Proceedings of the Workshop of the
ACL Special Interest Group on Computational Phonology (SIGPHON)
Association for Computations Linguistics
Barcelona, July 2004

Segment Predictability as a Cue in Word Segmentation: Application to Modern Greek

C. Anton Rytting
Department of Linguistics
The Ohio State University
Columbus, Ohio, U.S.A. 43201
rytting@ling.ohio-state.edu

Abstract

Several computational simulations of how children solve the word segmentation problem have been proposed, but most have been applied only to a limited number of languages. One model with some experimental support uses distributional statistics of sound sequence predictability (Saffran et al. 1996). However, the experimental design does not fully specify how predictability is best measured or modeled in a simulation. Saffran et al. (1996) assume transitional probability, but Brent (1999a) claims mutual information (MI) is more appropriate. Both assume predictability is measured locally, relative to neighboring segment-pairs.

This paper replicates Brent's (1999a) mutual-information model on a corpus of child-directed speech in Modern Greek, and introduces a variant model using a global threshold. Brent's finding regarding the superiority of MI is confirmed; the relative performance of local comparisons and global thresholds depends on the evaluation metric.

1 Introduction

A substantial portion of research in child language acquisition focuses on the *word segmentation problem*—how children learn to extract words (or word candidates) from a continuous speech signal prior to having acquired a substantial vocabulary. While a number of robust strategies have been proposed and tested for infants learning English and a few other languages (discussed in Section 1.1), it is not clear whether or how these apply to all or most languages. In addition, experiments on infants often leave undetermined many details of how particular cues are actually used. Computational simulations of word segmentation have also focused mainly on data from English corpora, and should also be extended to cover a broader range of the corpora available.

The line of research proposed here is twofold: on the one hand we wish to understand the nature of the cues present in Modern Greek, on the other we wish to establish a framework for orderly comparison of word segmentation algorithms across the desired broad range of languages. Finite-state techniques, used by e.g., Belz (1998) in modeling phonotactic constraints and syllable within various languages, provide one straightforward way to formulate some of these comparisons, and may be useful in future testing of multiple cues.

Previous research (Rytting, 2004) examined the role of utterance-boundary information in Modern Greek, implementing a variant of Aslin and colleagues' (1996) model within a finite-state framework. The present paper examines more closely the proposed cue of segment predictability. These two studies lay the groundwork for examining the relative worth of various cues, separately and as an ensemble.

1.1 Infant Studies

Studies of English-learning infants find the earliest evidence for word segmentation and acquisition between 6 and 7.5 months (Jusczyk and Aslin, 1995) although many of the relevant cues and strategies seem not to be learned until much later.

Several types of information in the speech signal have been identified as likely cues for infants, including lexical stress, co-articulation, and phonotactic constraints (see e.g., Johnson & Jusczyk, 2001 for a review). In addition, certain heuristics using statistical patterns over (strings of) segments have also been shown to be helpful in the absence of other cues.

One of these (mentioned above) is extrapolation from the segmental context near utterance boundaries to predict word boundaries (Aslin et al., 1996). Another proposed heuristic utilizes the relative predictability of the following segment or syllable. For example, Saffran et al. (1996) have confirmed the usefulness of distributional cues for 8-month-olds on artificially designed micro-

languages—albeit with English-learning infants only.

The exact details of how infants use these cues are unknown, since the patterns in their stimuli fit several distinct models (see Section 1.2). Only further research will tell how and to what degree these strategies are actually useful in the context of natural language-learning settings—particularly for a broad range of languages. However, what is not in doubt is that infants are sensitive to the cues in question, and that this sensitivity begins well before the infant has acquired a large vocabulary.

1.2 Implementations and Ambiguities

While the infant studies discussed above focus primarily on the properties of particular cues, computational studies of word-segmentation must also choose between various implementations, which further complicates comparisons. Several models (e.g., Batchelder, 2002; Brent's (1999a) MBDP-1 model; Davis, 2000; de Marcken, 1996; Olivier, 1968) simultaneously address the question of vocabulary acquisition, using previously learned word-candidates to bootstrap later segmentations. (It is beyond the scope of this paper to discuss these in detail; see Brent 1999a,b for a review.)

