Assessing the Efficacy of Transitional Probabilities for Learning Syntactic Categories

Erin Conwell (Erin_Conwell@brown.edu)
Brown University
Department of Cognitive and Linguistic Sciences, Box 1978
Providence, RI 02912 USA

Benjamin J. Balas (bjbalas@mit.edu)
Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences, 43 Vassar Street
Cambridge, MA 02139 USA

Abstract

While research on both infant language abilities and the informativeness of natural language for the formation of grammatical categories has advanced considerably, the extent to which these two fields inform each other is limited. To address this issue, we ask whether tracking transitional probabilities, a skill which infants are known to apply to language learning, is useful for learning grammatical categories from natural child-directed speech. We systematically remove subsets of the data to assess the relative contributions of several potential sources of information. Our analysis finds that immediately following a high frequency function word provides considerable information about whether a word is a noun or a verb. However, in unsupervised clustering, this information alone does not result in highly accurate categorization of nouns and verbs.

Keywords: Language learning; grammatical categories; unsupervised learning; supervised learning.

Introduction

Central to an understanding of how children learn language is a theory of how they learn the syntactic categories of individual lexical items. Early work on this topic concluded that children must have an innate predisposition to learn categories such as noun and verb as well as a set of innate rules linking individual words to these categories via meaning (e.g., Pinker, 1984). Such theories claim that children cannot possibly learn grammatical categories from syntactic distribution because the stimulus is impoverished and noisy (Chomsky, 1965). Not only do children not receive enough positive evidence to support category learning, they also hear words used across category boundaries or without any syntactic support at all (e.g., isolated words). This characterization of the input to children's language learning presents a grim outlook for the possibility that children, like adult linguists, categorize words based on their syntactic privileges.

However, empirical research with both adult and child language learners, as well as closer examination of the language children hear, suggests that these theories may underestimate both the abilities of language learners and the informativeness of a child's linguistic environment. Artificial language studies demonstrate that learners are highly sensitive to the statistics of the language that they hear and can use that information to find word boundaries and learn word classes (Gerken, Wilson & Lewis, 2005; Gómez & Lakusta, 2004; Saffran, Aslin & Newport, 1996). Models of language learning indicate that the language that children hear appears to contain a number of cues that could be useful for learning syntactic categories (Mintz, Newport & Bever, 2002; Monaghan, Chater & Christiansen, 2005; Redington, Chater & Finch, 1998).

However, these two lines of work inform one another to a limited degree. That is to say that while our understanding of infants' linguistic abilities has advanced considerably, this knowledge is not used to inform models of category learning. Likewise, although recent models of category learning are very successful at distinguishing between grammatical categories, little empirical research is conducted to determine whether infants can make use of the kinds of cues that these models exploit (for exceptions, see Mintz, 2003; 2006; Shi, Morgan & Allopenna, 1998 and Shi, Werker & Morgan, 1999).

Improved communication between these two lines of research is critical. For a model of language acquisition to be psychologically plausible, it must make use of information that is actually available to a language learner. Conversely, infants may show evidence of certain linguistic abilities, but if they are to use them for language acquisition, these abilities must be relevant to learning not only the carefully controlled language they hear in the laboratory, but also the uncontrolled corpus of speech that they hear in the real world. Our model evaluates whether the kinds of statistics that infants track in laboratory studies are useful for the acquisition of grammatical categories over a natural speech corpus.

While function words are often absent from children's earliest language *productions*, studies of infants' language *perception* indicate that they are sensitive to the syntactic properties of function words from a very early age (Shady, 1996; Soderstrom, White, Conwell & Morgan, *in press*), which suggests that these words could play a critical role in the acquisition of more advanced syntax. Höhle, Weissenborn, Kiefer, Schulz and Schmitz (2004)

demonstrate that German-learning infants can use familiar determiners to categorize novel words as nouns. These findings dovetail with work by Valian and Coulson (1988) which shows that increasing the frequency of category markers improves adults' learning of the syntax of a miniature artificial language. The frequency of function words in speech to children and findings that infants have some knowledge of function word syntax both suggest that these words might be particularly relevant to the learning of grammatical categories.

