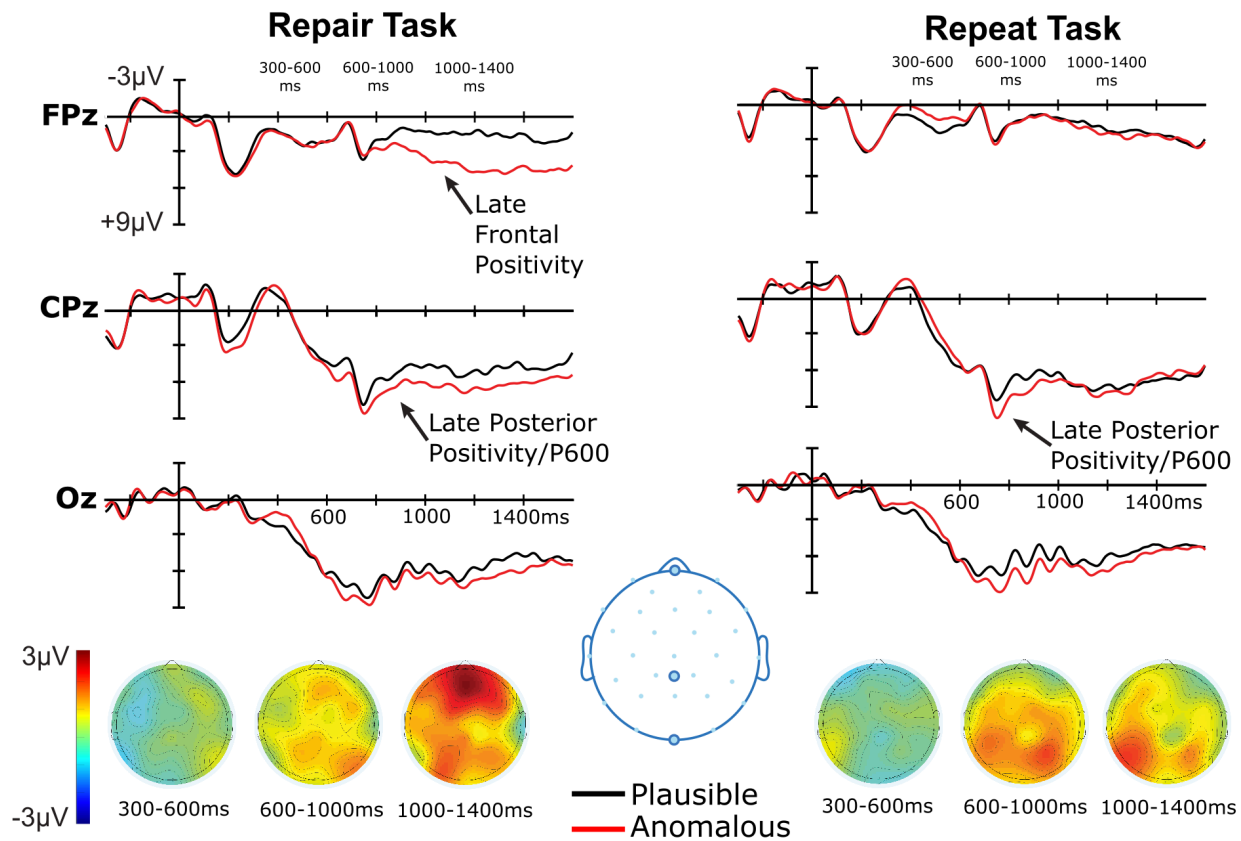
¹Tufts University, Medford, MA 02155

Language comprehension requires us to infer the underlying message being communicated. However, this message can be complicated by the presence of errors, ambiguities and misperceptions. Rather than passively accepting these errors, comprehenders sometimes engage in additional analysis of the input, which manifests as a late posteriorly distributed positive-going waveform, known as the P600 [1,2]. However, it has been unclear whether the late posterior positivity/P600 simply reflects a re-analysis of the input (reprocessing in attempt to gather more information), or whether it additionally reflects attempts to actively repair the input (i.e. actively change its surface features) to re-establish coherence. To distinguish between these accounts, we manipulated task requirements as participants read sentences, which were either plausible or anomalous (*The judge's gavel was banged/*pardoned*). Based on the prior literature, these semantically attracted anomalies were likely to elicit a late posterior positivity/P600 effect [2,3]. In both tasks, participants indicated after each sentence whether it was plausible or anomalous. In the 'Repeat' task, they then repeated the sentence exactly as it was presented, making repair difficult. In the 'Repair' task, they repaired any errors (if present) and spoke the corrected versions of the sentences aloud. This resulted in a 2 x 2 within-participants design that crossed Task (Repeat or Repair; blocked with order counterbalanced) and Plausibility (Plausible or Anomalous). If the late posterior positivity/P600 effect only reflects detection of conflict and re-analysis, then its magnitude should be the same in both the Repeat and the Repair tasks. Conversely, if it is only elicited when participants engage in linguistic repair, then it should be seen in the Repair but not in the Repeat task. Finally, if linguistic re-analysis and repair processes involve distinct cognitive operations, we may observe two separate neural components across tasks, potentially with different time-courses or scalp topographies.

Methods: 21 participants read 192 scenarios (96 anomalous and 96 plausible) while EEG was recorded. All sentences followed the form "[article/pronoun] [adjective] [noun] [was/were/had been] [verb]", and all nouns were semantically attracted to the preceding verb. To assess differences across conditions, we extracted ERPs to the critical verbs, and carried out 2x2 repeated measures cluster mass univariate ANOVAs across all scalp electrodes within a common P600 time window (600-1000ms) as well as within a later 1000-1400ms window. We also examined effects within an earlier N400 time window (300-600ms).

