Speech segmentation and word discovery: A computational perspective

Michael R. Brent

Johns Hopkins University

Address correspondence to: Michael R. Brent
430 E. 63rd St, #10H
New York, NY 10021

email: brent@jhu.edu
phone: 212-980-4825
fax: 212-753-5510

Summary

The segmentation and word discovery problem arises because speech does not contain any reliable acoustic analog of the blank spaces between words of printed English. As a result, children must segment the utterances they hear in order to discover the sound patterns of individual words in their language. A number of computational models have been proposed to explain how children segment speech and discover words, including ten new models in the last five years. This paper reviews all proposed models and organizes them according to their fundamental segmentation strategies, their processing characteristics, and the ways in which they use memory. All proposed models are found to use one of three fundamental strategies: the utterance-boundary strategy, the predictability strategy, or the word-recognition strategy. Selected predictions of the models are explained, their performance in computer simulations is summarized, and behavioral evidence bearing on them is discussed. Finally, ideas about how these diverse models might be synthesized into one comprehensive model are offered.

Keywords: Language acquisition, computational models, segmentation, statistical models, machine learning, natural language processing, speech processing

We experience speech in our native language as a sequence of discrete words. This can lead to the impression that speech must contain some acoustic analog of the blank spaces that appear between printed words in many languages. When listening to an unfamiliar language, however, we generally hear a continuous stream of speech broken only by the silences at the ends of utterances. *Utterances,* the units of speech that *are* delimited by easily recognized acoustic boundaries, typically consist of multiple words. This dichotomy between the experiences of hearing known and unknown languages raises two questions. First, how do speakers with native mastery of a language effortlessly and unconsciously transform a stimulus consisting of continuous speech into a percept consisting of a sequence of discrete words? That is, how is knowledge of a particular language brought to bear on the problem of segmenting speech into words? Second, how do infants and toddlers segment speech and learn new words despite their limited knowledge of the ambient language? In particular, how do they overcome the handicap of their regular encounters with utterances that contain one or more unfamiliar words?

At a fundamental level, the tasks that adults and young children face are the same — adults sometimes encounter and learn novel words, and children often recognize familiar words within utterances. However, adults are already familiar with a vastly larger proportion of the words they encounter than children are. As a result, research on adult speech segmentation has focused on questions of on-line lexical access[1,2,3,4,5] and cues that may help to limit the number of unsuccessful lexical access attempts,[6,7] setting aside the issue of novel words. Research on language learning, on the other hand, has focused primarily on cues to help limit the number of hypothesized novel words, and is only beginning to address on-line lexical access.[8] This article reviews recent work on speech segmentation from the perspective of language learning, focusing particularly on computational models.

A language learner hears utterances that may contain one or more unfamiliar words, each of which may or may not refer to some observable object, action, or property. For example, let *abcde* be an utterance, where each letter stands for a perceptual unit of speech, such as a phoneme or syllable. There are 15 sub-sequences of *abcde*: *a, b, c, d, e, ab, bc, cd, de, abc, bcd, cde, abcd, bcde*, and *abcde* itself. One possible learning strategy would be to store all of these in memory, as candidate words, in the hopes of eventually figuring out a meaning and a syntactic function for one or more of them. However, this would impose a considerable memory burden. Further, the problem of matching sounds up with meanings would be enormously complicated by the presence of so many candidates for the sounds of words. From the computational perspective, the aim of research in segmentation and word discovery is to identify mechanisms that children use to reduce these computational burdens by reducing the number of candidate word sounds.

The segmentation and word discovery problem has received considerable attention dating back to the work of Roger Brown[9] in the 1960's (see also Zelig Harris[10]). Recent years have seen a surge of interest in this topic, leading to a number of proposed cues and strategies that children might use. Among these are three types of phonological cues. First, some languages provide what Cutler and colleagues have called *rhythmic* cues.[11,12] For example, most stressed syllables in English are word-initial,[7] and adult speakers of English find it more natural to segment connected speech in such a way that stressed syllables occur at the beginnings of words.[13,14] Language learners following such a strategy might eliminate candidate words containing non-initial stressed syllables, thereby reducing the number of candidates. Indeed, Jusczyk and colleagues have shown that 7.5-month-old (but not 10.5-month-old) American infants have difficulty recognizing the sound-patterns of words with non-initial stress.[15] A second type of cue that children might use is *allophonic variation* — the fact that some speech sounds are pronounced

differently when they occur word-finally as opposed to word-initially.[16,17,18] A third type of phonological cue derives from the phonotactic properties of a language — which sequences of phonemes are common in words of the language, which are rare, and which are not permitted at all. For example, a learner who knew that English does not allow words beginning in two stop consonants could avoid mis-segmenting /bɪgkæt/ (*bigcat*) into /bɪ/ and /gkæt/. There is evidence that both adults[6,19] and infants[20] can make use of language-particular phonotactic cues. While rhythmic cues have been minor players in computational models of language learning and allophonic cues have not appeared at all, phonotactic cues in various forms have featured prominently in several models.[21,22,23]

