Developmental Science

SPECIAL SECTION: COMPUTATIONAL PRINCIPLES OF LANGUAGE ACQUISITION

The secret is in the sound: from unsegmented speech to lexical categories

Morten H. Christiansen,[1] Luca Onnis[2] and Stephen A. Hockema[3]

1. Department of Psychology, Cornell University, Ithaca, USA
2. Department of Second Language Studies, University of Hawaii, Honolulu, USA
3. Faculty of Information Studies, University of Toronto, Toronto, Canada

Abstract

When learning language, young children are faced with many seemingly 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 determine their lexical category, and we propose an integrated model of how children might go from unsegmented speech to lexical categories. We corroborated this theoretical model using a two-stage computational 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 from the first stage were nouns, verbs, or something else. The 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. Thus, we suggest that a core computational principle in language acquisition is that the same source of information is used to learn about different aspects of linguistic structure.

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 words, infants appear to be utilizing several different cues, including lexical stress (e.g. Curtin, Mintz & Christiansen, 2005; Jusczyk, Cutler & Redanz, 1993; Jusczyk, Houston & Newsome, 1999), transitional probabilities between syllables (e.g. Aslin, Saffran & Newport, 1998; Saffran, Aslin & Newport, 1996), and phonotactic constraints on phoneme combinations in words (e.g. Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud & Jusczyk, 1993; Mattys & Jusczyk, 2001). 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; Cairns, Shillcock, Chater & Levy, 1997; Hockema, 2006; see Brent, 1999, for a review), especially when combined with infor-mation 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 deter-mining 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. Cartwright & Brent, 1997; Mintz, Newport & Bever, 2002; Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007; Redington, Chater & Finch, 1998; for a review, see Redington & Chater, 1998), frequent word frames (e.g. *I X it*; Mintz, 2003; Monaghan & Christiansen, 2004; see also Chemla, Cristophe, Bernal & Mintz, this issue), and phonological cues (e.g. Cassidy & Kelly, 1991, 2001; Christiansen & Monaghan, 2006; Durieux & Gillis, 2001; Monaghan *et al.*, 2005, 2007; Shi, Morgan & Allopenna, 1998; see Kelly, 1992; Monaghan & Christiansen, 2008, for reviews). Indeed, merely pay-ing attention to the first and last phoneme of a word has been shown to be useful for predicting lexical categories

across various languages, such as English, Dutch, French and Japanese (Onnis & Christiansen, 2008).

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 present an integrated model to test this hypothesis through a two-stage corpus analysis. In Stage 1, we first use information about phoneme distributions to segment words out of a large corpus of phonologically transcribed child-directed speech. The output – including errors – from the first stage then provides the input for Stage 2, in which phoneme-distributional information is used to predict the lexical category (noun, verb, or other) of the words segmented in Stage 1. Finally, we discuss the limitations of the current model and how infants may utilize the information that our model shows is inherent in phoneme distributions.

Our results provide the first demonstration of an integrated model in which it is possible to get from un-segmented speech to lexical categories using only information about the distribution of phonemes in the input. Thus, as a core computational principle, we suggest that the child may be using the same source of information (e.g. phoneme distributions) to learn about different aspects of linguistic structure (e.g. word segmentation and lexical-category discovery).

Stage 1: Discovering words

Infants are proficient statistical learners, sensitive to sequential sound probabilities in artificial (e.g. Aslin *et al.*, 1998; Saffran *et al.*, 1996) and natural (e.g. Friederici & Wessels, 1993; Jusczyk *et al.*, 1993; Mattys & Jusczyk, 2001) language. Such statistical learning abilities would be most useful for word segmentation if natural speech was made up primarily 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 /ɪŋ/ (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 adult utterances spoken in the presence of children from all the English corpora

in the CHILDES database (MacWhinney, 2000). 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 pronunciation. Another 9,117 non-standard 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, eliminating 124,189 utterances containing 537,083 words. The resulting corpus contained 4,933,794 words distributed over 1,369,574 utterances.

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. 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 corpus and used this information in a simple procedure 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 a boundary occurring for a phoneme transition pair was greater than .5.