ERP Results: Between 600-1000ms, we observed a main effect of Plausibility due to a larger P600 to anomalous than plausible completions (spatial mass peak: P4, extent: 627-1000ms, $p < 0.001$). However, there was no main effect of Task, or Task x Plausibility interaction. This positive-going effect continued into the later 1000-1400ms time window in both the Repeat and the Repair tasks (a main effect of Plausibility, Spatial mass peak: AF3, extent: 1000-1400ms, $p < 0.001$). At frontal sites, however, the effect appeared to be much more robust in the Repair than the Repeat task. Statistically, this was reflected by a cluster that showed an interaction between Plausibility and Task, which was limited to frontal sites (Spatial mass peak: AF4, extent: 1000-1400ms, $p = .03$). Follow-up analyses confirmed that the cluster showing a main effect of Plausibility extended to all these frontal sites in the Repair task, but not the Repeat task. No significant clusters were observed within the 300-600ms time window.

Discussion: The presence of a late posterior positivity/P600 between 600-1000ms in both the Repeat and Repair tasks suggest that this component does not reflect repair processes, but instead reflects the diagnosis of a comprehension error and re-analysis of the input [4]. In contrast, in the current paradigm, linguistic repair processes were associated with a still later frontally distributed positivity (1000-1400ms). We suggest that this reflected the re-establishment of coherence after comprehenders successfully repaired the anomalies, following previous work linking late frontal positivities to successful shifts in the discourse model [5,6].



References:

- [1] Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785-806.
- [2] Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1), 117-129.
- [3] Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2), 205-225.
- [4] Van de Meerendonk, N., Kolk, H. H., Chwilla, D. J., & Vissers, C. T. W. (2009). Monitoring in language perception. *Language and linguistics compass*, 3(5), 1211-1224.
- [5] Kuperberg, G. R., Brothers, T., & Wlotko, E. (2020). A Tale of Two Positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience*, 32(1), 12-35.
- [6] Brothers, T., Wlotko, E. W., Warnke, L., & Kuperberg, G. R. (2020). Going the extra mile: Effects of discourse context on two late positivities during language comprehension. *Neurobiology of Language*, 1(1), 135-160.