Segmentation strategies

With one exception,[24] existing computational models of segmentation and word discovery do not address the interaction between segmentation and the mapping of hypothesized word forms to their meanings or their syntactic privileges. As a result, the input they take consists entirely of representations of speech sounds. The term *phoneme* will be used to describe the fundamental units of the input representation, since most computer simulations reported in the literature use phonemic representations. However, many of the models described below could easily be applied to representations based on other units, such as syllables (see Box 1).

Insert Box 1 about here

All existing models implicitly impose a constraint against hypothesizing overlapping words. For example, no model would hypothesize that two words, *ab* and *bc*, both occur in a single instance of the utterance *abcde*. Most models also forbid embedded words. For example, they would not hypothesize that two words, *ab* and *abc*, both occur in a single instance of the utterance *abcde*. (For exceptions, see refs. 24 and 34). In addition to these structural principles,

all existing computational models of segmentation and word discovery are based on one of three fundamental strategies:

1. Hypothesize word boundaries after phoneme sequences that are characteristic of the ends of utterances (*utterance-boundary strategy*).
2. Hypothesize word boundaries before phonemes that would not have been predicted on the basis of the preceding phonemes (*predictability strategy*).
3. Hypothesize whole words and recognize them when they occur in utterances (*word-recognition strategy*).

The remainder of this section discusses each of these strategies in turn.

Computational models relying primarily on the utterance-boundary strategy have been described by both Aslin et al.[22] and Christiansen et al.[23] (Table 1).

Insert Table 1 about here

Both groups implemented this strategy using neural networks. The networks were trained to predict, among other things, the locations of *utterance* boundaries in sequences of phonemes transcribed from spontaneous speech to young children. The networks were interpreted as predicting a word boundary when the activation on the utterance-boundary unit exceeded its mean activation. This method is motivated by the notion that the phoneme sequences immediately preceding utterance boundaries will bear some statistical similarity to those immediately preceding word boundaries. This notion is plausible because the ends of utterances are also the ends of words. As a result, phonemes or phoneme sequences that are rare at the ends of words will also be rare at the ends of utterances. Conversely, sequences that are common at the ends of words will also be common at the ends of utterances, provided that the words they occur in can appear at the ends of utterances. However, the fact that /ə/ occurs very often at the

ends of words like *the* and *a* does not imply that /ə/ occurs frequently at the ends of utterances, since the words *the* and *a* are themselves extremely rare at the ends of utterances.

The utterance-boundary strategy relies on learning and exploiting certain phonotactic regularities of the input language — those governing the ends of words. Thus, it is a special case of the more general class of phonotactically driven strategies. However, no models have yet been proposed that rely primarily on learning more general phonotactic regularities.

The second segmentation strategy does not rely on utterance boundaries at all. Instead, it relies on the fact that guessing an unknown phoneme based on adjacent phonemes in the same word is easier than guessing on the basis of adjacent phonemes in different words. For example, most occurrences of the phoneme /ð/ are followed by vowels, as in the very frequent words *the*, *this*, *that*, and *them*, and this implies logically that only a small percentage are followed by other phonemes — /m/, for instance. When /m/ does follow /ð/, as in *bathe more*, its very surprisingness suggests that it is likely to be the first phoneme of a new word. This effect forms the basis for a number of segmentation models. For instance, Saffran and colleagues[25] — treating syllables rather than phonemes as the fundamental units of input — have proposed that children might estimate the probability of each syllable in the language conditioned on its predecessor as follows:

$$\Pr(y|x) \approx \frac{\mathrm{freq}(xy)}{\mathrm{freq}(x)}$$

This is the same computation that was illustrated informally in the /ðm/ example — the conditional probability of *y* given *x* is estimated by the proportion of *x*'s that have been followed by *y*'s in the learner's experience so far. This conditional probability estimate is commonly called the *transitional probability*. Saffran et al. suggest that children may segment utterances at low points of the transitional probability between adjacent syllables — that is, when a syllable occurs

that is surprising given its predecessor.

The predictability strategy has also been implemented using neural networks by Elman[26] and Cairns et al.[27] (Table 1). In this approach the network is trained to predict the next phoneme. The network's prediction for each phoneme is compared to the actual phoneme. The more surprising the actual phoneme is, in view of the network's prediction, the more likely it is to be word-initial.