Results and discussion

Of the 3,025 possible phoneme transition pairs, 1,119 (37%) never occurred in the corpus. Figure 1 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. Each phoneme transition pair was weighted by its frequency of occurrence across the corpus in order to approximate the distribution of the phoneme-transition-pair tokens that a child might actually come across in the input. The bar height indicates the percentage of phoneme transition pairs with a given probability of having a word boundary between them. There are 50 bins in the histogram, so each bin accounts for a probability range of .02. Figure 1 illustrates that the distribution of used phoneme transition pairs was strongly bimodal. Most phoneme transitions either were associated only with a word boundary or occurred only within a word, but not both. Indeed, the left- and right-most bins account for 56% of the transitions heard in everyday speech. The fact that the left-most bin is 3.2 times as high as the right-most bin is because that only 1 in every 3.6 phoneme transitions involved a word boundary.[1]

---

[1] Words were, on average, 3.0 phonemes long (*SD* = 1.2), but not all words were preceded or followed by a boundary transition (because some occurred on utterance breaks).
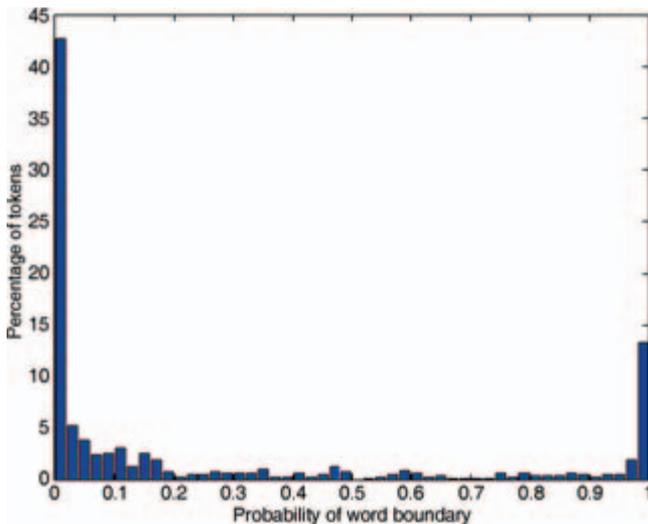
Figure 1 *The distribution of phoneme transition pairs given the probability of encountering a word boundary between the two phonemes in the corpus of child-directed speech. A probability of 1 indicates that the two phonemes never occur together as a pair inside a word but always straddle a word boundary, whereas a probability of 0 implies that the phoneme pair always occurs inside a word and is never separated by a word boundary.*

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, 3,152,842 word boundaries were inserted within the 1,369,574 utterances, yielding 4,522,416 potential words. To determine how well complete words could be discovered using this simple model, we used a conservative measure of word segmentation in which a word is considered to be correctly segmented only 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 could be useful for learning morphological structure). Using this measure, the model discovered 3,413,064 actual words.

We used two measures – accuracy and completeness – to gauge word segmentation reliability. Accuracy is computed as the number of correctly segmented words (hits) in proportion to all predicted lexical candidates, both correct word candidates (hits) and incorrectly segmented candidates (false alarms). Completeness is calculated as the number of correctly segmented words (hits) in proportion to the total number of words in the corpus, that is, the correct words (hits) and the words that the model failed to segment out (misses). Thus, accuracy

Table 1 *The type and token distributions of the lexical candidates, words, fragments and combo-words from Stage 1*

	Lexical candidates	Words	Fragments	Combo-words
Types	117,472	9,263	31,627	76,582
Tokens	4,522,416	3,413,064	614,931	494,421

provides an estimation of the percentage of the segmented lexical candidates that were actual words, whereas completeness indicates the percentage of words that the model actually found out of all the words in the corpus.

