The Richness of Distributional Cues to Word Boundaries in Speech to Young Children

Gaja Jarosz and J. Alex Johnson

Yale University

Contact address: Department of Linguistics

Yale University

370 Temple St., Room 204

P.O. Box 208366

New Haven, CT 06520-8366

USA

Email: gaja.jarosz@yale.edu

We would like to thank Daniel McClory and Ahmet Aktay for help with data processing in the early stages of the this work. We are very grateful to Kemal Oflazer for providing the tools and assistance to process the Turkish data. We would also like to thank audiences at BUCLD 33 and NECPHON 2 for helpful comments as well as LouAnn Gerken and three anonymous reviewers for constructive critiques.

Abstract

This study is a systematic analysis of the information content of a wide range of distributional cues to word boundaries, individually and in combination, in naturally occurring child-directed speech across three languages (English, Polish, and Turkish). The paper presents a series of statistical analyses examining the relative predictive strength of these cues, the overlap in the information about word boundaries they contain, and the variability in their relative strengths and interactions across the languages. We find that the information content of individual distributional cues is not constant across languages, with relative reliability of cues varying across languages and with individual cues providing much less information in Polish and Turkish than in English. However, we also find that when these cues are combined, the cumulative information content of a diverse array of distributional cues provides a significant source of information about word boundaries across all three languages.

Introduction

One of the first language learning tasks infants must solve is the segmentation of fluent speech into individual words. Speech signal is a continuous stream that does not reliably contain pauses or other language-independent cues to word boundaries. Nevertheless, based on exposure to the ambient language, infants are able to use a variety of acoustic and statistical cues to extract words from fluent speech by the time they are one year old. In particular, experimental work has shown that infants are capable of using phonotactics (Mattys & Jusczyk, 2001; Mattys, Jusczyk, Luce & Morgan, 1999), prosody (Jusczyk, Houston & Newsome, 1999; Mattys et al., 1999; Myers et al., 1996; Saffran, Newport & Aslin, 1996; Thiessen & Saffran, 2004), allophony (Jusczyk, Hohne & Bauman, 1999), coarticulation (Johnson & Jusczyk, 2001), and statistical regularities (Saffran, Aslin & Newport, 1996; Saffran et al., 1996) to segment fluent speech. Since the work of Saffran et al. (1996), research on infant speech segmentation has documented the abilities of infants to use a variety of statistical regularities in the speech signal to segment fluent speech (Jusczyk et al., 1999; Pelucchi, Hay & Saffran, 2009a; Thiessen & Saffran, 2003; Weiss, Gerfen & Mitchel, 2010). These **distributional cues** include transitional probabilities as well as related statistics calculated over sequences of units at various levels of linguistic representation, such as phonemes or syllables.

Although distributional cues have played a prominent role in experimental work with infants, the extent to which distributional cues provide reliable information about word boundaries in spontaneous child-directed speech remains controversial. Based on corpus analyses of child-directed speech, some authors have argued that distributional cues provide sufficient information for building an initial lexicon (Swingley, 2005), while others have argued that distributional cues are unreliable and that accurate segmentation requires strategies relying on principles or constraints derived from Universal Grammar (Gambell & Yang, 2006; Yang, 2004). Results of computational modeling work raise further questions about the reliability of distributional cues. Specifically, state-of-art segmentation models make limited use of the kinds of distributional cues explored in the experimental literature, relying primarily on other sources of information (Batchelder, 2002; Blanchard, Heinz & Golinkoff, 2010; Brent, 1999; Goldwater,

Griffiths & Johnson, 2009; Johnson, 2008b; Liang & Klein, 2009; Venkataraman, 2001). Additionally, with a few notable exceptions (Batchelder, 2002; Blanchard et al., 2010; Fleck, 2008), many models have been developed and tested only on individual languages, usually English. Indeed, performance on languages besides English has generally been markedly worse. In order to better understand the role that distributional cues play in infant speech segmentation across languages, it is crucial to determine how the availability of distributional information in spontaneous speech varies cross-linguistically.