Other models do not accumulate a stored vocabulary, but instead rely on the degree of predictability of the next syllable (e.g., Saffran et al., 1996) or segment (e.g., Christiansen et al., 1998). The intuition here, first articulated by Harris (1954), is that word boundaries are marked by a spike in unpredictability of the following phoneme. The results from Saffran et al. (1996) show that English-learning infants do respond to areas of unpredictability; however, it is not clear from the experiment how this unpredictability is best measured. Two specific ambiguities in measuring (un)predictability are examined here.

Brent (1999a) points out one type of ambiguity, namely that Saffran and colleagues' (1996) results can be modeled as favoring word-breaks at points of either low transitional probability or low mutual information. Brent reports results for models relying on each of these measures. It should be noted that these models are *not* the main focus of his paper, but provided for illustrative purposes; nevertheless, these models provide the best comparison to Saffran and colleagues' experiment, and may be regarded as an implementation of the same.

Brent (1999a) compares these two models in terms of word tokens correctly segmented (see Section 3 for exact criteria), reporting approximately 40% precision and 45% recall for transitional probability (TP) and 50% precision and 53% recall for mutual information (MI) on the first 1000 utterances of his corpus (with improvements given larger corpora). Indeed, their performance on word tokens is surpassed only by Brent's main model (MBDP-1), which seems to have about 73% precision and 67% recall for the same range.[1]

Another question which Saffran et al. (1996) leave unanswered is whether the segmentation depends on local or global comparisons of predictability. Saffran et al. assume implicitly, and Brent (1999a) explicitly, that the proper comparison is local—in Brent, dependent solely on the adjacent pairs of segments. However, predictability measures for segmental bigrams (whether TP or MI) may be compared in any number of ways. One straightforward alternative to the local comparison is to compare the predictability measures compare to some global threshold. Indeed, Aslin et al. (1996) and Christiansen et al. (1998) simply assumed the mean activation level as a global activation threshold within their neural network framework.[2]

1.3 Global and Local Comparisons

The global comparison, taken on its own, seems a rather simplistic and inflexible heuristic: for any pair of phonemes *xy*, either a word boundary is always hypothesized between *x* and *y*, or it never is. Clearly, there are many cases where *x* and *y* sometimes straddle a word boundary and sometimes do not. The heuristic also takes no account of lengths of possible words. However, the local comparison may take length into account too much, disallowing words of certain lengths. In order to see that, we must examine Brent's (1999a) suggested implementation of Saffran et al. (1996) more closely.

In the local comparison, given some string …*wxyz*…, in order for a word boundary to be inserted between *x* and *y*, the predictability measure for *xy* must be lower than both that of *wx* and of *yz*. It follows that neither *wx* nor *yz* can have word boundaries between them, since they cannot simultaneously have a lower predictability measure than *xy*. This means that, within an utterance, word boundaries must have at least two segments between them, so this heuristic will not correctly segment utterance-internal one-phoneme

[1] The specific percentages are not reported in the text, but have been read off his graph. Brent does not report precision or recall for utterance boundaries; those percentages would undoubtedly be higher.

[2] These methodologies did not ignore local information, but encoded it within the feature vector. However, Rytting (2004) showed that this extra context, while certainly helpful, is not strictly necessary in the Greek corpus under question. A context of just one phoneme yielded better-than-chance results.

words.[3] Granted, only a few one-phoneme word types exist in either English or Greek (or other languages). However, these words are often function words and so are less likely to appear at edges of utterances (e.g., ends of utterances for articles and prepositions; beginnings for postposed elements). Neither Brent's (1999a) implementation of Saffran's et al. (1996) heuristic nor Aslin's et al. (1996) utterance-boundary heuristic can explain how these might be learned.

Brent (1999a) himself points out another length-related limitation—namely, the relative difficulty that the 'local comparison' heuristic has in segmenting learning longer words. The bigram MI frequencies may be most strongly influenced by—and thus as an aggregate largely encode—the most frequent, shorter words. Longer words cannot be memorized in this representation (although common ends of words such as prefixes and suffixes might be).

In order to test for this, Brent proposes that precision for word types (which he calls "lexicon precision") be measured as well as for word tokens. While the word-token metric emphasizes the correct segmentation of frequent words, the word-type metric does not share this bias. Brent defines this metric as follows: "After each block [of 500 utterances], each word type that the algorithm produced was labeled a true positive if that word type had occurred anywhere in the portion of the corpus processed so far; otherwise it is labeled a false positive." Measured this way, MI yields a word type precision of only about 27%; transitional probability yields a precision of approximately 24% for the first 1000 utterances, compared to 42% for MBDP-1. He does not measure word type recall.