Using this conservative measure we computed segmentation accuracy and completeness for segmented words. Overall, the model identified 69.2% of the words in our corpus (completeness), and 75.5% of the lexical candidates it identified were valid words (accuracy). The missegmented words were classified into word fragments (where a boundary had erroneously been inserted within a word; for example, 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; for example, the boundary between *come* and *on* was missed, yielding a single lexical candidate, *comeon*). There were 614,931 fragments and 494,421 combo-words, of which 31,627 and 76,582 were unique, respectively (see Table 1 for additional information). Interestingly, the top-three most frequent fragments were /d/, /s/ and /t/ (29,142, 25,759 and 16,269 occurrences respectively), all of which are very common morphological suffixes. Meanwhile, the top-five most frequent combo-words were *that's_a* (6,210), *this_is* (6,179), *look_at* (4,667), *I_know* (3,865), and *it's_a* (3,558), which arguably all represent atomic, deictic concepts or speech acts. These intriguing results invite more exploration into possible interactions among the processes of learning to segment speech, learning morphology and word learning. As a first step, we treat some of the combo-words in Stage 2 as actual words when analysing the usefulness of phoneme distribution information for discovering lexical categories.

Stage 2: Discovering lexical categories

The results from Stage 1 replicate what was found in previous work (Hockema, 2006), this time using a larger inventory 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 that a simple model that attends only to these can do well enough to bootstrap the word segmentation process. However, performance was not perfect, as evidenced by the considerable number of word fragments and combo-words. The question thus remains whether the imperfect output of our segmentation procedure can be used in Stage 2 to learn about higher-level properties of language.

Experimental evidence suggests that both children (Slobin, 1973) and adults (Gupta, 2005) are particularly sensitive to the beginnings and endings of words. From previous work, we know that beginning and ending phonemes can be used cross-linguistically to discriminate the lexical categories of words from pre-segmented input (Onnis & Christiansen, 2008). In Stage 2, we explore whether such word-edge cues can still lead to reliable lexical classification when applied as part of an integrated model to the noisy output of our word segmentation procedure. We hypothesized that missegmented phoneme strings would not cause too much difficulty because such phoneme sequences are more likely to have less coherent combinations of word-edge cues compared with lexical categories such as nouns and verbs.

Method

Corpus preparation

The imperfectly segmented corpus produced by the segmentation procedure in Stage 1 was used for the word-edge analyses. The lexical category for each word was obtained from CELEX (Baayen *et al.*, 1995). Several words 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. Moreover, research in computational linguistics has shown that a procedure that simply picks the most frequent syntactic category for each word in a corpus is able to tag about 90% of the words correctly (Charniak, Hendrickson, Jacobson & Perkowitz, 1993). We therefore assigned dual-category words their most frequent lexical category from CELEX. In total, there were 117,472 distinct lexical-candidate types, of which 9,263 were words, with the reminder being combo-words and fragments (see Table 1). Among words, 4,783 were nouns (447,658 tokens) and 1,727 were verbs (667,401 tokens).

Cue derivation

Given that the CELEX DISC encoding used in Stage 1 employed 55 phonemes, we represented each lexical item as a vector containing 110 (55 beginning + 55 ending) bits. The bits in the vector that corresponded to beginning and ending phonemes were assigned 1; all others were assigned 0. Thus, the encoding of each word in the corpus consisted of a 110-bit vector, with most bits having the value 0 and two having a value of 1, along with the word's associated lexical category.

Analyses

We considered the 5,000 most frequent lexical candidates from the segmented output of Stage 1. There were 2,117 unique words, whose summed frequencies accounted for 98.7% of word tokens in the whole corpus; there were 1,620 unique combo-words, which accounted for 61.8% of combo-word tokens in the whole corpus; and 1,263 unique fragments, which accounted for 86% of fragment tokens in the whole corpus. In total, the 5,000 most frequent lexical candidates from the segmented corpus accounted for 92.9% of the corpus.

