Variation Sets Facilitate Artificial Language Learning

Luca Onnis (lo35@cornell.edu) **Heidi Waterfall (he32@cornell.edu)**[1]

Shimon Edelman (se37@cornell.edu)

Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

Variation set structure — partial alignment of successive utterances in child-directed speech — has been shown to correlate with progress in the acquisition of syntax by children. The present study demonstrates that arranging a certain proportion of utterances in a training corpus in variation sets facilitates word segmentation and phrase structure learning in miniature artificial languages by adults. Our findings have implications for understanding the mechanisms of L1 acquisition by children, and for the development of more efficient algorithms for automatic language acquisition, as well as better methods for L2 instruction.

Keywords: language acquisition, grammar inference, artificial language learning, implicit learning, variation sets, child-directed speech.

Variation sets in language learning

Imagine entering a room and overhearing the following round of conversation in an unfamiliar language:

(1a) `kedmalburafuloropesai`
(2a) `gianaber`
(3a) `manadukbiunel`
(4a) `kiciorudanamjeisulcaz`

Not surprisingly, you cannot even make out the individual words in any of those utterances (represented here in print by unbroken sequences of letters). Now, suppose the utterances you overheard were these:

(1b) `kedmalburafuloropesai`
(2b) `rafuloro`
(3b) `manaloropesai`
(4b) `kedmalbumanaloropesai`

Because these four utterances appear related to each other, this sample (unlike the previous one) affords a glimpse of some of the structures behind the unfamiliar language. These structures are readily revealed by the two computational operations proposed by Zellig Harris (1946) for language discovery: *alignment* of utterances and their subsequent *comparison* (when applied recursively along with statistical control over structure inference, this approach proved effective in unsupervised language acquisition from raw corpora; Solan, Horn, Ruppin, and Edelman, 2005). In the present example, simple, "local" alignment of successive utterances immediately reveals that some sequences of letters repeat, while others change. Specifically, local alignment of (1b) and (2b) suggests that (1b) is composed of at least three elements:

(1c) `kedmalbu rafuloro pesai`
(2c) `-------- rafuloro -----`

Aligning the next pair yields new elements:

(2c) `rafu loro -----`
(3c) `mana loro pesai`

Further,

(3c) `-------- mana loro pesai`
(4c) `kedmalbu mana loro pesai`

The property of sample (1b-4b) that afford this discovery, which is absent from sample (1a-4a), is that its successive utterances form partial self-repetitions, or *variation sets*.

Variation sets are a prominent feature of child-directed speech:[2] about 20% of utterances in child-directed speech appear within variation sets, whose prevalence and composition has been shown to facilitate lexical and syntactic development (Hoff-Ginsberg, 1986; Küntay and Slobin, 1996; for recent reviews and results, see Waterfall, 2007a,b). Indeed, our second example (1b-4b) comes from a snippet of a real corpus — a variation set addressed to a 14 month old child studied by Waterfall, in which we replaced the English words with nonce strings:

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

From a cognitive computational standpoint, the key characteristic of variation sets is that the structure they contain can be revealed by a *local* mechanism that aligns and compares adjacent utterances. This characteristic allows even memory-limited learners to discover structure that they would miss, if the relatable utterances were scattered over a longer exchange.

In addition to facilitating the segmentation of utterances into lexical elements, variation sets can yield higher-order structural properties of the language. For

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

[2] There are indications that variation sets are also prevalent in adult conversations (Pickering and Garrod, 2004; Szmrecsanyi, 2005).

example, the material replicated across the first two sentences, push them, is a verb phrase. Thus, alignment and comparison would break the first sentence into three constituents, corresponding respectively to the main clause (you got to), verb phrase (push them), and participial phrase (to school).

To assure safe generalization, any corpus-based inference about structure needs to pass a test of statistical significance (Edelman and Waterfall, 2007). Given a variation set, the null hypothesis is that of chance partial alignment of the utterances. The learner may test it by comparing the dissimilarity between the utterances to a baseline value — e.g., the cumulative average dissimilarity for the corpus at hand. A convenient measure of the dissimilarity between two strings of words is their Levenshtein (edit) distance, defined as the smallest number of (possibly individually weighted) elementary edit operations — insertions, deletions, and substitutions of words — that transform one string into another.

