Outputs as inputs: Sequential Models of the Products of Infant 'Statistical Learning' of Language

Angelica Buerkin-Pontrelli[1] (abuerk@sas.upenn.edu), **Joseph Coffey**[2] (jrcoffey@g.harvard.edu), **Daniel Swingley**[1] (swingley@psych.upenn.edu)

[1]Department of Psychology, University of Pennsylvania, 425 S. University Ave., Philadelphia, PA 1910, USA
[2]Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, MA 02138, USA

Abstract

To explore whether current notions of statistically-based language learning could successfully scale to infants' linguistic experiences "in the wild", we implemented a statistical-clustering word-segmentation model (Saffran et al., 1997) and sent its outputs to an implementation of a "frame" based form class tagger (Mintz, 2003) and, separately, to a simple word-order heuristic parser (Gervain et al., 2008). We tested this pipeline model on various input types, ranging from quite idealized (orthographic words) to more naturalistic resyllabified corpora. We ask how these modeled capacities work together when they receive the noisy outputs of upstream word finding processes as input, which more closely resembles the scenario infants face in language acquisition.

Keywords: language acquisition; distributional analysis; word segmentation; word class acquisition, word order acquisition

Introduction

In about one year, infants progress from knowing very little about their native language to having learned about its sounds, its phonotactic properties, dozens of its words, and even certain elements of syntax (for example that nouns tend to follow determiners and verbs tend to follow pronouns; e.g., Shi & Melançon, 2010, Cauvet et al., 2014). Computational models of these achievements mostly examine each linguistic level in isolation. The researcher conceptualizes the problem, presupposes the infant's access to the linguistic or perceptual elements that constitute the problem, and (usually) attempts to find relatively simple learning skills that infants might use to solve the problem. Such skills might be demonstrated using experimental tests of artificial-language learning, and then offered as an important part of how infants acquire their native language.

This is an appropriate research strategy, but it is not without its limits. One major limit is that although some acquisitions might seem logically prior to others (phones before words, words before form-class categories) infants might learn parts of every linguistic level of description in parallel (e.g., Frost & Monaghan, 2016). Another limit, which we address here, is that even under the assumption of sequential learning at each linguistic level, it is certainly too optimistic to suppose that the inputs to each process are perfect, rather than being the noisy products of the prior process.

Some researchers have proposed sequential models to connect word segmentation abilities to other downstream capacities such as word form learning and lexical categorization (e.g., Phillips & Pearl, 2015; Christiansen et al., 2009). Agreeing with the spirit of these proposals, we also depart from the one-capacity-at-a-time strategy by investigating three parts of the infant's learning problem together: word segmentation, form-class categorization, and word-order learning. To explore whether current notions of statistically-based language learning could successfully scale to infants' linguistic experiences "in the wild", we test our model on three different input types, ranging from entirely idealized (orthographic words) to relatively naturalistic (though phonologically idealized) resyllabified corpora.

Our strategy is not to engineer the most successful possible model of the infant language learner. Rather, we attempt to create a simple and ecologically valid characterization of an infant using only abilities which experimental studies have identified as apparently available to infants and potentially useful for language learning. In essence, we created a statistical clustering implementation of Saffran et al. (1997), and sent its outputs to an implementation of a "frame" based form class tagger (Mintz, 2003) and, separately, to a simple word-order heuristic parser (Gervain et al., 2008).

The question we pursue is how these modeled capacities could work together when they receive the noisy outputs of upstream word finding processes as input, which may more closely resemble the scenario infants face "in the wild".

Methods

We syllabified a dictionary-based phonological version of the Brent & Siskind corpus (English IDS, 14 mothers' transcribed speech; 2001) using two strategies. The first, *within-word* strategy, left monosyllables unmodified, but split polysyllabic words into syllables according to the maximum onset principle

(McCarthy & Prince, 1994, Prince & Smolensky, 1993). The second, *across-word* strategy assigned consonants to syllables (including across word boundaries) according to a set of probabilities biased toward maximal onset with some attractive influence of stress and sensitivity to sonority ordering (Swingley, 2005, Appendix B). The syllabified corpora were then passed through a statistical clustering algorithm roughly similar to that of Swingley (2005), in which adjacent units with high mutual information (MI) and frequency were iteratively bound together into new units. Under all parametrizations this yielded outputs of English words, part-words (under-combinations), and non-words (over-combinations).

