Statistical Speech Segmentation and Word Learning in Parallel

Daniel Yurovsky, Chen Yu, and Linda B. Smith
Indiana University

Children's language learning is remarkable not only for the speed and apparent ease with which it happens, but also for the complexity of the input on which learners must operate. Consider the case of learning concrete nouns, which make up the preponderance of early vocabularies (Macnamara, 1972). Children must solve at least two significant problems from scratch: **speech segmentation** – identifying the individual words in continuous speech, and **word-object mapping** – determining which of the segmented words should be mapped to which candidate referent in the environment.

In both problems, learners must contend with significant ambiguity of the information available in natural learning environments. For instance, the majority of everyday spoken language, even child-directed speech, is comprised of multi-word utterances (Brent & Siskind, 2001). Because spoken utterances lack an acoustic analog of white spaces to mark word boundaries, the segmentation task is nontrivial. Similarly, a single word-learning context often contains multiple candidate words and multiple object referents. In order to learn the meanings of these words, young word learners must determine which of the word-object pairs are correct mappings and which are spurious.

One way that human learners could contend with the ambiguity of a *single* instance is through accumulation of statistical evidence *across* instances. Saffran, Aslin, and Newport (1996) showed evidence for such a mechanism in 8-month-old infants, who are sensitive to transitional probabilities between syllables in continual speech. This sensitivity supports statistical speech segmentation. Recently, Smith and Yu (2008) demonstrated the availability of a statistical solution to the word-object mapping problem in 12 and 14-month-old infants (see also Gleitman, 1990). Infants at this age show sensitivity to co-occurrence frequencies between words and objects, a sensitivity which may underlie early word learning.

Although such mechanisms are known to be available to young infants, the exact nature of their interaction remains an open question. Graf Estes, Evans, Alibali, and Saffran (2007) demonstrated a potential serial link by exposing 17-month-old infants to a speech segmentation task followed by word-

* This research was supported by a National Science Foundation Graduate Research Fellowship to the first author. The authors wish to thank Damian Fricker for collecting much of the data, and Morten Christiansen, Jenny Saffran, Michael Frank, and Roberta Golinkoff for their insightful feedback.

learning task. They showed that word-object mappings in the second task were learned only when candidate labels were consistent with transitional probability statistics from the segmentation task. But, a more efficient learning mechanism might instead accumulate information for both tasks in parallel, allowing the processes to mutually constrain each other. Of course, such efficiency gains are contingent on the ability of human learners to cope with a potentially prohibitive cognitive load. Frank, Mansingkha, Gibson, and Tenenbaum (2006) used a small artificial language to probe for a parallel solution to speech segmentation and word-object mapping in adults, finding success only in very limited conditions. Given the simplicity of their language relative to real language input, these results could be interpreted as support for an impossible cognitive load hypothesis. Here, we explore the opposite possibility.

It is possible, instead, that the very simplifications introduced in the creation of Frank et al.'s (2006) artificial language rendered the task impossible. The language which serves as input to real learners engaged in speech segmentation and word-object mapping may contain additional regularities that support parallel solution of these problems. We address this question in two steps. First, we analyze a corpus of naturalistic child-directed speech for the presence of potential regularities. Second, we encode these regularities in a new artificial language to which adult learners are exposed. By manipulating the regularities present in the artificial language, we can determine which natural regularities are necessary for parallel segmentation and word learning.

1. Corpus Analysis

In statistical segmentation studies, continuous speech streams are typically composed so that each word is equally likely to follow each other word. This minimizes the transitional probability of syllables across words, and thus supports segmentation via transitional probability. However, this construction does not conform to natural speech, which contains a high degree of dependence between words. Such dependence makes general segmentation more difficult, but may make segmentation of object labels easier. If object words were reliably flanked by a consistent set of phrases, learners could infer the boundaries of new object words from the boundaries of flanking phrases without correct internal segmentation of the phrases themselves. The importance of naming phrases is highlighted by Fernald and Hurtado's (2006) study of infant orientation responses to labels. The authors found faster orientation to a label's target when the label was presented in a carrier phrase (e.g. "look at the __") than when the same referent was presented alone.