Children's early syntactic development is perhaps best characterized as involving fragmentary and coarse-grained knowledge of linguistic regularities and constraints (e.g. Tomasello, 2003). Thus, 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, 1973). In addition, the adult-like lexical categories likely to emerge first will be the ones most relevant to children's early syntactic productions. For example, noun and verb categories are learned earlier than mappings to conjunctions and prepositions (Gentner, 1982). Our analyses therefore focus on three broad lexical categories: NOUNS, VERBS and OTHER, plausibly reflecting early stages of lexical acquisition, in which OTHER forms an amalgamated 'super-category' incorporating all lexical items that are not nouns or verbs. Given that many combo-words correspond to word combinations that a child may plausibly treat as a single lexical unit (e.g. *look_at*, *show_me*, *want_to*), we treated all combo-words beginning or ending with a verb as belonging to the category of VERBS for the purpose of classification, and similarly combo-words beginning or ending with a noun were treated as being NOUNS. Combo-words that included both a noun and a verb were designated as belonging to OTHER. Words that had a lexical category other than noun or verb were assigned to OTHER, along with fragments and combo-words that did not include nouns or verbs.

To assess the extent to which word-edge cues can be used reliably for this three-way lexical-category classification, we performed a linear discriminant analysis dividing words into NOUNS, VERBS or OTHER. Discriminant analyses provide a supervised classification of items into categories based on a set of predictor variables. The chosen classification maximizes the correct classification of all members of the predicted groups. In essence, a discriminant analysis inserts a hyper-plane 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 hyper-plane. To assess effectiveness, we used a 'leave-one-out cross-validation' method, which provides a conservative measure of classification performance, and works by predicting the classification of words that are not used in positioning the hyper-plane. This means that the hyper-plane is constructed on the basis of the information from all words except one, and then used to determine the classification of the omitted word. This is repeated for every word, and the overall classification performance can then be determined.

Previous analyses of the potential usefulness of phonological cues for lexical-category discovery have tended to focus on analyses of word types (e.g. Cassidy & Kelly, 1991; Durieux & Gillis, 2001; Monaghan *et al.*, 2005). However, children are not exposed to word types but have to learn about their native language from tokens that occur with varying frequency. For example, in our corpus of child-directed speech the most frequent word, *you*, occurs 234,744 times, whereas *acrobats* occurs only once. As demonstrated in the Appendix, log frequency provides a reasonable approximation of the word token statistics to which a child is likely to be sensitive. In our discriminant analyses, we therefore weighted each word-edge vector by its log frequency.

To establish chance-level performance, a baseline condition was 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 contained dummy variables, 1, 2, or 3, for the three lexical categories (dependent variable). This last column contained 1,549 values of 1 (NOUNS), 1,018 values of 2 (VERBS) and 2,433 values of 3 (OTHER). We randomly scrambled the order of the entries in the lexical-category column while leaving the other 110 columns (the word-edge predictors) unchanged. Such scrambling 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 scramblings and tested the ability of the 110 word-edge cues to predict the scrambled lexical categories in 100 separate discriminant analyses. In this way, it was possible to test whether the actual distribution of beginning and ending phonemes within nouns and verbs in the experimental condition provided for better lexical-category classification than did the randomly scrambled baseline condition.

Results and discussion

Using the word-edge cues, 53% of the cross-validated lexical tokens were classified correctly.[2] This result compared well with 33% overall baseline classification. The results for each lexical category are illustrated in Figure 2 (left). For nouns, word-edge cues yielded 63% correct classification, compared with 34% for the baseline. Verb classification was 55%, compared with a baseline of 34%, and other classification was 48%, compared with 33% for the baseline.

These results provide an estimate of the completeness of the classification procedure; that is, of how many of the words belonging to a given category were classified
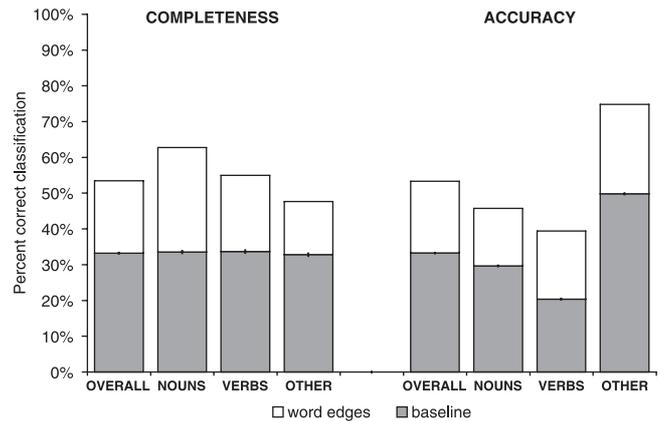