The two resulting output corpora (one from each syllabifier) and a third corpus of the orthographic words served as inputs for two separate downstream models. The first model, a form-class categorizer, which tracks non-adjacent patterns while grouping the word forms that intervene, has been proposed as a potential source of information for early syntactic categorization (e.g., Mintz, 2003, Chemla et al., 2009., Weisleder & Waxman, 2010, Mintz et al., 2011, Moran et al., 2016, but see also Stumper, 2011). The second, word-order finding model, capitalizes on the relative order of frequent and infrequent elements at utterance boundaries as a potential cue to more general word-order principles of the language (Gervain et al., 2008).

By passing two separate syllabified versions of the output corpus to the two downstream models, we were able to compare both models' results across inputs which ranged from entirely idealized to relatively naturalistic. In the following sections, we present three illustrative input model parameterizations varying in the word-finding clusterer's mutual information and frequency thresholds for assuming wordhood (95th, 85th, and 75th percentiles) for each input corpus (orthographic, within-word syllabification, and across-word syllabification).

Results

Word Finding Success

Before passing the outputs of the word finder to the downstream models, we briefly examine its ability to identify words in the input corpora. **Table 1** shows the word finding model's accuracy and recall for both versions of the corpus (*within-word, across word*) at three frequency & MI thresholds (75th, 85th, and 95th).

We considered the total number of correctly identified words (*hits*) as well as the total number of incorrectly identified words (*false alarms*). Accuracy was calculated as hits divided by (hits + false alarms), while recall represented the number of hits divided by all the words in the orthographic corpus.

Despite having sampled from relatively stringent frequency and MI criteria, the word finder's accuracy and recall scores varied substantially (**Table 1**). As the percentile-based thresholds for the word finder were increased, accuracy improved while recall decreased, presenting a more precise but sparse picture of word-finding.

Corpus	*Frequency & MI %ile*	*Accuracy*	*Recall*	*Hits*	*False Alarms*
Within Word	75th	0.35	0.20	1260	2306
Within Word	85th	0.57	0.13	812	612
Within Word	95th	0.72	0.03	196	76
Across Word	75th	0.23	0.21	1301	4359
Across Word	85th	0.33	0.13	820	1643
Across Word	95th	0.54	0.04	280	242

Table 1: Accuracy and recall for the six input model parameterizations presented throughout this paper.

Word Class Identification

Following Mintz (2003) we identified the 50 most frequent frames in our corpora, considering as a frame any sequence of ordered units, or word candidates, which could be actual, part, or non words, with exactly one intervening unit. For instance in the sentence "Put the bottle on the table" all of the following sequences are frames: [put_bottle], [the_on], [bottle_the], [on_table]. In the resyllabified corpora, frame elements could also include non and part words, so for instance [put_bot] or [le_the] from the example above.

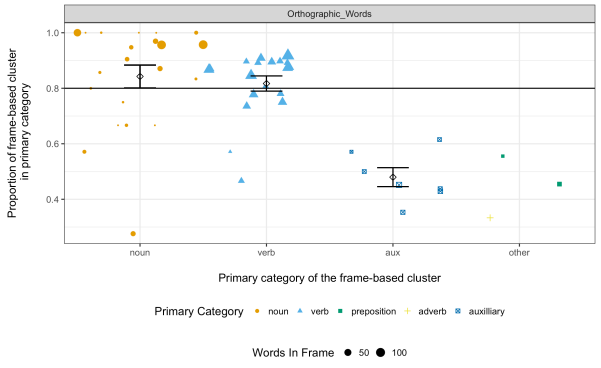

Figure 1a: Accuracies of the 50 most frequent frames in the Orthographic corpus. Point size represents number of words per frame.

Here we explore the word-class identification performance of the frequent frames. We present results for the orthographic corpus as a gold standard (**Fig 1a**). "Primary category" in the plots is our label of the

predominant word class in each plotted frame; not all analyses found frames of the same categories. Our accuracy measures in this section represent the proportion of words in each frame that shared that particular frame's primary category.