This same limitation in finding longer, less frequent types may apply to comparisons against a global threshold as well. This is also in need of testing. It seems that both global and local comparisons, used on their own as sole or decisive heuristics, may have serious limitations. It is not clear *a priori* which limitation is most serious; hence both comparisons are tested here.

2 Constructing a Finite-State Model

2.1 Outline of current research

While in its general approach the study reported here replicates the mutual-information and transitional-probability models in Brent (1999a), it differs slightly in the details of their use. First, whereas Brent dynamically updated his measures over a single corpus, and thus blurred the line between training and testing data, our model pre-compiles statistics for each distinct bigram-type offline, over a separate training corpus.[4] Secondly, we compare the use of a global threshold (described in more detail in Section 2.3, below) to Brent's (1999a) use of the local context (as described in Section 1.3 above).

Like (Brent, 1999a), but unlike Saffran et al. (1996), our model focuses on pairs of segments, not on pairs of syllables. While Modern Greek syllabic structure is not as complicated as English's, it is still more complicated than the CV structure assumed in Saffran et al. (1996); hence, access to syllabification cannot be assumed.[5]

2.2 Corpus Data

In addition to the technical differences discussed above, this replication breaks new ground in terms of the language from which the training and test corpora are drawn. Modern Greek differs from English in having only five vowels, generally simpler syllable structures, and a substantial amount of inflectional morphology, particularly at the ends of words. It also contains not only preposed function words (e.g., determiners) but postposed ones as well, such as the possessive pronoun, which cannot appear utterance-initially. For an in-depth discussion of Modern Greek, see (Holton et al., 1997). While it is not anticipated that Modern Greek will be substantially more challenging to segment than English, the choice does serve as an additional check on current assumptions.

The Stephany corpus (Stephany, 1995) is a database of conversations between children and caretakers, broadly transcribed, currently with no notations for lexical stress, included as part of the CHILDES database (MacWhinney, 2000). In order to preserve adequate unseen data for future simulations and experiments, and also to use data most closely approximating children of a very

[3] At the edges of utterances, this restriction will not apply, since word boundaries are automatically inserted at utterance boundaries, while still allowing the possibility of a boundary insertion at the next position.

[4] While this difference is not intended as a strong theoretical claim, it can be seen as reflecting the fact that even before infants seem to begin the word segmentation process, they have already been exposed to a substantial amount of linguistic material. However, it is not anticipated to affect the general pattern of results.

[5] Furthermore, if Brent's 'local comparison' implementation were based on syllables to more closely coincide with Saffran's et al. (1996) experiment (not something Brent ever suggests), it would fail to detect any one-syllable words, clearly problematic for both Greek and English, and many languages besides.

young age, files from the youngest child only were used in this study. However, since the heuristics and cues used are very simple compared to vocabulary-learning models such as Brent's MDLP-1, it is anticipated that they will require relatively little context, and so the small size of the training and testing corpora will not adversely effect the results to a great degree.

As in other studies, only adult input was used for training and testing. In addition, non-segmental information such as punctuation, dysfluencies, parenthetical references to real-world objects, etc. were removed. Spaces were taken to represent word boundaries without comment or correction; however, it is worth noting that the transcribers sometimes departed from standard orthographic practice when transcribing certain types of word-clitic combinations. The text also contains a significant number of unrealized vowels, such as [ap] for /apo/ 'from', or [in] or even [n] for /ine/ 'is'. Such variation was not regularized, but treated as part of the learning task.

The training corpus contains 367 utterance tokens with a total of 1066 word tokens (319 types). Whereas the average number of words per utterance (2.9) is almost identical to that in the Korman (1984) corpus used by Christiansen et al. (1998), utterances and words were slightly longer in terms of phonemes (12.8 and 4.4 phonemes respectively, compared to 9.0 and 3.0 in Korman).

The test corpus consists of 373 utterance tokens with a total of 980 words (306 types). All utterances were uttered by adults to the same child as in the training corpus. As with the training corpus, dysfluencies, missing words, or other irregularities were removed; the word boundaries were kept as given by the annotators, even when this disagreed with standard orthographic word breaks.

