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≤N/lnN** 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—á 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/ln59=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/ln43=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. Vater*s* Buch ('father's book'), *Mann*s* Buch ('man's book')
 c. *das Mann*s* Nachbar*s* 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.