One remaining question, however, is how infants might use their knowledge of high frequency function words to learn categories of content words (e.g., noun vs. verb). Infants are known to track transitional probabilities (TPs) between syllables, which could be useful for segmenting words from fluent speech (Saffran, et al., 1996). This kind of learning relies on the ability to track which sound sequences follow and precede which other sound sequences and how often. When adults are exposed to artificial languages, they can learn the phrase structure of these languages via transitional probabilities, especially when these probabilities are less than 1 (Thompson & Newport, 2007). These results suggest that learners are adept at tracking TPs in language and that they can exploit these statistics to learn about the structure of the language.

There are also clear cues to the beginnings and ends of utterances, including pauses and changes in pitch. Not only are infants are sensitive to these cues (Nazzi, Kemler Nelson, Jusczyk, & Jusczyk, 2000), but research also shows that the presence of an utterance boundary facilitates infants' segmentation of words from fluent speech (Fernald, McRoberts, & Herrera, 1992). Occurring in utterance initial or final position can be described in terms of transitional probabilities as well: the probability that a word immediately precedes or follows an utterance boundary. This information may also be relevant to categorization of nouns and verbs.

We know that these sources of information are available to infants during the language learning process. The next question is whether such information is useful in a natural language context. Other research with infants indicates that the presence of a highly familiar word facilitates the segmentation of words from fluent speech (Bortfeld, Morgan, Golinkoff & Rathbun, 2005) and that infants are sensitive to the syntactic privileges of high frequency function words (Höhle, et al., 2004; Soderstrom, et al., *in press*). If infants can rely only on highly frequent words as a cue to categorization, this considerably lessens the memory load required by other models (Redington, et al., 1998). We now evaluate whether transitional probabilities with high frequency words and the likelihood of appearing at the beginnings or ends of utterances are useful cues for categorizing nouns and verbs in natural child-directed speech. We also assess the relative contributions of each information source by systematically removing subsets of the data. In so doing, we are able to identify which aspects of the data make the greatest contribution to the categorization process.

Method

Corpus Preparation

To determine whether transitional probabilities are useful for learning grammatical categories from natural language, we begin with the maternal speech to two children, Lily and William, from the Demuth Providence Corpus (Demuth, Culbertson & Alter, 2006). These children were recorded at home for an hour every other week for two years, beginning with their first words. Because of some interesting patterns in her development, the Lily corpus contains almost twice as many recordings as the William corpus. To make the two corpora comparable, we use only the first 20 files from each. During this period, William was 16-26 months old and Lily was 13-21 months old. These data, therefore, represent the linguistic input to learners before and during the onset of combinatory speech.

We extract all maternal utterances from the first 20 files in each corpus and remove the CHAT coding conventions from them to leave only the words in each utterance. Common contractions (e.g., *gonna*, *don't* and *I'm*) are left in place, while less common contractions (e.g., *d'you*, *dontcha*) are spelled out to ameliorate variation in spelling among transcribers. We use the first 5 files from each corpus to create the set of high frequency words, while the transitional probabilities are obtained from the next 15 files. The logic is that very young learners may need time to learn which words are highly frequent before they begin using those words to learn categories. In the Lily corpus, the first 5 files contain 20,110 words in 5,579 maternal utterances and the next 15 files contain 67,526 words in 14,322 maternal utterances. For the William corpus, these numbers are 20,084 words in 4,805 utterances in the first 5 files and 47,911 words in 10,821 utterances in the next 15 files.

Calculating Transitional Probabilities

To determine which words should be included as highly frequent, we count the number of times each word occurs in the first five files and divide by the total number of words in those files to calculate the percentage of the total number of tokens accounted for by each word. Those words comprising more than 1% of all tokens are considered to be highly frequent. This cut-off is arbitrary, but it is probable that a word which accounts for more than 1% of all tokens would be familiar to the learner. In the Lily corpus, these words are *the, you, a, that, and, that's, your, in, is, I, yeah* and *little*. For the William corpus, these words are *you, the, that, is, and, a, what, it, this, that's, on, do, oh, I, right, to, can, what's* and *here*. Notice that there is considerable overlap between the two sets of words and that most of them are closed-class function items.