2.3 Model Design

Used as a solitary cue (as it is in the tests run here), comparison against a global threshold may be implemented within the same framework as Brent's (1999) TP and MI heuristics. However, it may be implemented within a finite-state framework as well, with equivalent behavior. This section will describe how the 'global comparison' heuristic is modeled within a finite-state framework.

While such an implementation is not technically necessary here, one advantage of the finite-state framework is the compositionality of finite state machines, which allows for later composition of this approach with other heuristics depending on other cues, analogous to Christiansen et al. (1998). Since the finite-state framework selects the best path over the whole utterance, it also allows for optimization over a sequence of decisions, rather than optimizing each local decision separately.[6]

Unlike Belz (1998), where the actual FSM structure (including classes of phonemes that could be group onto one arc) was learned, here the structure of each FSM is determined in advance. Only the weight on each arc is derived from data. No attempt is made to combine phonemes to produce more minimal FSMs; each phoneme (and phoneme-pair) is modeled separately.

Like Brent (1999a) and indeed most models in the literature, this model assumes (for sake of convenience and simplicity) that the child hears each segment produced within an utterance without error. This assumption translates into the finite-state domain as a simple acceptor (or equivalently, an identity transducer) over the segment sequence for a given utterance.[7]

Word boundaries are inserted by means of a transducer that computes the cost of word boundary insertion from the predictability scores. In the MI model, the cost of inserting a word boundary is proportional to the mutual information. For ease in modeling, this was represented with a finite state transducer with two paths between every pair of phonemes *(x,y)*, with zero-counts modeled with a maximum weight of 99. The direct path, representing a path with no word boundary inserted, costs $-MI(x,y)$, which is positive for bigrams of low predictability (negative MI), where word boundaries are more likely. The other path, representing a word boundary insertion, carries the cost of the global threshold, in this case arbitrarily set to zero (although it could be optimized with held-out data). A small subset of the resulting FST, representing the connections over the alphabet {ab} is illustrated in Figure 1, below:

[6] See Rabiner (1989) for a discussion of choosing optimization criteria. It is worth noting that this distinction does not come into play in the one-cue model reported here, as all decisions are modeled as independent of one another. However, it is expected to take on some importance in models combining multiple cues, such as those proposed in Section 4 of this paper.

[7] While modeling the mishearing of segments would be more realistic and highly interesting, it is beyond the scope of this study. However, a weighted transducer representing a segmental confusion matrix could in principle replace the current identity transducer, without disrupting the general framework of the model.

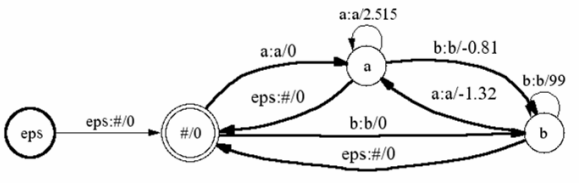

Figure 1: The MI model over the alphabet {ab}

The best (least-cost) path over this subset model inserts boundaries between two adjacent *a*'s and two adjacent *b*'s, but not between *ab* or *ba*; thus the (non-Greek) string *...ababaabbaaa...* would be segmented *...ababa#ab#ba#a#a...* by the FSM.

The FSM for transitional probability has the same structure as that of MI, but with different weights on each path. For each pair of phonemes *xy*, the cost for the direct path from *x* to *y* is $-log(P(y|x))$. The global threshold cost of inserting a word boundary was set (again, arbitrarily) as the negative log of the mean of all TP values. In the two-phoneme subset (shown in Figure 2), the only change is that the direct pathway from *a* to *b* is now more expensive than the threshold path, so the best path over the FSM will insert word boundaries between *a* and *b* as well. Hence our example string *...ababaabbaaa...* would be segmented *...a#ba#ba#a#b#ba#a#a...* by the FSM. (The stranded 'word' *#b#* would of course be an error, but this problem does not arise in actual Greek input, since two adjacent *b*'s, like all geminate consonants, are ruled out by Greek phonotactics.)

