Word Segmentation: Quick but not Dirty

Timothy Gambell
1814 Clover Lane
Fort Worth, TX 76107
timothy.gambell@aya.yale.edu

Charles Yang*
Department of Linguistics
Yale University
New Haven, CT 06511
charles.yang@yale.edu

June 2005

Acknowledgments

Portions of this work were presented at the 34th Northeaster Linguistic Society meeting, the 2004 Annual Meeting of the Linguistic Society of America, the 20th International Conference on Computational Linguistics, Massachusetts Institute of Technology, Yale University, University of Delaware, University of Southern California, University of Michigan, University of Illinois. We thank these audiences for useful comments. In addition, we are grateful to Steve Anderson, Noam Chomsky, Morris Halle, Bill Idsardi, Julie Legate, Massimo Piatelli-Palmarini, Jenny Saffran and Brian Scholl for discussions on the materials presented here.

*Corresponding author.

1 Introduction

When we listen to speech, we hear a sequence of words, but when we speak, we-do-not-separate-words-by-pauses. A first step to learn the words of a language, then, is to extract words from continuous speech. The current study presents a series of computational models that may shed light on the precise mechanisms of word segmentation.

We shall begin with a brief review of the literature on word segmentation by enumerating several well-supported strategies that the child may use to extract words. We note that, however, the underlying assumptions of some of these strategies are not always spelled out, and moreover, relative contributions of these strategies to the successful word segmentation remain somewhat obscure. And it is still an open question how such strategies, which are primarily established in the laboratory, would scale up in a realistic setting of language acquisition. The computational models in the present study aim to address these questions. Specifically, by using data from child-directed English speech, we demonstrate the inadequacies of several strategies for word segmentation. More positively, we demonstrate how some of these strategies can in fact lead to high quality segmentation results when complemented by linguistic constraints and/or additional learning mechanisms. We conclude with some general remarks on the interaction between experience-based learning and innate linguistic knowledge in language acquisition.

2 Strategies for Word Segmentation

Remarkably, 7.5 month-old infants are already extracting words from speech (Jusczyk & Aslin, 1995). The problem of word segmentation has been one of the most important and fruitful research areas in developmental psychology, and our brief review here cannot do justice to the vast range of empirical studies. In what follows, we will outline several proposed strategies for word segmentation but that is simply for the convenience of exposition: these strategies are not mutually exclusive, and they have been proposed to be jointly responsible for word discovery (Jusczyk, 1999).

2.1 Isolated Words

It appears that the problem of word segmentation would go simply away if all utterances consist of only isolated words; the child could simply file these away into the memory. Indeed, earlier proposals (Peters, 1983; Pinker, 1984) hypothesize that the child may use isolated words to bootstrap for novel words. Recent corpus analysis (Brent & Siskind, 2001;

cf. Aslin, Woodward, LaMendola, & Bever, 1996; van de Weijer, 1998) has provided quantitative measures of isolated words in the input. For instance, Brent & Siskind (ibid) found that in English mother-to-child speech, an average 9% of all utterances are isolated words. Moreover, for a given child, the frequency with which a given word is used in isolation by the mother strongly correlates with the timing of the child learning that word. Clearly, isolated words are abundant in the learning data and children do make use of them.

The question is How. What would lead a child to recognize a given segment of speech to be an isolated word, which then would come for free. In other words, how does the child distinguish single-word utterances from multiple-word utterances? The length of the utterance, for instance, is not a reliable cue: the short utterance "I-see" consists of two words while the longer "spaghetti" is a single word. We are aware of no proposal in the literature on how isolated words can recognized and, cosequently, extracted. Unless the mechanisms for identifying isolated words are made clear, it remains an open question how these freebies actually help the child despite the corpus studies. We return to this issue in section 5.1.

2.2 Statistical Learning

Another traditional idea for word segmentation is to use statistical correlates in the sound patterns of words (Chomsky, 1955; Harris, 1955; Hayes & Clark, 1970; Wolff, 1977; Pinker, 1984; Goodsitt, Morgan, & Kuhl, 1993; etc.).[1] The insight is that syllables within a word tend to co-occur more frequently than those across word boundaries. Specifically, word segmentation may be achieved by using the *transitional probability* (TP) between adjacent syllables A and B, i.e.,

$$\text{TP}(A \rightarrow B) = \frac{\text{Pr}(AB)}{\text{Pr}(A)}$$

where where P(AB) is the frequency of B following A, and P(A) is the total frequency of A.

Word boundaries are postulated at the points of *local minima*, where the TP is lower than its neighbors. For example, given sufficient amount of exposure to English, the learner may establish that, in the four-syllable sequence "prettybaby", TP(pre→tty) and TP(ba→by) are both higher than TP(tty→ba), thus making "tty-ba" a place of local minimum: a word boundary can be (correctly) identified. It is remarkable that, based on only two minutes of exposure, 8-month-old infants are capable of identify TP local minima among a sequence

[1] It may be worth pointing out that Harris (1955) attempts to establish morpheme boundaries rather than word boundaries. Moreover, his method is not statistical but algebraic.

of three-syllable pseudo-words in the continuous speech of an artificial language (Saffran, Aslin, & Newport, 1996; Aslin, Saffran, & Newport, 1998).

Statistical learning using local minima has been observed in other domains of cognition and perception (Saffran, Johnson, Alsin, & Newport, 1999; Gomez & Gerken, 1999; Hunt & Aslin, 2001; Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002) as well as in tamarin monkeys (Hauser, Newport, & Aslin, 2001). These findings suggest that statistical learning is a domain general and possibly evolutionarily ancient mechanism that may have been co-opted for language acquisition. Statistical learning has also been viewed by some researchers as a challenge to Universal Grammar, the domain-specific knowledge of language (Bates & Elman, 1996; Seidenberg, 1997, etc.). However, to the best of our knowledge, the effectiveness of statistical learning in actual language acquisition has not be tested. Much of the experimental studies used artificial languages with synthesized syllables, with the exception of Johnson & Jusczyk (2001), who also used artificial languages but with natural speech syllables. A primary purpose of the present paper is to give some reasonable estimate on the utility of statistical learning in a realistic setting of language acquisitio.

2.3 Metrical Segmentation Strategy

Another useful source of information for word segmentation is the dominant metrical pattern of the target language, which the child may be able to extract on a (presumably) statistical basis. For instance, about 90% of English content words in conversational speech are stress initial (Cutler & Carter, 1987), and this has led some researchers to postulate the Metrical Segmentation Strategy whereby the learner treats the stressed syllable as the beginning of a word (Cutler & Norris, 1988).

There is a considerable body of evidence that supports the Metrical Segmentation Strategy. For instance, 7.5-month-old infants do better at recognizing words with the strong/weak pattern heard in fluent English speech than those with the weak/strong pattern. Nine-month-old English infants prefer words with the strong/weak stress pattern over those with the weak/strong pattern (Jusczyk, Cutler, & Redanz, 1993). Moreover, the use of the Metrical Segmentation Strategy is robust that it may even lead to segmentation errors. Jusczyk, Houston, & Newsome (1999) found that 7.5-month-old infants may treat the sequence “taris” in“guitar is” as a word. Since “tar” is a strong syllable, this finding can be explained if the infant is extracting words by looking for the dominant stress pattern in her language.

However, a number of questions remain. To use the Metrical Segmentation Strategy, the learner must be able to identify the language-specific stress pattern, for the metrical systems

in the world's languages differ considerably. This can only be achieved after the learner has accumulated a sufficient and representative sample of words to begin with–but where do *these* words come from? There appears to be a chicken-and-egg problem at hand. It is suggested that infants may use isolated words to bootstrap for the Metrical Segmentation Strategy (Johnson & Jusczyk, 2001), but this may not be easy as it looks: as noted earlier, there has been no proposal on how infants may recognize isolated words as such.

Furthermore, even if a sample of seed words is readily available, it is not clear how the infant may learn the dominant prosodic pattern, for whatever mechanism the child uses to do so must in principle generalize to the complex metrical systems in the world's language (Halle & Vergnaud, 1987; Idsardi, 1992; Halle, 1997). While the Metrical Segmentation Strategy works very well–90%–for languages like English, there may be languages where even the most frequent metrical pattern is not dominant, thereby rendering the Metrical Segmentation Strategy less effective. We do not doubt the usefulness of a stress-based strategy, but we do wish to point out that, because it is a language-specific strategy, how children can get this strategy off the ground warrants some discussion. In section 5.1, we propose a weaker but likely universal strategy of how to use stress information for word segmentation.