For every word in the remaining 15 files, we calculate the forward and backward transitional probabilities between that word and each high frequency word. Transitional probability is the number of times a word appears

immediately adjacent to the high frequency word divided by the total number of times that word appears in the corpus. For example, the word *baseball* appears 4 times in the William corpus and 2 of those times it appears immediately after *the*. Therefore, the backward TP between *baseball* and *the* is 0.5.

Likewise, the probability of every word in each corpus occurring at the beginning of an utterance is the number of times the word appears at the beginning of an utterance divided by the total number of times the word appears in the corpus. We perform the same calculation with the number of times a word appears at the end of an utterance to determine the probability of that word occurring utterance-finally. Single word utterances are excluded from this analysis. We then place all of these probabilities into a matrix such that each row represents a word and each column represents a feature (i.e., the TP between that word and one of the high frequency words).

Modeling the Data

We carry out two analyses on the transitional probabilities obtained from the corpora: unsupervised clustering and supervised classification. In both cases, we ultimately describe each target word type as a point in a high-dimensional space, each axis of which denotes the value of a unique transitional probability. By systematically eliminating subsets of features (TPs) from consideration, we can perform both procedures in distinct subspaces to determine how particular families of transitional probabilities contribute to noun/verb categorization. In particular, we consider (1) the full set of TPs, (2) all TPs, except first word or last word information, (3) only forward TPs, and (4) only backward TPs. A description of each subset and the number of components analyzed for each is provided in Table 1. We proceed by describing the computational details of our clustering and classification procedures.

Preprocessing Given the matrix of TPs for all target words, we first eliminate any words that have zero values for all TPs we consider. We also remove all words that are not nouns or verbs. Next, given the high dimensionality of our raw data (26 features for Lily, 40 features for William) we carry out Principal Components Analysis (PCA) to find a low-dimensional embedding of the target words in feature space. PCA is a commonly used technique that finds a rigid rotation of the original coordinate frame such that the maximum amount of variance in the data is captured by the first new axis, and decreasing amounts are captured by each subsequent axis. To determine the best number of components for describing each data set with, we perform a graphical analysis on the plots of explained variance vs. component number to find an "elbow" in each plot. At this point, the addition of a new component does not substantially increase the amount of variance explained by the new axes. Table 1 shows the number of components used in each analysis. Although the number of components varies for each subset, we find that repeating our analysis with different choices has a negligible effect on the outcomes. This procedure is carried out anew for each feature subset we consider.

Clustering With each target word now embedded in a lower-dimensional space, we use the *k*-means algorithm to do unsupervised clustering of the data (MacQueen, 1967). The algorithm requires that the user specify the number of clusters *k* to search for. The result is a solution for each value of *k* in which each target word is assigned to a unique cluster. For each feature subset, we report the characteristics of the 2-cluster solution, since ideally this would produce one category of nouns and one of verbs.

***k*-Nearest-Neighbor Classification** Finally, given the same embedding of the data we use for clustering, we classify each target word as noun or verb in a supervised manner. We accomplish this using a "leave-one-out" procedure, in which each target word is removed from the data set one at a time and the remaining labeled items are used to determine its category. We use the *k*-nearest neighbor procedure to assign labels. In this algorithm the user specifies the number of "neighbors" that get to vote for category membership of the target according to their own label. Given a value for *k*, the *k* closest data points to the target in feature space are identified, and the target is labeled according to the majority label of these neighbors (Duda, Hart, & Stork, 2001).

We set $k=1$ for each feature set and also determine the maximum performance across values of *k* less than 50.

Table 1: Composition of each subset used for analysis.