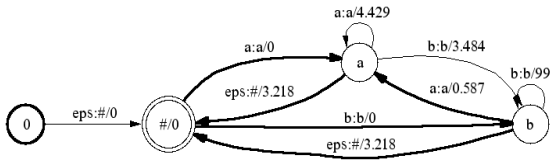

Figure 2: The TP model over the alphabet {ab}

During testing each FST model was composed (separately) with the segment identity transducer for the utterance under consideration. A short sample section of such a composition, with the best path in bold, is shown in Figure 3.

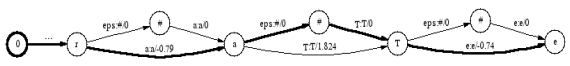

Figure 3: A section of the composition of the MI model and an utterance acceptor

The output projection of the best path from the resulting FST was converted back into text and compared to the text of the original utterance. These compositions, best-path projections, and conversions were performed using the AT&T finite state toolkit (Mohri et al., 1998).[8]

2.4 A Concrete Example

Take, for example, an utterance from the test corpus /tora#Telis#na#aniksume#afto/ 'now you want us to open this.' The mutual information and transitional probability figures for this utterance are given in Table 1.

Context		Predictability	
Left	Right	MI	TP
#	t	0.000	3.219
t	o	−1.661	0.781
o	r	−1.018	*2.350*
r	a	−0.800	1.113
a	T	***1.824***	***6.375***
T	e	−0.744	1.540
e	l	*−0.225*	*3.059*
l	i	−0.903	1.197
i	s	−0.491	2.382
s	n	***1.555***	***4.317***
n	a	−0.300	1.613
a	a	***2.516***	***4.429***
a	n	−0.339	2.424
n	i	*−0.071*	2.029
i	k	−0.337	*2.633*
k	s	−0.444	2.428
s	u	*−0.172*	*3.219*
u	m	−1.387	2.413
m	e	−1.230	1.055
e	a	***1.095***	*3.008*
a	f	−1.473	2.525
f	t	−2.068	0.484
t	o	−1.661	0.781
o	#	0.000	3.219

Table 1: MI and TP values for bigrams in the test utterance /tora#Telis#na#aniksume#afto/. Values above threshold are bold; local maxima italicized.

In this example, the correct boundaries fall between the pairs (a,T), (s,n), (a,a), and (e,a). Both the mutual information and the transitional probability for the first three of these pairs are above the global mean, so word boundaries are posited under both global models.[9] (Since each of these is also a local maximum, the local models also posit boundaries between these three pairs.) The pair (e,a) is above threshold for MI but not for

[8] FSM Library Version 3.7, freely available from http://www.research.att.com/sw/tools/fsm/

[9] Since all values are given in terms of negative MI and negative log probability, high values for both measures indicate relatively improbable pairings.

TP, so the global TP model fails to posit a boundary here. Finally, the two local models posit a number of spurious boundaries at the other local maxima, shown by the italic numbers in the table. The resulting predictions for each model are:

Global MI: #tora#Telis#na#aniksume#afto#
Global TP: #tora#Telis#na#aniksumeafto#
Local MI: #tora#Te#lis#na#an#iks#ume#afto#
Local TP: #to#ra#Te#lis#na#ani#ks#ume#afto#

3 Results

The four model variants (global MI, global TP, local MI, and local TP) were each evaluated on three metrics: word boundaries, word tokens, and word types. Note that the first metric reported, simple boundary placement, considers only utterance-internal word-boundaries, rather than including those word boundaries which are detected 'for free' by virtue of being utterance-boundaries also. This boundary measure may be more conservative than that reported by other authors, but is easily convertible into other metrics.

The second metric, the percentage of word tokens detected, is the same as Brent (1999a). In order for a word to be counted as correctly found, three conditions must be met: (a) the word's beginning (left boundary) is correctly detected, (b) the word's ending (right boundary) is correctly detected, and (c) these two are consecutive (i.e., no false boundaries are posited within the word).

The last metric (word type) is slightly more conservative than Brent's (1999a) in that the word type must have been actually spoken in the same utterance (not the same block of 500 utterances) in which it was detected to count as a match. This lessens the possibility that a mismatch that happens to be segmentally identical to an actual word (but whose semantic context may not be conducive to learning its correct meaning) is counted as a match. However, this situation is presumably rather rare.

Tables 2 and 3 present the results over the test set for both the global and the local comparisons of the predictability statistics proposed by Saffran et al. (1996) and Brent (1999a).