Modeling subcategorical information maintenance in spoken word recognition

Wednesday Bushong (University of Hartford; bushong@hartford.edu) and T. Florian Jaeger (University of Rochester; fjaeger@ur.rochester.edu)

Language understanding requires listeners to integrate large amounts of perceptual information before it overwhelms sensory memory. However, cues to linguistic units (sounds, words, etc.) are *distributed* across the speech signal. How then do listeners manage what kinds of subcategorical information they maintain in memory and for how long? Consider the bolded word in Table 1: previous work has found that both the initial acoustic cues on the word itself (e.g., voice onset time) and later context (e.g., forest/fender) affect listeners' interpretation of the word as "tent" or "dent" [1-3,7]. If such evidence reflects subcategorical information maintenance beyond the word boundary, this challenges assumptions in models of word recognition (e.g., [4-6]). Formal models of cue integration across time that can address this question have, however, been lacking. We develop four competing models with different levels of information maintenance, and test them against data like [1-3,7].

Models. We develop four computational models of information maintenance. Figure 1(A-B) displays the formalization and behavioral predictions of each model. The *ideal integration* model assumes that listeners maintain subcategorical information about all cues in the signal over time, and thus optimally integrate those cues (proposed in [7]). The *ambiguity-only* model assumes that listeners are more likely to maintain subcategorical information over time when that information is perceptually ambiguous, and less likely when it is unambiguous (proposed in [1]). The *categorize-discard* model assumes that listeners maintain no subcategorical information about cues over time (proposed in [4]). Finally, we introduce a novel model, *categorize-discard-switch*, which assumes that listeners do not maintain subcategorical information about cues over time, but may change their categorization decisions based on subsequent cues. Notably, several of these models make similar *qualitative* predictions about human behavior, *despite the fact that they make different assumptions about information maintenance*. This highlights the importance of testing these models quantitatively by fitting them directly to behavioral data.

Experiments. We fit all models to four different behavioral experiments ($N_s = 39, 37, 48, 51$), three of which are from published sources [7-9]. Participants listened to sentences like those in Table 1 and responded whether they heard "tent" or "dent". Both voice-onset time (VOT) of the target and the bias of the later context were manipulated. **Analysis.** All four models are non-linear mixture models. We implemented hierarchical (mixed-effects) instances of these models using the `brms` package in R, and fit them against the data from the behavioral experiments. For each experiment, we measure the performance of the four models as the estimated log predictive density ($elpd_{waic}$)—a measure suitable for non-nested comparison of models with inherently different functional flexibility. **Results.** Figures 1(B) and 2 display model performance for all experiments. In all experiments, the ideal integration model outperformed the ambiguity-only model (Experiments 1, 2, 4: $\Delta elpd_{waic} < 2.5$ SEs \rightarrow "weak" evidence; Experiment 3: $\Delta elpd_{waic} > 5$ SEs \rightarrow "strong" evidence). Further, both the ideal integration and ambiguity-only models strongly outperformed the categorize-discard and categorize-discard-switch models ($\Delta elpd_{waic} > 5$ SEs).

Conclusions. We find consistently strong evidence in favor of the models which posit maintenance of subcategorical information over time. This suggests that listeners are able to maintain subcategorical information about prior linguistic input even beyond the word boundary, in contrast to theories which posit that listeners must immediately discard such information due to memory bottlenecks (e.g., [4-6]). That listeners have much more information available to them over time highlights the need for new theories of speech recognition. This work also demonstrates the importance of formalizing quantitative models of behavior to distinguish between different theories.

References. [1] Connine et al. (1991) *JML* [2] McMurray et al. (2009) *JML* [3] Brown-Schmidt & Toscano

(2017) *LCN* [4] Christiansen & Chater (2016) *BBS* [5] McClelland & Elman (1986) *Cog Psych* [6] Luce & Pisoni (1998) *Ear & Hearing* [7] Bicknell et al. (submitted) [8] Bushong & Jaeger (2017) *CogSci* [9] Bushong & Jaeger 2019 *JASA-EL*

Context	Sentence
Tent-biasing	When the ?ent in the forest was well camouflaged, ...
Dent-biasing	When the ?ent in the fender was well camouflaged, ...

Table 1: Example stimuli from Experiments 1-4. “?” indicates a sound along the /t/-/d/ continuum with varying voice onset time (VOT).

