Syntactic categories derived from frequent frames benefit early language processing in English and ASL

Early acquisition strategies likely don't yield adult knowledge directly, but instead provide a stepping stone to later knowledge (Frank et al. 2009, Phillips & Pearl 2015). One beneficial effect of preliminary knowledge could be that language input becomes easier for children to process, given their limited cognitive resources. We leverage the idea that more predictable language is easier to process (Levy 2008) and assess language predictability (and thus language processing ease) with the information-theoretic measure perplexity (Goodman 2001). As a case study, we investigate early syntactic categorization occurring around fourteen months (Waxman & Booth 2001), when children have minimal structural knowledge about their language. We evaluate the frequent frames (FFs) categorization strategy (Mintz 2003, Mintz 2006) on English and American Sign Language (ASL), finding that FFs derive preliminary "proto-categories" in both languages that make the child's language input easier to process (i.e., more predictable) than adult syntactic categories do. This suggests early acquisition strategies may yield knowledge that not only scaffolds children's future language acquisition but also their current language comprehension.

We selected perplexity as a measure of very early language processing because it quantifies how probable the word sequences are that children encounter in their input (Eq 1). Knowledge that makes these input data more probable also makes them more predictable and so easier to process at this stage of development -- and this makes their perplexity score lower.

The FFs strategy has had considerable success identifying adult syntactic categories for many spoken languages (Chemla et al. 2009, Weisleder & Waxman 2010, Wang et al. 2011). FFs rely on frequently encountered -- and thus likely salient -- framing units to group words together that appear in the same context. Following previous implementations for English, we derive the FF categories using word-level frames, including utterance boundaries, for both English and ASL (e.g., the frame for *READ* in *FINISH READ BOOK* = FINISH__BOOK and the frame for *BOOK* is READ__#). We use the same criterion as Mintz (2003) for determining which frames are frequent enough to be salient to a learner: a frame must capture 5% of the word types and 1% of the word tokens in the training corpus. We then apply these frames to corpora of naturalistic speech (Table 1).

We find that FF-based categories do not match adult syntactic categories very well for either English or ASL (Table 2) -- they are neither very accurate (low precision) nor very complete (low recall) when compared to adult categories. Yet, the FF-based categories for both languages make the learner's input *more* predictable than the adult categories do (lower perplexity), given the learner's minimal structural knowledge. This suggests that FFs can be used for early categorization across language modalities if the goal is to learn knowledge that is useful for a very young child. More generally, these results underscore the importance of evaluating the output of early acquisition strategies as a stepping stone to future knowledge and learning, rather than only by how well the output matches adult knowledge..

(498 words)

Eq 1: Perplexity for a sequence of words, which predicts how (un)surprising these words are, given syntactic category knowledge and a simple bigram generative model for how word sequences are produced.

$$Perplexity(Word\ sequence = w_1 w_2 ... w_N) \approx \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

$$P(w_x) = p(w_x | \text{ category}_x) * p(\text{category}_x | \text{ category}_{x-1})$$

Corpus	# utterances	avg words/utt	# tokens	# types	type/token ratio
ASLLRP	1641	6.6	10820	2321	0.215
Peter	3484	5.27	13039	930	0.071

Table 1. Description of the ASL ASLLRP corpus (Neidle & Vogler, 2012) and the English Peter corpus from CHILDES (MacWhinney 2000) used as input.

	FF-based: How close to adult categories		**Perplexity: Language processing ease**	
	Pairwise precision	***Pairwise recall***	***Adult categories***	***FF-based categories***
ASL	0.415	0.005	45.5	9.8
English	0.248	0.0164	607.9	122.6

Table 2. Categorization results, using both traditional evaluation metrics comparing the FF-based categories against the adult syntactic categories (FF-based: pairwise precision and recall) and our language processing metric (Perplexity). Precision and recall scores range from 0 to 1, with pairwise precision close to 1 indicating high accuracy and pairwise recall close to 1 indicating high completeness. Perplexity scores range from 1 to positive infinity, and lower scores indicate the data are more predictable, given a particular set of categories.

References

Chemla, E., Mintz, T., Bernal, S., & Christophe, A. 2009. Categorizing Words Using "Frequent Frames": What Cross-Linguistic Analyses Reveal About Distributional Acquisition Strategies. *Developmental Science*, 2(3), 396-406.

Frank, S. Goldwater, S. Keller,F. 2009. Evaluating models of syntactic category acquisition without using a gold standard, In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Goodman, J. T. (2001). A bit of progress in language modeling. Computer Speech & Language, 15(4), 403-434.

Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition, 106(3)*, 1126-1177.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

Mintz, T. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91-117.

Mintz, T. 2006. Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs*, 31-63.

Neidle, C. & Vogler, C. 2012. "A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface," *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012,* Istanbul, Turkey. http://www.bu.edu/asllrp/

Phillips, L. & Pearl, L. 2015. Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. *Workshop on Cognitive Modeling and Computational Linguistics 2015*, NAACL.

Sandler, W. & Lillo-Martin, D. 2006. *Sign Language and Linguistics Universals*. Cambridge: Cambridge University Press.

Wang, H., Höhle, B., Ketrez, N. F., Küntay, A. C., & Mintz, T. H. 2011. Cross-linguistic Distributional Analyses with Frequent Frames: The Cases of German and Turkish. In N. Danis, K. Mesh, & H. Sung (Eds.), *Proceedings of the 35th annual Boston University Conference on Language Development* (pp. 628-640). Somerville, MA: Cascadilla Press.

Weisleder, A. & Waxman, S. (2010). What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English, *Journal of Child Language,* 37, 1089-1108.