Global Comparison		Bound-aries	Word Tokens	Word Types
MI	Precision	43.9%	30.8%	22.3%
	Recall	54.4%	35.3%	29.7%
	F-Score	48.6%	32.9%	25.5%
TP	Precision	40.4%	28.4%	20.0%
	Recall	41.7%	29.0%	28.4%
	F-Score	41.0%	28.7%	23.5%

Table 2: Global Comparison: FST best paths with bigrams compared to a global threshold only

Local Comparison		Bound-aries	Word Tokens	Word Types
MI	Precision	42.0%	31.5%	20.1%
	Recall	62.6%	41.1%	27.8%
	F-Score	50.3%	35.7%	23.4%
TP	Precision	41.5%	28.0%	20.2%
	Recall	74.1%	41.6%	22.9%
	F-Score	53.2%	33.5%	21.4%

Table 3: Local Comparison: Replication of Brent (1999a); each bigram compared to both neighbors

4 Conclusion

4.1 Comparing the Four Variants

The findings here confirm Brent's (1999a) contention that mutual information is a better measure of predictability than is transitional probability—at least for the task of identifying words, not just boundaries. This is particularly true in the global comparison. Transitional probability finds more word boundaries in the 'local comparison' model, but this does not carry over to the task of pulling out the word themselves, which is arguably the infant's main concern. This result should be kept in mind when interpreting or replicating (Saffran et al., 1996) or similar studies.

While Brent's 'local comparison' heuristic was unable to pull out one-phoneme-long words, as predicted above, this did not adversely affect it as much as anticipated. On the contrary, both the local and global comparison heuristics tended to postulate too many word boundaries, as Brent had observed. This is not necessarily a bad thing for infants, for several reasons.

First, infants may have a preference for finding short words, since these will presumably be easier to remember and learn, particularly if the child's phonetic memory is limited. Second, it is probably easier to reject a hypothesized word (for example, on failing to find a consistent semantic cue for it) than to obtain a word not correctly segmented; hence false positives are less of a problem than false negatives for the child. Third and most importantly, this cue is not likely to operate on its own, but rather as one among many contributing cues. Other cues may act as filters on the boundaries suggested by this cue. One example of this is the distribution of segments before utterance edges, as used by e.g., Aslin et al. (1996) and Christiansen et al. (1998) which indicate the set of possible word-final segments in the language.

However, as far as these results go, the word type metric shows that the finite-state model using a global threshold suffered slightly less from this problem than the local comparison model. For the MI variants, both recall and precision for word type were about 2% higher on the global threshold variant. For transitional probability, the precision of the local and global models was roughly equal, but recall for the global comparison model was 5.5% higher. Not only were the global models better at pulling out a variety of words, but they also managed to learn longer ones (especially the global TP variant), including a few four-syllable words. The local model learned no four-syllable words, and relatively few three-syllable words.

The mixed nature of these results suggests that evaluation depends fairly crucially on what performance metric needs to be optimized. This demands stronger prior hypotheses regarding the process and needed input of a vocabulary-acquiring child. However, it cannot be blindly assumed that children are selecting low points over as short a window as Brent's (1999a) MI and TP models suggest. Quite possibly the best model would involve either a hybrid of local and global comparisons, or a longer window, or even a 'gradient' window where far neighbors count less than near ones in a computed average.

However, further speculation on point this of less importance than considering how this cue interacts with others known experimentally to be salient to infants. Christiansen et al. (1998) and Johnson and Jusczyk (2001) have already began simulating and testing these interactions in English. However, more work needs to be done to understand better the nature of these interactions cross-linguistically.

4.2 Further Research

As mentioned above, one obvious area for future research is the interaction between predictability cues like MI and utterance-final information; this is one of the cue combinations explored in Christiansen et al. (1998) in English. Previous research (Rytting, 2004) examined the role of utterance-final information in Greek, and found that this cue performs better than chance on its own. However, it seems that utterance-final information would be more useful as a filter on the heuristics explored here to restrain them from oversegmenting the utterance. Since nearly all Greek words end in /a/, /e/, /i/, /o/, /u/, /n/, or /s/, just restricting word boundaries to positions after these seven phonemes boosts boundary precision considerably with little effect on recall.[10] Preliminary testing suggests that this filter boosts both precision and recall at the word level. However, a model that incorporates the likelihoods of word boundaries after each of these final segments, properly weighted, may be even more helpful than this simple, unweighted filter.