2.4 Phonotactic Constraints

Phonotactic constraints refer to, among other things, the structural restrictions on what forms a well-formed syllable in a particular language. For instance, although "pight", "clight" and "zight" are not actual English words, they could *in principle* be English words, in a way that "vlight", "dnight", "ptight" could never be. This is because only certain consonant clusters can serve as onsets for a valid English syllable (Halle, 1978). Note that phonotactic constraints are language specific and must be learned on the basis of experience. Remarkably, 9-month-old infants have been shown to be sensitive to the phonotactic constraints of their native languages (Jusczyk, Friederici, et al. 1993; Jusczyk, Luca, & Charles-Luce, 1994; Mattys, Jusczyk, Luce, & Morgan, 1999; Mattys & Jusczyk, 2001).

Phonotactic knowledge may be useful for word segmentation in two ways. First, the infant may directly use phonotactic constraints to segment words: e.g., in a sound sequence that contains "vt", which is not a possible English onset or coda, the learner may conclude that 'a word boundary must be postulated between "v" and "t". Some complications arise, though, for the learner must be able to distinguish consonant sequences that belong to two adjacent syllables within a same word from those that belong to two words altogether. For

example, "mb" is not a valid English onset or coda, but it does not necessarily indicate word boundary: a word such as "embed" consists of two syllables that span over the consonant sequence.

We believe that the use of phonotactic knowledge in word segmentation is less direct. The tacit assumption underlying all discussion on word segmentation is that words consist of syllables, which in turn consist of more primitive units of onset (a sequence of consonants) and rime, which in turn consists of the nucleus vowel and coda (also a sequence of consonants). Phonotactic constraints specify the well-formedness of syllables in a particular language, and they enable the child to parse speech segments (consonants and vowels) into syllables. This step of syllabification is necessary for further development of word segmentation and other aspects of the phonological system. In this sense, phonotactic constraints are logical priors to strategies such as Metrical Stress and statistical learning, as both require learner to treat the syllable as the basic unit of information. For example, the infant learner apparently keeps track of transitional probabilities over adjacent syllables, rather than arbitrary phonological units. Syllabification also allows the learner to identify the metrical pattern of her language, as it is standardly assumed that the syllable is the stress-bearing unit.

It seems rather straightforward to learn phonotactic constraints if the learner has knowledge of syllabic structures–and therefore knows where to find phonotactic constraints. It suffices to pay attention to utterance-initial (before the first vowel) and utterance-final (after the last vowel) consonant sequences to learn the valid onsets and codas respectively. Eventually, and probably quickly, the infant will have seen all–in English, a few dozens–possible forms of onsets and codas. This simple method of learning may explain children's rapid acquisition of non-native phonotactic regularities from brief exposure in laboratory settings (Onishi, Chambers, & Fisher, 2002; Chambers, Onishi, & Fisher, 2003)

2.5 Allophonic and articulatory cues

A somewhat more subtle cue for word segmentation comes from the context-dependent allophonic variations. For instance, the allophone /t/ in English is aspirated at the beginning of a word such as "tab", but is unaspirated at the end of a word such as "cat". If the child is sensitive to how allophonic distributions correlate with wordhood, this could be a useful guide for finding word boundaries. Also of use is the degree of coarticulation of adjacent phonemes (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), which varies as a complex function of their positions within or across syllable/word boundaries (Ladefoged,

1993; Krakow, 1999), and other factors. There is developmental evidence (Jusczyk, Hohne, & Bauman, 1999; Johnson & Jusczyk, 2001) that young infants could use allophonic as well as articulatory cues for word segmentation.

How do children come to know that word boundaries have articulatory correlates? One possibility is that some aspects of such knowledge are innate and fall out of the organization of speech articulation (Brownman & Goldstein, 1992). But experience must also play a role in the mastery of these cues, as different languages have different articulatory patterns. For instance, Jusczyk, Hohne, & Bauman (ibid) show that 9-month-old infants do not seem to use the allophonic variation between "nitrates" and "night rates" to find word boundaries,[2] while 10.5-month-old infants can. On the other hand, if the infant extracts allophonic cues from the linguistic data, it seems, as in the case of the Metrical Stress Strategy, that she must have extracted a set of words to begin with. Hence, The assumptions and mechanisms required for the successful application of articulatory cues remain somewhat unclear.

2.6 Memory

Last but not least, memory must play an important role in word segmentation, for a word isn't learned until it is filed away in the lexicon. However, it seems that the sound patterns of words may be extracted and stored independently of–and prior to–and learning of word meanings. In an important study (Jusczyk & Hohne, 1997), 8-month-old infants were familiarized with novel words embedded in stories with speaker as well as order variations. Even weeks later, they listened to the previously heard words significantly longer than foil words that they had not been exposed to. This apparently takes place well before the word learning stage: it is highly unlikely that the infants understood the meanings of the words such as "python", "vine", "peccaries", etc. that are used in this study. Once familiar words–or familiar sound patterns that are potentially words–are filed into the memory, the learner may use them to extract new words (Peters, 1983; Pinker, 1984); we return to the role of memory in word segmentation in section 5.3.2.

2.7 Discussion

Although researchers may stress specific segmentation strategies, there appears to be a consensus that no single factor alone is sufficient for word segmentation (Jusczyk, 1999). If so, the question arises, How does the learner choose the appropriate segmentation when/if

[2]The problem is not perceptual: even younger infants are perfectly capable to detecting their acoustic differences (Hohne & Jusczyk, 1994).

multiple strategies are available? Consider a hypothetical scenario, where we have three segmentation strategies A, B, and C, and infants have been shown to make use of them experimental settings. Moreover, let us suppose that from corpus studies, A, B, and C can extract 30%, 30%, and 40% of (non-overlapping sets of) words, respectively. Yet one is not warranted to claim that the problem of word segmentation is therefore solved; the details of how such strategies interact must still be spelled out as concrete mechanisms.[3] Which strategy, or which strategies, and in which order, would the child employ to analyze a specific stream of speech? Are A, B, and C universal or fitted for particular languages? (Some of the proposals such as the Metrical Segmentation Strategy clearly are.) If the latter, what kind of learning data and what kind of learning algorithm would lead to the successful acquisition of these strategies before they can be applied to segmentation? It seems, then, that only language-independent strategies can set word segmentation in motion before the establishment and application of language-specific strategies. Finally, current research on segmentation is almost always conducted an experimental setting with artificial languages; it remains to be seen how segmentation strategies scale up in a realistic environment of language acquisition.

3 Modeling Word Segmentation: Preliminaries

In the rest of this paper, we will present a series of computational models that directly addresses the issues of strategy interaction and scalability in word segmentation. Before diving into the detail, it is perhaps useful to state up front what our computational model is *not*.

First and foremost, what we present is not a computational study of child-directed linguistic data. We are interested in the psychologically plausible algorithms that the child may use for word segmentation in an online (real-time) fashion. This is to be distinguished from recent work in the distributional properties of the learning data (Cutler & Norris, 1987; Redington, Chater, & Finch, 1998; Mintz, Newport, & Bever, 2002; Swingley, 2005, etc.), which, to quote one of these studies, "is not to model the actual procedure a child might use, but rather to examine *the information available in children's input*." (Mintz et al. ibid; p396. emphasis original). The two lines of work are complementary but distinct. If distributional

[3]The lack of an account of how various segmentation strategies work together may have to do with the methodologies in word segmentation research. In order to establish the effectiveness of a particular segmentation strategy, one needs to neutralize other potential cues for word boundaries. Only recently have we seen works that address this issue by pitting competing strategies–stress vs. statistical learning, specifically–against each other (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003).

regularity is available in the linguistic input, it remains to be shown that the child could in principle make use of it in a plausible fashion (see Yang, 2002 for extensive discussion). Moreover, the study of corpus statistics is carried out by researchers, who have prior knowledge about the kind of statistical pattern to look for. Whether children enter into the task of language learning with similar preconceptions is a different question. To make an analogy, statistical regularities surely exist in a consequence of integers emitted by a pseudo-random number generator (and hence "pseudo"). With the aid of sufficient computing power, and familiarity with the design of such systems, one may extract the underlying regularities from a sufficiently large sample of numbers. Yet it is a different question whether this deciphering process can be accomplished with psychologically plausible means. Hence, the mere existence of corpus statistics may tell us nothing about a human learner's ability to use it (cf., Legate & Yang, 2002). In the present work, by contrast, we are only interested in the segmentation mechanisms that the child can plausibly use. In section 4.3, we will return to this issue with a comparison of our model and a corpus study of word segmentation (Swingley, 2005),

Second, our computational model differs from most traditional work in word segmentation. Finding structure in linguistic data is a central problem for computational linguistics, and word segmentation has had an important presence (Olivier, 1968; Wolff, 1977; de Marcken, 1996; Brent & Cartwright, 1996; Brent, 1999a; Batchelder, 2002; see Brent, 1999b for summary). Characteristic of these works is the conception of word segmentation as an optimization problem: the model's task is to induce a lexicon that describes the observed utterances in some information-theoretic sense. Because of the different assumptions and methodologies for optimization problems, most of which come from computer science and engineering, these models are not best suited as tools for understanding how human infants segment words.

