Processing causatives in first language acquisition: a computational approach

Guanghao You, Moritz Daum, Sabine Stoll
guanghao.you@uzh.ch

One of the most challenging tasks of infant language learners is to extract meaning from the speech stream they hear. Currently, it is not yet clear (i) how a child's expressive speech is related to the information distribution in child-directed speech and (ii) how such distributions in child-directed speech, in contrast to adult-directed speech, are shaped to facilitate meaning extraction. To answer these questions, **we simulate speech processing** with a computational approach probing the distributional features in the different speech genres, namely child speech, child-directed speech, and, as a baseline, adult-directed speech. **The goal of this study** is to investigate the semantic representations of lexical causatives (e.g. "break" and "open" in English) in these different genres, so as to probe how child speech and child-directed speech interact over time and whether adaptation occurs between them. Here we focus on lexical causatives because they are semantically very salient but at the same time they lack explicit markers to ease processing. The semantics of lexical causatives may thus be greatly characterized by their distributional patterns of their surrounding linguistic units.

Our **data** comes from the Manchester language acquisition corpus (Theakston et al., 2001). This longitudinal corpus includes naturalistic speech of 12 children (age range 20-36 months) and their caretakers (child speech: 249,574, child-directed speech: 374,226 utterances). To access the semantics of lexical causatives **we employ three steps**: (i) we use the skip-gram word embeddings (word2vec), an algorithm inferring from co-occurrence patterns (Mikolov et al., 2013), to obtain the vectors of semantic representations of words in high-dimensional space and (ii) we construct an unweighted and undirected graph, with available prototypical causatives (Haspelmath, 1993) and their *n* most similar words in the word2vec model as the vertices (*n* determined by a fixed ratio 0.01 to the vocabulary size). The edges of the graph are the links between these similar pairs of words. Lastly, (iii) we remove the leaf vertices (with degree one), as they do not bridge any words in the graph and hence are of little relevance. This constructed graph represents the causative network with the most relevant semantics. We apply this approach to the accumulated speech by age (monthly intervals) for each child. Subsequently, we employ the average degree of vertices in the graph as the metric to indicate the connectivity of the causative network. As a baseline of adult-directed speech, we conduct the same steps for a subset (326,359 utterances) of the spoken corpus in the British National Corpus. **The results** show: **first**, all 12 children have developed a causative network with a moderate average node degree, showing an increase over time (ranging from 2.00 to 3.54; see Figure 1 and Figure 2). **Second**, we find that the networks in child and child-directed speech are positively correlated ($r(148) = .71$, $p < .001$) and the connectivity in both networks increases over time towards the network we find in adult-directed speech. **Third**, the networks in child-directed speech are always slightly more complex than the networks in child speech and this difference remains constant over time (95% CI: (-0.035, 0.032)). **Fourth**, in the latest recordings child-directed speech approaches adult-directed level (Figure 1).

To summarize, our results show that not only child speech but also child-directed speech to children under age 3 changes over time. We show that while remaining less complex than adult-directed speech, the speech of adults addressed to children strongly adapts to the increase of linguistic abilities of the learners they interact with. This adaptation with reduced complexity might be useful for children to ease meaning extraction and support development. Around age 3 the complexity of causative networks used by adults approaches the level of adult-directed speech, which suggests that the development of semantic complexity in child speech will be on a similar stage with a slight time delay of several months.

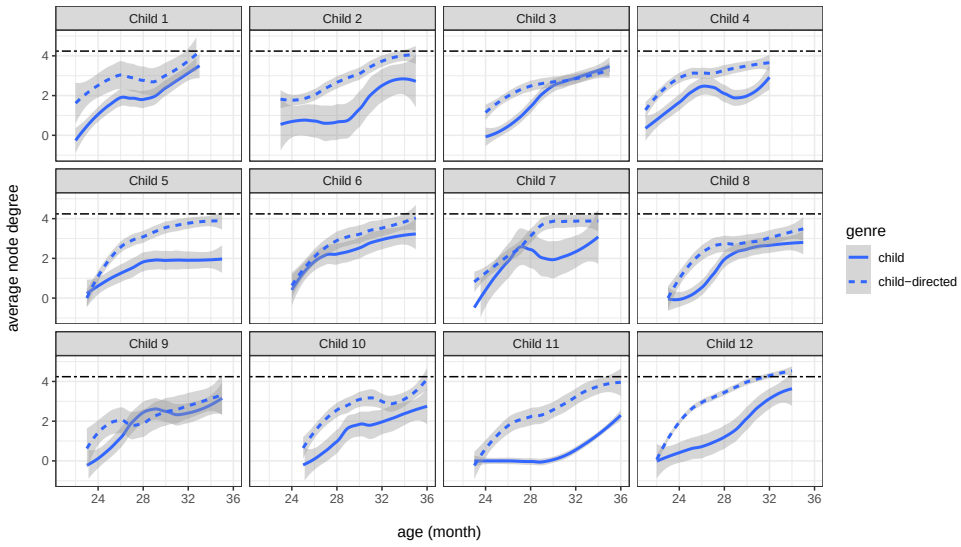

Figure 1: Average node degree of causative network in child and child-directed speech in all 12 dyads of the Manchester corpus. The black dashed line is the performance by adult-directed speech (BNC).

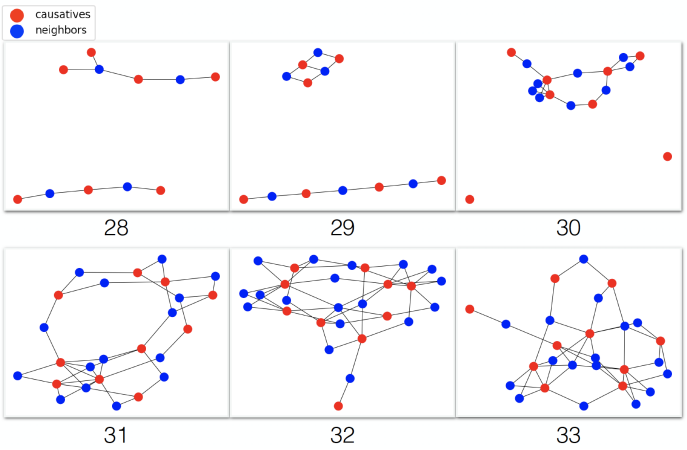

Figure 2: Visualization of the development of causative network in child speech: Child 1 from 28 to 33 months.

References

Haspelmath, M. (1993). More on the typology of inchoative/causative verb alternations. *Causatives and transitivity*, *23*, 87–121.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, *28*(1), 127–152.