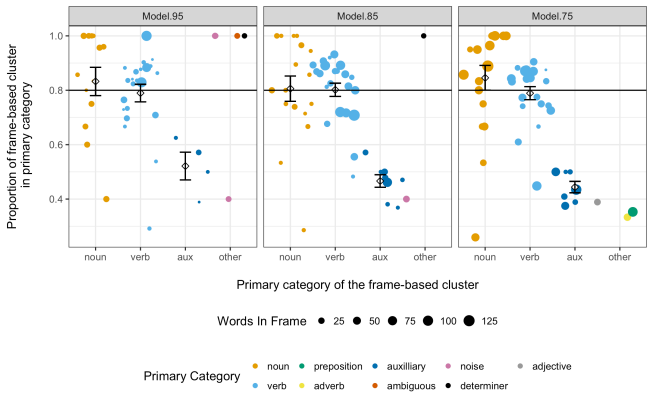

Figure 1b: Accuracies of the 50 most frequent frames in the within-word corpus. The horizontal line indicates 80% accuracy. Point size represents number of words per frame.

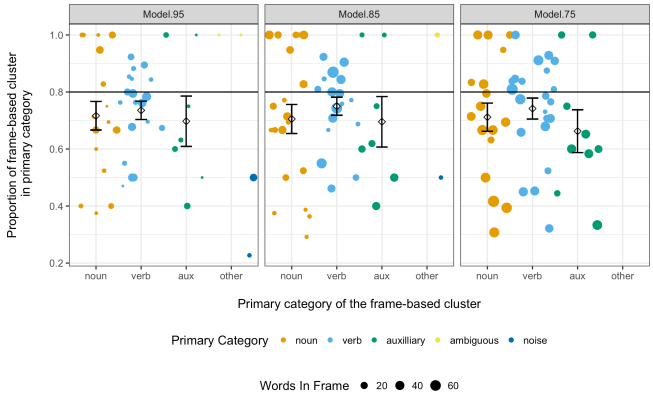

Figure 1c: Accuracies of the 50 most Frequent Frames in the Across Word Corpus. The horizontal line indicates 80% accuracy. Point size represents number of words per frame.

Within-Word Corpus We found that the outputs from our word-finder, though noisier than orthographic word input, led to decent form-class-identification performance for many nouns and verbs, but not other categories (**Fig 1b**). Nouns achieved the highest accuracy scores (Ms=84.6% to 80.6% across input models), followed by verbs (Ms= 80.2% to 78.9% across input models). Other categories performed significantly worse (i.e., frame-based clusters had highly non uniform contents) across input models.

Word-class identification declined when part-words and non-words were included (as impurities) in the accuracy computations shown in **Figure 1b** for nouns (Ms=74.4% to 66.9% across input models) and verbs (Ms= 72.4% to 61.2% across input models). All other categories, except determiners, (M = 1, n=2 frames), showed accuracy levels under 45%. Comparison with the orthographic standard, though, showed that nonwords were not numerous enough to corrupt the utility of frequent frames for identifying noun and verb categories.

Across-Word Corpus Word-class identification success declined moderately in the across-word corpus. Here, verbs achieved the highest accuracy scores (Ms=75.0% to 73.6% across input models), followed by nouns (Ms=71.7% to 70.5% across input models). Other categories performed significantly worse across input models. See **Figure 1c**.

In this analysis, counting the non-words in computing the accuracy of the frame-derived categories reduced performance more substantially. Verbs achieved the highest accuracy scores (Ms=61.3% to 49.2% across input models), followed by nouns (Ms=52.6% to 45.0% across input models), and other categories (Ms=47.9% to 11.1% across input models). Thus, the discovered categories were substantially impure with real words of the wrong category, and with mis-segmented nonwords.

Corpus	*Frequency & MI %ile*	*Non Words*	*Actual Words*	*% Non Words*
Within Word	75th	99	1190	0.08
Within Word	85th	280	1638	0.15
Within Word	95th	578	1610	0.26
Across Word	75th	397	940	0.30
Across Word	85th	676	1009	0.40
Across Word	95th	842	1182	0.42

Table 2: The raw counts and percentage of non words and actual words that appeared in the 50 most Frequent Frames for each version of the corpus and input model.

