Characterizing Motherese: On the Computational Structure of Child-Directed Language

Peter Brodsky (pb86@cornell.edu) **Heidi Waterfall (he32@cornell.edu)**[1]

Shimon Edelman (se37@cornell.edu)

Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

We report a quantitative analysis of the cross-utterance coordination observed in child-directed language, where successive utterances often overlap in a manner that makes their constituent structure more prominent, and describe the application of a recently published unsupervised algorithm for grammar induction to the largest available corpus of such language, producing a grammar capable of accepting and generating novel well-formed sentences. We also introduce a new corpus-based method for assessing the precision and recall of an automatically acquired generative grammar without recourse to human judgment. The present work sets the stage for the eventual development of more powerful unsupervised algorithms for language acquisition, which would make use of the coordination structures present in natural child-directed speech.

Keywords: Language acquisition; grammar inference; computational linguistics.

Introduction

Does child-directed speech — what Newport, Gleitman, and Gleitman (1977) called "Motherese" — possess special characteristics that make it easier to learn from? In this paper, we present two kinds of corpus-based evidence that should be useful in addressing this question. First, we report a quantitative analysis of the cross-utterance coordination observed in child-directed language, where successive utterances often overlap in a manner that makes their constituent structure more prominent. Second, we describe the application of a recently published unsupervised algorithm for grammar induction to the largest available corpus of child-directed language, and the performance of the resulting grammar in accepting and generating novel well-formed sentences. This work sets the stage for the development of more powerful unsupervised algorithms for language acquisition, which would make use of the coordinated structures present in natural child-directed speech.

Cross-utterance coordination in Motherese

There is a great deal of evidence suggesting that parents produce structured dialogues when talking with very young children. Parents' speech to young children is highly repetitive and often includes clusters of partial self-repetitions — *variation sets* — when speaking to young children acquiring language (Furrow, Nelson, & Benedict, 1979; Kavanaugh & Jirovsky, 1982; Kaye, 1980; Snow, 1972; Hoff-Ginsberg, 1985, 1986, 1990; Küntay & Slobin, 1996; Waterfall, 2006).

Variation sets

Hoff-Ginsberg (1985) conducted one of the initial examinations of the effect of maternal self-repetitions on children's progress in language acquisition. She showed that alternations in maternal self-repetitions that conformed to major constituent boundaries were related to growth in children's verb use, while those repetitions that altered material within a phrasal constituent aided in noun-phrase growth. In a subsequent study, Hoff-Ginsberg (1986) found that the frequency of self-repetitions and expansions was positively correlated with child verb phrase development. Similarly, Hoff-Ginsberg (1990) confirmed that maternal self-repetitions and expansions were positively correlated with the average number of verbs per utterance in child speech.

Hoff-Ginsberg's analyses, however, concentrated on the corpus as a whole and did not examine the contingent nature of clusters of such repetitions. Küntay and Slobin (1996) pioneered the research into variation sets, conducting the first longitudinal study specifically analyzing the effect of local clusters of partial repetitions in child-directed speech on language development. Focusing on the acquisition of Turkish, they found that variation sets made up approximately 20% of child-directed speech. The use of variation sets was positively associated with children's acquisition of specific verbs.

In sum, variation sets seem to be ideal environments for learning lexical items and constituent structures. By holding most of the utterance constant, while altering it slightly (see Table 1 for an example), parents may allow children to *discover* lexical items, syntactic constituents, and their place in the syntax, vis-à-vis comparison and contrast, as envisaged (in the context of the discovery of grammar by linguists) by Zellig Harris (1946).

Waterfall (2006) conducted the first longitudinal study of variation sets in English. We briefly mention here some of her findings (Waterfall, 2007). The participants were twelve parent-child dyads (ages 14-30 months). The subjects were balanced for child gender, child birth or-

[1]Also with the Department of Psychology, University of Chicago, Chicago, IL 60637 USA.

Table 1: A variation set addressed to a 14 month old.

```
You got to push them to school.
Push them.
Push them to school.
Take them to school.
You got to take them to school.
```

der, and for maternal educational level as a measure of socio-economic status.[2] The subjects were videotaped in their homes for 90 minutes every four months starting when the child was approximately 14 months old and continuing to 30 months. There were five observations in total. The data for (Waterfall, 2007) come from transcripts made from those videotapes.

