Using Phoneme Distributions to Discover Words and Lexical Categories in Unsegmented Speech

Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Stephen A. Hockema (shockema@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
Bloomington, IN 47405 USA

Luca Onnis (lo35@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

When learning language young children are faced with many formidable challenges, including discovering words embedded in a continuous stream of sounds and determining what role these words play in syntactic constructions. We suggest that knowledge of phoneme distributions may play a crucial part in helping children segment words and determining their lexical category. We performed a two-step analysis of a large corpus of English child-directed speech. First, we used transition probabilities between phonemes to find words in unsegmented speech. Second, we used distributional information about word edges—the beginning and ending phonemes of words—to predict whether the segmented words were nouns, verbs, or something else. These results indicate that discovering lexical units and their associated syntactic category in child-directed speech is possible by attending to the statistics of single phoneme transitions and word-initial and final phonemes.

Introduction

One of the first tasks facing an infant embarking on language development is to discover where the words are in fluent speech. This is not a trivial problem because there are no acoustic equivalents in speech of the white spaces placed between words in written text. To find the words, infants appear to be utilizing several different cues, including lexical stress (Curtin, Mintz & Christiansen, 2005), transitional probabilities between syllables (Saffran, Aslin & Newport, 1996), and phonotactic constraints on phoneme combinations in words (Jusczyk, Friederici & Svenkerud, 1993). Among these word segmentation cues, computational models and statistical analyses have indicated that, at least in English, phoneme distributions may be the single most useful source of information for the discovery of word boundaries (e.g., Brent & Cartwright, 1996; Hockema, 2006), especially when combined with information about lexical stress patterns (Christiansen, Allen & Seidenberg, 1998).

Discovering words is, however, only one of the first steps in language acquisition. The child also needs to discover how words are put together to form meaningful sentences. An initial step in this direction involves determining what syntactic roles individual words may play in sentences. Several types of information may be useful for the discovery of lexical categories, such as nouns and verbs, including distributions of word co-occurrences (e.g., Redington, Chater & Finch, 1998), frequent word frames (e.g., *I X it*; Mintz, 2003), and phonological cues (Kelly, 1992; Monaghan, Chater & Christiansen, 2005). Indeed, merely paying attention to the first and last phoneme of a word has been shown to be useful for predicting lexical categories across different language such as English, Dutch, French and Japanese (Onnis & Christiansen, 2005).

During the first year of life, infants become perceptually attuned to the sound structure of their native language (see e.g., Jusczyk, 1997; Kuhl, 1999, for reviews). We suggest that this attunement to native phonology is crucial not only for word segmentation but also for the discovery of syntactic structure. Specifically, we hypothesize that phoneme distributions may be a highly useful source of information that a child is likely to utilize in both tasks. In this paper, we test this hypothesis by carrying out a two-step corpus analysis in which information about phoneme distribution is used first in Experiment 1 to segment words out of a large corpus of phonologically-transcribed child-directed speech and then in Experiment 2 to predict the lexical category of these words (noun, verb, or other). The results show that it is possible to get from unsegmented speech to lexical categories with a reasonably high accuracy and completeness using only information about the distribution of phonemes in the input.

Experiment 1: Discovering Words

Infants are proficient statistical learners, sensitive to sequential sound probabilities in artificial (Saffran et al., 1996) and natural language (Jusczyk et al., 1993). Such statistical learning abilities would be most useful for word segmentation if natural speech was primarily made up of two types of sound sequences: ones that occur within words

and others that occur at word boundaries. Fortunately, natural language does appear to have such bimodal tendencies (Hockema, 2006). For example, in English /tg/ rarely, if ever, occurs inside a word and thus is likely to straddle the boundary between a word ending in /t/ and another beginning with /g/. On the other hand, the transition /Iŋ/ (the two phonemes making up *–ing*) almost always occurs word internally. Here we demonstrate that sensitivity to such phoneme transitions provides reliable statistical information for word segmentation in English child-directed speech.