Figure 2 *The completeness (left) and accuracy (right) of classification into lexical categories of the top-5000 lexical candidates from the segmentation procedure using the first and last phoneme in each lexical candidate (white bars) compared with baseline classifications (grey bars – error bars indicate standard error of the mean).*

correctly as being in that category. We further measured the accuracy of the classifications for each of the three categories; that is, how many of the lexical candidates classified as being in a given category actually belonged to that category. The lexical-classification accuracy is reported in Figure 2 (right), and includes a comparison with the baseline condition. These results show that more than 50% of both nouns and verbs can be classified correctly using the word-edge cues alone, and that such classifications are reasonably accurate: approximately 40% of words classified as nouns and verbs were classified correctly as such. For all classifications, word-edge cues provided for significantly better classification than the baseline (*p*'s < .001; see also Figure 2). This suggests that nouns and verbs utilize separate and fairly coherent clusters of word-edge cues, indicating that word-edge cues are useful for the discovery of nouns and verbs even when provided with suboptimally segmented input. Moreover, the results compare well with those of Onnis and Christiansen (2008), who used a perfectly segmented corpus as input.

General discussion

In this paper, we have presented a two-stage integrated model of the usefulness of information about phoneme distributions for word segmentation and lexical-category discovery. To our knowledge, this is the first time that a combined approach has demonstrated how a single probabilistic cue (i.e. phoneme distributions) can be used to get from unsegmented speech to broad lexical categories. Crucially, both stages utilized very simple computational principles to take advantage of the phoneme distributional cues, requiring sensitivity only to phoneme transitions and word edges. Importantly, these two sensitivities are in place in infants (transitional probabilities: Aslin *et al.*,

---

[2] A three-way discriminant analysis inserts two distinct hyper-planes to divide up the word space, each described by a separate function. In the current analyses, Function 1 explained 55.4% of the variance, Wilk's lambda = .719, χ^2 = 8295, *p* < .001; Function 2 explained 44.6% of the variance, Wilk's lambda = .862, χ^2 = 3732 *p* < .001.

1998; Saffran *et al.*, 1996) and young children (word edges: Slobin, 1973). The integrated two-stage model also demonstrates that segmentation does not have to be perfect for it to be useful for learning other aspects of language. Thus, we propose that a core computational principle in language acquisition is the use of the same source of probabilistic information to learn about different aspects of language structure; here, the use of phoneme distributions to inform word segmentation and lexical-category discovery.

A limitation of the current work is that we have not presented a complete developmental model showing how information about phoneme distributions may be utilized to get the child from unsegmented speech to lexical categories; rather, we have presented analyses of the potential usefulness of such information. Thus, it is an open question as to how infants might make use of the phoneme-transition-pair regularities demonstrated in our Stage 1 analysis. One possibility is 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 biphone transition probabilities and the actual probability of finding a word boundary across phoneme pairs. As expected, this was significantly negative ($r = -.25$, $p < .00001$), but perhaps not strong enough to completely 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 lexical 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 at 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.

A related issue arises with regard to the use of supervised discriminant analyses in our Stage 2 model of lexical-category discovery. Nonetheless, despite its seeming statistical complexity, a linear discriminant analysis is a simple procedure that can be approximated by simple learning devices such as two-layer 'perceptron' neural networks (Murtagh, 1992). Onnis and Christiansen (2008) therefore trained perceptrons to predict the lexical category (NOUNS, VERBS and OTHER) given word-edge vectors as input for the top-500 most frequent words. The networks were then tested on their ability to generalize from these 500 words to a new set of 4,230 words, and demonstrated a reasonably high level of performance (43.5% overall correct classification). The underlying theoretical idea is that the child may use a variety of cues to learn an initial set of words, including approximations of how they may be used syntactically, and would then be able to use word-edge cues to help determine the lexical category of subsequently encountered words. This perspective is consistent with data indicating that 4-year-olds are able to use phonological information to help them learn novel nouns and verbs (Cassidy & Kelly, 2001).