Categorization Performance

Next, we assessed whether the frame-based clusters served as decent foundations for syntactic category learning. It is, after all, implausible that children have as many form-class categories as frames; thus, we explored how infants could combine the fifty frame-based categories into actual syntactic categories. To this end, we performed a post-hoc hierarchical cluster analysis using the *h-clust* package in R (R Core Team, 2018).

We clustered word types according to their (binary) vector of appearances in each of the frequent frames. This process created a dendogram which represented all the possible groupings of categories of the words (ranging from one to the total number of words in the frame-based clusters). Each dendrogram was cut to

yield five groups. Non- and part-words were excluded from the analyses below.

This clustering process is not meant as a model of the infant learner. But, if our analysis were able to group, say, most of the noun word types across frames into one coherent cluster, this sort of result would suggest that the information that frequent frames provide could, in principle, pave the way for syntactic category learning.

Again, we present categorization performance for the orthographic corpus as a gold standard (**Fig 2a**) for comparison, and show the best performing model for the within-word corpus (**Fig 2b**), and the best model for the across-word corpus input (**Fig 2c**).

Figure 2a: Clusters formed in orthographic Corpus

Figure 2b: Clusters formed in within-word Corpus

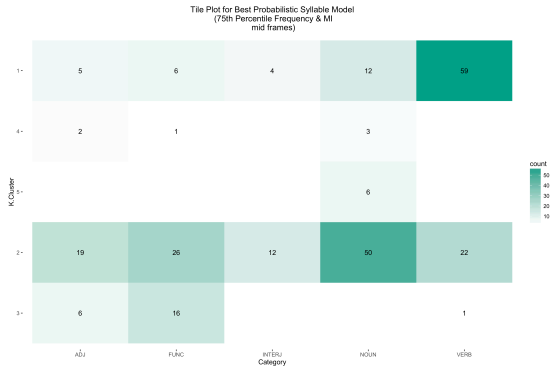

Figure 2c: Clusters formed in across-word corpus

Within-Word Corpus The input model which achieved the highest categorization success required MI scores in the 95th percentile or higher for word segmentation (**Fig 2b**). Here, 52/74 verb types identified by the frequent frame analysis clustered together (i.e., appeared in a similar set of frames). On the other hand, only 25/50 noun types appeared in the same cluster, while the other 25 noun types were scattered across three separate clusters. Categories other than {noun, verb} did not form distinct clusters.

Across-Word Corpus For the across-word corpus, the most successful model assumed 7^{5th} percentile threshold or higher for word segmentation. In this analysis, both verbs nouns and verbs tended to reliably form clusters. Overall, 59/82 verb types and 50/71 noun types appeared in single clusters respectively. Again, other categories failed to form distinguishable clusters.

Word Order Identification

Next, we calculated how often frequent and infrequent elements appeared at the utterance boundaries of each of our three input corpora. Following Gervain et al. (2008), an element in the corpus was considered frequent (FW) if its relative frequency of occurrence exceeded one of two predetermined thresholds: .01 or .0025. All other elements were considered infrequent (IW). Using these criteria for classifying frequent and infrequent elements, we identified all two-"word" sequences at the beginning and end of the utterances in the three corpora and tagged them according to four possible word orders: FW-IW, or *frequent-first*, IW-FW, or *frequent-last*, and IW-IW or FW-FW (*equal frequency*). We considered the two relative frequency thresholds (.01, .0025) separately in our analyses. As a reference, results for the orthographic words corpus are shown in **Figure 3a**.