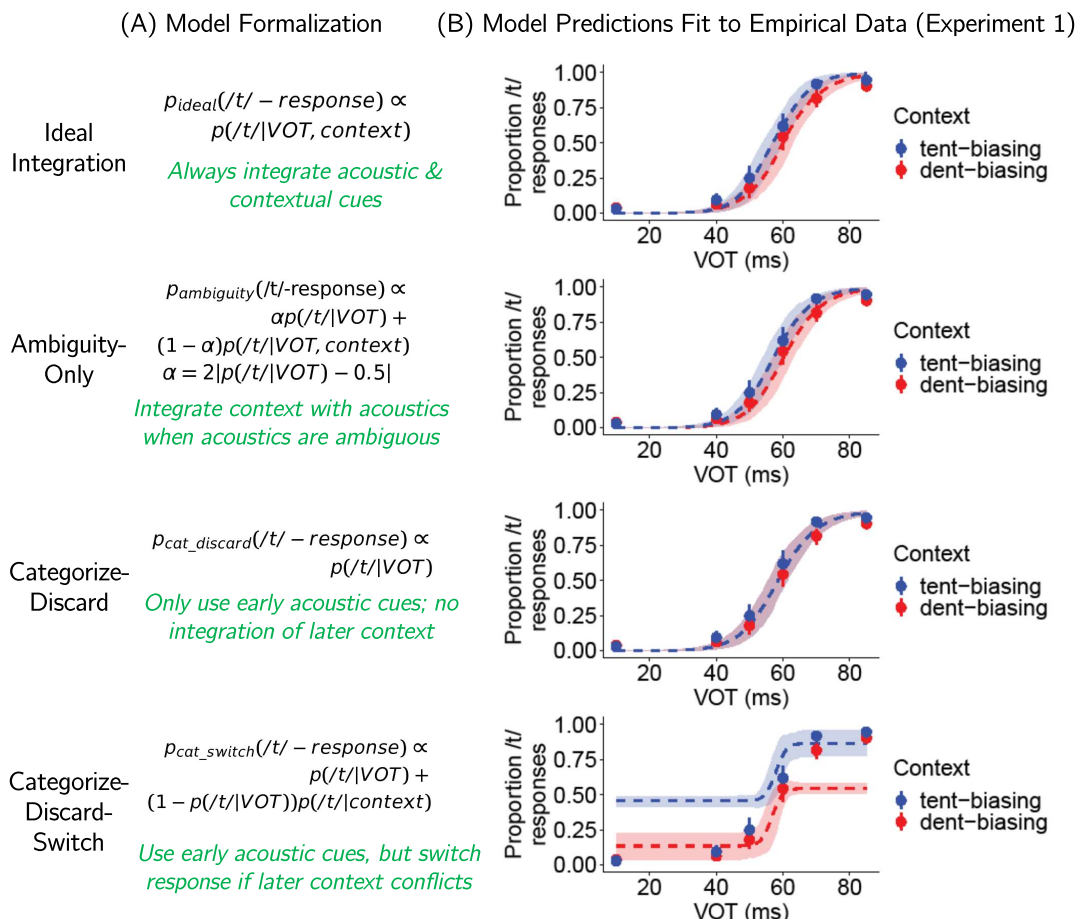


Figure 1: (A): Formalization of each model. (B): Model predictions (dashed lines) fit to empirical data (points; identical across rows). Shaded intervals are 95% confidence intervals.

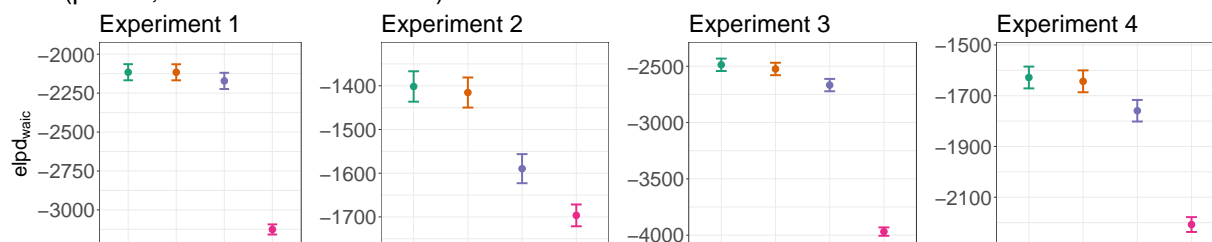


Figure 2: Model fits ($elpd_{waic}$) for Experiments 1-4 (higher values \rightarrow better fit): **ideal integration**, **ambiguity-only**, **categorize-discard**, **categorize-discard-switch**.