The primary goal of the present work is to contribute to the debate regarding the availability of distributional cues in naturally occurring speech to young children by examining interactions of multiple cues across multiple languages. We use the term 'distributional cues' broadly to include any statistical regularities at the phonological level, such as regularities in relative stress and consonant phonotactics, not just regularities of syllable or phoneme sequences. We examine the extent to which distributional cues of the sort explored in the experimental literature are able to predict word boundaries in child-directed speech across three languages (English, Polish, and Turkish). All the cues we consider are statistics that can be associated with positions between adjacent phonemes in a corpus of transcribed speech. Crucially, as in the experimental studies, the distributional cues we consider are statistics that can be readily calculated from (transcripts of) continuous speech without prior knowledge of word boundaries. We investigate the reliability and the variation in the relative strengths of a variety of sequential statistics calculated at several levels of linguistic representation across the three languages. Much experimental work has investigated the effects that the presence of multiple cues to word boundaries has on infant speech segmentation (Johnson & Jusczyk, 2001; Mattys, 2004; Mattys et al., 1999; Mattys, White & Melhorn, 2005; Morgan & Saffran, 1995; Thiessen & Saffran, 2003; Thiessen & Saffran, 2004). However, the extent to which multiple coinciding cues to word boundaries exist in the primary language data and the extent to which distinct cues capture complementary information are unclear. Accordingly, a central focus of the present study is to determine the extent to which distinct cues reflect complementary information about word boundaries by examining the interaction of multiple cues in spontaneous spoken language.

In sum, the goals of the present study are to examine the *availability* and the *richness* of distributional cues to word boundaries in naturally occurring child-directed speech as well as the variability in the interaction and strengths of these cues across languages. We approach these questions by systematically analyzing the predictive power of a large set of distributional cues to word boundaries individually and in combination in child-directed speech across the three languages. Specifically, in a series of statistical analyses, we use logistic regression modeling to predict word boundaries on the basis of various distributional cues. Our analyses investigate how much information about word boundaries could in principle be extracted by learners with no prior knowledge of word boundaries. Our findings indicate that distributional cues are a rich source of information about word boundaries across languages when the combined contribution of many diverse cues is considered. Although there is overlap in information across cues, there is also a significant amount of complementary information available when cues are combined. Our results suggest that computational models of segmentation have yet to fully utilize the cumulative information available via the array of distributional cues. We also find that cue reliability across languages is variable and reflects the phonological structures of the languages.

Distributional Cues to Word Boundaries in Spontaneous Speech

As discussed above, experimental findings demonstrate that infants are capable of relying on a variety of distributional cues, alone or in combination, for word segmentation. Most of these results have been obtained on the basis of exposure to artificial languages carefully constructed to contain the particular regularities in question while controlling for other factors. The complementary questions investigated in the present work are whether the sorts of regularities used in these experiments are present in spontaneous speech to young children and how they vary and interact across languages.

One way to measure the availability of various sources of information in naturally occurring child-directed speech is by applying computational models of word segmentation to such data. Although there is a large literature on computational approaches to word segmentation, modeling work can be divided into two main approaches: *boundary-finding* models that focus on identifying word boundaries, and *lexicon-building* models that involve the learning of a lexicon of words together with the

segmentation of the input corpus (for similar discussion, see Daland & Pierrehumbert, 2011; Frank, Goldwater, Griffiths & Tenenbaum, 2010). There are many important differences between models within these broad classes (for extensive discussion see e.g. Brent, 1999; Goldwater et al., 2009). However, the distinction between these approaches most relevant to the present goals is that boundary-finding approaches rely exclusively on distributional cues similar to those explored in behavioral studies, while lexicon-building approaches necessarily rely on additional information and biases associated with extracting a lexicon from the speech input. As a result, performance of existing boundary-finding models provides the best estimate of the information content of distributional cues because performance of lexicon-building models reflects additional learner biases and information sources.