To determine whether or not variation sets foster the acquisition of syntactic constituents, child-directed speech was analyzed for the manipulation of multi-word constituents in variation sets (e.g., `You can sit [on this]. You can sit [up here].`). Children's speech was then examined for the use of those structures. Lastly, children's production of a constituent (e.g., direct objects) was correlated with the manipulation of that constituent in variation sets.

Table 2: Correlation results for variation set use and children's constituent structure production (reproduced from Waterfall, 2007). Significance levels: * – $p < .05$; ** – $p < .01$; *** – $p < .0001$.

Prepositional Phrase Adjuncts	.58*
Entire clause	.67**
Subjects	.68**
Direct objects	.91***

When examining multi-word constituents,[3] Waterfall (2006) found that children's production of a structure is highly correlated with parents' manipulation of that structure in variation sets. The results form a fairly natural scale. Variation sets are most beneficial for constituent structures that are typically considered obligatory (e.g., direct objects) and are less so for those items that are typically considered optional (e.g., prepositional phrase adjuncts). Constituents that are obligatory in some cases but that can also be omitted or not used in others fall somewhere in the middle (e.g., subjects can be omitted in commands or conjoined clauses; an entire subordinate clause can be an adjunct or a complement of the verb). We note that it may be possible to rephrase "obligatory" in linguistic terminology as "there is a very high probability that two elements will co-occur in the data." Thus, it may be the case that children use statistical information when acquiring constituents — a notion that is compatible with the computational approach of Solan, Horn, Ruppin, and Edelman (2005).

Finding variation sets

The search for variation sets in a corpus can be easily automated. The program we wrote for that purpose scans the corpus, opening a new variation set record when a non-stoplisted[4] word appears for a second time on the working set queue. The first sentence of the variation set is the one in which the repeated word first occurs. The variation set is closed with the last sentence that contained a word also contained by any of the other sentences in the candidate set, so long as the first penultimate occurrence of the repeating word is also on still on the queue.[5] The simplest case results in a variation set consisting of two sentences, both of which share at least one word in common. Interleaved variation sets are also possible — sentences sharing words with other, non-adjacent sentences.[6] The variation sets detected in the corpus are then displayed to the user, along with their main computational characteristics, computed according to methods explained below.

Diameter of variation sets

One key computational characteristic of a variation set is its *diameter*, or equivalently the maximal dissimilarity between utterances that comprise it. We define dissimilarity between two strings of words in terms of the Levenshtein (edit) distance: the smallest number of individually weighted elementary edit operations (insertions, deletions, and substitutions on a per word basis) that transform one string into another. The mean values of variation set diameter for four corpora are shown in Figure 1.

Prevalence of variation sets

How surprised should one be to find a variation set in a corpus? To estimate the significance of a series of utterances forming a variation set, we would need to know the probability distribution over utterances, which is, of course, unavailable. We can, however, try to approximate that distribution using a statistical *language model* derived from the available corpus. Given a sequence

[2]The data mentioned here are a subset of a larger, unrelated, longitudinal study conducted by Goldin-Meadow, Hutenlocher, & Levine, under NIH Grant # PO1 HD40605, 2002-2007.

[3]For some single word constituents (e.g. wh-items) and for manipulations that occur within a constituent (e.g. manipulation of a definite article within a noun phrase), variation set use is not significantly correlated with production. For a detailed explanation of why this might be the case, see (Waterfall, 2006).

[4]Words that are to be excluded from consideration, such as `the` or `and`, are put on a stop list.

[5]The working set queue is bounded by two limits: the number of words and the number of lines. When either is reached, words are taken off the head of the queue until neither limit is exceeded.

[6]An example of an interleaved variation set addressed to a 14-month old: `Piggies / You want to read that? / Oh that is piggies. / You want to read this one?`

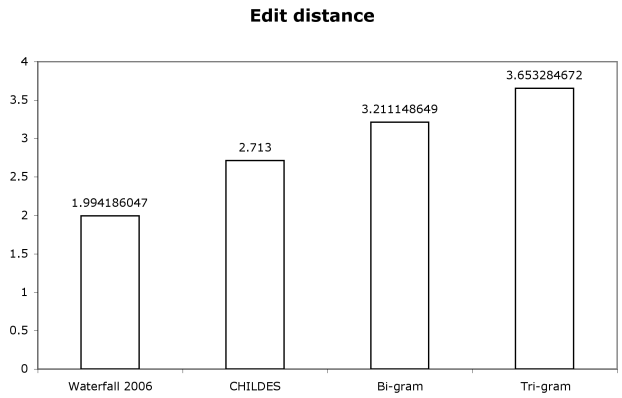

Figure 1: Mean edit-distance diameters of variation sets from four sources: a 42,530-utterance corpus of children-directed language (mean words per utterance WPU=4.32) from Waterfall (2006), the 300,000-utterance English CHILDES collection (MacWhinney, 2000), and bi- and tri-gram level statistical models of the CHILDES corpus, generated by an algorithm described below.