Interpreting implausible sentences: The role of phonological similarity

Jianyue BAI, Zhenguang Cai (The Chinese University of Hong Kong)

People sometimes believe that an implausible sentence (*The father handed the gun the son*) results from mis-perception/production (e.g., from the omission of *to*) and would thus edit the sentence by inserting *to* (or deleting *to* in the case of *The father handed the son to the gun*) to arrive at a plausible interpretation (Gibson et al., 2013). Both the insertion and deletion edits require a change to the syntax of the original sentence. However, a third edit is possible: people can assume that *the son* and *the gun* were accidentally swapped in perception/production and can thus arrive at a plausible interpretation by exchanging the positions of the two nouns without changing the syntax.

We examined the use of the exchange edit by manipulating the phonological similarity of nouns. Phonologically similar words, compared to dissimilar ones, are more likely to be exchanged in their position in both speech production (Dell & Reich, 1981) and in memory retrieval (Baddeley, 1966). Thus, at hearing *The father handed the gun the son*, people may believe that *son* and *gun* were accidentally swapped, either due to speech error by a speaker or due to mis-perception by themselves.

The experiment (38 participants, 40 target items, 60 fillers) adopted a design of 2 (plausibility: plausible vs. implausible) \times 2 (Structure: double-object [DO] vs. prepositional object [PO] dative) \times 2 (Similarity: dissimilar vs similar theme and recipient nouns) (see Table 1). The theme and recipient nouns differed in the onset in the similar condition but had no phonological overlap in the dissimilar condition (e.g., *son-gun* and *son-bill*). The theme nouns (*gun* and *bill*) were the same in syllable number and comparable in frequency. A pretest revealed no difference in plausibility between plausible sentences with similar nouns and those with dissimilar nouns. In the experiment, participants listened to a sentence and then answered a yes/no comprehension question which helped to determine whether participants literally interpreted or re-interpreted a sentence.

Logit mixed-effect results showed that, consistent with Gibson et al. (2013), DO sentences were more likely to be re-interpreted than PO sentences ($\beta = 1.52$, $SE = 0.32$, $z = 4.76$, $p < .001$). More critically, sentences with similar nouns were re-interpreted more often than those with dissimilar nouns ($\beta = -1.33$, $SE = 0.54$, $z = -2.48$, $p = .013$). There is also a marginally significant interaction between structure and similarity ($\beta = 1.89$, $SE = 1.05$, $z = 1.79$, $p = .073$). Separate analyse showed that the phonological similarity effect emerged in PO sentences ($\beta = 0.82$, $SE = 0.42$, $z = -1.95$, $p = .051$) but not in DO sentences ($\beta = 0.15$, $SE = 0.80$, $z = 0.19$, $p = .85$).

If people only apply insertion or deletion on the preposition *to* in the comprehension of implausible dative sentences, we should not expect sentences with similar nouns to be more often re-interpreted than those with dissimilar nouns. The results thus suggest an additional edit on the nouns themselves. We believe that people sometimes arrive at a plausible re-interpretation of an implausible sentence with similar theme and recipient nouns by exchanging positions of the nouns. We are in the process of doing a structural priming experiment to further test this exchange account.