Method

Corpus preparation. For our analysis we extracted all the speech directed by adults to children from all the English corpora in the CHILDES database (MacWhinney, 2000). The resulting corpus contained 5,470,877 words distributed over 1,369,574 utterances. Because most of these corpora are only transcribed orthographically, we obtained citation phonological forms for each word from the CELEX database (Baayen, Pipenbrock & Gulikers, 1995) using the DISC encoding that employs 55 phonemes for English. In the case of homographs (e.g., *record*), we used the most frequent of the pronunciations. Moreover, recent detailed analyses indicate that dual-category words are consistently in one category only in child-directed speech (Jim Morgan, personal communication). Another 9,117 nonstandard word type forms (e.g., *ain't*) and misspellings in CHILDES were coded phonetically by hand. Sentences in which one or more words did not have a phonetic transcription were excluded.

Analyses. We first computed the probability of encountering a word boundary between each possible phoneme transition pair in the corpus. There were 3,025 (55^2) possible phoneme transition pairs (types). Transitions across utterance boundaries were not included in the analyses. Having obtained the type probability of word boundary between each pair of phonemes, we made another pass over the CHILDES corpora phoneme stream and used this information in a simple system that inserted word boundaries in any transition token whose type probability was greater than .5. That is, we went through the unsegmented stream of phonemes and inserted a word boundary whenever the probability of such boundary occurring for a phoneme transition pair (token) was greater than .5.

Results and Discussion

Of the 3,025 possible phoneme transition pairs, 954 (35%) never occurred in the corpus. Figure 1.a provides a histogram showing the distribution of phoneme transition pairs as a function of how likely they are to have a word boundary between them, given the proportion of occurrences in our corpus for which a boundary was found. The bar height indicates the percentage of phoneme transition pairs with a given probability of having a word boundary between them. The separate column on the right indicates the percentage of phoneme transition pairs that never occurred in the corpus. Figure 1.a clearly illustrates