The poor performance of certain boundary-finding models has lead some authors to conclude that distributional cues are an unreliable source of information about word boundaries and that learning must rely on principles of Universal Grammar (Gambell & Yang, 2006; Yang, 2004). Specifically, Yang and Gambell tested the segmentation strategy suggested by Saffran et al. (1996), placing word boundaries between adjacent syllables whose transitional probabilities were lower than the transitional probabilities of surrounding syllable transitions. When applied to transcribed English child-directed speech, this strategy correctly identified just 23 percent of the target words and just 42 percent of the words predicted by the model were actual words[1]. Some models relying on statistical regularities perform better, especially when they rely on multiple distributional cues (Aslin, Woodward, LaMendola & Bever, 1996; Cairns, Shillcock, Chater & Levy, 1997; Christiansen, Allen & Seidenberg, 1998; Daland & Pierrehumbert, 2011; Swingley, 2005; Xanthos, 2004). For example, Christiansen et al. (1998) found that a simple recurrent network (Elman, 1990) relying on several kinds of sequential statistics correctly identified about 43 percent of the target words and posited actual words about 45 percent of the time. Brent (1999) tested several boundary-finding strategies using distributional cues and found somewhat

[1] The performance measures reported in this section correspond to *precision* (accuracy) and *recall* (completeness), respectively, (or their harmonic mean, f-score) calculated over word tokens; these measures are defined in (2) below.

higher performance, in the range of 45 to 55 percent, on spontaneous child-directed speech in English.

However, even this performance does not rival the performance of state-of-the-art lexicon-building segmentation models (Batchelder, 2002; Blanchard et al., 2010; Brent, 1999; Fleck, 2008; Goldwater et al., 2009; Johnson, 2008b; Johnson & Goldwater, 2009; Liang & Klein, 2009; Monaghan & Christiansen, 2010; Venkataraman, 2001). Testing on the same corpus as Brent (1999), Johnson and Goldwater (2009) report the highest unsupervised segmentation performance to date (around 88). Earlier segmentation results relying on similar generative models also perform well above the distributional models discussed above (Batchelder, 2002; Blanchard et al., 2010; Brent, 1999; Goldwater et al., 2009; Johnson, 2008b; Liang & Klein, 2009; Venkataraman, 2001). For example, the algorithm proposed by Brent (1999) correctly identifies about 69 percent of the target words and predicts actual words about 67 percent of the time (Goldwater et al., 2009). As already mentioned, these models all involve the learning of a lexicon of words together with the segmentation of the input corpus[2]. From a computational perspective, this lexicon-building approach has clear advantages over purely boundary-finding approaches relying on distributional cues. A lexicon-building learner is able to rely directly on the regularities created by the concatenations of a fixed number of strings (words) in order to extract those strings that recur most regularly and are most likely to be words. Also, as the learner's lexicon grows, the learner can rely increasingly on lexical knowledge in segmenting novel utterances, resulting in improved segmentation over time. However, it is unclear how much of the lexicon-building models' performance can be attributed to distributional cues. While these models certainly rely on distributional regularities to identify viable lexical entries (some more directly than others), the lexicon-building process is also guided by other biases, especially preferences for a lexicon with fewer or shorter words. The lexicon-building strategy and additional biases of the learner thus obscure the relative importance of distributional cues. Indeed, the significant difference in performance between the earliest lexicon-building approaches and the

[2] The model proposed by Fleck (2008) relies on learning extended sequences of word endings and beginnings rather than full-fledged words.

best boundary-finding approaches suggests that it is lexical knowledge, along with the corresponding learning biases, that is responsible for the performance gain. In sum, despite the success of lexicon-building models, the additional information employed by these learners makes it difficult to evaluate the relative contribution of distributional cues to their performance.

Thus, prior computational modeling work raises questions about the availability and effectiveness of distributional cues to word boundaries in spontaneous child-directed speech. Comparing the performance of existing computational models of word segmentation reveals that models relying primarily on distributional cues perform poorly compared to state-of-the-art segmentation models, which do not rely primarily on distributional cues. Analyses of the information content of distributional cues employing supervised techniques yield higher performance than the boundary-finding models discussed above. For example, supervised performance for one kind of distributional cue (the diphone probability, probability of a word boundary falling between two phonemes) on child-directed speech in English is around 70% (Christiansen, Onnis & Hockema, 2009; Daland & Pierrehumbert, 2011; Hockema, 2006; see also Cairns et al., 1997). See Daland & Pierrehumbert (2011) for an unsupervised model based on diphone probabilities. Performance of supervised models with access to word boundaries during learning provides an upper bound on the model's performance. So while these results suggest unsupervised distributional learners may yet be able to extract more information from distributional cues, this upper bound for distributional cues based on diphone probabilities is still substantially lower than performance of (unsupervised) state-of-the-art lexicon-building models.