of words that form a partial utterance $w_1, w_2, \ldots, w_k$, a language model (Goodman, 2001) assigns to each of its n possible continuations $w_{k+1}^{(n)}$ — that is, to each of the words that may appear in the next place in a well-formed utterance in the given language — a probability $P(w_{k+1}|w_1, w_2, \ldots, w_k)$. Once "trained" on a corpus, a language model can be used to estimate the probability of given utterances, or to generate new ones according to the probability distribution it embodies, the latter use being related to the bootstrap methods in mathematical statistics (Efron & Tibshirani, 1993).

We estimated the significance of variation sets in the Waterfall (2006) and CHILDES data, by (1) generating artificial corpora with simple bi- and trigram language models,[7] and (2) comparing the prevalence of variation sets in those corpora and in real data. Our stochastic algorithm for reproducing n-gram distributions from the training corpus while generating novel utterances uses the ADIOS graph data structure (Solan et al., 2005), annotated with the probabilities of the various arcs (as estimated from a training corpus). Given a new sentence, the algorithm instantiates each word as a node in a double-ended queue (deque), one per sentence. Each deque is assigned a unique ID, which is added to its node's ID set. Because each distinct word has one and only one node associated with it in the graph, the number of sentence IDs in the nodes' ID sets increases during training. The links between nodes are stored in a hash table where the key is the sentence ID and the value is the node. When generating sentences, an n-gram of arbitrary length is used to produce an intersection of the sentence ID sets from every element in the n-gram. Each sentence ID is then used to access the right-linked list hash table. If a node is recovered, it is added to a counted collection, from which a node is drawn at random to produce the next element of the generated utterance.[8]

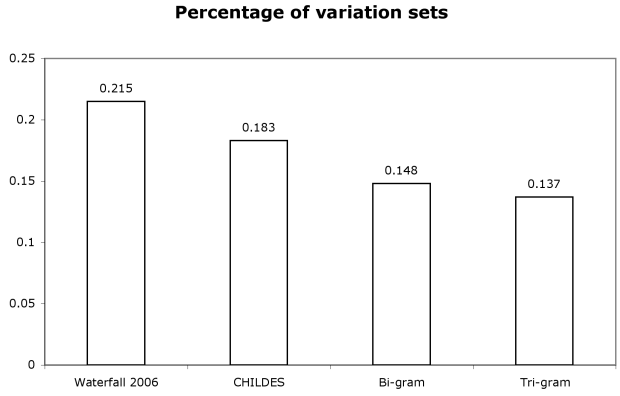

Figure 2: Percentage of words in variation sets in the (Waterfall, 2006) and CHILDES data (left two bars) and in artificial corpora generated by language models matching the bi- and trigram statistics of the Waterfall corpus (right two bars).

Figure 2 shows a comparison of the prevalence of variation sets in two natural corpora, and in two artificially generated ones that match the bi- and trigram statistics of (Waterfall, 2006). Variation sets are seen to occur more often in natural corpora (where they also have a lower raw Levenshtein-distance diameter and a higher informational value; not plotted), indicating that this hallmark of Motherese cannot be due simply to its bi- or trigram statistics.

Informativeness of variation sets

How useful is a variation set for the learner? A pair of utterances that have nothing in common (a fact represented by some arbitrary maximal value of edit distance) is not informative, and neither is a pair of identical utterances. An optimally informative pair would therefore balance overlap and change; we denote the normalized distance between members of such a pair by $0 < \beta < 1$. To compute the average informational value of an entire variation set, every utterance in the variation set is compared to every other utterance (pairs that are identical or share no non-stoplisted words are not compared; this prevents interleaved variation sets from artificially depressing the average informational content of the variation set sample; is also prevents highly repetitive variation sets from generating outliers).

Following this line of reasoning, we define the information-theoretic similarity between two utterances

[7] An n-gram language model conditions the probability of a word on n preceding words in the utterance.

[8] This approach does not handle loops – recurring nodes in the same path. This option will be added in the future.

within a variation set to be: $1 - \frac{|L(\vec{u}_1,\vec{u}_2)-\beta|}{\beta}$, where L takes two utterances (with stoplisted elements, such as closed-class words, removed) and returns their Levenshtein distance normalized to lie between 0 and 1, and β is a baseline reference value between 0 and 1 which turns the distance into an information-theoretic measure.