In our corpus analyses, we look for consistent regularities, or frames, in naming phrases in naturalistic child-directed speech. Such regularities may hold the key to understanding if and when speech segmentation and word learning proceed in parallel (for frequent frames in grammatical category induction, see Mintz, 2003).

1.1 Method

Data. The corpus consisted of transcripts of child-directed speech from 17 parent-child pairs engaged in three sessions of free-play (Yu, Smith, & Pereira, 2008). In each session, parents and children - seated on opposite sides of a table - were given three objects with which to play. Parents were taught labels (e.g. ‘dax’, ‘toma’) for each of the objects, and asked to use them whenever they wished to refer to the objects. No other instructions were given. The parent-child dyads played with three novel objects in each of the minute-long sessions, seeing a total of 9 objects over the course of the experiment.

Analysis. Audio recordings of each parent’s speech were automatically partitioned into individual utterances using speech silence, and these utterances were then transcribed by human coders. In total, the corpus contained 3165 individual utterances. Any utterance containing one of the novel labels was considered a naming event, resulting in 1624 such events. Approximately 20% (672) of the utterances consisted of a bare label (somewhat higher than reported by Brent & Siskind, 2001). These were excluded from subsequent analysis as they contained no local contextual linguistic information.

The statistical patterns in the remaining naming events were extracted using a six-word window (or frame) made up of the three words on either side of an object label. If fewer than three words preceded or followed a label in any given utterance, blanks were inserted to fill out the window (e.g. “__ __ the toma is blue __”). Next, individual object labels were replaced with a common token (OBJ), and the frequency of each resulting multi-word frame was computed.

1.2 Results

Naming phrases used by parents showed a high degree of consistency, such that the 21 most frequent frames accounted for 52.4% of all naming events containing a single object label. These 21 frequent frames demonstrated two significant regularities (Table 1). First, object labels occurred consistently in the final position of each phrase (see also Aslin, Woodward, LaMendola, & Bever, 1996). Second, the words which immediately preceded each label were highly predictable, coming from a very small set made up mostly of articles (Shafer, Schucard, Schucard, & Gerken, 1998).

Taken together, these results demonstrate that naming events in parent-child free play contain consistent, reliable structure. Child-directed speech thus includes local contexts (frames) which are informative about the probable positions of object labels. More specifically, this contextual information may allow learners to become aware of an impending naming event before the label is spoken, and thus to focus their segmentation effort accordingly.

Table 1: The 21 most frequent naming frames. Two regularities are apparent. First, object labels occur reliably in final frame position. Second, labels are reliably preceded by a small set of onset cues (a, the, and, say).

Phrase	# in corp.	Phrase	# in corp.
…the OBJ…	60	…that is the OBJ…	17
…that is a OBJ…	45	…look at the OBJ…	17
…and the OBJ…	41	…I have the OBJ…	14
…a OBJ…	39	…you want the OBJ…	11
…it is a OBJ…	36	…color is the OBJ…	11
…this is a OBJ…	34	…is that the OBJ…	11
…and a OBJ…	31	…there is the OBJ…	10
…can you say OBJ…	28	…you put the OBJ…	10
…here is the OBJ…	25	…to put the OBJ…	9
…and OBJ…	23	…one is the OBJ…	9
…where is the OBJ…	18		**52.4%**

1.3 Discussion

Analyzing the structure of natural naming events is an important step towards modeling children’s word-learning. Because consistency in naming event structure constrains the space of potential solutions, the same mechanism which fails in an unstructured environment may successfully extract words from fluent speech and map them to their referent objects when additional regularity is present. Specifically, the information encoded in frequent naming frames may allow young learners to identify the utterances most likely to be naming events, and to spot the label within each frame without completely segmenting the other words. Encoding these regularities into an artificial language, we test this hypothesis empirically.