In the present project, we set out to study the effects of variation sets on language acquisition in a controlled situation involving miniature artificial languages. Small-scale artificial languages generated by simple grammars have long been used in controlled studies of language acquisition (Reber, 1967; Miller, 1968). Because the utterances generated by an artificial grammar from a nonce lexicon are novel for the learners, who are also unaware of the experimental manipulation, it is possible to gauge the learnability of various properties of the language after a relatively brief exposure to samples drawn from it. This paradigm proved useful in studying aspects of infant language learning, including segmentation (Saffran, Aslin, and Newport, 1996), and sensitivity to the ordering or words and to abstract patterns (Gómez and Gerken, 2000). Moreover, neural signatures of grammatical violations in artificial languages are similar to those evoked by structural violations in natural language (Christiansen, Conway, and Onnis, 2007).

In the two experiments reported below, subjects learned, respectively, to segment continuous utterances into word-like units, and to group these into phrasal categories. We hypothesized that learning would be significantly more efficient when some of the utterances in the learning phase are arranged in variation sets.

Experiment 1: learning word segmentation

Experiment 1 tested the subjects' ability to segment continuous speech into word-like units.

Participants and materials. The subjects, 31 Cornell students, were paid $4. In the learning phase, participants listened to a sequence of utterances ("sentences") consisting of concatenated "words." The sentences were generated by a simple rewrite rule:

S → A B C

where A = {da, kozi, spinose}, B = {pera, kadro, fama, zupa}, and C = {piu, prati, guklozi}. In this miniature language, the classes A, B, and C can be conceived of as lexical categories containing respectively three, four, and three lexical items. A sample sentence from this language, which consists of 3 × 4 × 3 = 36 unique sentences, is S → da pera guklozi.

The actual sound sequence presented to participants in the learning phase was generated as follows. For each sentence, white spaces were removed (e.g., da pera guklozi → daperaguklozi) and each letter was mapped into a single phoneme. We used the MBROLA speech synthesizer (Dutoit, 1997) to convert the resulting sequences into sound files, using a constant length of 80 *ms* for consonants and 260 *ms* for vowels. We selected the Italian diphone set of phonemes in MBROLA for two reasons. First, we intended to give participants the impression that they were engaged in a foreign language learning task. Second, the Italian diphone set appears to provide clearer and cleaner phonetic realizations of phonemes than the English one, when instantiated in flat-prosody artificial words like ours. All phonemes had an equivalent phonemic realization in English and were thus familiar enough to English speakers.

In this manner, the text-to-speech conversion procedure generated for each sentence a seamless stream of phonemes in which no acoustic property signaled the beginning and end of a word, except at the sentence boundaries, where 800 *ms* pauses were inserted. Sentences were presented to the subject via headphones. A total of 106 sentences were presented during the learning phase. Because the A and C words varied in syllable length (1-3 syllables), sentence length varied from a minimum of 4 syllables to a maximum of 10 syllables. This variability ensured that participants did not adopt segmentation strategies based on perceiving words as regular patterns of equal length.

In the test phase, we administered a forced-choice test between words and part-words, where part-words were defined as syllable sequences straddling word boundaries. A participant who succeeded to individuate the words in the learning phase should reliably prefer words over part-words. As test words we used the four bisyllabic words that appeared always in the sentence-middle position (B words: {pera, kadro, fama, zupa}). As test part-words, we chose 8 out of the 20 bisyllabic segments that straddled word boundaries. For instance, in the segment kozipera, zipe was a part-word formed by the last syllable of kozi and the first syllable of pera. Acoustically, both words and part-words were equally potentially good word candidates, because (i) during the learning phase they had been generated as a seamless chain of phonemes in each sentence, and (ii) both words and part-words were synthesized anew rather than being sliced up as fragment recombinations of syllables from the sentences.

Words and part-words had different statistical properties. Words had high word-internal transitional probabilities between syllables (e.g., $TP(\texttt{ra}|\texttt{pe}) = 1$), while part-words had low word-internal transitional probabilities (e.g., $TP(\texttt{pe}|\texttt{zi}) = 0.25$). In addition, the test words had a much higher frequency (mean frequency = 25.75, sd = 0.96) than the part-words (mean frequency = 8.5, sd = 0.51). As suggested by the finding that both adults and infants are sensitive to loci of high and low transitional probability and use them to group syllables into word-like units (Saffran et al., 1996), we anticipated that the words in this language could be discovered and preferred over part-words because both their word-internal TPs and raw frequencies were higher than those of part-words.