Despite this clear pattern of results, there are a number of reasons to suspect that existing models do not reflect the full potential of distributional sources of information. First, the experimental results reviewed above have shown that infants are sensitive to a rich set of distributional cues at various levels of linguistic structure, yet many models, including the supervised analyses of the input discussed above, have investigated the segmentation capacities of individual cues. Boundary-predicting models that do rely on multiple distributional cues perform better than similar models relying on single cues (Aslin et al., 1996; Christiansen et al., 1998; Swingley, 2005; Xanthos, 2004), suggesting that multiple cues can

combine productively to provide additional information. However, even these results do not necessarily represent the full potential of distributional information because, like all formal models of acquisition, these models must make specific assumptions about how cues are calculated and used by the learner. Also, these models consider at most a handful of distributional cues so it is possible that incorporating a richer set of cues or combining cues differently could provide more information. Furthermore, while state-of-the-art models perform very well on English, performance of the same models on child-directed speech in other languages is less impressive. For example, performance of several state-of-the-art models on child directed speech in Sesotho is between 40 and 55% (Blanchard et al., 2010; Johnson, 2008a). Thus, information sources that work well in English are not as reliable in other languages. Nevertheless, children do learn to segment successfully in other languages, which suggests they may be relying on additional cues or combinations of cues.

These considerations motivate the present study, which is a systematic examination of the availability of rich distributional cues to word boundaries in spontaneous child-directed speech across languages. In contrast to the unsupervised models discussed above, the goal here is not to model the process by which children segment speech. Rather, the goal is a complementary one, aiming to estimate how much information about word boundaries can in principle be extracted from the speech signal using distributional cues alone. In this respect, the approach pursued here provides a more direct measure of the reliability of distributional cues: it does not presuppose a particular segmentation strategy but instead relies on standard statistical techniques to identify the best way to extract information from each cue (or combination thereof). As a result, in addition to addressing the theoretical debate concerning the availability of distributional cues in the signal, our analyses also examine how a learner could combine distributional cues in order to capture maximal information, identifying avenues for further modeling work. Thus, our approach is comparable to the supervised analyses discussed above, except our primary focus is on the information content of multiple cues and how their strengths and interactions vary across languages. Our approach is to consider a large and varied array of possible distributional cues and let the fitting process select the best way of using the cues in each language in order to determine their potential.

The remainder of the paper is organized as follows. The following section introduces our basic methodology and data, including the use of logistic regression for evaluating the informativeness of cues (and cue combinations) for word segmentation. We then present the set of distributional cues we investigate. Next, we present a series of four statistical analyses examining this set of distributional cues. Finally, we present general discussion and conclusions.

General Method

This section presents an overview of the methodology and describes the data and its transcription.

Evaluating Distributional Cues Using Logistic Regression

The analyses discussed below rely on logistic regression to evaluate the capacities of a variety of distributional cues, individually and in combination, in predicting word boundaries. Logistic regression is a standard statistical approach for binary classification (see e.g. Hastie, Tibshirani & Friedman, 2009). It is a generalized linear model that is used to predict the probability of some event Y in terms of the logistic of the weighted sum of independent variables X_i, as shown in (1).

(1) Logistic Regression Curve: $$p(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}}$$

The coefficients β_i of the model are fitted so as to maximize data likelihood. All the regressions presented here are performed using the statistical computing package R (R Development Core Team, 2008).