On the one hand, previous computational models often *over*-estimate the computational capacity of human learners. For example, the algorithm of Brent & Cartwright (1996) produces a succession of lexicons, each of which is associated with an evaluation metric that is calculated over the entire learning corpus. A general optimization algorithm ensures that each iteration yields a better lexicon until no further improvement is possible (which may not produce the target lexicon).[4] It is unlikely that algorithms of such complexity are something a human learner is capable of using.

[4]Brent (1999a) presents a variant of this idea and improves on the computational efficiency by the use of dynamic programming, which involves additional assumptions about the order in which words are presented in the learning data.

On the other hand, previous computational models often *under*-estimate the human learner's knowledge of linguistic representations. Most of these models are "synthetic" in the sense of Brent (1999b): the raw material for segmentation is a stream of segments, which are then successively grouped into larger units and eventually, conjectured words. This assumption probably makes the child's job unnecessarily hard in light of the evidence that it is the syllable, rather than the segment, that makes up the primary units of speech perception (Bertocini & Mehler, 1981; Bijeljac-Babic, Bertocini, & Mehler, 1993; Jusczyk, 1997; Jusczyk, Kennedy, & Jusczyk, 1998; Eimas, 1999). The very existence of the Metrical Segmentation Strategy suggests that infants treat the syllable as the unit of prosodic marking. In addition, when infants compute transitional probabilities, they apparently do so over successive syllables, rather than over segments or any other logically possible units of speech; see section 6.2 for additional discussion. Since syllables are hierarchical structures consisting of segments, treating the linguistic data as segment sequences as in previous segmentation models makes the problem harder than it actually is: for a given utterance, there are fewer syllables than segments, and hence fewer segmentation possibilities. In line with the empirical findings, our model treats the syllable as the relevant primitive unit of phonological information.

The performance of word segmentation models is evaluated following the conventional methods in information retrieval: both *precision* and *recall* must be reported. These performance measures are defined as follows:

(1) a. $$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

b. $$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

For instance, if the target segmentation is "big bad wolf", and the model outputs "bigbad wolf", then precision is 1/2 ("wolf" out of "bigbad" and "wolf", and recall is 1/3 ("wolf" out "big", "bad", and "wolf").

We would like to stress the importance of *both* precision and recall as appropriate quality assessment for word segmentation; not all segmentation models or corpus studies report these figures. In general, it is easy to obtain high performance for one of the two measures but relatively difficult to obtain high performance for both. Take, for example, a search engine that looks for documents that are relevant to the keyword " iPod". It is trivial to achieve high precision–100%, in fact–by extracting precisely *one* web page of, say, `http://www.apple.com/ipod/`, but that is obviously missing out many more hits, resulting in an extremely low recall. Alternatively, recall can just easily be boosted to perfection. One can extract every web page on the Internet, which surely contain all those related to the

iPod, albeit along with numerous false alarms. In general, there is a precision vs. recall tradeoff: lowering one usually boosts the other. In the information retrieval literature, the so-called F-measure is often used to combine precision and recall:

(2)

$$F = \frac{1}{\alpha\frac{1}{p} + (1-\alpha)\frac{1}{r}}$$

where p is precision, r is recall, and α is a factor that weighs the relative importance of p and r (and is often chosen to be 0.5 in practice).

According to the most extensive quantitative comparisons of word segmentation models (Brent, 1999b), the highest performance comes from Brent (1999a), a modification of Brent & Cartwright (1996), with the precision and recall in the range of 70%-80%. Other models are considerably lower, hovering around 40-50% (Elman, 1990; Olivier, 1968; Christiansen, Allen, & Seidenberg, 1998).

Like certain previous models, our computational model uses actual child-directed speech as segmentation materials. Specifically, we have taken the adult utterances from the Brown files (1973) in the CHILDES corpus (MacWhinney, 1995). We obtained the phonetic transcriptions of words from the CMU Pronunciation Dictionary (Version 0.6).[5] In the CMU Pronunciation Dictionary, lexical stress information is preserved by numbers: 0 for stressless, 1 for primary stress, 2 for secondary stress, etc. For instance, "cat" is represented as "K AE1 T", "catalog" is "K AE1 T AH0 L AO0 G", and "catapult" is "K AE1 T AH0 P AH2 L T". For each word, we grouped the phonetic segments into syllables. This process is straightforward, at least for English, by the use of the principle "Maximize Onset", which maximizes the length of the onset as long as it is valid consonant cluster of English, i.e., it conforms to the phonotactic constraints of English. For example, "Einstein" is "AY1 N S AY0 N" as segments and parsed into "AY1N STAY0N" as syllables: this is because /st/ is the longest valid onset for the second syllable containing "AY0" while /nst/ is longer but violates the English phonotactics.

Finally, we removed the spaces (and punctuation) between words, but the boundaries between utterances–as indicated by line breaks in CHIDLES–are retained. Altogether, there are 226,178 words, consisting of 263,660 syllables. The learning material is therefore a list of unsegmented syllable sequences, and the learner's task is to find word boundaries that group substrings of syllables together.

[5] Some words have multiple pronunciations in the Dictionary; in these cases, we consistently used the first entry.

4 Statistical Learning is Ineffective

Our approach to modeling is a modular one. We attempt to implement a succession of models that incorporates, one at a time, the cues for word segmentation reviewed in section 2. In our view, implementing multiple cues simultaneously could obscure their respective contributions to word segmentation, as the current research does not give a clear guide on how to "weigh" the competing strategies. For the purpose of the present study, implementation of strategies is halted if the performance of the model is deemed satisfactory. If none of the strategies proves effective, then we consider alternative strategies that have not been suggested in the existing literature.

We start with an evaluation of statistical learning with local minima (Chomsky, 1955; Saffran et al. 1996).

4.1 Modeling Statistical Learning

The simplest cue to implement is statistical learning using transitional probabilities, and there are good reasons to believe that statistical learning is also the first strategy available to a child learner. This choice is both logical and empirical. Among the strategies reviewed in section 2, only statistical learning avoids the chicken-and-egg problem, as it is the only language-independent strategy for finding words. (We will return to the case of isolated words in section 5.1.) In addition, recent work has turned up empirical evidence that statistical learning may be the very process that gets word segmentation off the ground. For instance, Johnson & Jusczyk (2001) show that, in word segmentation task using artificial language, 9-month-old infants use stress cues over statistical cues when both types of information are available. However, Thiessen & Saffran (2003) showed that younger (7-month-old) infants show the opposite pattern, where statistical cues take priority. And they conclude that statistical learning may provide the seed words from which language particular stress patterns may be derived.

The modeling of statistical learning is straightforward, though it may be useful to make the details of our implementation clear. The model consists of two stages: training and testing. During the training stage, the learner gathers transitional probabilities over adjacent syllables in the learning data. The testing stage does not start until the entire learning data has been processed, and statistical learning is applied to the same data used in the training stage. This is markedly different from the standard methodology in computational linguistics, where the corpus data is typically divided into two separate portions for training and testing respectively: that is, the model is tested on a *different* data set from the one on

which it is trained. Our approach may be seen as giving the statistical learner some undue advantage. However, we believe our approach is justified, if only as a matter of necessity. If one were to use novel data for testing, there is a high likelihood that many syllable pairs in testing are never attested in training: this is known as the "sparse data" problem in computational linguistics (Jelinek & Mercer, 1980). How should the statistical learner proceed when confronted with a syllable pair never seen before? Computational linguistics offers a variety of techniques for smoothing over the missing probability mass (Chen & Goodman, 1996) though it is not clear how these techniques are applicable as a model of human language processing. Moreover, a poor choice of smoothing technique could unfairly affect the performance of statistical learning.

Another technical detail also needs to be spelled out: the TPs are gathered without stress information. That is, when counting syllable frequencies, the learner does not distinguish, say, a stressed syllable /b**a**/ from among the unstressed one.[6] This assumption is again out of necessity: as it stands, the corpus contains 58,448 unique syllable pairs according to this counting procedure. If the stress levels of syllables are taken into consideration, the statistical learner must keep multiple copies of a syllable, which leads to the explosion of possibility transitional probability pairs.