The four test words were presented twice in counterbalanced order, each time with a different part-word: one that contained the first syllable of the word, and one that contained its second syllable. For instance, the word `pera` was paired with both `zipe` and `ragu`. For each pair, participants had to chose which one was a word.

Procedure. Participants were told they were going to listen to a miniature language containing new words, a situation akin to a child learning its first language, or to an adult trying to figure out a foreign language. They were encouraged to listen attentively and find words in the speech. Participants were randomly and blindly assigned to one of two learning conditions, Varset and Scrambled, which consisted of exactly the same sentences that differed in the order of presentation. In the Scrambled condition, which served as the control, sentences were presented in pseudo-random order such that no two adjacent sentences shared any lexical items. This condition established a baseline for how much learning would take place in the absence of variation set structure. A sample of the sequence from the Scrambled condition appears below:

`kosifamapiu`
`spinozeperaguklozi`
`kosifamapiu`
`daperaguklozi`
`spinozefamaprati`
`daperapiu`

In contrast, sentences in the Varset condition were pseudo-randomly ordered such that 20% of adjacent sentences contained one overlapping lexical item. The remaining 80% satisfied the same criterion as in the Scrambled condition (i.e., no lexical overlap between adjacent sentences). Varset and Scrambled sentences alternated in blocks: 6 blocks of 4 Varset sentences alternated with 6 blocks of 13 or 14 Scrambled sentences. A sample of the Varset condition is given below:

`daperapiu`
`kosifamapiu` (Varset block starts here)
`kosizupaguklozi`
`kosiperapiu`
`kosizupaguklozi`
`daperaprati` (Scrambled block starts here)
`kosifamapiu`
`dazupaprati`
`spinozekadroguklozi`

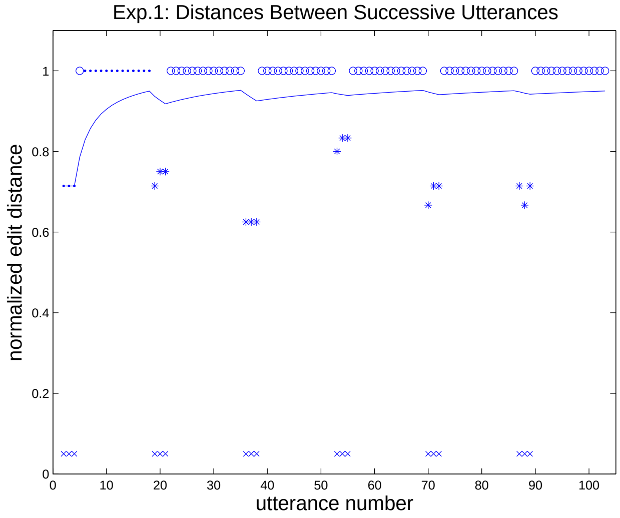

Figure 1: Edit distances d_n between successive utterances (n and $n+1$) in the Varset training data of Experiment 1, plotted against n. Solid line: cumulative average $d_{avg} = (1/n)\sum_{i=1}^{n} d_i$. $(\cdot)$: pairs for which d_n and d_{avg} do not differ significantly according to a 2-sided t-test. $(*)$: $d_n < d_{avg}$. $(\circ)$: $d_n > d_{avg}$. $(\times)$ at the bottom of the plot denotes alignable pair. Note that the cumulative statistics of the edit distance values reveal most of the alignments, where they exist, to be significant. A learner can rely on this feature of the training corpus in distinguishing between significant and spurious patterns in structure discovery.

An analysis of edit distances between successive sentences in the training data in the Varset condition (Figure 1) reveals that in almost every variation set, the edit distance between the two sentences is significantly smaller than the baseline provided by the cumulative average. We note that a learner sensitive to this statistic could use it to distinguish between significant and spurious patterns in structure discovery.

The learning phase lasted 5 minutes. Before the test, participants were told that they would have to choose between two sounds, one of which would be a word and the other a part-word.

