Greek Word Segmentation Using Minimal Information

C. Anton Rytting
Department of Linguistics
The Ohio State University
Columbus, Ohio 43210
rytting@ling.ohio-state.edu

Abstract

Several computational simulations have been proposed for how children solve the word segmentation problem, but most have been tested only on a limited number of languages, often only English. In order to extend the cross-linguistic dimension of word segmentation research, a finite-state framework for testing various models of word segmentation is sketched, and a very simple cue is tested in this framework. Data is taken from Modern Greek, a language with phonological patterns distinct from English. A small-scale simulation shows using this cue performs significantly better than chance. The utility and flexibility of the finite-state approach is confirmed; suggestions for improvement are noted and directions for future work outlined.

1 Introduction

A substantial portion of research in first-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. Note that the hardware and software constraints on the human learner are very different from those faced by a speech recognition system, and hence strategies appropriate for one may be irrelevant or disastrously inappropriate for the other.

While a number of robust strategies have been proposed and tested for English and a few other languages (discussed below), it is not clear whether or how these apply to other languages. For example, the Metrical Segmentation Strategy (e.g., Cutler & Norris 1988) turns out to be very robust for English, but is not necessarily applicable to other languages, simply because not all languages share English's predilection for strong word-initial syllables (though language-appropriate variants of the strategy (stress-based for English) have been proposed, e.g., using the syllable in French (Cutler & Mehler, 1993) and the mora in Japanese (Otake, Hatano, Cutler, & Mehler, 1993)).

Some more generic strategies (e.g., the Possible Word Constraint: see e.g., Norris et al. 1997, 2001) have been proposed and tested, primarily on English, but also on typologically distinct languages such as Sesotho (Cutler, Demuth, & McQueen, 2002). Nevertheless, rigorous testing in a larger sample of languages seems advisable before making strong claims of universal applicability. One interesting strategy explored in e.g., (Aslin et al., 1996) is the use of context around (and particularly before) utterance boundaries to predict word boundaries. The applicability of this cue is discussed for both English and Turkish; a simulation on English data is reported. One goal of the research presented here is to further explore that strategy on a different data set, taken from a language with phonological patterns quite different from English or Turkish.

The work presented here is intended as a small part of a more general line of research, whose purpose is twofold: on the one hand I wish to understand the nature of the cues present in Modern Greek, on the other I wish to establish a framework for orderly comparison of word segmentation algorithms across the desired broad range of languages.

1.1 Infant Studies

At least four types of information in the speech signal have been identified as likely cues for infants: (1) supersegmental cues (e.g., stress) which begins to play a role in (English-learning) 7.5 month-olds (Jusczyk, Houston, et al., 1999); (2) sub-segmental cues such as coarticulation and allophonic alternations, which infants

begin using between 7.5 and 10.5 months of age (Jusczyk, Hohne, et al., 1999); (3) segmental cues, such as wordlikeness and phonotactic constraints, which seem to be available by 9 months of age (e.g., Jusczyk, Luce, et al., 1994; Mattys and Jusczyk, 2001), and (4) statistical cues from recurrent patterns e.g., of syllables, evident in English-learning 8-month-olds on an artificial micro-language of 4 words (Saffran et al. 1996).[1]

1.2 Computational Models

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. In addition, several models (e.g., Batchelder, 2002; Brent's MLDP-1, 1999a; Davis, 2000; de Marcken, 1996; Olivier, 1968) simultaneously address the question of vocabulary acquisition, using previously learned word-candidates to bootstrap later segmentations. While these models are highly interesting both from their view of the long-term process of language acquisition and their high success rate, they are hard to relate to the infant studies discussed above. Hence, it is beyond the scope of this paper to discuss them at length.[2]

Rather, this paper focuses on models that do not accumulate a stored vocabulary, but rely on either on statistics derived from utterance boundaries (typically generalized over feature matrices, as in (Aslin et al., 1996; Christiansen et al., 1998)) or from the degree of predictability of the next syllable (e.g., Saffran et al., 1996) or segment (Christiansen et al., 1998). The intuition here, first articulated by Harris (1954) is that word boundaries will be marked by a spike in unpredictability of the following phoneme. Christiansen et al. (1998) also test the contribution of stress and phonemic information in addition to that of utterance boundaries, and show that while stress contributes in certain circumstances, it is not as crucial as featural information near utterance boundaries.