Subset	Corpus	Nouns	Verbs	Components
All TPs	Lily	1323	702	5
	William	747	501	7
No First/Last	Lily	1046	535	7
	William	605	438	5
Forward	Lily	502	391	4
	William	334	316	8
Backward	Lily	919	378	4
	William	510	357	4

Results

The results of all cluster analyses are summarized in Figure 1 and all *k*-nearest neighbor analyses in Figure 2. For the cluster analyses, accuracy is the percentage of words in a particular group that are members of the majority category. For the *k*-nearest neighbor analyses, accuracy is the percentage of trials on which a word is correctly classified in a leave-one-out procedure.

When all potential factors are included in the analysis, two-cluster analysis results in an average of 67.3% correct classification for the Lily corpus and 66.8% correct classification for the William corpus. The *k*-nearest

neighbor evaluation indicates that, for one neighbor, the Lily corpus produces 67.8% correct classification and the William corpus has 74% correct classification. Maximum correct classification is obtained with 31 neighbors in the Lily corpus, raising accuracy to 73.9%. In the William corpus, maximum correct classification of 78.8% is obtained with 11 neighbors.

We next ask whether information about how often a word appears first or last in an utterance is useful for categorization by removing this information from the data and re-examining the accuracy of both the cluster analysis and the *k*-nearest neighbor analysis. In the Lily corpus, this improves the 2-cluster accuracy to 77.7%. In the William corpus, the improvement is less striking, with 2-cluster accuracy now 72.4%. For *k*-nearest neighbor, using only a single neighbor, the Lily corpus is 78.3% accurate and the William corpus is 75.1% accurate. Maximum correct classification of 85.8% is achieved with 7 neighbors in the Lily corpus and maximum correct classification of 79.1% is obtained with 41 neighbors in the William corpus. Since removing information about sentence position improves accuracy of categorization in one corpus and does not change the accuracy for the other corpus, we conclude that this information is not useful for the formation of noun and verb categories. These data are excluded from further analyses.

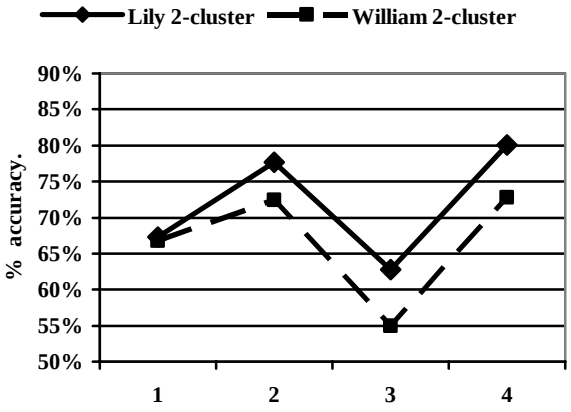

Figure 1: Mean accuracy of for *k*-means clustering for each subset of the corpus. (1) contains all information, (2) includes all TPs, but no first or last word information, (3) contains only forward TPs and (4) contains only backward TPs.

We now turn to the issue of the relative contributions of preceding and following high frequency words. Because function words often mark the beginnings of phrases, it is likely that following a function word provides more information about a category than preceding a function word (e.g., nouns follow determiners, verbs follow auxiliaries, etc.). To test this, we once again run clustering and *k*-nearest neighbor analyses on the data, this time separating forward and backward transitional probabilities. Forward transitional probability is the likelihood that a target word will immediately precede a high frequency word. Backward transitional probability is the likelihood that a word immediately follows a high frequency word.

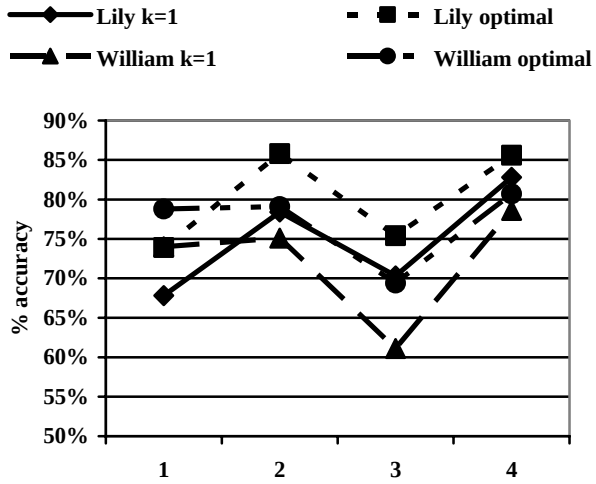

Figure 2: Accuracy for *k*-nearest-neighbor classification for each subset of the corpus. (1) contains all information, (2) includes all TPs, but no information about utterance position, (3) includes only forward TPs and (4) includes only backward TPs.