Results. The subjects' performance in Experiment 1 is summarized in the form of a box-and-whiskers plot in Figure 2, left. Subjects in the Varset condition preferred words over part-words on the average 5.74 times out of 8, which is significantly better than chance ($t(18) = 6.34, p < .001$). In contrast, subjects in the

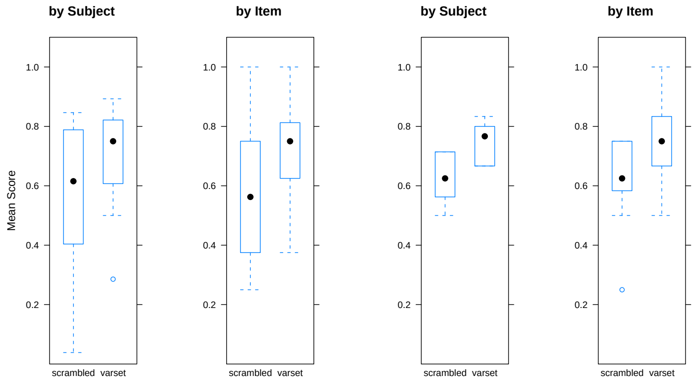

Figure 2: distribution of mean scores by subject and by item in Experiments 1 (left two plots) and 2.

Scrambled condition preferred words over part-words on the average 4.47 out of 8 ($t(16) = 1.19, p = .25$, n.s.). In addition, Varset scores were significantly better than Scrambled ($t(34) = 2.68, p = .011$). This difference was confirmed by a nonparametric Kruskal-Wallis rank sum test, which yielded $\chi^2 = 15.628, p < 0.000077$.

Because effects that turn out significant in separate by-subject and by-item analyses may still be unreliable when all the random effects are considered jointly (Baayen, 2006), we also fit a mixed linear model to the data using the lme4 package (Bates, 2005). In addition to offering a more reliable picture of the data by accommodating crossed subject and item random effects, lmer tolerates unbalanced data (as when the numbers of subjects per condition differ), and also allows one to specify a distribution other than normal. A binomial logit-link mixed linear model fit to the scores yielded a significant effect of condition, $z = 2.688, p < 0.00719$, confirming the outcome of the t-tests reported above.

Thus, subjects failed to find words in unsegmented speech in the Scrambled condition, despite transitional probabilities supporting segmentation. At the same time, in the Varset condition, in which 20% of the sentences formed variation sets in the learning phase, subjects performed significantly better, and better than chance.

Experiment 2: learning phrase structure

Identifying word boundaries in continuous speech allows the learner to acquire the lexicon. This ability, in turn, sets the stage for discovering in sentences patterns such as phrase structure. In Experiment 2, we asked whether the presence of variation sets facilitates the learning of phrase structure. In particular, we were interested in testing whether the ability to discover phrases improves when the partial lexical variation between adjacent sentences is consistent with phrasal constituency structure.

Participants and materials. The subjects, 29 Cornell students, were paid $6 for their participation. In the learning phase, subjects listened to sentences containing pseudo-words. The sequences were generated by the following phrase structure rules:

S1 (.70) → Phrase1 Phrase2 Phrase3
S2 (.25) → Phrase1 Phrase2
S3 (.05) → Phrase3 Phrase2 Phrase1
Phrase1 (.92) → A B
Phrase1 (.08) → G
Phrase2 (1.0) → C D
Phrase3 (1.0) → E F

As in the description of Experiment 1, capital letters stand for lexical categories containing one or three words (A = {a1,a2,a3}, B = {b1,b2,b3}, C = {c1,c2,c3}, D = {d1,d2,d3}, E = {e1,e2,e3}, F = {f1,f2,f3}, and G = {g1}).[3] The resulting language consisted of sentences with two or three phrases, with Phrase3 being optional at the end of the sentence or being moved in first position. Phrase1 contained either two words from categories A and B, or a substituting word g1. Sentences and phrases were generated according to the probabilities indicated in parentheses next to the each rule. Sentence length ranged from 3 to 6 words. No feature of individual words other than their distribution in the sentences signaled their class membership. The actual lexical items were the following 19 monosyllabic pseudo-words: `arv, bim, skiv, cree, dro, goz, heeb, irg, tood, kleep, kuhl, larp, mib, nerk, tiv, plam, yent, quive, roo, boont, silg, slar, smir, nork, vit, whap, plid, ziln, ziz`. Words for each participant were randomly assigned to the lexical items a1, a2, . . . , g1, and were recorded by a trained female voice.