The general line of research herein proposed focuses on the same cues as (Christiansen et al., 1998) beginning (in the work reported here) with segmental probability distributions at utterance boundaries. This first step corresponds most closely with (Aslin et al., 1996), where utterance boundaries were treated as a cue on their own. Aslin and his colleagues propose that "even the most minimal assumption about what an infant can recognize as a word boundary--namely, the pause after an utterance--is sufficient, in principle, for the learning the word boundaries within an utterance" (p. 133). In that study, however, a considerable amount of context before such an utterance was given, namely 18-bit feature vectors of one, two, or three phonemes immediately preceding the final utterance boundary. For a model with 30 hidden units, the following results for boundary detection are reported (as estimated from their bar graph, fig.8.8, p. 132), accompanied by the claim that only with two- and three-phoneme sequences is their system capable of learning boundary locations:

	Hits	False Alarms	Precision (H/(H+FA))
3 phones	62%	22%	74%
2 phones	53%	23%	70%
1 phone	45%	44%	51%
Random	5%	15%	25%

Table 1. Results reported in Aslin et al. (1996, Fig. 8.8, p. 132).

They further claim that feature vectors are necessary for learning: a string of three phonemes (where each phoneme is represented as an atomistic unit) is not sufficient information, although no comparative figures are listed for this condition.

This study may be seen as a replication of (Aslin et al., 1996); however, it differs in several crucial respects—not with an eye toward improving upon their results, but rather on examining further their definition of "minimal necessary cues." First, instead of training the transitional probabilities indirectly with connectionist networks, the probabilities are encoded directly within a finite-state framework. Secondly, actual phone identities (rather than feature bundles) are used as symbols. Finally, information about a single segment is used. While this very austere use of minimal information is surely inadequate to the full task of segmentation, it nevertheless serves to demonstrate the gains even a very small amount of information can give. Any evidence of better-than-chance results would suggest that, for Modern Greek at least, even more minimal cues are possible than those Aslin et al. (1996) propose.

The results of this study may be taken as a rough approximation of how predictable word boundaries are from (unigram) segmental information alone in the subset of Modern Greek experienced by young children. These findings may provide an additional baseline for measuring and comparing the relative contributions of other cues such as stress as a word segmentation cue.

2 Constructing a Finite-State Model

2.1 Data

The Greek CHILDES corpus (Stephany, 1995) is a database of conversations between children and caretak-

[1] Full mention of all the studies done is not possible here; for a fuller review see e.g., (Johnson & Jusczyk, 2001).

[2] For useful reviews of various computational models, see (Brent, 1999a,b).

ers, broadly transcribed, currently with no notations for lexical stress. Audio tapes exist, but are currently unavailable for general use (Stephany, p.c.). However, the transcriptions themselves give an indication at the phonemic level of the sort of input Greek children are likely to have in learning their language. In order to preserve adequate unseen data for future simulations and experiments, only a small subset of the total Greek CHILDES corpus was used.

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. Word boundaries are represented by the symbol #, utterance boundaries by $, following Brent (1999a). Each line of the file was assumed to be an independent utterance. Spaces were assumed to represent word boundaries without comment or correction; however, it is worth noting that the transcribers sometimes departed from standard orthographic practice with respect to certain types of word-clitic combinations. The text also contains a significant number of unrealized final vowels (apocopy), such as [in] 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 comparable to the Korman (1984) corpus used by Christiansen et al. (3.0), utterances and words were slightly longer in terms of phonemes (12.8 and 4.4 phonemes respectively, compared to 9.0 and 3.0). (Statistics on the corpus used in (Aslin et al., 1996) were not provided.)

The test corpus consists of utterances by adults to the same child as in the training corpus. Utterances with dysfluencies, missing words, or other irregularities were discarded; the remaining utterances include 273 utterance tokens with a total of 699 words (229 types).

2.2 Model Design

This model differs from incremental models such as (Brent 1999a) in that it pre-compiles statistics for the candidate word-final phonemes off-line, over the entire corpus. These probabilities are thus static. While this difference is not intended as a strong theoretical claim, it reflects the fact that even before infants seem to be learning the word segmentation process, they have already been exposed to a large amount of linguistic material. The information gleaned from the corpus is represented in three separate (but composible) finite-state machines:

(1) Like most models in the literature, this model assumes (for sake of convenience and simplicity) that the child hears the correct sequence of the actual segments produced within an utterance. Hence, the model does not take into account the possibility of mishearing a segment, as that would add undue complication at this stage. 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.[3]

(2) An optional source of knowledge used is the number of words in a given utterance. This is naturally a strong assumption to make; it is included primarily to provide comparisons with baselines used by Brent (1999a) and Christiansen et al. (1998), which provide pseudo-random baselines that make reference either to number of boundaries directly or information concerning average word length. Results are given both with and without this constraint.

(3) The main item under examination is naturally the relative likelihood of breaking the word after a given segment S.