In a preliminary study, we found that the mean informational value of the maternal speech corpus (Waterfall, 2006) on a variation set by variation set basis correlates with the child's vocabulary size (Figure 3). A value of β ("bias" or baseline, explained above) of 0.487 yields the strongest correlation ($R^2 = 0.21$). These data suggest that a fine balance between change and overlap between sentences in variation sets (about 50% overlap and 50% new material, with a slight preference towards the latter) may be the most conducive to vocabulary growth.

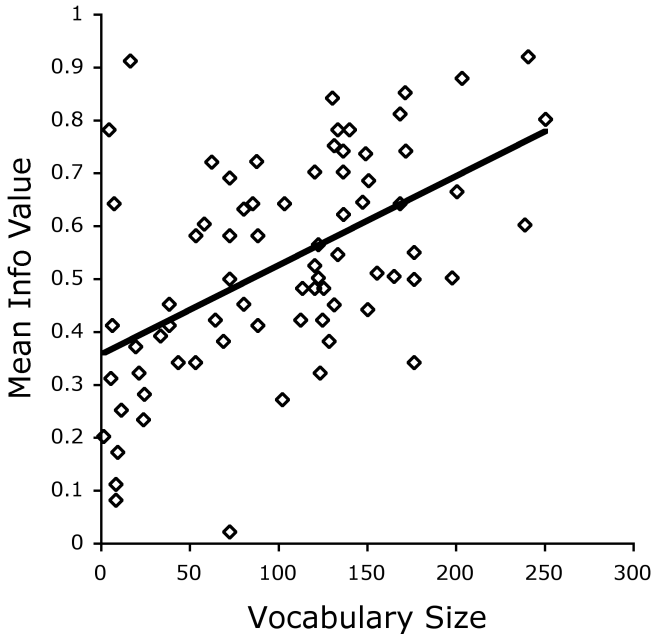

Figure 3: Information value of variation sets in the corpus of Waterfall (2006). The mean information value of the maternal variation sets is most correlated with the size of child's productive vocabulary ($R^2 = 0.21$) for $\beta = 0.487$. For points with vocabulary size > 25, the regression is more pronounced ($R^2 = 0.40$).

Language acquisition from Motherese

Very little work has been carried out to date on the acquisition of a generative grammar from real child-directed language (Solan et al., 2005; in comparison, the CHILDES data are often used to test algorithms that address specific problems in language acquisition, such as word segmentation or auxiliary fronting in forming polar interrogatives in English). In this section, we report the first quantitative results on learning a highly productive construction-like grammar from the CHILDES corpora, using ADIOS (Automatic DIstillation Of Structure), a batch algorithm that learns phrase structure rules from raw corpus data by recursively aligning utterances while abstracting any patterns (Solan et al., 2005). As customary in computational linguistics, we describe the performance of the learned grammar in terms of *recall* (defined as the proportion of the sentences in a withheld test corpus that can be generated by the grammar), and *precision* (the proportion of the sentences generated by it that are acceptable).

Recall and precision on CHILDES

Because of the greedy nature of the ADIOS algorithm, the learned grammar depends on the order of the sentences in the corpus, which is why the results from several learners trained on permuted versions of the corpus are usually pooled. We have trained 30 learners on permutations of most of the English portion of CHILDES (approximately 300,000 sentences), reserving 500 sentences for testing recall performance; the recall level was 0.50. To test precision, we had each of 10 learners generate 100 sentences, which were then manually judged as grammatical or not; the precision level was 0.63.[9]

These levels of recall and precision are much higher than those achieved by the ADIOS algorithm on the Wall Street Journal corpus used, e.g., by Pullum and Scholz (2002) in their assessment of the "poverty of the stimulus" argument. This, of course, merely confirms that the language of CHILDES is structurally simpler than that of the WSJ. More generally, the present results exceed the performance of other unsupervised algorithms that can learn from raw text, but fall somewhat short of the parsing performance achieved by algorithms that work with hand-tagged part of speech data (Klein & Manning, 2002).