2. Artificial Language Experiment

To study joint speech segmentation and word-object mapping, we exposed adult participants to a series of individually ambiguous training trials based on the cross-situational learning paradigm (Yu & Smith, 2007). On each trial, adult learners saw two objects and heard two phrases of continuous speech from an artificial language. In order to learn word-object mappings, they had to determine which phrase referred to which object, where the word boundaries were, and finally which words were object labels and which word were function words. Crucially, local contextual patterns found in the child-directed speech corpus were encoded into the artificial language presented to participants.

Table 2: The 2x2 design of the artificial language experiment. Phrasal position of the object label varies on the horizontal, presence of the onset cue varies on the vertical

	Final Position	**Middle Position**
Preceding Cue	*Full Language* "Look at *the* OBJ"	*Onset-Only Language* "At *the* OBJ look
No Cue	*Position-Only Language* "*The* look at OBJ"	*Control Language* *"The* look OBJ at"

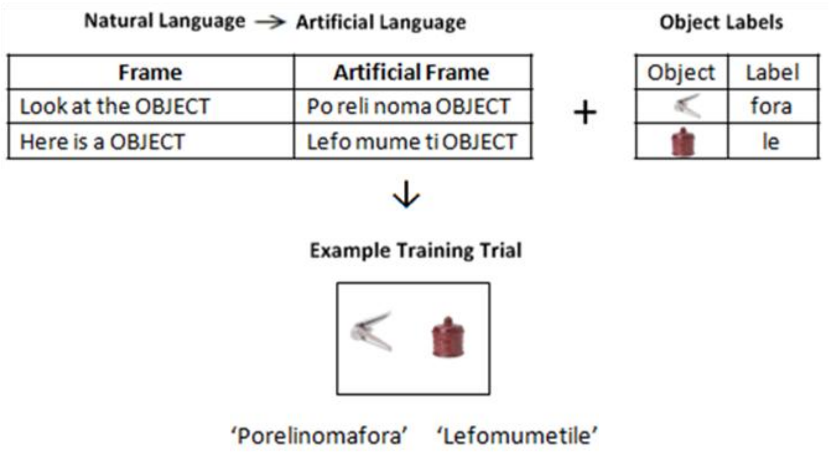
Using a 2x2 design, we tested the impact each of the two regularities found in the corpus on simultaneous speech segmentation and word learning. We thus created four artificial languages. Phrases from the *full* language encoded the 21 most frequent naming frames directly, such that object labels always occurred in final position and were always preceded by a member of the set of onset cues (see Table 1). The *position-only* language encoded the first regularity but not the second, the *onset-only* language encoded the second regularity but not the first, and the *control* language encoded neither. Importantly, the languages were identical in all aspects of their construction (described below) except for the position of words within naming phrases.

2.1 Method

Participants. 92 undergraduate students from Indiana University participated in exchange for course credit. These participants were divided into four approximately equal groups, each exposed to one of the artificial languages.

Materials. Stimuli for the experiment consisted of 18 unique objects (from Yu & Smith, 2007), and 38 unique words. Eighteen of these words acted as labels for the novel objects, and the other 20 were mapped onto the words contained in the 21 most frequent frames found in the corpus analysis. Half of the words of each type were one syllable (CV) in length, and the other half were two syllables (CVCV) long, necessitating the construction of 57 unique syllables. These syllables were created by sampling 57 of the 60 possible combinations of 12 constants and 5 vowels. Syllables were assigned to words randomly, so that nothing about a word's phonetic properties could be used to distinguish object labels from other words in the language.

Words were then concatenated together without intervening pauses to create artificial language equivalents of each of the 21 frequent phrases in the corpus. Participants were exposed to synthesized versions of these phrases constructed using MBROLA (Dutoit, Pagel, Pierret, Bataille, & van der Wrecken, 1996). No prosodic or phonetic properties could be used to determine word boundaries, forcing participants to rely on statistical information.



Frame	Artificial Frame
Look at the OBJECT	Po reli noma OBJECT
Here is a OBJECT	Lefo mume ti OBJECT

Object	Label
	fora
	le

Figure 1: An example training trial from the *full language* condition. Trials were constructed by mapping naming event patterns from the child-directed speech into the artificial language.