Another fruitful direction is the exploration of prosodic information such as lexical stress. With the exception of a certain class of clitic groups, Greek words have at most one stress. Hence, at least one word boundary must occur between two stressed vowels. Relations between stress and the beginnings and endings of words, while not predicted to be as robust a cue as in English (see e.g., Cutler, 1996), should also provide useful information, both alone and in combination with segmental cues.

Finally, the relationship between these more 'static' cues and the cues that emerge as vocabulary begins to be acquired (as in Brent's main MBDP-1 model and others discussed above) seems not to have received much attention in the literature. As vocabulary is learned, it can help bootstrap these cues by augmenting heuristic cues with actual probabilities derived from its parses. Hence, the combination of e.g., MLDP-1 and these heuristics may prove more powerful than either approach alone.

5 Acknowledgements

This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Sincere thanks go to the NSF for their financial support, and to Chris Brew, Eric Fosler-Lussier, Brian Joseph, members of the computational linguistics and phonology discussion groups at the Ohio State University, and to anonymous reviewers of previous versions of this paper for their helpful comments and encouragement.

References

Richard N. Aslin, Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In *Signal to syntax*, James L. Morgan and Katherine Demuth, ed., pages 117-

[10] Naturally, in unrestricted speech the characteristics of word boundaries diverge from those of utterance boundaries. For example, final vowels may delete from utterance-medial words; instances such as [in] for /ine/ 'is' and [ap] for /apo/ 'from' were already mentioned. However, if we assume that the canonical forms of these words occur frequently enough to be acquired normally, then knowledge of these canonical forms may assist the acquisition of variant forms (as well as the phrasal phonological processes that give rise to them) later on.

134, Lawrence Erlbaum Associates, Mahwah, New Jersey.

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

Anja Belz. 1998. An Approach to the Automatic Acquisition of Phonotactic Constraints. In *Proceedings of SIGPHON '98: The Computation of Phonological Constraints,* T. Mark Ellison, ed., pages 35-44

Michael R. Brent. 1999a. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71-105.

Michael R. Brent. 1999b. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, 3(8):294-301.

Morton H. Christiansen, Joseph Allen, and Mark S. Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2/3):221-268.

Anne Cutler. 1996. Prosody and the word boundary problem. In *Signal to syntax*, James L. Morgan and Katherine Demuth, ed., pages 87-100, Lawrence Erlbaum Associates, Mahwah, New Jersey.

Matt H. Davis. 2000. *Lexical segmentation in spoken word recognition.* Unpublished PhD thesis, Birkbeck College, University of London. Available: http://www.mrc-cbu.cam.ac.uk/personal/matt.davis/thesis/index.html

Carl G. de Marcken. 1996. *Unsupervised language acquisition*. PhD dissertation, MIT, Cambridge, MA. Available: http://xxx.lanl.gov/abs/cmp-lg/9611002

Zelig S. Harris. 1954. Distributional structure. *Word*, 10:146-162.

David Holton, Peter Mackridge and Irene Philippaki-Warburton. 1997. *Greek: A Comprehensive Grammar of the Modern Language*, Routledge, London and New York.

Elizabeth K. Johnson and Peter W. Jusczyk. 2001. Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, 44 (4), 548 567.

Myron Korman. 1984. Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, 5:44-45.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. 1998. A Rational Design for a Weighted Finite-State Transducer Library. *Lecture Notes in Computer Science*, 1436.

D. C. Olivier. 1968. Stochastic grammars and language acquisition mechanisms. PhD dissertation, Harvard University, Cambridge, Massachusetts.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* 77:2, pages 257-285.

C. Anton Rytting. 2004. Greek word segmentation using minimal information. In *Proceedings of the Student Research Workshop at HLT/NAACL 2004*, pages 207-212, Association for Computational Linguistics, Boston, Massachusetts. Available: http://acl.ldc.upenn.edu/hlt-naacl2004/studws/pdf/sw-8.pdf

Jenny R. Saffran, Richard N. Aslin and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science,* 274, 1926-1928.

Ursula Stephany. 1995. The acquisition of Greek. In *The crosslinguistic study of language acquisition.* D. I. Slobin, ed., Vol. 4.