The third information source was tested in three variants. The first one is of course the approximation suggested by Aslin et al. (1996), that P(#|S) may be approximated by using P($|S), the probability of an utterance-break given the segment. This approximation yields the ranking e>s>o>u>i>a>m>n, with /e/ most likely to end an utterance. This information source was compared to two related alternatives, which were used as upper and lower bounds to measure the effectiveness of the utterance-boundary approximation of word boundaries. As an upper bound (3_U), the true value for P(#|S) is used, corresponding to training on labeled data, or a store of already-learned vocabulary.[4] The lower bound (3_L) consists of the seven final segments {a,e,i,o,u,n,s}, but the frequency ranking replaced by an equi-probable assumption. In a sense, this is equivalent to a grammar book listing the possible final segments of Greek without regard to their actual likelihood. Finally, these three variants are compared with a random walk for which no information is used, but boundaries are inserted completely by chance. Each of these three types of knowledge was modeled by means of a finite state machine, using the AT&T finite-state tools.[5]

[3] While modeling the mishearing of segments is beyond the scope of this study, a weighted transducer could in principle represent a segmental confusion matrix in a modular way and augment the current identity transducer. For further discussion of issues in using "unsanitized data," (Sundaram, 2003) may be helpful.

[4] The resulting ranking, o>i>e>s>a>u>n>j>m>p, is rather different than the one above, reflecting the frequency of masculine and feminine articles /o/ and /i/, which are never utterance-final.

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

(1) Segments: Linear FSA (trivially equivalent to an identity transducer).

(2) Number of words: Unweighted FSA.

(3_U) Upper bound: Weighted FST, with weights corresponding to -Log(P(#|S)) for a word boundary and -Log(1-P(#|S)) for an arc with no word boundary.

(3) Utterance-Boundary Probabilities: Same as (3_U), with weights corresponding to -Log(P($|S)) for a word boundary and -Log(1-P($|S)) for no word boundary. In the condition where (2) was not used, a weight of -1.7 (determined empirically on the training data) was added to the word-boundary arc, to offset the tendency of P($|S) to underestimate P(#|S).

(3_L) Unweighted (or equally weighted) version of (3). In the condition where (2) was not used, a weight of (-0.5) was added to the arc that adds boundaries, which caused the FST to insert word boundaries after every instance of a vowel, /n/, or /s/.

3 Results

Six different conditions were tested, corresponding to the three variants FSTs (3), (3_U), and (3_L), both with and without the exact-word constraint in FSM (2). Each of these were composed (separately) with the "segment identity" acceptor (1) for a given utterance. The output-projection of the best path from each resulting FST was converted back into text and compared to the text of the original utterance. Scores for both boundaries and words are reported (where a word is counted as correctly segmented only if both its left and right boundaries are correctly placed). In the case where several best-paths of equal cost exist, the average scores for precision and recall are counted.

The results with and without the number of words known are shown in Tables 2 and 3, following. In both cases, the precision scores patterned as expected. The upper bound condition (representing a supervised case, where statistics on the word boundaries are available for the training data) proved the most accurate on the test data. This suggests (as has been confirmed for English in such studies as Brent 1999a) that the learning of patterns over already-acquired vocabulary has perhaps the largest effect in the acquisition of new vocabulary.

The utterance-based approximation, corresponding most closely to (Aslin et al., 1996), seems to be slightly better overall than the lower bound. Without the number of words known, (3) has an F-score of 20.2 for words and 70.2 for boundaries, whereas (3_L) has F-scores of only 17.0 (word) and 68.0 (boundaries), though this difference may not be significant. This difference was less than expected, given preliminary examination of the training data; it may be that once the set of allowable word-final phonemes is observed, the relative probabilities of those phonemes is not as usefully learned from utterance boundaries. However, the lower bound (corresponding to purely symbolic knowledge of the allowable word-final segments) is significantly better than the random walk, suggesting that any knowledge, no matter how rudimentary, begins to make a difference.

	Words		Word Boundaries	
	Precision	Recall	Precision	Recall
Upper bound	277/699 (39.6%)	277/699 (39.6%)	751/972 (77.3%)	751/972 (77.3%)
Utt-prob.	226/699 (32.4%)	226/699 (32.4%)	721/972 (74.2%)	721/972 (74.2%)
Lower bound	221/699 (31.6%)	221/699 (31.6%)	708/972 (72.9%)	708/972 (72.9%)
Random Walk	119/699 (17.0%)	119/699 (17.0%)	639/972 (65.7%)	639/972 (65.7%)

Table 2: Test Results with Constraint (2)