Each position between consecutive phonemes in a transcribed corpus of speech either corresponds to a word boundary or not. It is possible to use this binary variable as the dependent variable in a logistic regression model in order to evaluate the capacity of various independent variables at predicting word boundaries. In the present study the independent variables are distributional cues calculated at the corresponding positions in the corpus. For instance, in one of the analyses we examine a logistic regression model that predicts the probability of a word boundary at each position in the corpus based on the bigram transitional probabilities between the two phonemes on either side of that position. Given the bigram transitional probabilities for all the positions in the corpus and the binary vector that indicates for each position whether it is a boundary or not, logistic regression maps the values of the

bigram statistic to a probability of a boundary occurring in a way that best fits the data. In this case, low bigram transitional probabilities get mapped to high likelihoods of boundary occurrence since boundaries are more likely where the two phonemes on either side of the position are unlikely to occur together in connected speech. A major advantage of logistic regression, however, is that it straightforwardly extends to the case when there are multiple distributional cues. In this case, the fitting process determines each cue's association with word boundaries (positive or negative) and the relative weight each of the distributional cues should receive to best fit the data.

Evaluating Segmentation Performance

A fitted logistic regression model can be analyzed in various ways in order to determine how much information the independent variables capture. Our primary objective in the present work is to quantify the predictive content of different cues in such a way that their predictiveness can be compared across cues and languages. A secondary consideration is using a meaningful measure that can be related to previous work. We are able to meet both these goals by using the fitted regression models to predict word boundaries and evaluating the goodness of the resulting predictions. Specifically, in the analyses below we use fitted logistic regression models to predict word boundaries given some threshold of probability, and then we evaluate the resulting predictions using the standard f-score measure, which provides a measure of a model's ability to differentiate between boundaries and non-boundaries.

F-score is a standard measure used for evaluating performance of computational models, and it results in a value between 0 and 1, often expressed as a percentage. It is the harmonic mean of precision, which penalizes false positives and recall, which penalizes false negatives. Precision and recall are also known as accuracy and completeness, respectively. The three measures are defined below in (2).

(2) Evaluation Metrics

a. $$precision = \frac{\#\,true\ positives}{\#\,true\ positives + \#\,false\ positives}$$

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

c. $$f - score = \frac{2 * precision * recall}{precision + recall}$$

The f-scores of fitted regression models can be used to compare the predictiveness of distributional cues and their combinations within and across languages. They can also be compared to the performance of existing computational models of word segmentation. To facilitate comparison with the modeling work reviewed above, we report f-scores calculated over whole word tokens. This means a true positive corresponds to the correct segmentation of an entire word and is counted only when both word boundaries are correctly identified and no spurious boundaries are posited word-internally.

Before continuing, it is important to understand several properties of these logistic regression analyses. As shown above in (1), regression modeling assumes that cues combine additively via linear combination. Combination of cues by weighted sum is a simple and powerful way to model interactions, but it does mean that our analyses cannot extract information that requires more complex cue interactions. Thus, it is important to keep in mind that our analyses may still underestimate the information content of distributional cues. Since we are interested in the potential predictiveness of distributional cues infants could readily extract from connected speech, the cues themselves are calculated from transcriptions without word boundaries. However, in order to evaluate the predictiveness of a cue or cue combination, the logistic regression models in the first three analyses are fitted using word boundaries. This approach thus provides an estimate of the potential information content of distributional cues, estimating an upper-bound for segmentation based on a linear combination of distributional cues. It measures how much information about word boundaries could in principle be extracted by such a distributional learner. In Analysis 4, we show how weighting of cues could be accomplished without access to word boundaries.

The Data and Participants

The analyses discussed below are conducted on transcribed, child-directed speech in three languages: English, Polish, and Turkish. These three languages were chosen because they differ along a number of dimensions of potential relevance to word segmentation. A major goal of this work is to examine differences in the availability and strength of distributional cues to word boundaries across languages, and the differences between these languages enable an evaluation of the effect that such differences make for the effectiveness of distributional cues to word boundaries in word segmentation.