After the TPs are gathered over the entire learning data, the testing stage starts. For each utterance, the learner traverses the syllables from left to right and looks up the TPs between successive pairs. A word boundary is postulated at the place of a local minimum. That is, there is a word boundary AB and CD if if TP(A→B) >TP(B→C) < TP(C→D). The conjectured word boundaries are then compared against the target segmentation. Scoring is done for each utterance, using the definition of precision and recall in (1)

4.2 Results from Statistical Learning

Modeling shows that the statistical learning (Saffran et al., 1996) does not reliably segment words such as those in child-directed English. Specifically, precision is 41.6%, recall is 23.3%. In other words, about 60% of words postulated by the statistical learner are not English words, and almost 80% of actual English words are not extracted. This is so even under favorable learning conditions:

- the child has syllabified the speech perfectly,

[6] Which is of course not to say that infants cannot distinguish them perceptually. The fact that young learners can acquire the Metrical Stress Strategy suggests that they clearly are capable of recognizing stress levels.

- the child has neutralized the effect of stress among the variants of syllables, which reduces the sparse data problem,
- and the data for segmentation is the same as the data used in training, which eliminates the sparse data problem

We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason. A necessary condition on the use of TP local minima to extract words is that words must consist of multiple syllables. If the target sequence of segmentation contains only monosyllabic words, it is clear that statistical learning will fail. A sequence of monosyllabic words require a word boundary after each syllable; a statistical learner, on the other hand, will only place a word boundary between two sequences of syllables for which the TPs within are higher than that in the middle. Indeed, in the artificial language learning experiment of Saffran et al. (1996) and much subsequent work, the pseudowords are uniformly three syllables long. However, the case of child-directed English is quite different. The fact that the learning data consists of 226,178 words but only 263,660 syllables suggests that the overwhelming majority of word tokens are monosyllabic. More specifically, a monosyllabic word is followed by another monosyllabic word 85% of time. As long as this is the case, statistical learning cannot work.

One might contend that the performance of statistical learning would improve if the learner is given more training data to garnish more accurate measures of transitional probabilities. We doubt that. First, the problem of monosyllabic words, which causes the local minima method problems, will not go away. Second, statistics from large corpora may not help at that much. In realistic setting of language acquisition, the volume of learning data is surely greater than our sample but is not unlimited before children would use statistical learning to segment words by the 7-8th month (Jusczyk & Aslin, 1995). Empirically, we found that the TPs stablize fairly quickly, which means more data may not give much better TPs. A reasonable measure of the informativeness from the quantity of training data is to calculate the sum of the absolute value changes in the TPs ($\sum |\Delta_{\text{TP}}|$) over some time interval of say, every 1000 syllables processed during training. If this number changes very little, then we know that the TP values have stablized. As Figure 1 illustrates, the change in TPs is fairly rapid for the first 100,000 syllables processed: this is expected because many syllable pairs at this stage are new. However, $\sum |\Delta_{\text{TP}}|$ slows down considerably afterwards, indicating that the values of TPs are no longer changing very much. Indeed, after 100,000 syllables, $\sum |\Delta_{\text{TP}}|$ hovers around the neighborhood of 10 to 30 for every 1000 syllables processed. It is not straightforward to estimate precisely how much a particular TP changes

during each training interval. For instance, TP(A→B) will change as long as the learner sees AB occurring jointly as well as A occurring with something other syllable. In other words, though each training interval consists of 1,000 syllables, many thousands of TPs may be changing during this time. On average, then, each TP may change only by a truly miniscule value: 10–30 divided by 58,448, the total number of unique syllable pairs. When we expand the interval to 10,000 syllables, during which it is far more likely that the value of every TP would change, we obtain the average $|\Delta_{TP}|$ to be 0.0028: again, a minute adjustment of the TP statistics. These calculations suggest that by using a larger set of training data, the estimated TPs will certainly be closer to their realistic values, but the improvement is likely to be marginal.

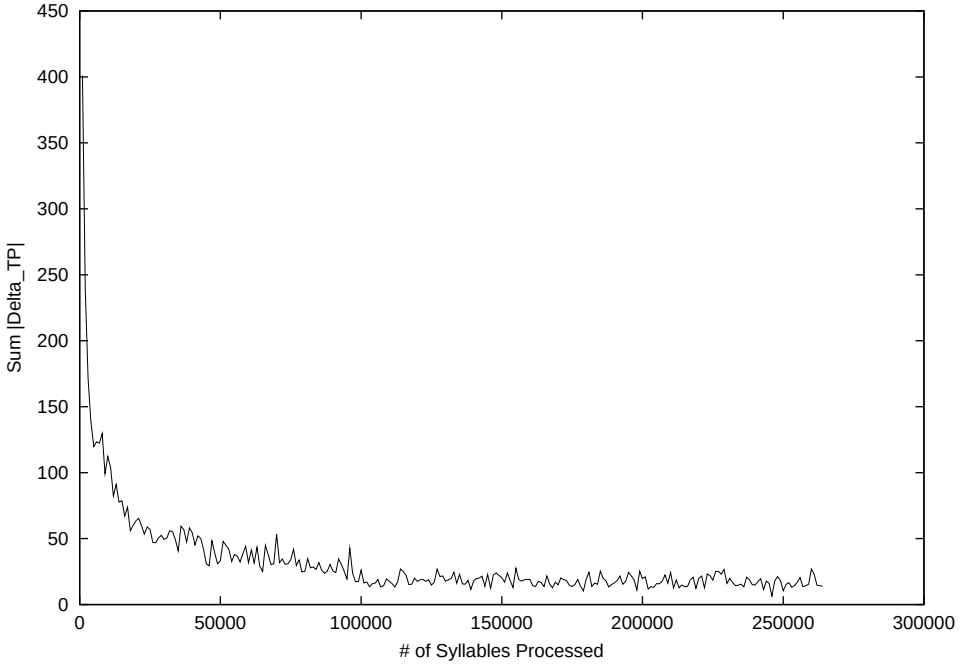

Figure 1: $\sum |\Delta_{TP}|$ during the course of training. Note the rapid stablization of TPs.

4.3 Corpus Statistics and Statistical Learning

Our modeling results differ significantly from those from a recent study of statistical learning (Swingley, 2005), where it is claimed that on the basis of English and Dutch child-directed corpora, statistical regularities do correlate with word boundaries. In this section, we highlight some differences between these two lines of work.

One source for such contrasting findings has to do with the methodology. Swingley (2005) carried out "off-line" study of statistical regularity in the input while the present work models the on-line process of segmentation using statistical regularities. But as noted earlier, the poor performance of statistical learning in our simulation has to do with the presence of many monosyllabic words, which cannot be reliably with the means of TP local minima. Unless the parents in Swingley's corpus spoke a great deal more "big" words, modeling results ought to be fairly consistent between the two models. It turns out that the performance differences lie in the segmentation mechanisms that these two models implement.

Swingley's corpus study makes use of multiple source of statistical information. Specifically, it maintains three kinds of information units: single syllables, adjacent syllable pairs (bigrams), and adjacent syllable triples (trigrams). Four types of statistical information are accumulated: the frequencies of these three units, in addition to he mutual information between adjacent syllable pairs (I_{AB}).[7] These numbers are then ranked along a percentile scale, much like standardized tests. For instance, if a syllable is at least as frequent as 70% of all syllables, it receives a percentile score of 70, and if the mutual information between a syllable pair is at least as high as those of 60% of all syllable pairs, it receives a percentile of 60%. Let R_A, R_{AB}, R_{ABC} be the percentile score of single, double, and triple syllables based on frequency, and RI_{AB} be that of mutual information between A and B. Now a decision procedure is applied according to a percentile cutoff threshold (θ), and units are considered to be words according to the following criteria:

(3) a. if $R_A > \theta$, then A is a word.
b. if $R_{AB} > \theta$ and $RI_{AB} > \theta$, then AB is a word.
c. if $R_{ABC} > \theta$, $R_{AB} > \theta$, and $R_{BC} > \theta$, then ABC is a word.

Roughly speaking, the syllables and syllable sequences whose presence dominte the learning data are conjectured to be words. By experimenting with a range of values for θ, Swingley reports fairly high precision when θ is in the range of 50%–100% percentile. While

[7] I_{AB} is defined as $\log_2 \frac{p(AB)}{p(A)p(B)}$, which is similar, though not equivalent, to TP(A→B); see Aslin, Saffran, & Newport (1998) and Swinney (1999).

our model of statistical learning is a direct implementation of a specific statistical learning model (Saffran et al, 1996) and much subsequent work, Swingley's statistical criteria for words are quite different and build into a number of assumptions and mechanisms, some of which have not been empirically tested. Ultimately, the results from the two lines of works do not form a useful comparison.