Procedure. Participants were told that they would be exposed to scenes consisting of two novel objects, and a phrase referring to each of them. Each phrase would contain exactly one word labeling an on-screen object, along with several function words corresponding to the grammar of the artificial language. Participants had to determine which phrase referred to which object, how the phrases they heard should be segmented into words, and which of these words referred to which of the objects. Next, participants observed an example trial using English words and familiar objects to demonstrate the task. Importantly, the example contained both an object-final phrase ("observe the tractor") and an object-medial phrase ("and the dog over there") to prevent the participants from expecting any particular positional regularity.

After the example, participants observed 108 training trials, each containing 2 objects and 2 spoken artificial language phrases (Figure 1). Trials began with two seconds of silence, each phrase was approximately two seconds in length, and 3 seconds of silence succeeded each phrase, resulting in trials approximately 12 seconds long. Each object appeared 12 times, and each naming frame occurred a number of times proportional to its appearance in the child-directed speech corpus. The entire training set ran just over 20 minutes.

After training, participants were tested first for speech segmentation and then word-object mapping. On each segmentation test trial, a participant heard 2 two-syllable words: a word from the experiment and a foil created by concatenating the first syllable of one word and the second syllable of another (following Fiser & Aslin, 2002). They were asked to indicate which of the words was more likely to be part of the artificial language. Six correct object labels were tested against 6 object foils, and 6 correct frame words were tested against 6 frame foils, resulting in 72 total segmentation trials. Each possible word occurred an equal number of times in testing, preventing participants from using test frequency as a cue to correctness.

Subsequently, participants were tested on their knowledge of word-object mappings. On each test trial, participants heard one of the object labels and were asked to select its correct referent from a set of four alternatives. All of the labels were tested once in random order.

To assess the independent and joint contribution of both the final position and onset cue regularities, one group of participants was exposed to each of the four possible presence/absence combinations of these cues. Materials and procedure were identical for each of the groups except for the order of words within each artificial language naming phrase.

2.2 Results

Full Language. Twenty-four participants were exposed to the *full* language, in which object labels always occurred in final phrasal position, and were also preceded by onset cues. When tested on segmentation performance, participants performed above chance for object-labels ($t_o = 2.69$, $p < .05$), but not frame words ($t_f = .51$, *n.s.*). This suggests that participants discovered the position regularity and focused their segmentation effort appropriately. Participants also successfully learned word-object mappings ($t = 4.98$, $p < .001$). What's more, across subjects, segmentation accuracy for a given object label was significantly correlated with mapping accuracy for that label ($r = .302$, $p < .001$). Finally, object-label segmentation and frame word segmentation were uncorrelated by participant ($r = -.224$, *n.s.*), suggesting that segmenting frame words did not help participants segment object labels in this task.

Position-Only Language. Twenty-two participants were exposed to the *position-only* language, in which object labels always occurred in final phrasal position but were not preceded by consistent onset cues. In this condition, participants successfully segmented object-labels ($t_o = 2.13$, $p < .05$), and were close to segmenting frame words at above chance levels ($t_f = 1.86$, $p = .07$). As in the *full* language, word-object mapping accuracy was also significantly above chance ($t = 4.12$, $p < .001$). Again, segmentation accuracy for an individual label was correlated with mapping accuracy for that label ($r = .262$, $p < .01$). This time, however, object-label and frame word segmentation were significantly correlated ($r = .476$, $p < .01$). Thus, when onset cues were absent, segmentation of frame words helped participants to segment object labels.