The three languages represent a range of morphological richness and syllable complexity. While English is rather impoverished morphologically, Turkish is an agglutinative language, with complex words formed via the combination of many, easily separable morphemes, each conveying particular meaning or information. Polish is a highly inflected language with a complex system of inflectional morphology that marks words' grammatical functions with fused morphemes expressing multiple pieces of information. Since both morphological boundaries and word boundaries influence phonotactics, it is possible that richer morphology may influence the effectiveness of distributional cues to word boundaries. With respect to syllable complexity, Turkish syllables are the simplest, followed by English, then by Polish. Syllable complexity corresponds to more permissive consonant cluster combinatorics, which may also influence the effectiveness of distributional cues to word boundaries. The languages also vary with respect to the regularity of word-level stress, which is highly regular in Polish but less so in the other two languages. Since experimental work has shown that stress patterns guide infants' segmentation strategies (Jusczyk et al., 1999; Mattys et al., 1999; Saffran et al., 1996), it is useful to examine the effectiveness of cues based on sequential stress regularities in languages with different degrees of stress regularity. Finally, the languages differ with respect to the presence of certain phonological dependencies. Specifically, Polish and Turkish exhibit voicing assimilation and vowel harmony, respectively, which are absent from English. Polish voicing assimilation affects sequences of obstruent consonants (consonants formed by obstructing airflow, such as [s z t d p b]), and causes these sequences to be pronounced with consistent voicing throughout, even when these sequences fall across word boundaries (Gussmann, 1992). The domain of Turkish vowel harmony, in contrast, is the word, and causes suffixes to alternate such that they match root vowels in frontness and rounding (Clements & Sezer, 1982). These dependencies may likewise influence the predictiveness of different distributional cues to word boundaries.

All the child-directed speech is extracted from corpora that are available from the CHILDES database in the form of orthographic transcripts (MacWhinney & Snow, 1985). The English data is extracted from the Bernstein-Ratner corpus (Bernstein-Ratner, 1987) and consists of spontaneous speech to nine children, ages ranging between 13 and 23 months. Transcriptions of the Bernstein-Ratner corpus

have been used for the evaluation of a number of segmentation models (Batchelder, 2002; Bernstein-Ratner, 1987; Blanchard et al., 2010; Brent, 1999; Brent & Cartwright, 1996; Goldwater et al., 2009; Johnson, 2008b; Johnson & Goldwater, 2009; Venkataraman, 2001), and we use it for consistency with previous work. The Turkish data (Slobin, 1982) consists of spontaneous speech to 33 children from 24 months to 56 months of age. Finally, the Polish data (Weist, Wysocka, Witkowska-Stadnik, Buczowska & Konieczna, 1984) consists of spontaneous speech to four children aged 19 months to 38 months. These corpora cover the youngest age ranges available for child-directed speech in Polish and Turkish, but the age ranges are nonetheless somewhat higher than for the English data. It is not clear what effects the difference in age ranges may have on the effectiveness of distributional cues to word boundaries, and this should be investigated in further work. Regardless of the age ranges, the data from each language reflects the phonological structure and general properties of spontaneous, child-directed speech in that language.

Data Transcription and Coding

The orthographic transcripts of child-directed speech were phonemically transcribed in order to arrive at an approximation of the speech signal to which the children were exposed. All phonemic transcripts were created automatically by replacing each orthographic word with its standard pronunciation, as described below. We aimed to be as consistent as possible in the transcription choices across the languages – for all languages we represented all segments, including complex segments like diphthongs, long vowels, and affricates, using single characters. Our automatic transcription methods were determined by the availability of phonetic dictionaries and similar resources for each of the languages. For all languages we used the orthographic word boundaries to represent word boundaries in the phonemic transcripts and the utterances coded in the transcriptions to identify utterance boundaries.