First, the limit of word length to 3 syllables is arbitrary and longer words cannot be extracted. Yet such a limit may be necessary for Swingley's corpus study. Severe sparse data problems will rise if length of syllable consequences increases to four or higher: vast majority of syllable four-grams will have zero occurrences in any speech corpus, and those that do occur will likely have very low number of occurrences. In engineering practice, virtually all statistical models of language stop at trigrams, even when the amount of training data is far higher than those used in Swingley's corpora, or what a child could conceivably likely encounter during language acquisition.

Second, though we find the percentile-based criteria in (3) to be interesting, we are not aware of any direct experimental evidence suggesting that they might be used in actual word segmentation. Mutual information, which is similar to transitional probability, may indeed be a cue for word boundary and one which the learner can probably exploit. However, whether high frequency alone (of syllables, or syllable sequences) correlates with word boundaries remain to be seen. Moreover, there is is little reason to suppose that the learner will treat frequency information and mutual information with the same percentile threshold (θ).

Third, the very status of θ remains unclear. Swingley (2005) does not provide the raw data but from the graph (his Figure 1: p100), it appears that θ=80% yields highest precision results. But how does the learner determine the best values of θ? So far as we can tell, there are two possibilities. One is that the optimal value of θ is innately available. The other possibility that the optimal θ is obtained via some sort of learning procedure as the result of word segmentation–which again seems to require a set of seed words to begin with, and such a procedure that determines the value of θ needs to be spelled out. In either case, independent motivation is required.

Finally, issues remain in the interpretation of Swingley's results. It is true that overall precision may be quite high for certain values of θ but it is worth noting that most of the three-syllable words determined by Swingley's criteria are wrong: the precision is consistently under 25-30% (Swingley, ibid; Figure 1) regardless the value of θ. Moreover, the statistical criteria in (3) produce very low recalls. Swingley does not provide raw data but the performance plots in his paper show that the maximum number of correctly extracted

words does not appear to exceed 400-500. Given that Swingley's corpus contains about 1,800 distinct word types (ibid; p96), the recall is at best 22-27%.

In sum, the corpus study of Swingley (2005) considers a number of statistical regularities that could be extracted in the linguistic data. The extraction of these regularities, and the criteria postulated for finding word boundaries, are not always supported by independent evidence. Even if these assumptions were motivated, the segmentation results remain poor, particularly for recall and longer words. We therefore do not consider this work to affect our conclusion that statistical learning fails to scale up in a realistic setting of language learning.

Before we present our own proposal for word segmentation, we would like to reiterate that our results strictly pertain to one specific type of statistical learning model, namely the local minima method of Saffran et al. (1996), by far the best known and best studied proposal of statistical learning in word segmentation. We are open to the possibility that some other, known or unknown, statistical learning approach may yield better or even perfect segmentation results, and we believe that computational modeling along the lines sketched out here may provide quantitative measures of their utility once these proposals are made explicit.

5 Segmentation under Linguistic Constraints

Now the segmentation problem appears to be in a somewhat precarious state. As discussed earlier, statistical learning appears to be the only language-independent way of extracting words from which language-particular strategies can be developed. However, modeling results show that statistical learning does not–and cannot, at least for English–scale up to a realistic setting of language learning. How *do* children segment words?

In this section, we consider some computationally simple, psychologically plausible, and linguistically motivated constraints that complement existing proposals of word segmentation. Some of these constraints may be innate and domain-specific knowledge of language, while others are probably general principles of symbolic information processing. They lead to marked improvement in performance and may contribute to the understanding of the word segmentation process.

5.1 Constraint on Word Stress

Modern machine learning research (Gold, 1967; Valiant, 1984; Vapnik, 1995), together with earlier observation on the limitations of associationist and inductive learning (Chomsky 1959, 1975), suggest that constraints on the learning space and the learning algorithm are essential for (realistically efficient) learning. When a domain neutral learning model–e.g., statistical learning–fails on a domain specific task–e.g., word segmentation–where children clearly succeed, it is likely that children are equipped with knowledge and constraints specific to the task at hand. It is then instructive to identify such constraints to see to what extent they complement, or even replace, domain neutral learning mechanisms.

Which brings us to one important source of information for word segmentation that has been left unexplored, namely, single word utterances. Though most researchers share the intuition that isolated words are easy to learn, we do not know of any proposal that can reliably identify isolated words as such. So consider the following self-evident linguistic principle:

(4) The Unique Stress Constraint (USC): A word can bear at most one primary stress (a strong syllable).

The USC virtually follows from the definition of the phonological word (Chomsky & Halle, 1968; Liberman & Prince, 1977). Its pervasiveness can be appreciated if we–at least those of us that are old enough–warped ourselves back to the 1977 premiere of *Star Wars*. Upon hearing "chewbacca" and "darthvader" for the very first time, it must have been immediately clear that the former utterance is one word, the latter is two (though whatever they meant was altogether a different matter). Both sequences are three syllables so length is not a useful guide. Yet "chewbacca" contains only one primary stress, which falls on /ba/, whereas "darthvader" contains two primary stresses, which fall on /darth/ and /va/ respectively: USC immediately segments the utterances correctly. Likewise, we believe that USC provides important clues for word boundaries for an infant learner whose situation is not much different from first time *Star Wars* viewers.

First, and most directly, USC may give the learner many isolated words for free. This, so far as we know, constitutes the only known mechanism that takes advantage of the abundance of single word utterances (Brent & Siskind, 2001). Specifically, if the learner hears an utterance that contains exactly one primary stress, she can immediately conclude that such utterance, regardless of its length, can and can only be a single word. Moreover, the segmentation for multiple word utterance can be equally straightforward under USC. (In section 6.1, we discuss a relaxation of this assumption.) Take a sequence $W_1 S_1 S_2 S_3 W_2$,

where W stands for a weak syllable and S stands for a strong syllable. A learner equipped with USC will immediately know that the sequence consists of three words: specifically, W_1S_1, S_2, and S_2W_2.

Second, and somewhat indirectly, USC can constrain the use of statistical learning. For example, the syllable consequence $S_1W_1W_2W_3S_2$ cannot be segmented by USC alone, but it may still provide highly informative cues that facilitate the application of other segmentation strategies. For instance, the learner knows that the sequence consists of two words, as indicated by two strong syllables.[8] Moreover, it also knows that in the window between S_1 and S_2 must lie a word boundary (or boundaries)–and *that* may be what statistical learning using local minima may be able to locate. As we show later, constrained application of statistical learning leads to enormous improvement on its effectiveness.

5.2 Remarks on Unique Stress Constraint

A number of remarks are in order before we present the modeling results using USC (in combination with other segmentation strategies). First, it is assumed that the learner be able to distinguish strong syllables and weak syllables. This is surely plausible in the light of the fact that the Metrical Segmentation Strategy is operative in 9-month-old, or perhaps even younger, infants. To find the dominant stress patterns involves at least (a) the recognition of strong vs. weak syllables, (b) a collection of reliably segmented words (through whatever means) and their respective stress patterns, and (c) a certain computational procedure that identifies the dominant pattern among the set of words. To make use of USC, only (a) is assumed. Hence, USC is a weaker assumption on the part of the learner than the Metrical Segmentation Strategy.

Which raises the second remark: *How* does the child identify metrical patterns in speech? That they do is beyond doubt, though it is by no means clear what kind of processes are involved. It is tempting to assume that pitch peaks directly pick out strong syllables (see Sluijter, van Heuven, & Pacily, 1996), but the search for direct acoustic correlates of stress has been largely illusive (Hayes, 1995). It is likely, then, that the identification of stresses involves, beyond an obvious perceptual component, certain cognitive/structural representation of speech and subsequent computations, which may be domain-specific phonological

[8] Perhaps more than two. Note that USC as formulated here does not assert that a word *must* have a primary stress. A language may contain a closed and fairly small set of functional words–some prepositions, auxiliaries, and determiners, in the case of English–that do not bear primary stresses. For instance, "drin-king-the-cham-pagne" is a sequence with two strong syllables at the two ends but nevertheless consists of three words ("drinking the champagne"). See section 5.3.2 for how functional words may be extracted from speech by statistical as well as non-statistical means.

knowledge. In any case, the identification of metrical pattern remains an open question though children have no trouble solving it at a very young age.

Third, USC, unlike the Metrical Segmentation Strategy, is a universal constraint on word boundaries rather than a language particular one. It therefore does not run into the chicken-and-egg dilemma noted earlier (cf., Thiessen & Saffran, 2003). In fact, along with statistical learning, USC is the only universally applicable procedure that can bootstrap a sufficiently reliable set of words from which language-particular strategies can be derived.

