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.

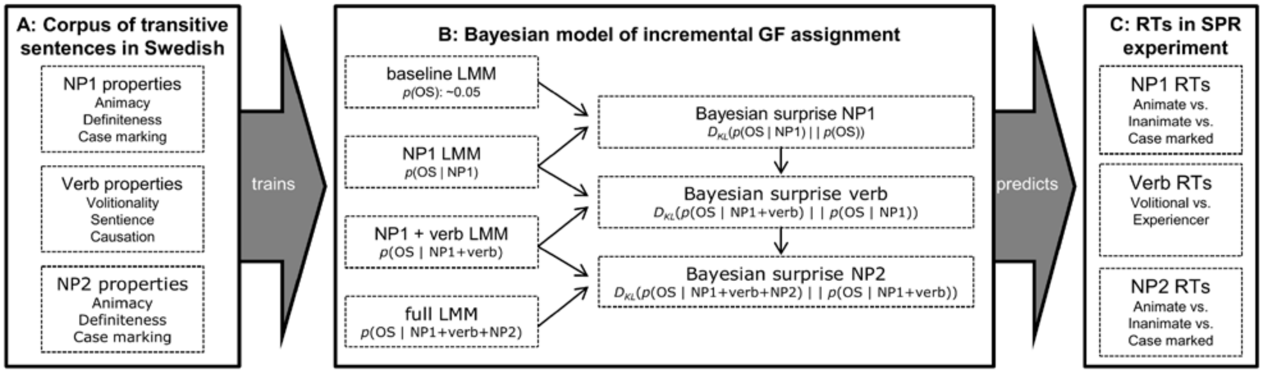

Figure 1. Illustration of the Bayesian model of incremental GF-assignment.

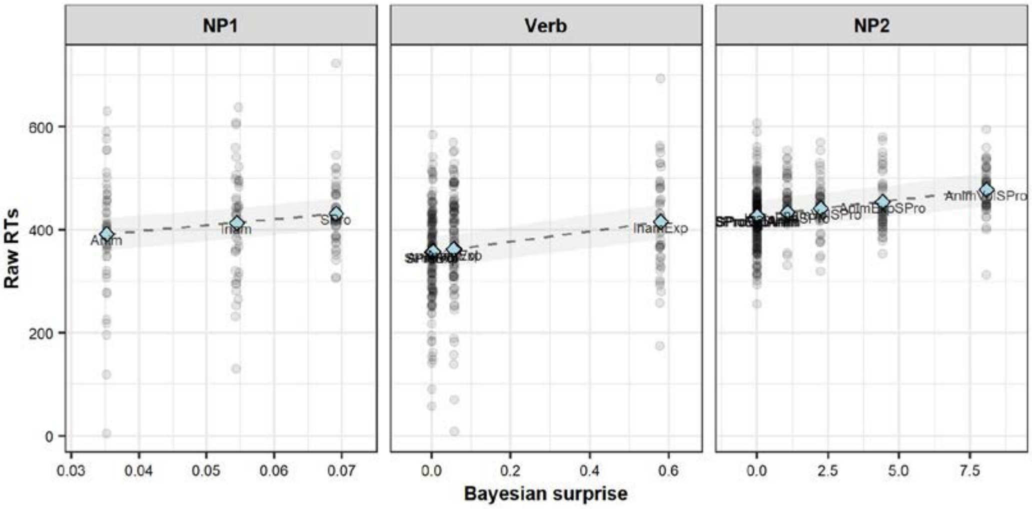

Figure 2. The relationship between Bayesian surprise as predicted by the model and raw reading times.