Brent (1999) created phonemic transcriptions of the Bernstein-Ratner English corpus, which was used to evaluate several segmentation models (Batchelder, 2002; Bernstein-Ratner, 1987; Blanchard et al., 2010; Brent, 1999; Brent & Cartwright, 1996; Goldwater et al., 2009; Johnson, 2008b; Johnson & Goldwater, 2009; Venkataraman, 2001). Brent removed disfluencies, nonwords, and utterances not directed at the children. All remaining words were broadly transcribed using a phonemic dictionary. We

were unable to use Brent's transcriptions directly because we are interested in examining word level stress as a distributional cue, and Brent's transcriptions do not include stress. To facilitate comparison with previous work, we wanted to remain as close as possible to Brent's transcriptions, however. We therefore used his orthographic transcripts but transcribed the words using the English phonetic transcriptions in CELEX, which encode stress (Baayen, Piepenbrock & Van Rijn, 1993). We discarded utterances that included words not found in CELEX, resulting in a loss of less than three percent of the utterances. Thus, other than the transcription differences, our English corpus is very close to the corpus used in previous work. The various characteristics of the resulting English corpus are summarized in the second column of Table 1.

For the Polish data, the utterances spoken by parents and grandparents were extracted and processed further (Weist orthographic transcriptions only include utterances directed at the children). We discarded utterances containing nonwords or disfluencies (marked with '@' in the transcriptions) as well as incomplete or interrupted utterances (marked with '+' in the transcriptions). The resulting utterances were automatically phonemicized, which can be reliably done based on the highly phonemic orthography. Each grapheme or digraph was translated into the phoneme corresponding to its standard pronunciation in the given context. Additionally, the phonemic transcripts were processed further to implement final devoicing and regressive voicing assimilation (Gussmann, 1992). Specifically, the voicing of all clusters of consecutive obstruent consonants within utterances was made to match the voicing of the final consonant of the cluster, and any word-final obstruents or clusters not followed by a consonant in the next word were made voiceless. As discussed above, this feature of Polish phonology may influence the effectiveness of distributional cues to word boundaries since voicing assimilation applies both within and across word boundaries. Finally, we were also interested in investigating stress as a cue to word boundaries. Polish lexical stress is very regular so we assigned lexical stress automatically by placing primary stress on the penultimate syllable (ultimate syllable in the case of a monosyllabic word) and placing secondary stresses left-to-right starting with the first syllable (Rubach & Booij, 1985). This automatic stress assignment misses some exceptional stress patterns; however, due to its regularity, it can

provide an upper-bound on the effectiveness of stress as a cue to word boundaries in languages, such as Polish, with predominantly regular stress. Characteristics of the resulting corpus are summarized in the third column of Table 1.

In the Turkish corpus there were very few utterances (about 400) spoken by the primary caretakers; most of the child-directed speech was spoken by the experimenters. Therefore for this data, we extracted the utterances spoken by both caretakers and experimenters for further processing. These orthographically transcribed utterances were automatically phonemicized using a full-scale finite-state implementation of Turkish phonology and morphology developed by Oflazer & Inkelas (2006). Their system provides a pronunciation of each word based on the SAMPA standard and relies on a full morphological analysis, which is essential for the correct placement of primary stress. We replaced the SAMPA multi-character transcriptions of long vowels and affricates with (unique) single characters in order to be consistent with the other two languages. The system does not encode secondary stress, the existence of which is controversial in Turkish (Oflazer & Inkelas, 2006). Any utterances containing a word that could not be phonemicized by the finite-state system were discarded, resulting in a loss of 27.5 percent of the utterances. These utterances were eliminated largely because they contained nonwords, disfluencies, or misspellings. However, this still left a slightly larger set of utterances than either the English or Polish data, as is shown in Table 1.

As shown in Table 1, the corpora are comparable with respect to the number of utterances and word tokens they contain. A number of interesting differences of potential significance to word segmentation performance across the languages are notable. While the number of word tokens across languages is comparable, the number of word types varies dramatically. This likely reflects the morphological complexity of the language, with more types corresponding to richer morphology. Also, while the three languages have a similar average number of words per utterance, the word lengths themselves vary, which may also reflect morphological complexity to some extent. Finally, the number of distinct cluster types (word-internal sequences of consonants), is an indicator of syllable complexity, with more types reflecting a more permissive system of consonant cluster phonotactics.