The word-recognition strategy works by hypothesizing word-like units, storing explicit representations of them, and attempting to recognize them when they occur in utterances. The term *word-like unit*, or simply *unit*, is used to emphasize the fact that hypothesized units are not necessarily actual words of the language — the hypothesis could be incorrect. When a unit that is hypothesized to occur in a particular utterance matches a unit that has previously been hypothesized and stored it is called a *familiar unit*; otherwise, it is called a *novel unit*. The recognition of a familiar unit in an utterance reduces the number of potential novel units because of the no-overlap principle. For example, if *cde* is recognized as an occurrence of a familiar unit in the utterance *abcde* then there are only three potential new units: *a*, *b*, and *ab*. This is a substantial reduction from the 15 possibilities when no occurrences of familiar units are recognized in *abcde*.

Most recognition based segmentation algorithms [21, 24, 28, 29, 30, 34] have been cast in terms of choosing a segmentation of the input in such a way as to *optimize* a set of criteria or, when they are stated more formally, to *optimize an objective function*. A simple example of optimization is fitting a straight line to a set of points in the plane, where the most commonly used objective function is the sum of squared deviations of the points from the line. The line that minimizes the sum of squared deviations is taken to the be the line that fits the points best. An example of the

optimization approach to segmentation and word discovery is the INCDROP model,[32,28,36] which can be characterized qualitatively as follows. Segment each utterance in such a way as to:

1. Minimize the sum of the lengths of all hypothesized novel units in the segmentation.
2. Minimize the number of hypothesized novel units in the segmentation.
3. Maximize the product of relative frequencies of the units in the segmentation. The relative frequency of a unit is the number of times that unit has occurred so far as a proportion of the total number of times all units have occurred so far.

Criterion 3 favors segmentations with fewer and hence longer units, all other things being equal. This is because each relative frequency is less than one, so multiplying more of them together leads to a smaller product, all other things being equal. However, the criteria balance each other. For example, analyzing every utterance as a single, long, novel unit would be favored by criterion 3, but it would be disfavored even more strongly by criterion 1. Conversely, analyzing each utterance as a sequence of short, familiar, one-phoneme words would be favored by criteria 1 and 2, but it would be disfavored even more strongly by criterion 3.[21]

The INCDROP optimization criteria can be derived rigorously from a probabilistic generative grammar. The grammar encodes the prior knowledge that sentences are constructed by selecting words from some finite, but initially unknown, lexicon, and stringing them together.[28] The INCDROP criteria also have a natural cognitive interpretation in terms of minimizing the burden of memorizing new words (by minimizing the number and length of new words) and minimizing the burden of accessing the memories of familiar words (by minimizing the number of accesses and maximizing the frequencies of the words to be accessed).

INCDROP makes a number of behavioral predictions, including these:

1. Utterances that contain no sequences matching stored units are analyzed as a single novel

unit. This unit is stored for recognition in later utterances.

2. A sequence that matches a stored unit and does not overlap any other sequences that match stored units will tend to be recognized as an instance of the stored unit, unless that unit is both very short and very rare. The contiguous sequences that remain after the recognized unit is extracted are segmented as though they were separate utterances.

(See ref. 32, 36 for additional predictions.) These two predictions can be derived directly from criteria 1-3 above.[31,32] The following utterances, which a mother was recorded saying to her child, illustrate the predictions:

That!…Isthatforthedoggy?

INCDROP predicts that a child who did not yet know any of the words in these utterances would treat the first utterance, *that*, as a single novel unit and store it for later recognition. *That* would then be extracted from *Isthatforthedoggy*. The remaining contiguous strings, *is* and *forthedoggy*, would be segmented as though they were separate utterances. Not recognizing any familiar units within them, the child would store them as novel units for later recognition. In the case of *is*, a very valuable new word would be stored. Although *forthedoggy* is not a word, its syntactic and semantic coherence suggest that storing it as a word would do little harm to a child's lexicon.

Under the INCDROP model, learners can make good use of isolated words without identifying isolated words as such. Instead, they assume that each utterance consists of a single word unless there is evidence that it contains a familiar unit within it. Furthermore, isolated words are not essential for bootstrapping segmentation. For example, if a learner with no knowledge of relevant words heard the utterances *getit? Igetit! Ican! canyou?* then INCDROP predicts the following segmentation: *getit? I_getit! I_can! can_you?* At the cost of mistaking *getit* for a single word, the learner is predicted to extract *I*, *can*, and *you* without ever hearing an

isolated word.