Onset-Only Language. Twenty-four participants were exposed to the *onset-only* language. In this language, object labels occurred in the middle of artificial language naming frames, but the labels were preceded by onset cues. In contrast to the previous conditions, participants now successfully segmented frame words ($t_f = 5.39$, $p < .001$) – at levels unparalleled in the previous conditions – but did not successfully segment object labels ($t_o = 1.34$, *n.s.*). Nonetheless, participants successfully learned word-object mappings ($t = 2.99$, $p < .01$), although accuracy was depressed relative to the previous conditions. However, segmentation and word-object mapping accuracy were no longer correlated (*r* = .029, *n.s.*), and neither were object-label segmentation and frame word segmentation (*r* = .185, *n.s.*). Thus, although participants learned word-object mappings in this condition, they did so through a qualitatively different strategy.

Control Language. Twenty-four participants were exposed to the *control* language which did not contain either regularity found in the corpus analysis. In this condition, participants did not successfully segment either object labels (t_o = 1.26, *n.s.*), or frame words (t_o = 1.93, *n.s.*). Neither did participants successfully learn word-object mappings (t_o = 1.78, *n.s.*). Perhaps unsurprisingly, segmentation and word-object mapping accuracy were uncorrelated (*r* = .006, *n.s.*), as were object-label segmentation and frame word segmentation (*r* = -.226, *n.s.*). With neither of the linguistic regularities present, participants failed to segment speech or to learn to word-object mappings in the joint task.

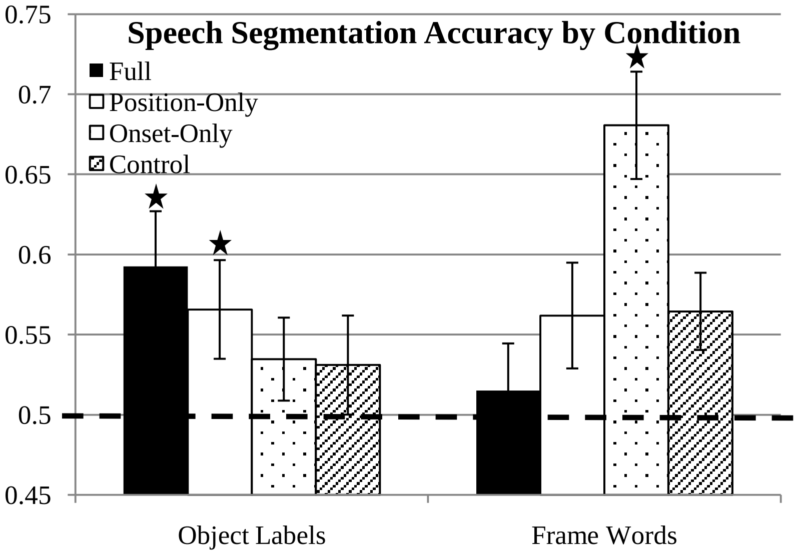


Figure 2: Speech segmentation accuracy by condition. Error bars indicate standard errors. Stars indicate above-chance levels of performance.

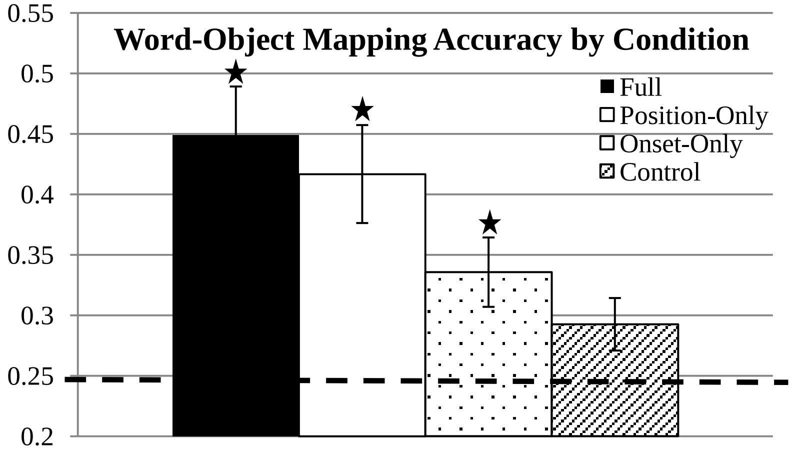


Figure 3: Word-object mapping accuracy by condition. Error bars indicate standard errors. Stars indicate above-chance levels of performance.