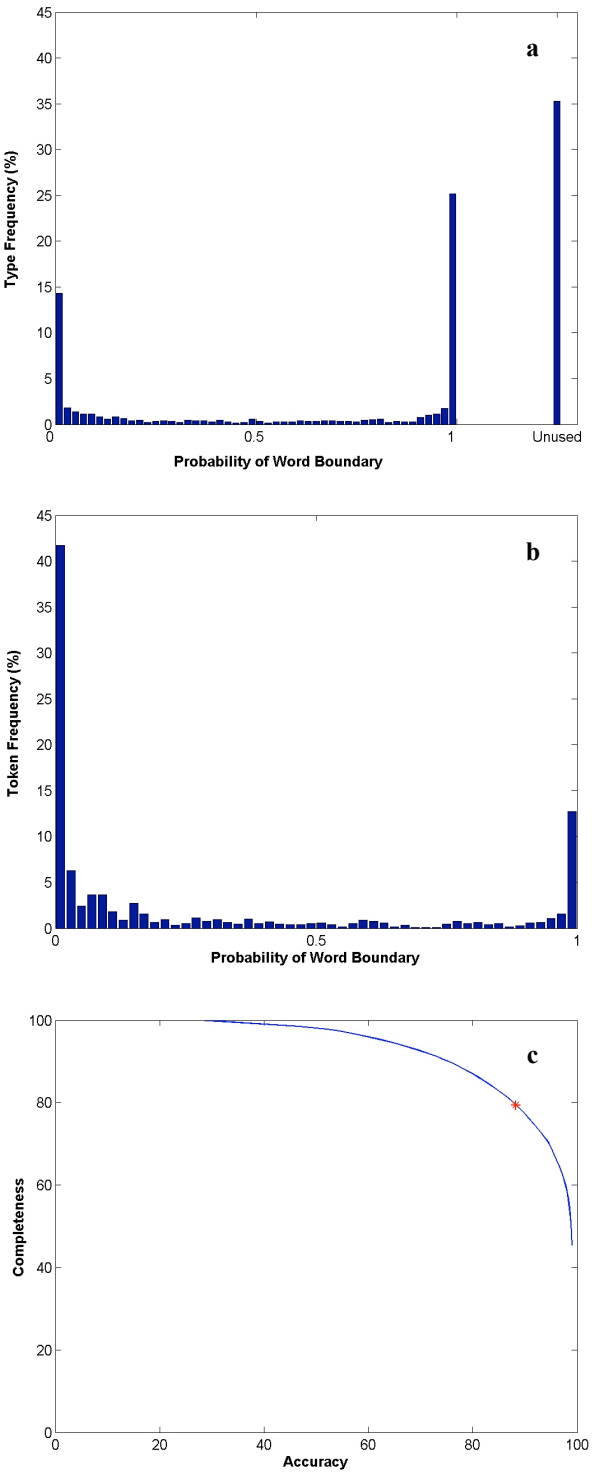

Figure 1: Distribution of phoneme transition pairs given the probability of encountering a word boundary between the two phonemes for types (a) and tokens (b), and a ROC curve (c) indicating the accuracy/completeness trade-off when predicting lexical boundaries using tokens.

that the distribution of used phoneme transition pairs was strongly bimodal. Most phoneme transitions were either associated only with a word boundary or occurred only inside a word, but not both. Indeed, 61% of the used phoneme transition pairs were in the right- or leftmost bin.

These data, however, show the distribution of phoneme transition pairs independently of whether they occur only once or many thousands of times. To get an idea of the distribution of the phoneme transition pair tokens that a child might actually come across in the input, we weighted each phoneme transition pair by its frequency of occurrence across the corpus. Figure 1.b shows the distribution of phoneme transition pairs that a child is likely to hear has a similar bimodal distribution as for the type analyses.

To assess the usefulness of this type of phoneme distribution information for lexical segmentation, we determined how well word boundaries can be predicted if inserted whenever the probability of boundary occurrence for a given phoneme transition pair is greater that .5. In all, 4,576,783 word boundaries were inserted. We used two measures—accuracy and completeness—to gauge the reliability of the lexical boundary predictions. Accuracy is computed as the number of correctly predicted boundaries (hits) in proportion to all predicted word boundaries, both correct (hits) and incorrect (false alarms). Completeness is calculated as the number of correct boundaries (hits) in proportion to the total number of boundaries; that is, the correct boundaries (hits) and the boundaries that the system failed to predict (misses). Thus, accuracy provides an estimation of the percentage of the predicted boundaries that were correct, whereas completeness indicates the percentage of boundaries actually found out of all the boundaries in the corpus. Figure 1.c shows an ROC curve, indicating the trade-off between accuracy and completeness given different cut-off points for when to predict a word boundary. The asterisk denotes the .5 cut-off point used in the current analyses, revealing an accuracy of 88% and a completeness of 79% for predicted word boundaries.

Predicting lexical boundaries is not the same as segmenting out complete words. We therefore used a conservative measure of word segmentation in which a word is only considered to be correctly segmented if a lexical boundary is predicted at the beginning and at the end of that word without any boundaries being predicted word-internally (Brent & Cartwright, 1996; Christiansen et al., 1998). For example, if lexical boundaries were predicted before /k/ and after /s/ for the word /kæts/ (*cats*), it would be considered correctly segmented; but if an additional boundary was predicted between /t/ and /s/ the word would be counted as missegmented (even though this segmentation would be useful for learning morphological structure). Using this conservative measure we computed segmentation accuracy and completeness for complete words. Overall, the model identified 70.2% of the words in our corpus (completeness), while 74.3% of the words it identified were valid (accuracy). The missegmented words were classified into word fragments (where a boundary had erroneously been inserted within a word; e.g., the word *picnic* got split into two fragments, /pIk/ and /nIk/) and combination words ("combo-words", where a boundary had been missed causing two words to be conjoined; e.g., the boundary between *come* and *on* was missed, yielding a single lexical unit, *comeon*). There were 558,707 fragments and 322,197 combo-words.