For the learning phase, we selected 365 sentences, which were arranged differently in Scrambled and Varset conditions. As in Experiment 1, no adjacent sentences in the Scrambled condition shared any lexical item. In the Varset condition, 20% of sentences contained partially overlapping lexical items that coincided with the phrases:

G C D
A B C D
A B C D E F
E F A B C D
G A B C D
A B C D G

Between the first and second sentences in the above list, the classes 'C D' (and their elements), which belong to Phrase2, remain constant, while 'A B' replaced 'G', which are both instantiations of Phrase1. There were 10 blocks of variation sets interleaved with 10 blocks of sentences arranged in scrambled order. Each variation

[3] In our notation, numbered lowercase letters are placeholders for pseudo-words that were selected in a different order by the software running the experiment.

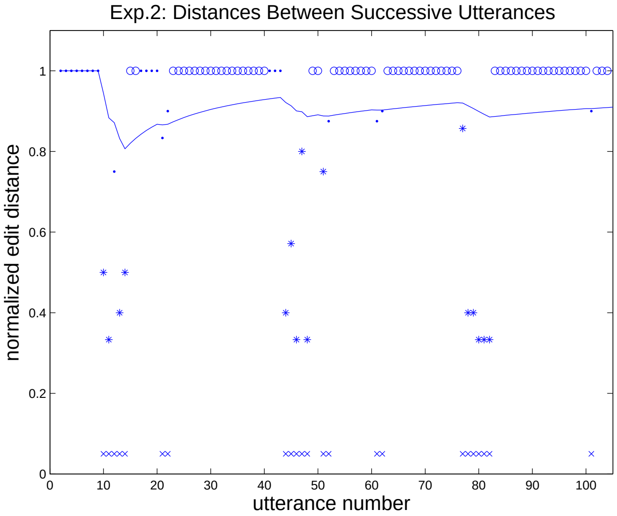

Figure 3: The first 100 edit distances between successive utterances in Experiment 2 (for the legend, see Figure 1). As in Experiment 1, the cumulative statistics of the edit distance values indicate that the alignments, where they exist, are significant, with a few exceptions.

set contained 6 sentences, with only one change between any two adjacent sentences. Unlike in Experiment 1, words were now separated by 300 *ms* pauses; sentences were separated by 750 *ms*. Because all words were separated by the same pause length and were generated by the MBROLA synthesizer without prosody or phoneme lengthening, no acoustic feature signaled the presence of phrase boundaries.

The research question was whether participants in the Varset condition would exploit variation sets to carry out a primitive alignment and grouping of words into phrasal constituents, e.g., whether they would judge a 'C D' pairing more likely than a 'D E' pairing. Consequently, the test phase was a forced-choice task consisting of 12 trials, with three trials testing each of the three phrase types ('A B', 'C D', 'E F'). A trial presented two pairs of words, one phrase pair (e.g., 'C D') and one pair that was a legal sequence in the language but straddled a phrase boundary (e.g., 'D E'). As in Experiment 1, an analysis of edit distances between successive sentences in the training data in the Varset condition (Figure 3) reveals that in most variation sets, the edit distance between the two sentences is significantly smaller than the baseline provided by the cumulative average.

Procedure. Subjects were randomly assigned to either the Varset or Scrambled condition. They were told that they were participating in an experiment about learning a new language, and that they should try to individuate the basic phrasal constituents of the sentences. As an example, the English sentence "My brother plays Nintendo at night" was described as having the following grouping: "(My brother) (plays Nintendo) (at night)." Learning lasted 18 minutes. In each trial, the subjects had to choose the stimulus that they deemed more likely to be a group or unit in the language.

Results. The subjects' performance in Experiment 2 is summarized in Figure 2. Subjects in the Varset condition preferred phrases over part-phrases with on the average 9.07 times out of 12, which is significantly better than chance ($t(14) = 6.35, p < 0.001$). Subjects in the Scrambled condition preferred phrases over part-phrases on the average 7.36 times out of 12, which is also better than chance ($t(13) = 3.085, p < .01$). In addition, learning in the Varset condition was significantly better than in the Scrambled condition ($t(27) = 2.60, p < 0.015$). This difference was confirmed by a Kruskal-Wallis test, $\chi^2 = 16.37, p < 0.00052$. Thus, while learning did occur in both conditions, it was significantly better when variation sets were present in the learning phase. A binomial logit-link mixed linear model fit to the scores yielded a significant effect of condition, $z = 2.678, p < 0.0074$, confirming this conclusion.