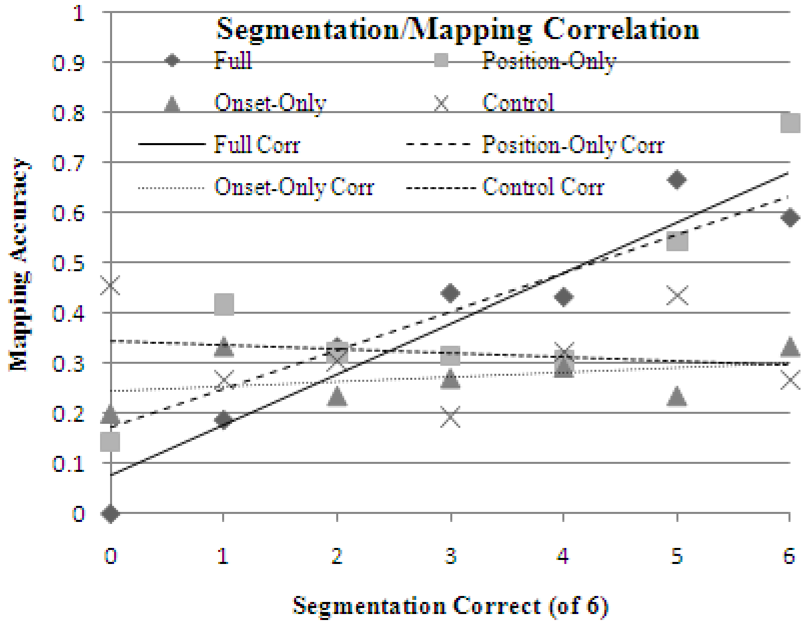


Figure 4: Correlation between segmentation and mapping accuracy by condition. Significant positive correlations were found in the *full* and *onset-only* languages.

2.3 Discussion

Exposing participants to artificial languages constructed from a corpus of child-directed speech, we were able to determine the independent and joint contributions of the two regularities apparent in the corpus. Keeping constant the words which made up naming phrases, we altered only their order across conditions. If parallel speech segmentation and word-object mapping rely on environmental cues to reduce cognitive load, this reliance should be reflected in the learning rates across our four conditions. In fact, this was precisely the case.

In the *full* language, which gave strong statistical cues about the phrasal position of object-words as well as cues to their onset, participants successfully segmented labels from continuous speech and mapped them to their referent objects. This success came in spite, or perhaps because, of chance-level performance on frame word segmentation. These results, along with the strong correlation between a word's probability of being segmented and the same word's probability of being correctly mapped, suggest that participants became attuned to the positional regularity and effectively ignored large portions of the speech input. This reduction in cognitive load may have supported learning.

The *position-only* language, in contrast, removed the onset cue by moving words in the cue set to the beginning of each sentence. In this condition, participants also successfully segmented object-labels from continuous speech, although at a reduced level. In trade, they performed at a near-significant level on frame-word segmentation. Also, unlike in the *full* language condition, segmentation of object labels and frame words were highly correlated, suggesting an interaction between the processes. Nonetheless, despite these differences, participants exposed to the *position-only* language performed well on the test of word-object mapping. Thus, removing the onset cue forced participants to actively process more of the speech stream, but the presence of the position cue kept cognitive load low enough to enable learning. These results are consistent with previous work by Frank et al. (2006).

Removing the position regularity from the *full* language yielded the *onset-only* language. In this condition, object-labels were preceded by a member of the small set of onset cues, but occurred always in medial phrasal position. Without labels in final position, participants performed at chance on tests of object-label segmentation, however performance on frame word segmentation reached levels unseen in the previous conditions. Surprisingly, although participants did not show knowledge of correct label segmentation, they did succeed in mapping words to objects at above chance (albeit reduced) levels. Thus, an onset cue alone was sufficient to enable word learning, a result perhaps anticipated by the work of Borteld, Morgan, Golinkoff, and Rathburn (2005).