These results replicate what was found in previous work (Hockema, 2006), this time using a larger alphabet of phonemes, a different lexicon for pronunciations, and an even larger, more diverse corpus of child-directed speech: phoneme transitions contain enough information about word boundaries such that a simple model that attends only to these can do well enough to bootstrap the word segmentation process. However, it is still an open empirical question as to how infants might actually make use of this regularity. Previous research has speculated that infants may attend to phoneme transition probabilities, with relatively infrequent transitions indicating word boundaries. We evaluated the potential of this strategy by computing the correlation between bigram transition probabilities and the actual probability of finding a word boundary across phoneme pairs. As expected, this was significantly negative ($r = -.25$), but perhaps not strong enough to wholly support the process, suggesting that infants relying on dips in transition probability to detect word boundaries would need to supplement this strategy with other cues (such as prosodic stress). This, however, does not rule out other strategies that could rely solely on pairwise phoneme statistics. For example, infants might bootstrap segmentation by building a repertoire of phonemes that frequently occur on word edges (first learned perhaps from isolated words). Our data show that transitions among these will very reliably indicate word boundaries. Note that for phoneme transition statistics to be useful, infants do *not* have to pick up on them directly, they just have to attend to word edges, which, given the regularity we found in the language, could be enough to bootstrap segmentation.

Experiment 2: Discovering Lexical Categories

In Experiment 1, we presented a simple phoneme-based model capable of reasonably accurate and complete segmentation of words from unsegmented speech. However, performance was not perfect as evidenced by the number of word fragments and combo-words. The question thus remains whether the imperfect output of our segmentation model can be used by another system to learn about higher-level properties of language. From previous work, we know that beginning and ending phonemes can be used to discriminate the lexical categories of words from pre-segmented input (Onnis & Christiansen, 2005). This is supported by evidence that both children (Slobin, 1973) and adults (Gupta, 2005) are particularly sensitive to the beginning and endings of words. In Experiment 2, we explore whether such word-edge cues can still lead to reliable lexical classification when applied to the noisy output of our word segmentation model. We hypothesized

that missegmented phoneme strings may not pose as much of a problem as one might expect because such phoneme sequences are more likely to have less coherent combinations of word-edge cues compared to lexical categories such as nouns and verbs.

Method

Corpus preparation. The segmented corpus produced by the segmentation model in Experiment 1 was used for the word-edge analyses. The lexical category for each word was obtained from CELEX (Baayen et al., 1995). Homophones were assigned the most frequent lexical class in CELEX. Several words also had more than one lexical category. Nelson (1995) showed that for these so-called dual-category words (e.g., *brush*, *kiss*, *bite*, *drink*, *walk*, *hug*, *help*, and *call*) no specific category is systematically learned before the other, but rather the frequency and salience of adult use are the most important factors. Dual-category words were therefore assigned their most frequent lexical category from CELEX. In total, there were 101,721 different lexical item types, of which 7,432 were words, and the remaining were combo-words and fragments. Among words, 4,530 were nouns, and 1,601 were verbs.

Cue derivation. The CELEX DISC phonetic code used in incorporates 55 phonemes to encode English phonology. Each lexical item was represented as a vector containing 110 (55 beginning + 55 ending) bits. If the word started and ended with one of the English phonemes, then its relevant bit in the vector was assigned a 1, otherwise a 0. Thus, the encoding of each word in the corpus consisted of a 110-bit vector with most bits having value 0 and two having value of 1. These 110 bits formed the Independent Variables to be entered in a discriminant analysis. The Dependent Variable was the lexical category of each item.