Fourth, the ideal application of USC presupposes that primary stresses of words are readily available in spoken language. In our modeling, we have assumed that every primary stress, which is obtained from the CMU Pronunciation Dictionary, is preserved in the learning data. However, complications may arise if the primary stresses of some words are lost in casual speech: for example, a single word utterance will not be recognized as such if the primary stress is not available. If so, the USC may not be as usual as in the ideal case. However, as discussed in section 5.3.2, this does not necessarily pose a serious problem if the application of USC is complemented by a simple "agnostic learning" strategy.

Finally, we will not speculate further on the functional motivations for USC other than pointing out that it is likely an innate constraint, perhaps reflecting the general principle that prosody marks linguistically significant units (morphemes, words, constituents, phrases, etc.) One reason for supposing so is that USC is a negative principle, which is known to create learnability problems under the standard assumption of no negative evidence (Brown & Hanlon, 1970; Berwick, 1985; Lasnik, 1989). Another reason is that if USC were learned from experience, the child must, at the minimum, have extracted a set of seed words already with language-independent means. However, statistical learning, which is the only other candidate strategy that would fit the bill, cannot generate such a set accurately, as the modeling results show. Hence another reason for supposing the innateness of USC.

5.3 Constraints are Effective for Segmentation

We now describe several variants of word segmentation models that make use of USC.

5.3.1 Statistical Learning under USC

In the first model, we apply statistical learning when USC does not automatically identify word boundaries. In the training stage, TPs are gathered as before. In the testing stage, the learner scans a sequence of input syllables from left to right:

(5) a. If two strong syllables are adjacent (i.e., "... S_1S_2 ..."), a word boundary is postulated in between.

b. If there are more than one (weak) syllables between two strong ones (i.e, $S_1W...WS_2$), then a word boundary is postulated where the pairwise TP is at the local mininum.

(5a) straightforwardly solves the monosyllablic word problem–by avoiding statistical learning altogether. Certain complications may arise in (5b). It is possible that multiple local minima exist between $S_1...S_2$, which would lead the model to postulate multiple word boundaries. This is sometimes justified, if the sequence of weak syllables happens to contain a stressless functional word; see footnote (8).

The improvement in segmentation results is remarkable: when constrained by USC, statistical learning with local minimum achieves precision of 73.5% and recall of 71.2%. In fact, these figures are comparable to the highest performance reported in the literature (Brent, 1999a), which nevertheless uses uses a computationally prohibitive algorithm that iteratively optimizes over the entire lexicon. By contrast, the computational complexity of the present model is exactly that of computation of transitional probabilities, which appears to be less costly but still leaves much to be desired.

5.3.2 Algebraic Learning: Quick but not dirty

Once the linguistic constraint USC is built in, we are in position to explore alternative segmentations that do not make use of statistical information at all. We do so out of the concern that the computational burden of statistical learning is no means trivial. English, for instance, has a few thousand syllables, and the number of transitional probabilities that the learner must keep track of is likely enormous. (In our relative small corpus, there are 58,448 unique syllable pairs.) Furthermore, given the definition of transitional probability—TP(A→B) = Pr(AB)/Pr(A)–it is clear that whenever the learner sees an occurrence of A, she would have adjust the values of all B's in TP(A→B). This means that, in a realistic setting of language acquisition, the learner must adjust the values of potentially *thousands* of TPs for *every* syllable processed in the input. Though we do not know how infants do so, we do consider it worthwhile to evaluate computationally less expensive but potentially more reliable segmentation strategies.

We pursue a number of variants of the traditional idea that recognition of known words may bootstrap for novel words (Peters, 1983; Pinker, 1984). This is a plausible strategy on a number of grounds. To begin with, young learners have memory for familiar sound

patterns, as shown by Juscyzk & Hohne (1997) that 8-month-old infants can retain sound patterns of words in memory. Therefore, if the child has learned the word "big", she might be able to recognize "big" in the utterance "bigsnake" and extract "snake" as a result. For concreteness, call this bootstrapping process *subtraction* (Gambell & Yang, 2003). Furthermore, the subtraction strategy is evidenced by familiar observations of young children's speech. The irresistible segmentation errors (e.g., "I was have" from *be-have*, "hiccing up" from *hicc-up*, "two dults" from *a-adult*) suggest that subtraction does take place (cf. Peters, ibid). Recent work (Bortfeld, Morgan, et al., 2005) demonstrates that infants as young as 6 months old may use this bootstrapping strategy. For word sequences such as XY, where Y is a novel word, infants prefer those that are paired with a familiar X, such as "Mommy", the child's name, and others that may be developmentally appropriate for this stage.

Under algebraic learning, the learner has a lexicon which stores previously segmented words. No statistical training of the TPs is used. As before, the learner scans the input from left to right. If it recognizes a word that has been stored in the lexicon, it puts the word aside and proceeds to the remainder of the string. Again, the learner will use USC to segment words in the manner of (5a): in our modeling, this constraint handles most cases of segmentation. However, USC may not resolve word boundaries conclusively. This happens when the learner encounters $S_1 W_1^n S_2$: the two S's stand for strong syllables, and there are n syllables in between, where W_i^j stands for the substring that spans from the ith to the jth weak syllable. In the window of W_1^n, two possibilities may arise.

(6) a. If both $S_1 W_1^{i-1}$ and $W_{j+1}^n S_2$ ($i < j$) are, or are part of, known words on both sides of $S_1 W_1^n S_2$, then W_i^j must be a word,[9] and the learner adds W_i^j as a new word into the lexicon. This is straightforward.

b. Otherwise, a word boundary lies somewhere in W_1^n, and USC does not provide reliable information. This is somewhat more complicated.

To handle the case of (6b), we consider two variants of algebraic learning:

(7) a. **Agnostic**: the learner ignores the strings $S_1 W_1^n S_2$ altogether and proceeds to segment the rest of the utterance. No word is added to the lexicon.

b. **Random**: the learner picks a random position r ($1 \leq r \leq n$ and splits W_1^n into two substrings W_1^r and W_{r+1}^n as parts of the two words containing S_1 and S_2 respectively.[10] Again, no word is added to the lexicon.

[9] Since it does not contain a strong syllable, it is most likely a functional word.

[10] These two resulting words may include materials that precede S_1 and materials that follow S_2, should such segmentation not be prohibited by USC.

The logic behind the agnostic learner is that the learner is non-committal if the learning data contains uncertainty unresolvable by "hard" linguistic constraints such as USC.[11] This could arise for two adjacent long words such as "l**a**nguage**a**cquisition", where two primary stresses are separated by multiple weak syllables as in the case of (6b). It could also arise when the input data (casual speech) is somewhat degraded such that some primary stresses are not prominently pronounced, as discussed in 5.2. While the agnostic learner does not make a decision when such situations arise, it can be expected that the words in the sequence $S_1 W_1^n S_2$ will mostly like appear in combinations with other words in future utterances, where USC may directly segment them out. The random learner is implemented as a baseline comparison, though we suspect that in actual language acquisition, the learner may invoke the language-specific Metrical Segmentation Strategy, rather than choosing word boundaries randomly, in ambiguous contexts such as $S_1 W_1^n S_2$.

Note further that in both versions of the algebraic model, no word is added to the lexicon when the learner is unsure about the segmentation; that is, both algebraic learners are conservative and conjectures words only when they are certain. This is important because mis-segmented words, once added to the lexicon, may lead to many more mis-segmentations under the subtraction algorithm. In section 6.1, we discuss ways in which this assumption can be relaxed.

Table 1 summarizes the segmentation results from the two algebraic learners, along with those from earlier sections on statistical learning.

Model	Precision	Recall	F-measure ($\alpha = 0.5$)
SL	41.6%	23.3%	0.298
SL + USC (5)	73.5%	71.2%	0.723
Algebraic agnostic (7a)	85.9%	89.9%	0.879
Algebraic random (7b)	95.9%	93.4%	0.946

Table 1: Performance of four models of segmentation. SL stands for the statistical learning model of Saffran et al. (1996), while the other three models are described in the text.

It may seem a bit surprising that the random algebraic learner yields the best segmentation results but this is not unexpected. The performance of the agnostic learner suffers from deliberately avoiding segmentation in a substring where word boundaries lie. The random learner, by contrast, always picks out *some* word boundary, which is very often correct. And this is purely due to the fact that words in child-directed English are generally short. Taken

[11] A comparable case of this idea is the Structural Triggers Learner (Fodor, 1998) in syntactic parameter setting. We thank Kiel Christianson for pointing out this connection.

together, the performance of the algebraic learners entails that (a) USC and subtraction work for most of words, and (b) when they don't, there are only very few weak syllables in the window between two strong ones (n is small in W_1^n) such that a random guess is not far off. On the other hand, while scoring is used here to evaluate the performance of segmentation model, which punishes the agnostic learner. In real life, however, it is perfectly fine to move on the next utterance when the learner is unsure about the segmentation of the present utterance.