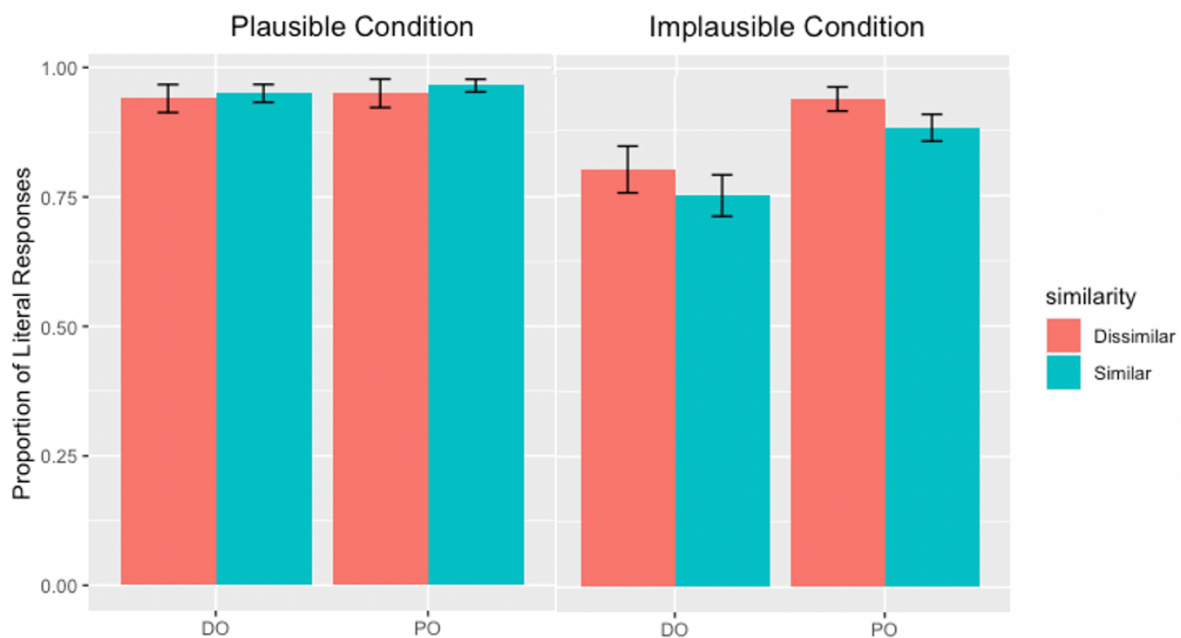
Baddeley, A. D. (1966). *Quarterly journal of experimental psychology*, 18, 362-365.

Dell, G. S., & Reich, P. A. (1981). *Journal of verbal learning and verbal behavior*, 20, 611-629.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). *PNAS*, 110, 8051-8056.

Table 1: Example item

Plausibility	Similarity	Structure	Sentence
plausible	similar	DO	The father handed the son the gun.
implausible	similar	DO	The father handed the gun the son.
plausible	dissimilar	DO	The father handed the son the bill.
implausible	dissimilar	DO	The father handed the bill the son.
plausible	similar	PO	The father handed the gun to the son.
implausible	similar	PO	The father handed the son to the gun.
plausible	dissimilar	PO	The father handed the bill to the son.
implausible	dissimilar	PO	The father handed the son to the bill.

**Figure 1. Proportion of literal interpretations across conditions. Error bars represent SEs.**

The Stability of Individual ERP Response Dominance Within and Across Conditions

Tamarae Hildebrandt & Jonathan R. Brennan (University of Michigan)

Introduction. Prior research shows individual differences in ERP responses, specifically in the relative prominence of the N400 and P600 components to agreement violations [1, 2]. Participants exhibited stability in their response dominance across conditions—reflecting a systematic positive, negative, or biphasic response to the agreement violations [1]. A significant question is whether response dominance is systematic within participants across other constructions that traditionally elicit a P600 response. If participants show a positive response dominance in one condition, will the response dominance remain positive in another condition. So, using magnitude and Response Dominance Index (RDI) [1], we explored (1) whether the dominance effect is stable across the different violations that traditionally elicit a P600 response and (2) if the dominance effect remains stable within participants across different violations? Preliminary results show that within a single condition the participants do show stable dominance effects to violations that are known to traditionally elicit a P600 response; however, within individual participants, this dominance effect is not stable across the other violations.

Methods. 520 sentence stimuli were divided into target (160), control (120), and filler (240) conditions. The stimuli were separated into two lists, such that each participant read and rated the acceptability of 260 American English sentences. The target word is bolded, and predicted violations (e.g., ungrammatical, dispreferred, or infelicitous) are marked with an asterisk (*).