More generally, evidence exists that infants can utilize the kind of phonological distributional information revealed by our analyses to learn about language. First, infants are able to use both transitional probabilities of syllables (Aslin *et al.*, 1998; Saffran *et al.*, 1996) and phonemes (Newport, Weiss, Wonnacott & Aslin, 2004) to do word segmentation. Second, 12-month-olds are capable of using the same source of information (syllable distributions) both to segment an artificial language and to learn about the possible ordering of words (Saffran & Wilson, 2003). Thus, infants are likely to take advantage of the probabilistic information inherent in phoneme distributions to help them get from unsegmented speech to broad lexical categories.

In future work, we plan to integrate segmentation and lexical-category discovery more closely in a developmental model. Children probably start working out how to use word forms while they are still honing their segmentation skills. A model that worked in a less serial fashion than the current two-stage one would perhaps be better at capturing developmental trends in both segmentation and lexical-category discovery. Such a model might also be useful for studying the effects of more coarse-grained probabilistic representations instead of the current categorical phonemic input. Children are sharpening their phoneme categories as they learn how to segment speech, and this may influence lexical-category discovery in important ways, perhaps resulting in specific developmental patterns of errors that can be the subject of further empirical studies.

Our results 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. Rather, young learners are likely to rely on many additional sources of probabilistic information (e.g. social, semantic, prosodic, word-distributional) to discover different aspects of the structure of their native language (e.g. Christiansen & Dale, 2001; Gleitman & Wanner, 1982; and contributions in Morgan & Demuth, 1996; Weissenborn & Höhle, 2001). 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 of syntactic development (e.g. word-distributional information: Monaghan *et al.*, 2005, 2007; Reali, Christiansen & Monaghan, 2003). This suggests that the phoneme distributional cues explored here could in future work be incorporated into a more comprehensive computational account of language development through multiple-cue integration.

Appendix

The use of log frequency is common in connectionist modelling (e.g. Harm & Seidenberg, 1999; Plaut, McClelland, Seidenberg & Patterson, 1996; Seidenberg & McClelland, 1989), and allows learning to be sensitive to token frequency information while preventing low-frequency tokens from being swamped by high-frequency items. Importantly, log frequency of word forms has also been shown to be an excellent predictor of the age at which words are acquired (e.g. Wijnen, Kempen & Gillis 2001; Zevin & Seidenberg, 2004). To establish whether raw frequency or log frequency best predicted age of acquisition for the words in our corpus of child-directed speech, we carried out regression analyses involving three different sets of age-of-acquisition norms: Zevin and Seidenberg (2004), the Bristol Norms (Stadthagen-Gonzalez & Davis, 2006), and Gilhooly and Logie (1980). As can be seen from Table A, word log frequency accounts for between 32.7 and 44.1% of the variance in age of acquisition, nearly 10 times more than the 3.0–4.9% obtained for raw word frequency. Thus, log frequency provides a reasonable approximation of the word token statistics to which a child is likely to be sensitive.

Table A *Variance in age of acquisition accounted for by raw and log frequency of word occurrence in the corpus of child-directed speech*

Predictor	Variance	Beta weight	*t*	*p* <
Zevin & Seidenberg Norms (*N* = 1,199)				
Raw frequency	.039	−.198	6.99	.0001
Log frequency	.327	−.572	24.15	.0001
Bristol Norms (*N* = 752)				
Raw frequency	.030	−.173	4.83	.0001
Log frequency	.345	−.587	19.89	.0001
Gilhooly & Logie Norms (*N* = 789)				
Raw frequency	.049	−.222	6.39	.0001
Log frequency	.441	−.664	24.94	.0001

Note: We thank Jason Zevin for suggesting these analyses.

Acknowledgements

References

Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, **9**, 321–324.

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. (1999). Speech segmentation and word discovery: a computational perspective. *Trends in Cognitive Science*, **3**, 294–301.