Based on these results, we conjecture that algebraic learning is a reasonable strategy for word segmentation and ought to be further evaluated in experimental settings. In the concluding section of this paper, we will suggest a number of ways in which the algebraic learner can be modified and improved.

6 Conclusion

Our computational models complement experimental research in word segmentation. The main results can be summarized as follows.

- The segmentation process can get off the ground only through the use of language-independent means: experience-independent linguistic constraints such as USC and experience-dependent statistical learning are the only candidates among the proposed strategies for word segmentation.
- Statistical learning does not scale up to realistic settings of language acquisition.
- Simple principles on phonological structures such as USC can constrain the applicability of statistical learning and improve its performance, though the computational cost of statistical learning may still be prohibitive.
- Algebraic learning under USC, which has trivial computational cost and is in principle universally applicable, outperforms all other segmentation models.

We conclude with some specific remarks on word segmentation followed by a general discussion on the role of statistical learning in language acquisition.

6.1 Directions for future work

One line of future work is to further disentangle the various segmentation strategies proposed in the literature. For example, the present models have not considered the role of

co-articulation cues in segmentation (Jusczyk, Houston, & Newsome, 1999; Johnson & Jusczyk, 2001) Our work has, however, helped to clarify some logical issues in the use of the Metrical Segmentation Strategy. The phonological principle USC is sufficient for generating a set of seed words, from which the language-specific Metrical Stress Strategy can be derived: just how children extract such statistical tendencies is still unknown, particularly when the full range of stress systems in the world's languages is taken into account. On the other hand, it may be useful to pitch USC against the Metrical Segmentation Strategy along the lines of Johnson & Jusczyk (2001; cf. Thiessen & Saffran, 2003). It would be interesting to see whether an appropriately aged learner, who has mastered the the Metrical Segmentation Strategy, favors language-specific or language-independent cues when both cues are available.

Another important question concerns how, or how well, infant learners can syllabify the input speech. Virtually all segmentation strategies (statistical learning, the Metrical Segmentation Strategy, USC, etc.) are dependent on the learner's ability to parse segments into syllables. It may be worthwhile to explore how syllabification can be achieved by phonetic and articulatory cues (Browman & Goldstein, 1995; Krakow, 1999) as well as the traditional conception of the syllable as a structured unit that surrounds the vowel with language-specific phonotactic knowledge, e.g., onset consonant clusters.

One of the potential problems with the algebraic learners is that they learn *too* fast. A learner equipped with USC can segment words reliably and very rapidly, and previously segmented words stored in the lexicon may lead to rapid segmentation of novel words under the subtraction strategy. In a human learner, however, reliable segmentation may take much longer. It is straightforward to augment our model to bring it a step closer to reality. For instance, one may add a frequency-dependent function that controls the construction of the lexicon. Specifically, addition of words to the lexicon as well as retrieval (in subtraction-based learning) from it are probabilistic. The retrieval of a word from the lexicon is determined by a function that increases when that word has been extracted by the input: the net effect is the learner may sometimes recognize a word but sometimes fail to do so. With a probabilistic function, one may remove the conservative requirement on the algebraic learners. A mis-segmented word can be added to the lexicon, but if it appears very infrequently, the negative effect of its use in subtraction is negligible. This then directly captures the frequency-dependent characteristics of word segmentation. We are currently exploring the behavior of the probabilistic learner in comparison to the time course of word segmentation in human learners.

Finally, it is important to examine the segmentation problem in a cross-linguistic setting.

In agglutinative languages (e.g., Turkish) and polysynthetic languages (e.g., Mohawk), the notion of "word" is inseparable from the morphosyntactic system, and is thus considerably different from the more clear-cut cases like Modern English. Consequently, word segmentation in these languages may involve simultaneous acquisition at other linguistic levels. It remains to be seen how any model of segmentation can generalize to these cases.

6.2 Statistical Learning and Language Acquisition

Statistical learning (Saffran et al., 1996) surely ranks among the most important discoveries of our cognitive abilities. Yet it remains to be seen, contrary to a number of claims (Bates & Elman, 1996; Seidenberg, 1997, etc.), whether statistical learning serves as an alternative to innate and domain-specific knowledge of language (Universal Grammar, broadly speaking). In addition, as the present study shows, it remains an open question whether statistical learning using local minima is used in actual word segmentation in the first place. In conclusion, we discuss a number of contemporary issues regarding domain specific knowledge and domain general learning mechanisms.

First, does the ability to learn diminish the need for Universal Grammar? Here we concur with Saffran et al. (1997), who are cautious about the interpretation of their results. Indeed, it seems that the success of learning strengthens, rather than weakens, the claim of Universal Grammar, or at least innate cognitive/perceptual constraints that must be in place in order for learning to occur.

Recall the classic poverty of stimulus argument for innateness (Chomsky, 1975): the case of auxiliary inversion in English interrogative questions (e.g., "John is nice" → "Is John nice?"). It is no doubt that this construction is learned, as not all language invert auxiliaries. But it is precisely the fact that auxiliary inversion *is* learned that establishes the argument for the innate principle of structure dependency in syntax. The learner could have learned many other logically possible transformations; the fact that they don't (Crain & Nakayama, 1987) suggests that they tacitly know what kind of regularities in question formation to keep track of.[12]

The same logic applies to the success of statistical learning in segmenting artificial language: it presupposes the learner knowing what kind of statistical information to keep track of. After all, an infinite range of statistical correlations exists: e.g., What is the probability

[12]Recent corpus studies (e.g., Pullum & Scholz, 2002) claim that the learner may have access to disconfirming evidence against incorrect hypotheses about question formation. First, there are various problems with how these studies gather corpus statistics. More important, even if the child does have access evidence for structure dependency, these studies fail to establish that such evidence is sufficient for eliminating incorrect hypotheses by the developmental stage they were tested for auxiliary inversion. See Legate & Yang (2002) for details.

of a syllable rhyming with the next? What is the probability of two adjacent vowels being both nasal? The fact that infants can use statistical learning in the first place entails that, at the minimum, they know the relevant unit of information over which correlative statistics is gathered: in this case, it is the syllables, rather than segments, or front vowels, or labial consonants.

A number of questions then arise. First, How do they know the primacy of syllables? It is at least plausible that the primacy of syllables as the basic unit of speech is innately available, as suggested in neonate speech perception studies (Bijeljiac-Babic, Bertoncini, & Mehler, 1993; cf., (Bertocini & Mehler, 1981; Jusczyk, 1997; Jusczyk, Kennedy, & Jusczyk, 1998; Eimas, 1999). Second, where do the syllables come from? While the experiments in Saffran et al. (1996) used uniformly CV syllables (and three syllables per word), many languages, including English, make use of a far more diverse range of syllabic types. Third, syllabification of speech is far from trivial, involving both innate knowledge of phonological structures as well as discovering language-specific phonotactic constraints. Finally, in a realistic learning environment, statistical information is bound to be far less perfect than the constructed in artificial language learning experiments, where within-word transitional probabilities are uniformly 1 and across-word transitional probabilities are uniformly 1/3. All of these are practical problems that the child has to solve before the claim of statistical learning is used in word segmentation can be established.

Indeed, we are not aware of any experimental study that uses realistic language input to test the effectiveness of statistical learning. (We do appreciate the likely practical difficulty with such experiments.) This leaves us with computational modeling–one which implements psychologically plausible mechanisms–as a reliable evaluation of the scope and limits of statistical learning. The fact that statistical learning does not extract words reliably, and the fact that simple linguistic constraints and algebraic learning do extract words reliably, raise the possibility that statistical learning may not be used in word segmentation at all. If an alternative learning strategy is simple, effective, and linguistically and developmentally motivated, it is reasonable to expect the child to use it too.

It is worth reiterating that our critical stance on statistical learning refers only to a specific kind of statistical learning that exploits local minima over adjacent linguistic units (Saffran et al., 1996). Rather, we simply wish to reiterate the conclusion from decades of machine learning research that no learning, statistical or otherwise, is possible without appropriate prior assumptions on the representation of the learning data and a constrained hypothesis space. Recent work on the statistical learning over non-adjacent phonological units has turned out some interesting limitations on the kind of learnable statistical correla-

tions (Newport & Aslin, 2004, Aslin, Newport, & Hauser, 2004; Toro, Sinnett, & Soto-Faraco, In press; Peña, Bonatti, Nespor, & Mehler, 2002; for visual learning tasks, see Tucker-Brown, Junge, & Scholl, submitted; Catena & Scholl, submitted). The present work, then, can be viewed as an attempt to articulate the specific linguistic constraints that might be built in for successful word segmentation to take place. Indeed, in other work (Yang 1999, 2002, 2004), we have incorporated domain-general probabilistic learning models (Bush & Mosteller, 1955; Atkinson, Bower, & Crouthers, 1965; Tversky & Edwards, 1966; Herrnstein & Loveland, 1975) into the bona fide problem of domain-specific learning, the setting of syntactic parameters in the Principles and Parameters framework (Chomsky, 1981). Even the algebraic learning we proposed here, which requires the learner to recognize identity of occurrences, resembles the pattern extraction process in early infant learning (Marcus, Vijayan, Rao, & Vishton, 1999), which may well be domain neutral as it has been replicated in other species (Hauser, Weiss, & Marcus, 2002). In all these cases, it is important to separate the learning mechanism, which may be domain general, from the learning constraints, which may domain specific.