(1) Example Stimuli from the Target, Control, and Filler Conditions [3,4]

- a. *Complementizer:* The belief **that** \emptyset_{Det} seven baristas are coffee snobs is widely accepted.
- b. *Without Comp:* The belief \emptyset_{comp} **these** seven baristas are coffee snobs is widely accepted.
- c. *Subject-Verb Agreement:* The cats **meow/*meows** by the window watching the birds.
- d. *Gender Reflexive:* The elderly gentleman fixed **himself/ *herself** up for the dance.
- e. *Lexical Semantic:* The child borrowed some **books/*conversations** from the library.

Procedure & Analysis. EEG data recorded from 22 adults is presented. Participants read sentences word-by-word (300ms word presentation, 200ms ISI) using the Rapid Serial Visual Presentation paradigm [1]. At the end of each trial, participants rated sentence acceptability using a four-point Likert-scale (1 *unacceptable* - 4 *acceptable*) [5]. Raw EEG data sampled at 500 Hz was band-pass filtered between 0.5–40 Hz and divided into 1300ms epochs around each target word in a sentence. Ocular signals were removed with ICA, and other artifacts were visually identified and excluded [6]. Two averaged amplitude time points were extracted—an N400 (300-500ms) and a P600 (500-800ms) timeframes—from a large centro-parietal ROI (C3, Cz, C4, CP1, CP2, P3, Pz, P4) [1]. Effect magnitudes were then calculated for the N400 (grammatical minus ungrammatical) and the P600 (ungrammatical minus grammatical) (Fig 1A). Using these effect magnitudes, the RDI metric was calculated $[((\text{P600 mag} - \text{N400 mag}) / (\text{sqrt } 2))]$, which assesses the relative prevalence of the ERP response (Fig 1B) [1, 2, 7].

Results. The results indicate a stable response within a single condition, such that individuals who show a large P600 effect in one condition tend to show little negativity in that same condition and vice versa. As shown in Fig 1(A), the correlations in all 4 conditions are negative and statistically significant (between -0.73 and -0.89; $p < 0.0001$). However, the results across conditions do not show the same stability. Fig 1(B) indicates that participants may not always show the same RDI across the conditions, which traditionally elicit a P600 response. Some participants remain positive-going, negative-going, or switch dominance responses as represented by the 4 symbols in select participants. Fig 1(C) shows the grand average ERP waveforms for all participants and the 3 separate RDI groupings for each condition. Graph C2 (P600 dominant) shows a statistically significant difference as compared to A2 (all participants), where the statistically significant effect has disappeared. **Conclusion.** More work is needed to understand the different processing strategies participants employ to process the “traditional P600 violations” to explain why the neural response differs from the “traditional” predictions.