It should be noted that no other method has to our knowledge been tested extensively on CHILDES. Moreover, grammars learned by the algorithms that rely on POS tags tend to result in low precision when the POS symbols in the generated sentences are substituted with actual words. In comparison, the level of precision attained by ADIOS (0.63) can be safely taken at face value. Some examples of incorrect and correct sentences it generates appear below:

```
I doesn't notice it if that's in your eye.
Out jump the tomato.
Mumma break it.
```

[9] We also tested the algorithm on the adult speech portions of the only two Hanja/Mandarin corpora from CHILDES: Chang and Zhou (2,000 and 8,000 sentences; in each case, 500 were reserved for testing). Single-learner recall was 0.31 and 0.32, respectively (comparable to that obtained for the much larger pooled English CHILDES corpus of 300,000 sentences). Five native speakers rated a sample of Zhou sentences at 0.93 precision, compared to 0.54 for novel ADIOS-generated sentences.

```
Wanna put some on your dress?
Shall we add another one like this?
It didn't make any noise.
```

A new method for estimating precision

The development of generative language models capable of learning from large, complex, real-life corpora such as CHILDES is hampered by the difficulty of estimating the precision of the models. To calculate precision — that is, the proportion of generated sentences that are acceptable — one needs either a reliable parser for the target grammar (which does not exist for realistic natural language data) or access to human subjects who would judge the acceptability of the generated sentences (an infeasible requirement in the development of large-scale learning models, where precision needs to be assessed repeatedly for thousands of sentences in each cycle).

We describe a novel method for estimating recall and precision, which bypasses these limitations. It relies on the observation that two highly constrained models trained on disjoint corpora are very unlikely to agree coincidentally about the acceptability of a given test sentence (operationalized as its probability given the grammar). Thus, any such agreement supports the hypothesis that the score given to a sentence is indeed valid. A precision-testing scheme based on this observation (see Figure 4) requires:

1. A language model, called the *processor*, which is trained on a part of the available corpus, and for which the precision needs to be estimated;
2. An auxiliary language model, called the *generator*, which is trained on another part of the corpus, and then used to generate sentences that have a high likelihood of being ill-formed.

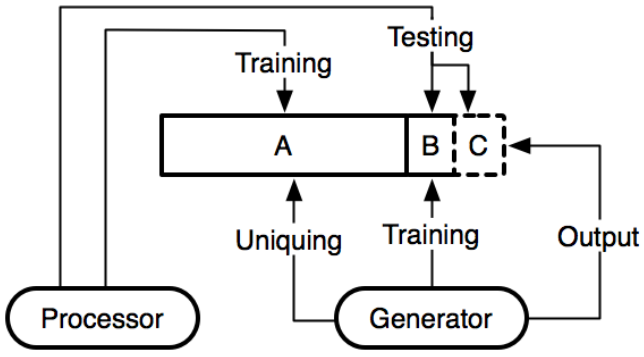

Figure 4: A scheme for testing precision without a parser or a human evaluator (see text for explanation).

The training corpus must be large enough so that less than half of it suffices to train both the processor and the generator. This corpus is split into two parts, A and B. The generator is trained on part B, and is used to produce sentences that follow the WPU distribution of that part. Every generated sentence that matches an utterance in either part A or part B is discarded; only novel (unique) sentences are retained. We note that the proportion of sentences produced by the generator that are found in part A is a pessimistic estimate of the generator's precision. For part A to be effective in "catching" most well-formed generated sentences, it is important that it be much larger than part B. When the number of unique sentences generated is equal to the number of sentences in part B (or when no new sentences can be created), the process is halted and the novel sentences, which for the above reason are more likely than the average generated sentence to be ill-formed, are placed in part C.

The processor is trained on part A; given a new sentence, it returns a score between 0 and 1 indicating its normalized probability. Computing the average score for sentences in part B would yield the processor's recall. However, by testing it on both parts B and C as described below, we can also estimate its precision. For every sentence in the union of B and C, the score is binned; a value of 0 is placed in the bin if the sentence comes from part C, and a value of 1 if it is from part B. The processor's precision is high insofar as the average score for sentences from part C (which, as explained above, are likely to be ill-formed) is low, and in particular significantly lower than that for sentences from part B. Figure 5 shows that this is indeed the case for two models trained on CHILDES: ADIOS interpolated with a bigram model (shaded bars) and a trigram model (open bars). A Fisher Exact test indicated that the difference between B and C average scores was highly significant for both models: $t = 1158.2$ ($p = 0.0000$) and $t = 136.8$ ($p = 0.0000$) respectively. A Friedman test of the difference between the distributions of scores generated by the two models showed the ADIOS+bigrams model to be significantly better: $\chi^2 = 17.8$ ($p < 0.001$).