Discussion

Our results, obtained with miniature language learning environments, indicate that the presence of variation sets in the learner's input, in the same proportion as in real child-directed speech (20%), facilitates the discovery of linguistic structure at two different levels of analysis: finding words in continuous speech, and identifying the phrasal constituents of sentences. Variation sets offer immediate and effective cues to linguistic structure by making it possible for the learner to resort to local (hence computationally inexpensive) and, crucially, statistically verifiable procedures based on alignment and comparison of successive utterances.

Current unsupervised computational approaches to finding structure typically rely on global cues, in that they amass statistical evidence over the entire learning experience (be it within an experimental session of six minutes, or over a sample corpus of language) to infer the reliability of candidate structures. This is true both in lexicon learning (e.g., Brent, 1999) and in syntax learning (e.g., Solan, Horn, Ruppin, and Edelman, 2005). This makes global approaches computationally costly (for example, requiring a word learning algorithm to maintain all possible candidate segmentations), as well as cognitively implausible.

Indeed, the lack of learning of our subjects in the Scrambled condition of Experiment 1 suggests that global, combinatorially promiscuous alignment is not resorted to even for a small lexicon. Given the small lexicon in Experiment 1, spurious variation sets interleaved by one or two sentences were likely to occur, and yet subjects did not seem to have used such non-local alignments. In contrast, in the Varset condition, in which local alignment cues were present, learning did occur, even for words that did not participate in variation sets in training. Presumably, once lexical candidates are revealed in a variation set, they are also more recognizable when they occur in other sentences, thus promoting in turn the segmentation of novel words.

The results of our Experiment 2 may be compared to earlier work in artificial language learning that used cross-sentential cues such as that of Morgan, Meier, and Newport (1989). These researchers found that when an artificial grammar was augmented with substitution phrases and variations in order of permutation between phrases, learning improved with respect to a baseline condition that contained no such variations, and whose adjacent sentences were merely repeated. The stimuli of Morgan et al. (1989), which included visual cues to category membership, consisted of pairs of aligned written sentences and geometrical figures on the screen. In our experiments, in comparison, sentences were presented sequentially in their natural auditory modality.

In a recent study, Thompson and Newport (2007) examined the effects on syntax learning of partially overlapping material between sentences, presented in the auditory modality. They did not, however, control the variation sets as such; rather, they constructed increasingly more complex languages that gave rise to more variation sets, and were able to show that more complex grammars could actually be easier to learn. In contrast, our experiments are the first ones to manipulate only the order of presentation of the stimuli (the variation sets), while maintaining the same complexity of the language across learning conditions (indeed, the very same sentences where used in both conditions in each experiment).

In summary, the positive effects of variation sets in the two experiments reported here suggest that learners can reuse the same algorithmic building blocks — alignment, comparison, and, presumably, significance assessment — at different levels of linguistic structure (here, lexical and phrasal units). We are presently extending our approach to investigate whether variation sets also facilitate the learning of other core features of language, such as lexical categorization, long-distance dependencies, and recursion.

Acknowledgment. LO was supported by NICHD Grant 5R03HD051671-02.

References

Baayen, R. H. (2006). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bates, D. (2005). Fitting linear mixed models in R. *R News 5*, 27–30.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning 34*, 71–105.

Christiansen, M. H., C. Conway, and L. Onnis (2007). Neural responses to structural incongruencies in language and statistical learning point to similar underlying mechanisms. In *Proc. of the 29th Annual Meeting of the Cognitive Science Society*, pp. –.

Dutoit, T. (1997). *An Introduction to Text-To-Speech Synthesis.* Dordrecht: Kluwer.

Edelman, S. and H. R. Waterfall (2007). Behavioral and computational aspects of language and its acquisition. *Physics of Life Reviews 4*, 253–277.

Gómez, R. L. and L. Gerken (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences 4*, 178–186.

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

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

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

Miller, G. A. (1968). *The psychology of communication: Seven essays.* Harmondworth, UK: Penguin Books.

Morgan, J. L., R. P. Meier, and E. L. Newport (1989). Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language 28*, 360–374.

Pickering, M. J. and S. Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*, 169–225.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior 6*, 855–863.

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

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

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory 1*, 113–149.

Thompson, S. P. and E. L. Newport (2007). Statistical learning of syntax: the role of transitional probability. *Language Learning and Development 3*, 1–42.

Waterfall, H. R. (2007a). Relation of variation sets to noun and verb development. Submitted.

Waterfall, H. R. (2007b). Relation of variation sets to syntactic development. Submitted.