Empirical tests of some of the INCROP predictions, as well as some predictions of the transitional probabilities model, are discussed below. It seems likely that testable predictions can be extracted from other segmentations models as well, but to the best of my knowledge no explicit behavioral predictions of other models have been published.

INCDROP and its predecessor DR Optimization[21] are *analytic* recognition models, meaning that they start with whole utterances as the default units and analyze them into smaller units as the evidence warrants. All other proposed word-recognition models are *synthetic*, meaning that they start with phonemes as the default units and join them into larger units as the evidence warrants.[24, 29, 30, 33, 34] Typically, synthetic algorithms only consider novel units that can be built up by combining either two or three familiar units; an utterance that consists of more than three phonemes but does not contain any other familiar units cannot be analyzed as a single novel unit.

The INCDROP model is *incremental*, meaning that it always finalizes the segmentation of an utterance by end of that utterance, without waiting to examine the next utterance. All other algorithms that implement the word recognition strategy using optimization are *batch*, meaning that they can store an unlimited amount of input, segment utterances out of order, and revise earlier decisions. Because humans segment each utterance as they hear it, batch algorithms are sometimes viewed as psychologically implausible. However, that is true only if these algorithms are interpreted as exact descriptions of cognitive models. An alternative interpretation is that writing about batch algorithms in the context of human segmentation and word discovery constitutes an implicit promissory note to the effect that there exists a closely related *incremental* algorithm. INCDROP fulfills that promise with respect to its predecessor, DR Optimization, which was implemented as a batch algorithm for optimizing the same criteria.

While the incremental processing of INCDROP constitutes progress over batch algorithms, humans segment *on-line*, meaning that they decide whether there is a word boundary between two phonemes within a fixed window of time after the phonemes are heard. Further, humans appear to use a *predictive on-line* algorithm, meaning that they guess at the identity of the current word before the end of the word has been heard.[1,3,5,8] Thus INCDROP, as a cognitive model, rests on the expectation that there exists a predictive on-line algorithm that closely resembles the current incremental implementation in both segmentation accuracy and behavioral predictions.

One proposed model, PARSER,[33] is based on the word-recognition strategy but not on optimization. Starting from the beginning of the input, PARSER repeatedly segments out (recognizes) the longest section of input that matches a stored unit. This longest-match approach, which is used instead of choosing the segmentation that maximizes the product of relative frequencies, makes PARSER nearly on-line. However, it seems likely that this approach will impair segmentation accuracy when PARSER is applied to natural language input. Nonetheless, PARSER is interesting because it is the only recognition-based model that was designed around psychological principles like on-line processing, decay of memory traces, and interference among similar memory traces. PARSER challenges to modelers starting from mathematical principles to fulfill the promise of on-line algorithms and to address the limitations of human memory.

No currently proposed model of segmentation and word discovery has been implemented as a predictive on-line algorithm. However, the algorithms based on the utterance-boundary strategy and the predictability strategy are *conservative on-line*, meaning that they are on-line but do not guess at the identity of the current word while it is still being read in. On the one hand, these algorithms are a step closer to the predictive on-line behavior of humans than INCDROP, and that constitutes a significant achievement. On the other hand, predictive on-line segmentation

requires stored representations of hypothesized words,[2,4] so utterance-boundary and predictability algorithms cannot be made predictive without, in effect, making them recognition algorithms (though not necessarily optimization algorithms). Recognition algorithms do have stored representations of hypothesized words and hence they can, in principle, be adapted to do predictive on-line segmentation.

Memory use in computational models

In the theory of computation, a fundamental distinction is made between *bounded memory* systems, which can use only a fixed amount of memory regardless of the input, and *unbounded memory* systems, whose use of memory can grow as needed with no fixed limit. Proposed algorithms based on the utterance-boundary and predictability strategies have bounded memory, while those based on recognizing hypothesized words use more memory as they hypothesize more words (Table 1). In human languages the number of distinct words encountered appears to grow without limit as more input is processed, since new words are continually being coined, so the word-recognition strategy will use more memory to store the new words as more input is processed. Of course, a bounded-memory system with a large enough memory could store all the words that one individual is likely to encounter in his or her lifetime.

Proposed segmentation models also differ in terms of what is represented in memory (Table 1). All currently proposed bounded-memory algorithms work primarily by representing sub-lexical units. Algorithms that are based on neural networks store implicit, distributed, connectionist-style representations, while the rest store explicit statistics about the frequency of individual phonemes and pairs of phonemes. Recognition-based models represent hypothesized units explicitly.