To assess the extent to which word-edge cues can be used for reliable lexical category classification, we performed a linear discriminant analysis dividing words into Nouns, Verbs, or Other. Discriminant analyses provide a classification of items into categories based on a set of independent variables. The chosen classification maximizes the correct classification of all members of the predicted groups. In essence, discriminant analysis inserts a hyperplane through the word space, based on the cues that most accurately reflect the actual category distinction. An effective discriminant analysis classifies words into their correct categories, with most words belonging to a given category separated from other words by the hyperplane. To assess this effectiveness, we used a "leave-one-out cross-validation" method, which is a conservative measure of classification accuracy, and works by calculating the accuracy of the classification of words that are not used in positioning the hyperplane. This means that the hyperplane is constructed on the basis of the information on all words except one, and then the classification of the omitted word is assessed. This is then repeated for each word, and the overall classification performance can then be determined.

Children's syntactic development is perhaps best characterized as involving fragmentary and coarse-grained knowledge of linguistic regularities and constraints (e.g., Tomasello, 2003). In this respect, it seems more reasonable to assume that the child will start assigning words to very broad categories that do not completely correspond to adult lexical categories (Nelson, 1995). In addition, the first adult-like lexical categories will be the most relevant to successful communication. For example, noun and verb categories will be learned earlier than mappings to conjunctions and prepositions (Gentner, 1982). Hence, the task of the discriminant analysis was to classify the whole corpus into three categories: Nouns, Verbs, and Other. This classification plausibly reflects the early stages of lexical acquisition, with Other being an amalgamated "super-category" incorporating all lexical items that are not nouns or verbs. Accordingly, the lexical category was derived from CELEX for all words. Words that had a lexical category other than noun or verb were assigned to Other, along with the combo-words and fragments.

To provide the best measure of the classification problem that a child faces during language learning, each case—that is, the 110-bit vector corresponding to each word, fragment, or combo-word—was weighted by its frequency. The resulting token-based discriminant analysis thus takes into account the frequency of occurrence of the lexical items in the corpus.

In evaluating the true contribution of word-edge cues to classification, it is important to take into account that a certain percentage of cases could be correctly classified simply by chance. To establish the chance-level of performance, a baseline condition was therefore generated using Monte Carlo simulations. The file containing the data from the corpus had 111 columns: the 110 columns of binary word edge predictors (Independent Variables), plus one column that had dummy variable scores of 1, 2, or 3 for the three lexical categories (Dependent Variable). This last column contained 4,530 values of 1 (Noun), 1601 values of 2 (Verb), and 95,590 values of 3 (Other). We randomly rescrambled the order of the entries in that column while leaving the other 110 columns (the word-edge predictors) unchanged. Thus, the new random column had the exactly same base rates as the old column in random order, while the first 110 columns were completely unchanged. The rescrambling maintains information available in the vector space, but removes potential correlations between specific word-edge cues and lexical categories, and thus represents an empirical baseline control. We created 100 different rescramblings for the Dependent Variable and tested the ability of the 110 word-edge cues to predict each one of the rescramblings in 100 separate discriminant analyses. The mean classification scores from the rescrambled analyses were then compared with the results from the word-edge analysis using standard t-tests. In this way, it was possible to determine whether in the experimental condition there was a significant phonological consistency within nouns, within verbs, and within other words or whether a three-way

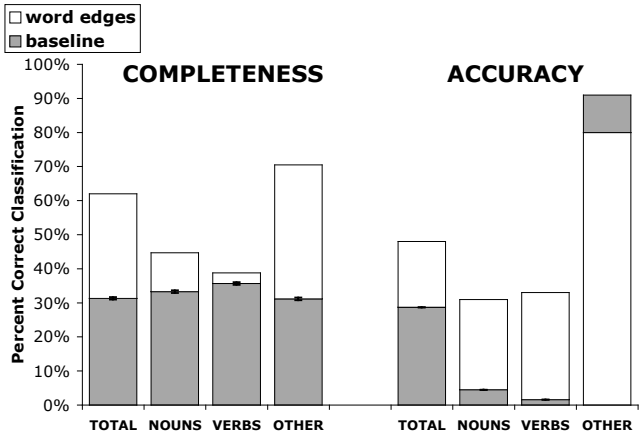

Figure 2: Classification completeness and accuracy of the lexical items from the segmentation model into lexical categories, based on first and last phoneme in each word.