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

Cairns, P., Shillcock, R.C., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up approach to speech segmentation. *Cognitive Psychology*, **33**, 111–153.

Cartwright, T.A., & Brent, M.R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, **63**, 121–170.

Cassidy, K.W., & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, **30**, 348–369.

Cassidy, K.W., & Kelly, M.H. (2001). Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic Bulletin and Review*, **8**, 519–523.

Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993). Equations for part-of-speech tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 784–789.

Christiansen, M.H., & Dale, R. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 220–225). Mahwah, NJ: Lawrence Erlbaum Associates.

Christiansen, M.H., & Monaghan, P. (2006). Discovering verbs through multiple-cue integration. In R.M. Golinkoff & K. Hirsh-Pasek (Eds.), *Action meets word: How children learn verbs* (pp. 88–107). Oxford: Oxford University Press.

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.

Durieux, G., & Gillis, S. (2001). Predicting grammatical classes from phonological cues: an empirical test. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1, pp. 189–229). Amsterdam: John Benjamins.

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development* (Vol. 2., pp. 301–344). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gilhooly, K.J., & Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behavior Research Methods and Instrumentation*, **12**, 395–427.

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* (pp. 3–48). Cambridge: Cambridge University Press.

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

Harm, M., & Seidenberg, M.S. (1999). Reading acquisition, phonology, and dyslexia: insights from a connectionist model. *Psychological Review*, **106**, 491–528.

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., Cutler, A., & Redanz, N.J. (1993) Infants' preference for the predominant stress patterns of English words. *Child Development*, **64**, 675–687.

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

Jusczyk, P.W., Houston, D.M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, **39**, 159–207.

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

Kelly, M.H. (1996). The role of phonology in grammatical category assignment. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 249–262). Mahwah, NJ: Lawrence Erlbaum Associates.

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 edn). Mahwah, NJ: Lawrence Erlbaum Associates.

Mattys, S.L., & Jusczyk, P.W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, **78**, 91–121.

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

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., & Christiansen, M.H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (TILAR Series) (pp. 139–163). Amsterdam: John Benjamins.

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

Monaghan, P., Christiansen, M.H., & Chater, N. (2007). The Phonological–Distributional Coherence Hypothesis: cross–linguistic evidence in language acquisition. *Cognitive Psychology*, **55**, 259–305.

Morgan, J.L., & Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Murtagh, F. (1992). The multilayer perceptron for discriminant analysis: two examples. In M. Schader (Ed.), *Analyzing and modeling data and knowledge* (pp. 305–314). Berlin: Springer-Verlag.

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: Lawrence Erlbaum Associates.

Newport, E.L., Weiss, D.J., Wonnacott, E., & Aslin, R.N. (2004). *Statistical learning in speech: Syllables or segments?* Paper presented at the Boston University Conference on Language Development.

Onnis, L., & Christiansen, M.H. (2008). Lexical categories at the edge of the word. *Cognitive Science*, **32**, 184–221.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains *Psychological Review*, **103**, 56–115.

Reali, F., Christiansen, M.H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: scaling up the connectionist approach to multiple-cue integration. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 970–975). Mahwah, NJ: Lawrence Erlbaum.

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: a distributional perspective. *Language and Cognitive Processes*, **13**, 129–191.

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

Saffran, J.R., & Wilson, D.P. (2003). From syllables to syntax: multilevel statistical learning by 12-month-old infants. *Infancy*, **4**, 273–284.

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

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review*, **96**, 523–568.

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.

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.

Stadthagen-Gonzalez, H., & Davis, C.J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, **38**, 598–605.

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

Weissenborn, J., & Höhle, B. (2001). *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition.* Amsterdam: John Benjamins.

Wijnen, F., Kempen, M., & Gillis, S. (2001). Root infinitives in Dutch early child language: an effect of input? *Journal of Child Language*, **28**, 629–660.

Zevin, J.D., & Seidenberg, M.S. (2004). Age of acquisition effects in reading aloud: tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, **32**, 31–38.