Different models also differ in whether their stored representations are stable or subject to

interference and/or decay (Table 1). This is important from the perspective of psychological plausibility, since human memory appears to be subject to interference and decay under certain conditions. All currently proposed models that use neural networks are subject to interference when their memory load approaches capacity. Models based on feed forward networks (e.g., ref. 22) are not subject to decay that depends only on time, but those based on recurrent networks (e.g., ref. 23, 26, and 27) can experience memory decay with time. The algorithms that store explicit representations, including the local statistics stored by transitional-probability models and the lexical representations stored by word-recognition systems, generally use stable memory. The exception is PARSER,[33] a word-recognition model with explicit representations of phoneme-strings that are subject to both interference and decay calculated according to explicit functions.

The development of new memory models for recognition-based algorithms is a promising avenue of research. While recognition-based algorithms use an unbounded amount of memory, reducing the rate at which memory usage grows by introducing forgetting mechanisms might enhance the psychological plausibility of these models.

Implementations and simulations

In principle, any fact about children's discovery of the sound patterns of words, including the types of errors they make, is relevant to deciding among alternative segmentation models. However, very little reliable, reproducible data about children's segmentation errors is available. But one fact is truly robust: Children eventually succeed at segmenting speech and acquiring a lexicon. Computer simulations have therefore focused on investigating the degree to which competing segmentation models can explain this central fact.

Computer simulations using phonemic transcripts of spontaneous, child-directed speech have been reported for a number of the algorithms described above. In the most extensive

published comparison to date, I found that MBDP-1, an implementation of the INCDROP model, yielded the most accurate segmentations and lexicons of the algorithms tested.[28] MBDP-1 had an average segmentation accuracy of about 70%, while two algorithms based on predictability (transitional probabilities and Elman's algorithm) had average accuracies in the 40%-45% range. A predictability algorithm based on mutual information (MI in Table 1), which I developed for comparison purposes, had an average accuracy of about 55%. MBDP-1 was the only algorithm for which the overall accuracy of the lexicon increased as more input was processed; for all other algorithms, the percentage of newly hypothesized words that had in fact occurred as words in the input declined as more input was processed.

Various combinations of cues have also been tested. In one simulation, input containing stress information was provided to an utterance-boundary segmenter.[23] The system learned not only which phonemes tend to occur at the ends of utterances but whether stressed or unstressed syllables are more likely at the ends of utterances. The results showed that, under certain assumptions about the stressing of function words, stress information can yield a statistically significant increase in the accuracy of an utterance-boundary segmenter, from 37% to 43%.

In another simulation, phonotactic constraints on the consonant clusters that can occur at English word boundaries were derived by explicitly memorizing the consonant clusters that occur at the beginnings and ends of *utterances* in the input. The results showed that this phonotactic knowledge boosted the performance of a batch recognition-based algorithm.[21]

Evidence from human subjects

Investigations involving human subjects have yielded several types of evidence about how people segment speech and discover the sound patterns of novel words. These investigations include studies of infants' perceptual abilities and knowledge of the ambient language, studies of

infants' and adults' patterns of inference using stimuli from auditory artificial languages, and studies of spontaneous speech by caretakers to children. Selected results of each type are reviewed briefly in this section.

Studies of infant speech perception have yielded a wealth of important results suggesting that substantial adaptation to the rhythmic, phonotactic, and allophonic properties of the ambient language occurs between the ages of 6 and 9 months.[15] Studies of infants' speech segmentation abilities using natural language stimuli have also shed light on the complex interplay of pattern recognition and rhythmic cues, and how this interplay changes during infancy.[15] However, these studies have not been aimed at differentiating among the proposed computational models.

Investigations of particular models have generally relied on auditory artificial languages. In one such study, Saffran and colleagues constructed a continuous stream of computer-synthesized nonsense syllables by concatenating, in random order, four three-syllable "words". They found that, after two minutes of exposure to this continuous stream, 8-month-old infants did not listen as long to the "words" from which it was constructed as they did to 3-syllable foils that had also occurred in the syllable stream but were not words.[35] Saffran et al. explained this result in terms of transitional probabilities, although Perruchet and Vintner[33] subsequently argued that it could also be explained by PARSER. Clearly, though, the absence of utterance boundaries in the stimulus implies that the utterance-boundary strategy cannot be invoked to explain this result. Similarly, an analytic recognition-based model like INCDROP cannot explain this result, since they bootstrap by treating whole utterances as words.