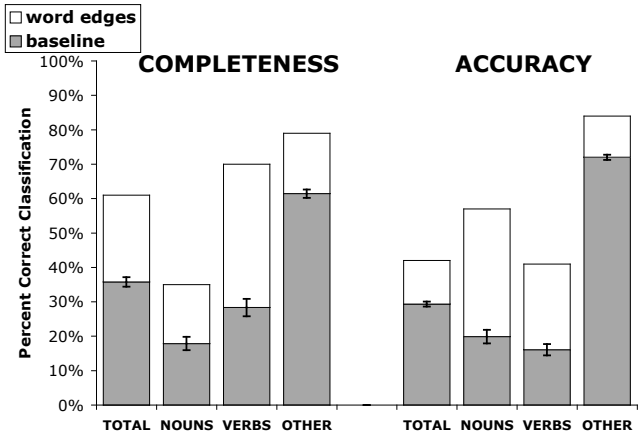

Figure 3: Classification performance of the model generalizing word-edge cues from 50 nouns and verbs to 101,565 novel lexical items.

classification of words randomly assigned to the three categories would result in the same level of classification.

Results and Discussion

Using the word-edge cues, 62.0% of the cross-validated lexical tokens were classified correctly, which was highly significant ($p < .001$). In particular, 44.7% of Nouns, 38.8% of Verbs, and 70.5% of Other words were correctly classified using word-edge cues. To test against chance levels 100 Monte Carlo discriminant analyses were run in the baseline condition where the 101,721 lexical item vectors were randomly assigned to one of the three categories, as described above. The baseline analyses yielded a mean correct classification of 31.3% (SD=3.7%). In particular, 33.3% (SD=3.9%) of nouns, 35.7% (SD=3.8%) of verbs, and 31.2% (SD=4.0%) of other words were correctly cross-classified. Word-edge classification was significantly higher than the baseline classification for nouns, verbs, and other items ($p < .001$).

The percentages reported above provide an estimate of the completeness of the classification procedure, i.e., how many words in a given category are classified correctly. We further measured the accuracy of the classifications for each of the three categories. Accuracy and completeness scores for both the word-edge and baseline analyses are shown in Figure 2. Both classification accuracy and completeness are high for Other items, though the baseline is higher for accuracy. This is not surprising, however, given the sheer disproportion between Nouns (694,796 tokens) and Verbs (665,658 tokens) on the one side, and Other items (3,216,329 tokens) on the other side. Nonetheless, the classification of Nouns and Verbs is both relatively accurate and complete, indicating that word-edge cues are useful for discovering the lexical categories of words.

A downside of the current analyses is that they are "supervised" in that the underlying discriminant analysis model is provided with both the word-edge cues and their lexical category when seeking to find the optimal mapping from the former to the latter. To determine whether supervised exposure to only a few words would allow for generalization to subsequent words using word-edge cues alone, we conducted additional discriminant analyses. We used the top-50 most frequent nouns (19) and verbs (31) along with 106 additional lexical items from the Other category with equally high frequency. This is meant to model the slow learning of the first approximately 50 words prior to the onset of the "vocabulary spurt" around 18 months (e.g., Nelson, 1973). These first words may be learned entirely through feedback and interactions with caregivers. In this context, we are further assuming that children would be sensitive to the repeated sound patterns of the Other items without necessarily having learned their meaning. A supervised model was created using the 156 lexical items and predictions made for the remaining 101,565 lexical item types, each weighted by frequency. We additionally ran 10 baseline models using the same procedure as before.