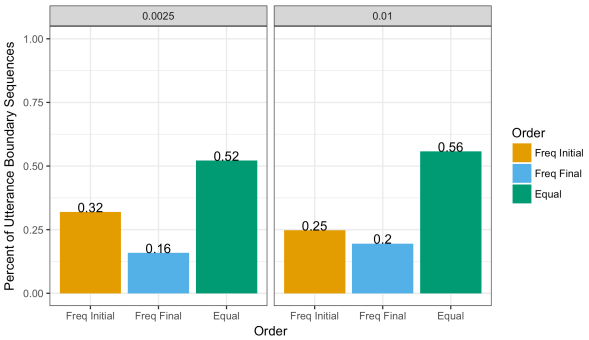

Figure 3a: Relative frequencies of orthographic words by frequency threshold

Within-Word Corpus Sequences following a frequent-first order outnumbered frequent-last sequences in all three input models. When the frequency threshold was set at .01 for the word-order-finder frequent-first orders

were only 1.48 to 1.28 times more frequent than frequent-last orders across input models. When the relative frequency threshold was relaxed to .0025, frequent-first orders appeared 2.0 to 1.4 times as often as frequent-last orders across input models (**Fig 3b**).

Across Word Corpus Running the word-order finding model on the probabilistically syllabified input substantially decreased the dominance of frequent-first sequences at utterance boundaries. For the .01 frequency threshold, frequent-first to frequent-last ratios spanned from 1.16 to 1.06 across input models. Relaxing the frequency thresholds slightly improved performance as the frequent-first to frequent-last ratios increased to 1.43 to 1.25 across input models. See **Figure 3c**.

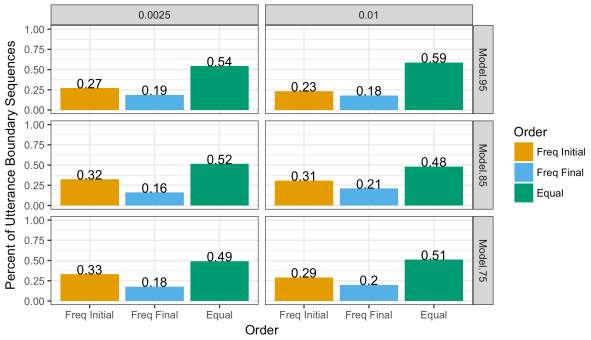

Figure 3b: Relative frequencies of sequences at utterance boundaries in the within-word corpus presented by frequency threshold and input model.

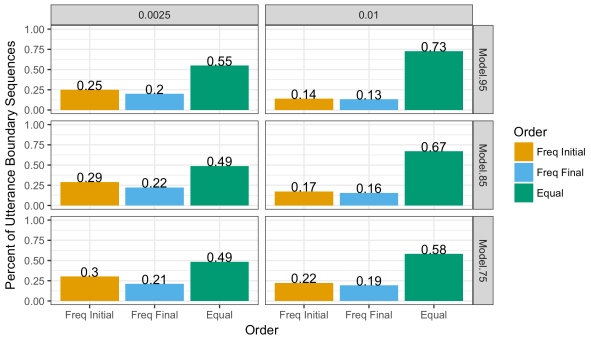

Figure 3c: Relative frequencies of sequences at utterance boundaries in the across-word corpus presented by frequency threshold and input model.

Thus, using the within-word corpus input we roughly replicated Gervain et al.'s (2008) findings in another VO language, at least in the overall ratios of frequent-initial to frequent-final sequences. However, while Gervain and colleagues found that equal frequency sequences (IW-IW, FW-FW) composed only 21% (.01 frequency threshold) and 8% (.0025 frequency threshold) of all utterance boundary sequences, such sequences were much more frequent in our analyses. In the within-word corpus, these equal frequency sequences composed 59% to 49% of all boundary sequences across input models. In the across-word input corpus, the number of equal frequency sequences increased, ranging from 73% to 49% across input models.

In our analyses we found stronger cues regarding the relative frequencies of frequent-initial to frequent-final sequences in the within-word corpus than in the across-word corpus. Such cues could, in theory, help infants derive word-order properties of their language.

Discussion

We found that noise introduced from imperfect word-finding did not prevent the frequent-frames analysis from achieving reasonably high accuracy in grouping some nouns and verbs, as long as the word-finder worked over the within-word input corpus. In this case, word classification remained somewhat robust for these form classes, even when non-words and part-words entered the analysis. When the word finder worked over a corpus of words syllabified across word boundaries, results of our downstream frequent-frames analysis declined. Furthermore, adding non-words and part-words to this noisier analysis greatly reduced the apparent success of frequent frames in form-class identification.