	Words		Word Boundaries	
	Precision	Recall	Precision	Recall
Upper bound	219/720 (30.4%)	219/699 (31.3%)	737/993 (74.2%)	737/972 (75.8%)
Utt-prob.	159/860 (18.5%)	159/699 (22.7%)	739/1133 (65.2%)	739/972 (76.0%)
Lower bound	195/1599 (12.2%)	195/699 (27.9%)	967/1872 (51.7%)	967/972 (99.5%)
Random Walk	39/1569 (2.5%)	39/699 (5.6%)	767/1842 (41.6%)	767/972 (78.9%)

Table 3: Test Results without Constraint (2)

4 Discussion

4.1 Comparisons with Aslin et al. (1996)

Obviously, the cues of preceding and following segments are in and of itself insufficient to predict a word boundary with any reasonable degree of accuracy, just as Christiansen et al. (1998) found that no one cue was sufficient for English. However, a few comparisons with Aslin's et al. (1996) data in Table 1 may be useful, although they should be interpreted cautiously given the differences in the training and testing corpora between their study and this one. Their results for the single-phoneme condition have nearly equal hits and false alarms—a precision of about 51%. They apparently do not consider this sufficient evidence of learning, although it is significantly better than their random baseline. Similarly, the worst non-random condition

reported here (lower bound without constraint (2)) also has a precision of 51%. This, too, is difficult to call "learning," as it represents the heuristic of always inserting a word boundary any time there could be one. The only fact that has been learned is which segments cannot be (excepting foreign loan-words) word-final.

However, if the criterion for learning (or at least satisfactory performance) is hits exceeding false alarms, then the utterance-boundary statistical heuristic, with 739 hits and only 396 false alarms, is nearly as accurate as Table 1's two-phoneme condition. While further information (whether phonological features, longer strings of phonemes, or some other cue) is needed to reach the 74% accuracy of Table 1's three-phoneme condition, it seems that even these very basic cues come closer to Aslin's et al. (1996) results than might be supposed. Importantly, the same general trend was shown--that utterance-final information translates into word-boundary information not only for English, but for other languages such as Modern Greek as well.

A number of further directions are possible under this framework, including:

(1) Using transitional probability ($P(S_{k+1}| S_k)$) and mutual information measures over two adjacent segments as cues to the likelihood of word boundaries between those two segments, as suggested in e.g., (Brent, 1999a).

(2) Developing more plausible models for approximating word-length distributions from utterance-length information, distances between stressed vowels, pause information, and other salient cues available to children.

(3) Incorporating stress cues (as potentially signaling both beginnings and approaching ends of content words) both alone and in combination with segmental cues.

Preliminary work on each of these avenues is currently underway. While some of these heuristics may require the use of other techniques in addition to finite-state techniques, the general finite-state framework is expected to prove useful as an organizing tool for comparing various cues in a simple, rational, and transparent way.

References

Aslin, Richard N., Woodward, Julide Z., LaMendola, Nicholas P., & Bever, Thomas G. 1996. Models of word segmentation in fluent maternal speech to infants. In James L. Morgan & Katherine Demuth , editors, *Signal to syntax*, pages 117-134. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

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

Davis, Matt H. (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

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

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

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

Joseph, Brian. 2001. 'Word' in Modern Greek. In R.M.W. Dixon & A. Aikhenvald (eds.) *Proceedings of the International Workshop on the Status of 'Word'*. Cambridge: Cambridge University Press (2001).

Jusczyk, Peter W., & Aslin, R. N. 1995. Infant's detection of sound patterns of words in fluent speech. *Cognitive Psychology*, 29:1-23.

Jusczyk, Peter W., Hohne, E. A., & Bauman, A. 1999. Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, 61:1465-1476.

Jusczyk, Peter W., Houston, Derek, & Newsome, Mary. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39:159-207.

Jusczyk, Peter W., Luce, Paul A., & Charles-Luce, Jan 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33:630-645.

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

Mattys, Sven L., Jusczyk, Peter W., Luce, Paul A., & Morgan, James L. 1999. Word segmentation in infants: How phonotactics and prosody combine. *Cognitive Psychology*, 38:465-494.

Mattys, Sven L. and Jusczyk, Peter W. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78:91-121.

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

Saffran, Jenny R., Aslin, Richard N., & Newport, Elissa L. 1996. Statistical cues in language acquisition: word segmentation by infants. In G.W. Cottrell, editor, *Proceedings of the 18th Annual Conference of the Cognitive Science Society*. pages 376-380. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Sundaram, Ramasubramanian. 2003. Effects of Transcription Errors on Supervised Learning in Speech Recognition. Unpublished Masters Thesis. Mississippi State University, Mississippi State, MS. http://www.isip.msstate.edu/publications/books/msstate_theses/2003/transcription_errors/.