Saffran et al.'s results suggest that, when confronted with long stretches of speech containing no familiar words and no utterance boundaries, infants can still discover novel words. The significance of this finding for language development depends on how often infants are

confronted with long utterances containing no familiar words and whether such utterances are used for word discovery. A number of studies have reported the that *average* length of an utterance in speech to young children is about 3.5 words. Using very conservative methods, Jeff Siskind and I estimated the average frequency of isolated words in the speech of 8 mothers to their infants (ages 9.5-12.5 months) to be about 7%, excluding interjections, onomatopoeia, social routines, and all words that did not also appear in multiword utterances. Thus, the language learning environment seems to afford plenty of short, easily remembered utterances that also occur as sub-sequences of longer utterances — just what would be needed to make an analytic word-recognition strategy such as INCDROP effective.

Inspired by the work of Saffran et al., Dahan and Brent carried out artificial language experiments in which stimuli consisting of 2, 3, and 5-syllable utterances were presented to adult subjects.[36] The subjects were exposed to both short (2 or 3 syllable) utterances and long (5 syllable) utterances that contained a short utterance within them. For example, subjects might hear *koshedi* and, after several unrelated items, *koshedifenu*. The results showed that the subjects tended to treat short utterances as a single unit by default — for example, subjects who had heard *koshedi* in isolation remembered *koshedi* better than they remembered *koshe,* and conversely for those who heard *koshe* in isolation. When a short utterance occurred within a longer utterance it was segmented out and the remainder was treated as a unit. For example, subjects who heard *koshedi* and *koshedifenu* remembered *fenu* better than they remembered *difenu*. These results and others reported in the paper are consistent with INCDROP. Further, transitional probabilities alone cannot explain the pattern of results (see ref. 36 for details).

The results of the Saffran et al. and the Brent and Dahan studies show two different behaviors that can be observed, depending on whether the materials consist of short utterances

with embedding or very long utterances. Although short utterances appear to be a better model of the language learning environment, both behaviors need to be explained. It would therefore constitute significant progress if a single model could explain both patterns, and could also perform as well as INCDROP in simulations on input transcribed from spontaneous speech.

Summary and conclusions

In the study of segmentation and word discovery we face a wealth of good intuitions that are ripe for integration into a more comprehensive model. Based on the simulations and empirical data cited above, I believe that a comprehensive model should be based on word recognition. Within a recognition-based model, both phonotactics (including the utterance-boundary strategy) and predictability can play a role in evaluating the probabilities that particular sound sequences are novel words of the language being heard, and hence should be stored in memory and recognized in future utterances. As the lexicon grows and familiar words come to dominate novel words, recognition will naturally come to dominate phonotactics and predictability in determining segmentation behavior.

In addition to modeling the information sources used for segmentation and word discovery, a comprehensive model should segment on-line and predict the completions of incoming words, consistent with the experimental data on humans. Further, it should attempt to model the limitations of human memory for speech, to the extent that these are understood.

The tools for constructing such a comprehensive model of segmentation and word discovery appear to be at hand. Probability, and generative probability models in particular, provide a universal scale for weighing information provided by sources such as word recognition, phonotactics, and predictability against one another. Optimization — the process of evaluating segmentations in order to find the most probable one — provides a framework within which

different processing models can paired with different probability models. Separating memory for hypothesized words from optimization processes that recognize words makes it possible to model decay and interference independently of information sources and processing considerations. Within the framework of probability, optimization, and memory that is separate from processing, it should be possible to create a single model that is suitable for studies of lexical access, speech segmentation, and word discovery.

Insert Box 2 about here

Acknowledgments

The writing of this paper was supported by grant DC03082 from the National Institutes of Health.

References

1.Marslen-Wilson, W. and Zwitserlood, P. (1989) Accessing spoken words: The importance of word onsets *JEP: Human Perception and Performance* 15, 576-585

2.McClelland, J. L., and Elman, J. L. (1986) The TRACE model of speech perception *Cog. Psych.* 18, 1-86

3.McQueen, J. M., Norris, D. G., and Cutler, A. (1994) Competition in spoken word recognition: Spotting words in other words *JEP: Learning, Memory, and Cognition* 20, 621-638

4.Norris, D. G. (1994) Shortlist: A connectionist model of continuous speech recognition *Cognition* 52, 189-234

5.Tabossi, P. Burani, C. and Scott, D. (1995) Word identification in fluent speech *J. Mem. Lang.* 34, 440-467

6.Norris, D., et al. (1997) The possible-word constraint in the segmentation of continuous speech *Cog. Psych.* 34, 191-243

7.Cutler, A., and Carter, D. M. (1987) The predominance of strong initial syllables in the English vocabulary *Comp. Speech Lang.* 2, 133-142

8.Swingley, D., Pinto, J. P., and Fernald, A. Continuous processing in word recognition at 24 months *Cognition* in press