In our word-order identification model we still found distinct relative frequency patterns at utterance boundaries in the within-word syllabified corpus. But this favorable pattern almost vanished once we removed the assumption that infants know word boundaries (across-word corpus). Though we cannot specify the learning consequences of the input having only a razor-thin advantage for the language-typical frequency pattern, it seems unlikely that infants would draw strong conclusions about their language's word order from such small margins.

Conclusion

Years of experiments in infant research have demonstrated a number of cognitive capacities that could, in principle, help perform the task of language learning. However, it is impossible to know how far such skills can take infants without modeling over corpora. Models often force us to make more realistic assumptions about the coverage that specific infant capacities offer. Here, we used a very simple implementation of infant capacities suggested by experiments that clearly demonstrate these capacities but that unquestionably underdetermine the quantitative characteristics of these abilities. Over a range of parameter values, though, we have shown that: (a) "frequent frames" made up of, and enclosing, units found by our word-finding algorithm tended to

correctly group together nouns and verbs, but not other categories; (b) units found by the algorithm tended to reveal the frequent-first, infrequent-second trend identified by Gervain et al. (2008) as a potential cue to word order; and (c) many of these informational gains were muddied considerably by making probably more realistic assumptions about the ambiguity of syllable boundaries.

Whether these results support a more or less optimistic stance concerning the potential for statistical learning cannot be stated conclusively, and it is to be expected that opinions will differ on this issue. Our assessment is that the present analyses, on balance, probably *underestimate* the difficulties infants face, because true phonetic variability is more severe than acknowledged by our input corpora. Even so, the results suggest that while some word-finding is feasible statistically, successful form-class identification cannot be accomplished in this pipeline using frequent frames, excepting a gross and still error-ridden division into unlabeled clusters of nouns, verbs, and miscellany. Frequency imbalances and their ordering might suggest aspects of word order to infants, but the imbalances are slight in our resyllabified data, so that one might question whether they are salient enough to drive infant intuitions. In our view, it is likely that the statistical outputs we modeled are somewhat informative, but clearly insufficient. The solution, we suggest, is probably not to reduce estimates of what infants know, but rather to find ways to incorporate knowledge of word meaning into the relevant computations.

Acknowledgements

This research was supported by the National Institutes of Health (NIH) Grant R01-HD049681 to Daniel Swingley.

References

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-B44.

Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development*, 10(1), 1-18.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'frequent frames': what cross- linguistic analyses reveal about distributional acquisition strategies. Developmental science, 12(3), 396-406.

Christensen, R. H. B., & Christensen, M. R. H. B. (2015). Package 'ordinal'. *Stand*, *19*, 2016.

Christiansen, M. H., Onnis, L., & Hockema, S. A. (2009). The secret is in the sound: From unsegmented speech to lexical categories. *Developmental science*, *12*(3), 388-395.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... & Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, i-185.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, *105*(37), 14222-14227.

MacWhinney, B. (2000). *The CHILDES project: The database*(Vol. 2). Psychology Press.

McCarthy, J. J., & Prince, A. S. (1994). The emergence of the unmarked: Optimality in prosodic morphology.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91-117.

Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. Cognitive psychology, 75, 1-27.

Moran, S., Blasi, D. E., Schikowski, R., Küntay, A. C., Pfeiler, B., Allen, S., & Stoll, S. (2018). A universal cue for grammatical categories in the input to children: Frequent frames. Cognition, 175, 131-140.

Phillips, L., & Pearl, L. (2015). Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. In *Proceedings of the 6th workshop on cognitive modeling and computational linguistics* (pp. 68-78).

Prince, A., & Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar. *Optimality Theory in phonology: A reader*, 3-71.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Shi, R., & Melançon, A. (2010). Syntactic Categorization in French- Learning Infants. *Infancy*, *15*(5), 517-533.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). "Frequent Frames" in German Child-Directed Speech: A Limited Cue to Grammatical Categories. Cognitive science, 35(6), 1190-1205.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology*, *50*(1), 86-132.

Weisleder, A., & Waxman, S. R. (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*(5), 1089-1108.