It is important to note that the generator and the processor are not trained on the same or even overlapping parts of the corpus. Training the generator on the same segment as the processor would create doubt as to whether the outcome is characteristic of the corpus in general or is specific to the common segment. By using disjoint training sets, we ensure that the only path to agreement between the two models is via the more abstract, general characteristics of the corpus. A large-scale Monte Carlo simulation study designed to validate this approach to precision estimation is currently under way. It involves artificial corpora generated by context-free grammars (allowing us to obtain the ground truth for precision using parsers for those grammars).

Conclusions

In this short paper, we described an initial quantitative investigation of a key characteristic of child-directed

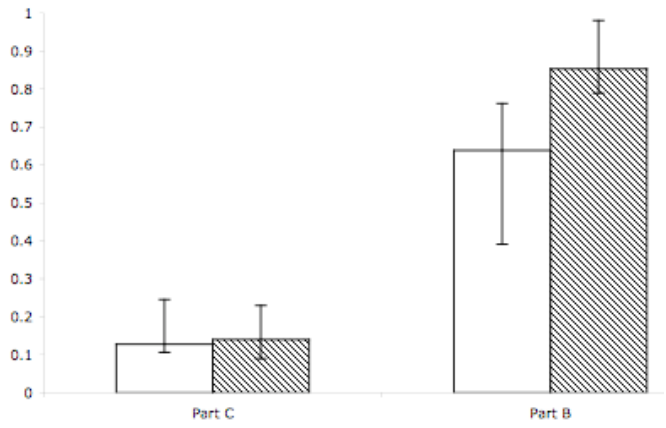

Figure 5: A large-scale estimate of the precision of two language models trained on the English CHILDES corpus. The plot shows the probabilities assigned by an interpolation of an ADIOS model with a bigram model (shaded bars) and a trigram model (open bars) to ill-formed test sentences newly generated by an auxiliary mechanism (left), and to withheld test sentences (right). As explained in the text, this result indicates that it is possible to estimate precision without recourse to a parser (which is not available for real natural language) or to human acceptability judgments.

language — its variation set structure — and demonstrated that our current algorithm for language acquisition, ADIOS, can learn a precise, relatively high-coverage generative grammar from the CHILDES corpus. We are presently working on developing a next version of the ADIOS algorithm, which will incorporate the insights from variation set studies, and thereby serve as a better model of human performance in language acquisition.

Acknowledgments

HRW's dissertation work was supported by NIH Grant # PO1 HD40605 to Susan Goldin-Meadow and Janellen Huttenlocher at the University of Chicago. The present project was supported in part by a seed grant from the Cornell Institute for the Social Sciences. The English CHILDES precision and recall data were generated by Morgan Ulinski. The Mandarin CHILDES data were generated by Shane Sniffen and Andrew Carr.

References

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. London: Chapman and Hall.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of Child Language*, *6*, 423-442.

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

Harris, Z. (1946). From morpheme to utterance. *Language*, *22*, 161-183.

Hoff-Ginsberg, E. (1985). Relations between discourse properties of mothers' speech and their children's syntactic growth. *Journal of Child Language*, *12*, 367-385.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: their relation to the child's development of syntax. *Developmental Psychology*, *22*, 155-163.

Hoff-Ginsberg, E. (1990). Maternal speech and the child's development of syntax: A further look. *Journal of Child Language*, *17*, 85-99.

Kavanaugh, R., & Jirovsky, A. (1982). Parental speech to young children: A longitudinal analysis. *Merrill-Palmer Quarterly*, *28*, 297-311.

Kaye, K. (1980). Why we don't talk "baby talk" to babies. *Journal of Child Language*, *7*, 498-507.

Klein, D., & Manning, C. D. (2002). Natural language grammar induction using a constituent-context model. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (p. 35-42). Cambridge, MA: MIT Press.

Küntay, A., & Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In D. Slobin & J. Gerhardt (Eds.), *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp* (p. 265-286). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum. (Volume 1: Transcription format and programs. Volume 2: The Database.)

Newport, E. L., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (p. 109-150). Cambridge: Cambridge University Press.

Pullum, G. K., & Scholz, B. (2002). Empirical assessment of poverty of the stimulus arguments. *The Linguistic Review*, *19*, 9-50.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, *43*, 549-565.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Science*, *102*, 11629-11634.

Waterfall, H. R. (2006). *A little change is a good thing: Feature theory, language acquisition and variation sets*. Unpublished doctoral dissertation, University of Chicago.

Waterfall, H. R. (2007). *The role of variation sets in verb learning*. (In preparation.)