When only forward transitional probabilities are included in the analysis, accuracy for 2 clusters decreases markedly in both the Lily corpus (62.8%) and the William corpus (55%). In a *k*-nearest neighbor analysis, accuracy when considering one neighbor is 70.3% for the Lily corpus and 61% for the William corpus. The maximum accuracy of *k*-nearest neighbor for the Lily corpus is 75.4% and occurs when 21 neighbors are considered. For the William corpus, the maximum accuracy is 69.4% and occurs when 11 neighbors are considered. Again, these accuracies decrease from the analysis in which both forward and backward transitional probabilities are considered. This suggests that forward transitional probabilities are not a good source of information for forming grammatical categories or, alternatively, that backward TPs are especially good for categorization.

When only backward transitional probabilities are considered, however, clustering accuracy improves relative to when forward TPs are considered. For the Lily corpus accuracy rises to 80.1% for two clusters. In the William corpus, accuracy remains unchanged at 72.8% for two clusters. Turning to the *k*-nearest neighbor analysis, accuracy when one neighbor is considered is 82.8% for the Lily corpus and 78.6% for the William corpus. The maximum accuracy of 85.6% for the Lily corpus is obtained when 11 neighbors are considered. For the William corpus, the maximum accuracy of 80.7% is obtained when 41 neighbors are considered. Because the accuracies of both the unsupervised and supervised methods improve when only backward transitional probabilities are considered, this

suggests that what information transitional probabilities contain for categorizing nouns and verbs is largely contained in backward transitional probabilities. That is to say, a high frequency word that immediately precedes a target word contains more information about that word's category than does a high frequency word that immediately follows the target word.

For all clustering analyses in all conditions, a similar pattern can be observed. While there is often one small, highly accurate cluster, typically composed of nouns, there is always one very large cluster which contains approximately equal numbers of each word category. Since all subsets of both corpora contain more nouns than verbs, this large category often contains all but a few of the verbs. Therefore, almost all of the verbs are distributed in a single location. While this does not allow a learner to discriminate between nouns and verbs, it does allow the learner to conclude that items outside of this area are probably *not* verbs. Furthermore, the success of the *k*-nearest neighbor analyses suggests that there may be some structure within this large cluster that might be revealed if a much larger number of clusters were to be considered.

General Discussion

This paper evaluates the effectiveness of transitional probabilities for learning the syntactic categories of words from a corpus of natural, child-directed speech. By removing subsets of data from the analysis, we are able to assess the relative contributions of each data source. While unsupervised clustering does not reach the levels of accuracy that some researchers find using other statistics (Mintz, 2003; Monaghan, et al., 2005), it is on par with some models (Redington, et al., 1998) and is motivated by empirical research into children's linguistic abilities. This work also uses two large corpora of speech and tries to categorize as many nouns and verbs as possible, not just the most frequent ones.

We examine the extent to which 3 subsets of these TPs are useful for categorizing nouns and verbs. The likelihood that a word appears at the beginning or at the end of an utterance probably contributes little information relevant to learning the differences between nouns and verbs. However, it is important to remember that this analysis includes disfluent utterances, false starts and interruptions, which may introduce noise into the data. Because these are natural corpora, we know that young language learners are exposed to such noise. Still, there is evidence that young learners are sensitive to disfluency as distinct from fluent speech (Soderstrom & Morgan, *in press*). If infants and the model were to exclude such utterances from the analysis, the efficacy of utterance initial and final statistics might improve.