Finally, when naming phrases contained all of the same words but neither of the cues found in the child-directed speech corpus, participants performed at chance on all tests. Thus, when exposed to the *control* language, participants were unable to cope with the cognitive load inherent in the simultaneous segmentation and word learning.

3. Conclusions

We began by considering the relationship between statistical speech segmentation and statistical word learning. While previous work has demonstrated a serial link (e.g. Graf Estes et al., 2007), in which word candidates generated via statistical segmentation are privileged by infants in statistical word learning, a robust parallel demonstration has remained elusive (Frank et al., 2006). Perhaps the computational resources required by the tasks are simply too costly to allow their simultaneous resolution. We proposed that construction of previous artificial languages may have averaged out the very regularities which support a parallel solution in naturalistic environments. To borrow from J. J. Gibson, 'it's not [just] what is inside the head that is important, it's what the head is inside of.'

Analysis of a corpus of child-directed speech from free-play found two potential sources of such scaffolding. First, object labels occurred consistently in the final position of naming phrases. Second, these labels were consistently preceded by one of a small set of onset cue words, predominantly articles. We constructed artificial languages following a 2x2 design to produce all possible presence/absence combinations of these regularities. Adult participants were exposed to an ambiguous word-object learning task in the cross-situational paradigm (Yu & Smith, 2007) in which labels were embedded within continuous speech phrases. This human simulation paradigm experiment (Gillette, Gleitman, Gleitman, & Lederer, 1990) allowed us to determine the independent and joint contributions of the two natural naming regularities.

The results showed that either regularity was independently sufficient to support learning, but that learning did not occur in the absence of both. This supports our hypothesis that environmental regularities play an important supporting role in parallel segmentation and word learning. Furthermore, participants in the successful conditions showed marked differences in their patterns of learning. In the *full* language, participants zeroed in on the object labels while ignoring frame words. Participants exposed to the *position-only* language learned to segment both object labels and frame words, and evidence suggests that knowledge about both word types supported each other. Finally, in contrast, participants in the *onset-only* condition learned to segment only the frame words, but nonetheless learned some word-object mappings.

There are, of course, further questions to be addressed. How different would the patterns of learning be in young infants? This is of interest both in the window in which statistical segmentation but not cross-situational learning has been observed, and after both have been documented. How much of the final-position benefit is the result of testing English-speaking participants? Aslin et al.'s (1996) results suggest that not all of the benefit is likely to be thus explained away. In either case, results presented here elucidate the link between segmentation and word-object mapping, and also suggest that there may be more than one route into word-learning. While participants in the *onset-only* condition did not successfully segment object words, they nevertheless learned their correct referents. It may thus be possible, as Peters (1977) suggested, to 'learn the tune before the words.'

References

Aslin, R.N., Woodward, J., LaMendola, N., & Bever, T.G. (1996). Models of word segmentation in fluent maternal speech to infants. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

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

Brent, M.R. & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33-B44.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458-467.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proccedings of the International Conference on Spoken Language Processing*, *3*, 1393-1396.

Frank, M. C., Mansinghka, V., Gibson, E., & Tenenbaum, J. (2006). Word segmentation as word learning: Integrating stress and meaning with distributional cues. In. H. Caunt-Nulton, S. Kulatilake, & I. Woo (Eds.) *Proceedings of the 31st Annual Boston University Conference on Language Development*. Boston, MA: Boston University.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 1–55.

Graf Estes, K., Evans, J.L., Alibali, M.W., & Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*, 254-260.

Macnamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review*, *79*, 1-13.

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

Peters, A. M. (1977). Language learning strategies: Does the whole equal the sum of the parts? *Language*, *53*, 560-573.

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

Shafer, V.L., Shucard, D.W., Shucard, J.L., & Gerken, L.A.(1998) An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language, and Hearing Research*, *41*, 874-886.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558-1568.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414-420.

Yu, C., Smith, L. B., & Pereira, A. F. (2008) Grounding word learning in multimodal sensorimotor interaction. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.) *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1017-1022). Austin, TX: Cognitive Science Society.