9.Brown, R., Cazden, C., and Bellugi, U. (1969) The Child's Grammar from I to III. In J. P. Hill, Ed., *Minnesota Symposium on Child Psychology*. Minneapolis: University of Minnesota Press

10.Harris, Z. S. (1955) From phoneme to morpheme *Language* 31, 190-222

11.Cutler, A., et al. (1992) The monolingual nature of speech segmentation by bilinguals *Cog. Psych.* 24, 381-410

12.Cutler, A., and Otake, T. (1994) Mora or phoneme? Further evidence for language-specific listening *J. Mem. Lang.* 33, 824-844

13.Cutler, A., and Norris, D. (1988) The role of strong syllables in segmentation for lexical access *JEP: Human Perception and Performance* 14, 113-121

14.Cutler, A., and Butterfield, S. (1992) Rhythmic cues to speech segmentation: Evidence from juncture misperception *J. Mem. Lang.* 31, 218-236

15.Jusczyk, P. W. (1997)*The Discovery of Spoken Language,* MIT Press.

16.Church, K. W. (1987) Phonological parsing and lexical retrieval *Cognition* 25, 53-69

17.Gow, D. W., and Gordon, P. C. (1995) Lexical and prelexical influences on word segmentation: Evidence from priming *JEP: Human Perception and Performance* 21, 344-359

18.Christophe, A., et al. (1994) Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition *J. Accoust. Soc. Am.* 95, 1570-1580

19.McQueen, J. M. Segmentation of continuous speech using phonotactics *J. Mem. Lang.* 39, 21-46

20.Mattys, S. L., et al. Phonotactics and Prosodic Effects on Word Segmentation in Infants *Cog. Psych.* (in press)

21.Brent, M. R. and Cartwright, T. A. (1996) Distributional regularity and phonotactics are useful for segmentation *Cognition* 61, 93-125

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

23.Christiansen, M. H., Allen, J., and M. Seidenberg (1998) Learning to segment speech using multiple cues *Lang. Cognit. Proc.* 13, 221-268

24.de Marcken, C. G. (1996) Linguistic structure as composition and perturbation, in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*

25.Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996) Word segmentation: The role of distributional cues *J. Mem. Lang.* 35, 606-621

26.Elman, J. L. (1990) Finding structure in time *Cog. Sci.* 14, 179-211

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

28.Brent, M. R. (1999) An efficient, probabilistically sound algorithm for segmentation and word discovery *Machine Learning Journal* 34, 71-106

29.Olivier, D. C. (1968) *Stochastic Grammars and Language Acquisition Mechanisms* Unpublished doctoral dissertation, Harvard University

30.Redlich, A. N. (1993) Redundancy reduction as a strategy for unsupervised learning *Neural Computation* 5, 289-304

31.Brent, M. R. (1996) Advances in the computational study of language acquisition *Cognition* 61, 1-37

32.Brent, M. R. (1997) Toward a unified model of lexical acquisition and lexical access *J. Psycholinguist. Res.* 26, 363-375

33.Perruchet, P., and Vintner, A. (1998) PARSER: A Model for Word Segmentation *J. Mem. Lang.* 39, 246-263

34.Wolff, J. G. (1977) The discovery of segments in natural language *Br. J. Psychol.* 68, 97-106

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

36.Dahan, D. and Brent, M. R. (1999) On the discovery of novel word-like units from utterances: An artificial-language study with implications for native-language acquisition *JEP: General* 129, in press

Box 1: Input representations

Computer simulations that take transcripts of spontaneous speech as input have all used transcription systems based on either atomic phonemes or phonemic features. In a transcription system based on atomic phonemes the first symbols of the representations of *dotty* and *dirty* are identical, while the first vowels are distinct — they are no more similar to each other than to consonants. In a representation based on phonemic features, phonemes are related by shared features that define a similarity metric. For example, vowels are more similar to one another than to consonants. A third possibility is to represent each syllable in the input with an atomic symbol that does not encode any information about its relation to other syllables. For example, the first symbols of the representations of *dotty* and *dirty* would stand for distinct syllables (pronounced /da/ and /dɚ/), which would be no closer to one another than to any other syllable.

Simulations of models based on the utterance-boundary strategy have used phonemic features because a featural representation makes it easier to learn generalizations about which sequences of phonemes tend to occur before utterance boundaries. (The utterance boundary strategy, discussed in the main text, is to hypothesize word boundaries after phoneme sequences that are characteristic of the ends of utterances.) The success of this strategy is expected to be sensitive to the input representation, so these models make some theoretical commitment to an input representation based on a featural decomposition of either phonemes or syllables. Featural representations may provide some robustness against the natural variability in the pronunciation of a word, since alternative pronunciations are likely to be nearer to one another by the feature metric than by the atomic phoneme metric. Simulations based on the predictability strategy and the word-recognition strategy have used input transcribed into atomic phonemes. However, these strategies do not imply any theoretical commitment about the input representation, since they

treat the input as a string of arbitrary symbols.