Furthermore, when only forward transitional probabilities between the target word and the high frequency words are considered, accuracy decreases. Conversely, when only backward transitional probabilities between the target word and a high frequency word are considered, accuracy improves. This suggests that the high frequency word preceding a target word is a good predictor of grammatical category. Höhle and colleagues (2004) show that German-learning infants do indeed use co-occurrence with determiners to categorize novel words as nouns. Our results complement the findings of their work by showing that information which is used by infants to learn categories in the laboratory is also useful over a corpus of natural child-directed speech.

Supervised learning via *k*-nearest neighbor analysis suggests that these cues are better for supervised learning than for unsupervised learning. This is problematic for a theory of natural language acquisition because infants do not learn language in a supervised way. The little feedback that they receive is often uninformative and there is good evidence that they do not make use of it (Morgan, Bonamo & Travis, 1995). The success of these methods does suggest, however, that infants might be able to use transitional probabilities to assign words to categories once rudimentary categories have been formed. Some theories of category acquisition propose that infants might begin with categories based on semantics (e.g., Pinker, 1989) or phonological properties (e.g., Gleitman & Wanner, 1982). One major obstacle for both types of theories is explaining how a learner would transition from such proto-categories to an adult-like system based on syntactic distribution. Perhaps transitional probabilities could be incorporated into these theories in a semi-supervised way to facilitate such a transition.

This paper finds that frequency and transitional probabilities, statistics which infants are known to compute in laboratory studies of language learning, are moderately effective for categorizing nouns and verbs in a corpus of child-directed speech. Systematic manipulation of the data indicates that the most powerful indicator of noun or verb status is the high frequency word that immediately precedes the target word. By better characterizing the relationship between the abilities of human learners and the richness of their linguistic environments, we can create a more accurate portrait of the language learning process. A continued exchange of ideas between modelers and experimental researchers will be vital to such a process.

Acknowledgments

BJB is funded by a National Defense Science and Engineering Graduate fellowship. The authors wish to thank Katherine Demuth for allowing us to use the Demuth Providence Corpus and Elizabeth McCullough for facilitating access to the corpus. We are also grateful to Naomi Feldman, Melanie Soderstrom and three anonymous reviewers for very helpful comments on an earlier version of this paper.

References

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science, 16*, 298-304.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Boston, MA: MIT Press.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language and Speech, 49*, 137-174.

Duda, R. O., Hart, P. E., & Stork, P. G. (2001). *Pattern classification*. New York: John Wiley and Sons, Inc.

Fernald, A., McRoberts, G. W., & Herrera, C. (1992). Prosodic features and early word recognition. Paper presented at the 8th International Conference on Infant Studies, Miami, FL.

Gerken, L. A., Wilson, R., & Lewis, W. (2005). 17-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249-268.

Gleitman, L. R. & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge, UK: Cambridge University Press.

Gómez, R. L. & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7*, 567-580.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy, 5*, 341-353.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA.

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

Mintz, T. H. (2006). Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs*. New York: Oxford University Press.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorization. *Cognition, 96*, 143-182.

Morgan, J. L., Bonamo, K. M., & Travis, L. L. (1995). Negative evidence on negative evidence. *Developmental Psychology, 31*, 180-197.

Nazzi, T., Kemler Nelson, D.G., Jusczyk, P.W., & Jusczyk, A.M. (2000). Six-month-olds' detection of clauses in continuous speech: Effects of prosodic well-formedness. *Infancy, 1*, 123-147.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: MIT Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

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

Shady, M. (1996). Infants' sensitivity to function morphemes. Unpublished dissertation, State University of New York, Buffalo, NY.

Shi, R., Morgan, J., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*, 169-201.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 27*, B11-B21.

Soderstrom, M. & Morgan, J. L. (*in press*). Twenty-two-month-olds discriminate fluent from disfluent adult-directed speech. *Developmental Science*.

Soderstrom, M., White, K. S., Conwell, E., & Morgan, J. L. (*in press*). Receptive grammatical knowledge of familiar content words and inflection in 16-month-olds. *Infancy*.

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

Valian, V. & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27*, 71-86.