Quantifying early syntactic productivity in child language corpora:
A critical evaluation of the Traceback method

A key question in language acquisition research is how children learn to use language productively. Usage-based approaches assume that children's early utterances are item-based, suggesting that productivity is emerging gradually. A number of influential corpus-based studies (e.g. Cameron-Faulkner et al. 2003, Dąbrowska & Lieven 2005, Vogt & Lieven 2010) have tested this hypothesis empirically using the so-called traceback method, which aims to account for the piecemeal emergence of constructional patterns. The underlying rationale of this method is to identify recurring chunks as well as slot-and-frame patterns in a small subsample of a language acquisition corpus, the so-called test corpus. These patterns are subsequently traced back to earlier utterances which constitute the so-called main corpus (Kol et al. 2014). If utterances occur verbatim in previous recordings, a chunk is established, e.g. *I want it*. If the target utterance is matched only partially, the procedure can yield a slot-and-frame pattern such as [*I want* X]. The method starts from the assumption that children's linguistic output can be attributed to a limited number of patterns which are reused with different slot fillers and thus contribute to the increasing productivity and creativity of children's linguistic output.

However, a number of important questions remain unclear. On the one hand, the individual parameters posited for the traceback studies could have an influence on the results. For example, the division into test and main corpus is relatively arbitrary (see e.g. Kol et al. 2014). Also, the composition of the main corpus differs between studies: Some take the parental input into account, others don't. Secondly, the threshold which establishes a sequence as an "entrenched" unit is different in each study (ranging from one to four occurrences, see e.g. Dąbrowska 2014). Thirdly, the operations used to identify schemas are also defined differently, and fourth, the constructions used by the child are highly context-dependent. Thus, factors such as the specific recording time can also significantly influence the traceback result. In addition, the question arises how informative the results of traceback studies are with regard to modelling an emerging network and to what extent such models can be considered cognitively realistic (see also Ibbotson et al. 2019).

To address these questions, we use four dense corpora of monolingual German-speaking children in the age span from 2;1 to 2;6 to replicate a traceback study (Author1 2019) with different parameter settings: We manipulate a) the frequency threshold and b) the composition of the main corpus (with or without input). In addition, we follow Kol et al. (2014) in using target corpora with reversed and randomized word order, and we check for the relevance of individual differences by tracing the utterances in each child's target corpus back to all other children's main corpora. This allows us to elaborate how the results vary and which consequences this has for future applications. Ultimately this also touches upon the question if abstract schemas exist at all, whether they are cognitively plausible, or, if not, whether purely exemplar-based models should be given preference instead (e.g. Ambridge 2019, Schmid 2020).

References

Ambridge, Ben. 2019. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*. doi 10.1177/0142723719869731

Lieven, Elena, Heike Behrens, Jennifer Speares & Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30. 333–370.

Dąbrowska, Ewa. 2014. Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics* 25(4). 615–653.

Dąbrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16(3). 437–474.

Ibbotson, Paul, Vsevolod Salnikov & Richard Walker. 2019. A dynamic network analysis of emergent grammar. *First Language* 39(6). 652–680.

Kol, Sheli, Bracha Nir & Shuly Wintner. 2014. Computational evaluation of the Traceback Method. *Journal of Child Language* 41(1). 176–199.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: usage, conventionalization, and entrenchment*. Oxford: Oxford University Press.

Vogt, Paul & Elena Lieven. 2010. Verifying Theories of Language Acquisition Using Computer Models of Language Evolution. *Adaptive Behavior* 18(1). 21–35.