There is substantial evidence that the atomic syllable is a salient perceptual unit for infants,[a,b] while evidence that the phoneme is also a salient unit for infants is, at present, lacking.[c] This would seem to suggest that simulations should use a syllabic representation. However, there are complications. The experiments with infants have been done primarily with consonant-vowel (CV) syllables, the canonical syllable type that occurs in all languages. In languages with more complex syllables, including English, words are typically resyllabified in context in such a way that the syllables can cross word boundaries. For example, the phrase *teak rail* in fluent speech arguably consists of the two syllables /ti/ (*tea*) and /krel/ (*krail*). A listener who hears the atomic syllables /ti/ and /krel/, but who represents the corresponding words in her lexicon using the atomic syllables /tik/ (*teak*) and /rel/ (*rail*), would fail to retrieve the correct lexical entries. If a word-recognition algorithm is to have any chance of succeeding on syllabic input, the syllabification must be consistent with word boundaries. But such a syllabification seems to encode a non-trivial portion of the word-boundary information that segmentation models are designed to uncover. Further, most of the experiments that failed to find evidence of phonemic representation have been done with infants that are much too young to learn words, and there is some evidence that finer-grained representations begin to develop even by six months.[d]

References

a Bijeljac-Babic, R., Bertoncini, J., and Mehler, J. (1993) How do four-day-old infants categorize multisyllabic utterances? *Dev. Psych.* 29, 711-721

b Jusczyk, P. W., Kennedy, L. J., and Jusczyk, A. M. (1995) Young infants' retention of information about syllables *Infant Behav. Dev.* 18, 27-42

c Jusczyk, P. W. *The Discovery of Spoken Language* (MIT Press, Cambridge, MA, 1997), pp.112-115, 124-127

d Hillenbrand, J. (1984) Speech perception by infants: Categorization based on nasal consonant place of articulation *J. Accoust. Soc. Am.* 75, 1613-1622

Box 2: Outstanding questions

- How can rhythmic and allophonic cues best be characterized and incorporated in segmentation models?
- How can the set of potential meanings that children entertain for each potential novel word best be characterized? How can the information provided by these sets of candidate meanings best be incorporated in segmentation models?
- How can the limitations of human memory for speech, both long term and working memory, best be characterized and incorporated in segmentation models?
- Is word discovery primarily analytic, starting with utterances as the default words, or synthetic, starting with small perceptual units as the default words?
- How can the utterance-boundary and predictability strategies best be used to assess the degree to which hypothesized novel words are plausible words of the language being learned?
- How can the word-recognition strategy be implemented as a predictive on-line algorithm, thereby making it more psychologically plausible?

Table 1: Characteristics of the computational models reviewed here. Dotted lines separate groups of models that do not differ in any of the characteristics listed here. For models that are stated abstractly the "Processing" characteristic is determined by the best implementation to date (e.g., the MBDP-1 implementation of INCDROP) which is described in one of the papers cited for the model. The notation "trees" in the "Memory Content" column indicates models that allow embedded word-like units and use parse trees rather than strings to represent units.

Model	Principle	Processing	Memory		
			Growth	Content	Stability
Aslin[22]	utt. boundary	on-line	bounded	implicit statistics	interference
CAS[23]	utt. boundary	on-line	bounded	implicit statistics	interference, decay
Elman[26]	predictability	on-line	bounded	implicit statistics	interference, decay
Cairns[27]	predictability	on-line	bounded	implicit statistics	interference, decay
T.P.[25,28]	predictability	on-line	bounded	explicit statistics	stable
M.I.[28]	predictability	on-line	bounded	explicit statistics	stable
PARSER[33]	recognition	on-line, synthetic	unbounded	strings, strengths	decay, interference
INCDROP[28]	recognition	incremental, analytic	unbounded	strings, frequencies	stable
DR Opt.[21]	recognition	batch, analytic	unbounded	strings, frequencies	stable
Olivier[29]	recognition	batch, synthetic	unbounded	strings, frequencies	stable
Redlich[30]	recognition	batch, synthetic	unbounded	strings, frequencies	stable
de Marcken[24]	recognition	batch, synthetic	unbounded	trees, frequencies	stable
MK10H[34]	recognition	batch, synthetic	unbounded	trees, frequencies	stable