The accuracy and completeness of the classifications for both the word-edge and baseline analyses can be seen in Figure 3. Classification based on word-edge cues was significantly higher than baseline classifications across all categories (p's $< .001$). Based on supervised exposure to only 50 nouns and verbs, the statistical model is able to generalize robustly to subsequent words based on word-edge cues alone. This kind of partial bootstrapping may help explain the vocabulary spurt: slow, supervised learning of the relationship between word-edge cues and lexical categories may be needed before it can be used to facilitate word learning. More broadly, the results not only compare well with those of Onnis and Christiansen (2005)—who used an optimally-segmented corpus as input—they also provide a first initial indication of how children might get from unsegmented speech to lexical categorization.

General Discussion

In this paper, we have presented a two-step analysis of the usefulness of information about phoneme distributions for the purpose of word segmentation and lexical category

discovery. To our knowledge, this is the first time that a combined approach has demonstrated how a single cue—phoneme distributions—can be used to get from unsegmented speech to broad lexical categories. Crucially, both steps utilized very simple computational principles to take advantage of the phoneme distributional cues, requiring only sensitivity to phoneme transitions and word edges. Importantly, these two sensitivities are in place in infants (transitional probabilities, see Saffran et al., 1996) and young children (word edges, see Slobin, 1973). Hence our analyses incorporate plausible developmental assumptions both about low computational complexity and about the type of information that might be perceptually available to infants and young children. The two experiments also demonstrate that segmentation does not have to be perfect for it to be useful for learning other aspects of language. Indeed, because word fragments and combo-words are likely to have less consistency in terms of their word-edge cues in comparison to nouns and verbs, missegmentations may even facilitate lexical-category discovery.

Our analyses have underscored the usefulness and potential importance of phoneme distributions for bootstrapping lexical categories from unsegmented speech. However, a complete model of language development cannot be based on this single source of input alone. Rather, young learners are likely to rely on many additional sources of probabilistic information (e.g., social, semantic, prosodic, word-distributional) to be able to discover different aspects of the structure of their native language. Our previous work has shown that the learning of linguistic structure is greatly facilitated when phonological cues are integrated with other types of cues, both at the level of speech segmentation (e.g., lexical stress and utterance boundary information, Christiansen et al., 1998; Hockema, 2006) and syntactic development (e.g., word-distributional information, Monaghan et al., 2005). This suggests that the phoneme distributional cues that we have explored here may in further work be incorporated into a more comprehensive computational account of language development through multiple-cue integration.

Acknowledgments

SAH was supported by a grant from the National Institute for Child Health and Human Development (T32 HD07475).

References

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Brent, M.R. & Cartwright, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221-268.

Curtin, S., Mintz, T.H. & Christiansen, M.H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96,* 233-262.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed*.) Language development, Vol. 2*. Hillsdale, NJ: Erlbaum.

Gupta, P. (2005). Primacy and recency in nonword repetition. *Memory*, *13*, 318-324.

Hockema, S.A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development*, *2*, 119-146.

Jusczyk, P.W. (1997). The discovery of spoken language. Cambridge, MA: MIT Press.

Jusczyk, P.W., Friederici, A.D. & Svenkerud, V.Y. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*, 402–420.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349-364.

Kuhl, P.K. (1999). Speech, language, and the brain: Innate preparation for learning. In M.D. Hauser & M. Konishi (Eds.), The design of animal communication (pp. 419-450). Cambridge, MA: MIT Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

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

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

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development, 38,* serial no. 149.

Nelson, K. (1995). The dual category problem in the acquisition of action words. . In M. Tomasello & W.E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 223-249). Hillsdale, NJ: Erlbaum.

Onnis, L. & Christiansen, M.H. (2005). Happy endings for absolute beginners: Psychological plausibility in computational models of language acquisition. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 1678-1683). Mahwah, NJ: Erlbaum.

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.

Slobin, D.I. (1973). Cognitive prerequisites for the development of grammar. In C.A. Ferguson & D.I. Slobin (Eds.), *Studies of child language development*. New York: Holt, Reinhart & Winston.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.