References

Atkinson, R., Bower, G, & Crothers, E. (1965). *An Introduction to Mathematical Learning Theory*. New York: Wiley.

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

Aslin, R., Woodward, J., LaMendola, N, & Bever, T. (1996). Models of word segmentation in speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to gramamr in early acquisition*. Hillsdale, NJ: Erlbaum. 117-134.

Bates, E., & Elman, J. (1996). Learning rediscovered. *Science*, 274, 1849-1850.

Batchelder, E. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83, 167-206.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General*, 117, 2133.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, 4, 247-260.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances?. *Developmental Psychology*, 29, 711721.

Berwick, R. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

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. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34, 71-106.

Brent, M. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science*, 3, 294-301.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93125.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, B33-44.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49, 155-180.

Browman, C., & Goldstein, L. (1995). Dynamics and articulatory phonology. In T. van Gelder & R. Port (Eds.) *Mind as Motion*. Cambridge, MA: MIT Press. 175-193.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley. 11-54.

Bush, R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 68, 313-323.

Catena, J., & Scholl, B. (Submitted). The onset and offset of visual statistical learning. Manuscript, Yale University.

Chambers, K., Onishi, K., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87, B69-B77.

Chen, S., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Santa Cruz, CA. 310-318.

Chomsky, N. (1955). The logical structure of linguistic theory. MIT Humanities Library. Microfilm. Published in 1977 by Plenum.

Chomsky, N. (1959). A review of B. F. Skinner's *Verbal Behavior*. *Language*, 35, 26-58.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

Chomsky, N. (1981). Lectures on Government and Binding Theory. Dordrecht: Foris.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221268.

Crain, S., & Nakayama, M. (1987). Structure dependency in grammar formation. *Language*, 63, 522-543.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133142.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113121.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, 105, 19011911.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

Fiser, J., & Aslin, R. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12, 499-504.

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, 29, 1-36.

Gambell, T., & Yang, C. (2003). Scope and limits of statistical learning in word segmentation. In *Proceedings of the 34th Northeastern Linguistic Society Meeting*. Stony Brook, NY. 29-30.

Gold, M. (1967). Language Identification in the Limit. *Information and Control*, 10, 447-74.

Gomez, R., & Gerken, LA (1999) Artificial grammar learning by 1-year-olds to specific and abstract knowledge. *Cognition*, 70, 109-133.

Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, 20, 229252.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan & G. Miller (Eds.) *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press. 294-303

Halle, M. (1997). On stress and accent in Indo-European. *Language*, 73, 275-313.

Halle, M., & Vergnaud, J.-R. (1987). *An essay on stress*. Cambridge, MA: MIT Press.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 31, 190222.

Hauser, M., Newport, E., & Aslin, R. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78, B41-B52.

Hauser, M., D. Weiss. D, & Marcus, G. (2002). Rule learning by cotton-top tamarins. *Cognition*, 86, B15-B22.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Hayes, J., & Clark, H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley. 221-234.

Herrnstein, R., & Loveland D. 1975. Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior*, 24, 107-116

Hunt, R., & Aslin, R. (2001). Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. *Journal of Experimental Psychology: General*, 130, 658-680.

Idsardi, W. (1992). The computation of prosody. Ph.D. dissertation, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA.

Jelinek, F., & Mercer, R. (1980). Interpolated estimation of markov source parameters from sparse data. In E. Gelsema & L. Kanal (Eds.) *Pattern recognition in practice*. Amsterdam: North-Holland. 381–402.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44, 548-567.

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

Jusczyk, P. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3, 323-328.

Jusczyk, P., & Aslin, R. (1995). Infant's detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 46, 65-97.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress patterns of English worlds. *Child Development*, 64, 675-687.

Juscyzk, P., & Derrah, C. (1987). Representation of speech sounds in infants. *Developmental Psychology*, 23, 648-654.

Jusczyk, P., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants's sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32, 402420.

Jusczyk, P., Goodman, M. B., & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language*, 40, 6282.

Jusczyk, P., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, 277, 19841986.

Jusczyk, P., Hohne, E. A., & Baumann, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61, 14651476.

Jusczyk, P., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39, 159207.

Jusczyk, P., Kennedy, L., & Jusczyk, A. M. (1995). Young infants' retention of information about syllables. *Infant Behavior and Development*, 18, 27-41

Jusczyk, P., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630645.

Krakow, R. (1999). Physiological organization of syllables: A review. *Journal of Phonetics*, 27, 2354.

Kirkham, N., Slemmer, J., & Johnson, S. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition*, 83, B34-B42.

Ladefoged, P. (1993). *A course in phonetics*. Fort Worth, TX: Harcourt Brace.

Lasnik, H. (1989). On certain substitutes for negative data. In R. Matthews & W. Demopoulos (eds.) *Learnability and Linguistic Theory*. Dordrecht: Reidel. 89-105.

Legate, J. A. & Yang, C. (2002). Empirical reassessment of poverty stimulus arguments. *Linguistic Review*, 19, 151-162.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431461.

Liberman, M., & Prince, A. On stress and linguistic rhythm. *Linguistic Inquiry*, 8, 249-336.

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd ed.). Hillsdale, NJ: Erlbaum.

de Marcken, C. (1996). The unsupervised acquisition of a lexicon from continuous speech. Unpublished Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Marcus, G., Vijayan, S., Rao, S., & Vishton, P. M. (1999). Rule-learning in seven-month-old infants. *Science*, 283, 77-80.

Mattys, S., & Jusczyk, P. (2001). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, 27, 644655.

Mattys, S., Jusczyk, P., Luce, P. A., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38, 465494.

Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393424.

Newport, E., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127162.

Newport, E., Hauser, M., Spaepen, G., & Aslin, R.N. (2004). Learning at a distance: II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, 49, 85-117.

Olivier, D. C. (1968). Stochastic grammars and language acquisition mechanisms. Unpublished Ph.D. dissertation, Harvard University, Cambridge, MA.

Onishi, K., Chambers, K., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory exposure. *Cognition*, 83, B13-B23.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002) Signal-driven computations in speech processing. *Science*, 298, 604-607.

Peters, A. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.

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

Pullum, G., & Scholz, B. (2002). Empirical assessment of poverty stimulus arguments. *Linguistic Review*, 19, 8-50.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425469.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-olds. *Science*, 274, 19261928.

Saffran, J., Aslin, R., & Newport, E. (1997). Letters. *Science*, 276, 11771181

Seidenberg, M. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275, 1599-1604.

Sluijter, A., van Heuven, V., & Pacily, J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 24712485.

Swingley, D. (1999). Conditional probability and word discovery: A corpus analysis of speech to infants. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st annual conference of the cognitive science society*. Mahwah, NJ: LEA. 724-729.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86-132.

Thiessen, E., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39, 706716.

Toro, J, Sinnett, S., & Soto-Faraco, S. (In press) Speech segmentation by statiscal learning depends on attention. *Cognition*.

Turk-Browne, N., Junge, J., & Scholl, B. (Submitted). The automaticity of visual statistical learning. Manuscript, Yale University.

Tversky, A. & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, 71, 680-683.

Valiant, L. (1984). A theory of the learnable. *Communication of the ACM*, 1134-1142.

Vapnik, V. (1995). *The Nature of statistical learning theory*. Berlin: Springer.

van de Weijer, J. (1998). Language input for word discovery. Unpublished doctoral dissertation, MPI Series in Psycholinguistics 9.

Wolff, J. G. (1977). The discovery of segments in natural language. *British Journal of Psychology*, 68, 97106.

Yang, C. (1999). A selectionist theory of language development. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. College Park, MD. 429-435.

Yang, C. (2002). *Knowledge and learning in natural language*. New York: Oxford University Press.

Yang, C. (2004). Universal grammar, statistics, or both. *Trends in Cognitive Sciences*, 8, 451-456.