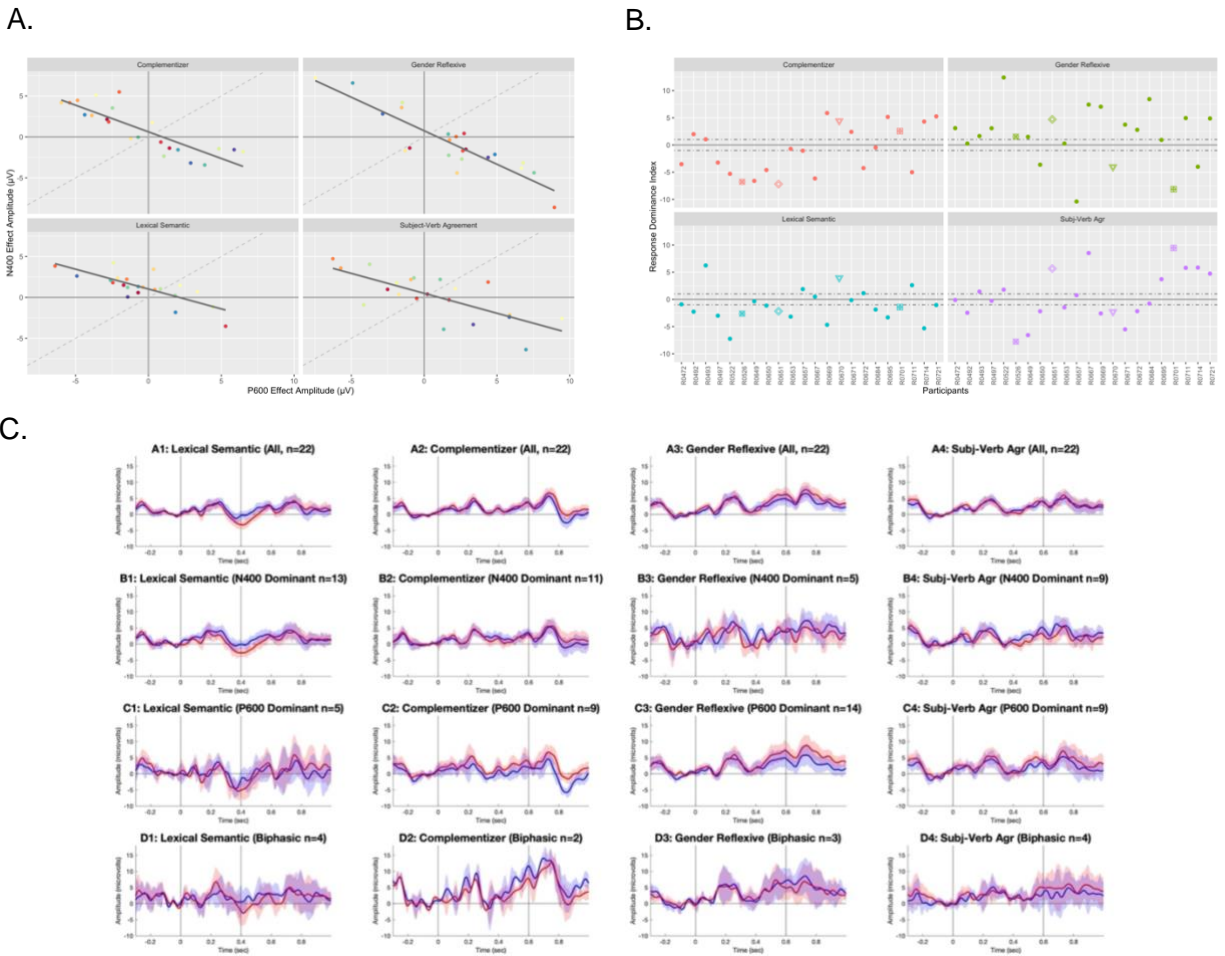


Figure 1. (A) N400 & P600 Magnitude Effects in each Condition. Participants are represented by the different colored dots. Dots above/to the left of the dashed line are individuals who show a primarily N400 effect, while dots below/to the right of the dashed line are individuals who show a primarily P600 effect. **(B) Individual Response Dominance Index (RDI) by Condition for each Participant.** Participants above 1 show a positive response, below -1 show a negative response, and between 1 and -1 show a biphasic response. 4 participants are shown in different symbols to show different dominances across conditions. **(C) ERP waveforms grouped by all, N400 dominant, P600 dominant, and biphasic.** The preferred conditions are shown in red (e.g., predicted grammatical), and the dispreferred conditions in blue (e.g., predicted ungrammatical).

Selected References. [1] Tanner, D. (2019). Robust neurocognitive individual differences in morphosyntactic processing: A latent variable approach. *Cortex* 111, 210-237. [2] Tanner, D., & Van Hell, J.G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia* 56, 289-301. [3] Tanner, D. (2018). "General files for "Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach"", <https://doi.org/10.7910/DVN/DKEKBH>, Harvard Dataverse, V1. [4] Martin, R. (2001). Null Case and the Distribution of PRO. *Linguistic Inquiry* 32(1), 141-166. [5] Dröge, A., Fleischer, J., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2016). Neural mechanisms of sentence comprehension based on predictive processes and decision certainty: Electrophysiological evidence from non-canonical linearizations in a flexible word order language. *Brain Research* 1633, 149-166. [6] Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, Mass.: MIT Press. [7] Tanner, D. (2018). "Analysis Scripts and Outputs for "Robust neurocognitive individual differences in grammatical agreement processing"", <https://doi.org/10.7910/DVN/031YTY>, Harvard Dataverse, V2.