Distributional Cues and Parameters

In the analyses below we examine a large set of distributional cues. In order to facilitate discussion and analysis, the cues are categorized according to their placement (setting) along a number of dimensions (parameters). Each individual cue reflects a combination of settings of these parameters. Although this set of cues is by no means an exhaustive list of possible distributional cues to word boundaries, we have included cues that vary across multiple dimensions in order to capture a range of distributional information. Specifically, we examine cues that vary between several levels of representation (**level**), several different kinds of statistics (**statistic**), and forward or backward calculation of the statistics (**direction**), where applicable. The final parameter determines whether the actual value of the statistic is used or whether the value is defined in relation to those that surround it in the transcript (**relation**). For each of these parameters, we consider a range of settings, guided by the kinds of distributional cues that have been tested in previous experimental and computational studies. A summary with examples for each of the cue parameter settings can be found in Table 2.

The first parameter according to which we organize the set of cues refers to the **level** of representation over which statistics are calculated. In prior experimental and computational work, statistical dependencies have been calculated at various levels, including the phoneme (see Mattys & Jusczyk (2001) and Brent (1999) for examples). At the phoneme (**P**) level, cues are simply calculated over the entire sequence of phoneme and utterance boundary symbols, and the values of the statistics are then associated with each position. Distributional cues to word boundaries can also be found across non-adjacent vowels and non−adjacent consonants (Newport & Aslin, 2004). We therefore include both consonant (**C**) and vowel (**V**) levels in our set. Vowel level cues are calculated on a version of the corpus with all consonants removed. In order to associate cues that reference only vowels with all positions in the original corpus, we simply repeat the same value for all positions between two vowels. For example, in a sequence such as $V_1C_1C_2C_3V_2$, the vowel-level bigram probability of V_2 given V_1 would be associated with each of the four positions between V_1 and V_2. Thus, vowel level cues alone are incapable of

distinguishing among the positions between two consecutive vowels, but they may still distinguish among positions in different vocalic contexts. Such coarse-grained cues may provide valuable information when considered in conjunction with finer-grained cues. The calculation of consonant level cues is analogous and involves repeating the value of the statistic across any intervening vowels. We focus our investigation on cues that could be extracted from the input data without prior language-specific knowledge. Therefore, we do not consider a syllable level since the relationship between syllable boundaries and word boundaries must be learned on a language-particular basis. Nonetheless, the consonant and vowel level cues capture some of the same sorts of long-distance dependencies that syllable level cues would. In particular, vowel-to-vowel dependencies are syllable-to-syllable dependencies that reference just the nuclei of syllables. We also examine stress (**S**) level cues, which are calculated just like vowel level cues except that calculations are made over degrees of stress (0, 1, or 2 in English and Polish, and 0 or 1 in Turkish) rather than distinct vowels. This allows statistical regularities in relative stress to be used in a language-independent fashion, just like statistical regularities referencing other units of sound structure.

Another parameter according to which we organize the set of cues refers to the kind of **statistic** being calculated. Although the most prominent statistics in the experimental literature are bigram transitional probabilities, both trigram transitional probabilities (Aslin et al., 1996; Blanchard et al., 2010; Cairns et al., 1997) and mutual information (Brent 1999) have been explored in the computational literature. All three measures are defined in (3):

(3) Statistics

a. Bigram: $\Pr(y \mid x) = \Pr(xy)/\Pr(x)$

b. Trigram: $\Pr(z \mid xy) = \Pr(xyz)/\Pr(xy)$

c. Mutual Information: $MI(x,y) = \log\left(\Pr(x,y)/\Pr(x)\Pr(y)\right)$

Since we define cues at various levels, we use the term ‘unit’ to refer to the individual elements, such as phonemes or stress levels, over which cues are calculated. Trigram transitional probabilities measure the probability of a unit given the two preceding elements and therefore capture somewhat longer distance statistical relationships, while mutual information provides a symmetric measure of the co-dependence

between two adjacent elements. We calculate all statistics using their relative frequency estimates. Transitional probabilities can be used in one of two ways for segmentation. One approach is to posit boundaries at positions of low probability where adjacent units are not statistically cohesive (Brent, 1999; Cairns et al., 1997; Elman, 1990). We call this approach 'unit-predicting' since the calculation of the cue references the unit following the position in question. Another approach, which we call 'boundary-predicting', posits boundaries at positions where the following symbol is likely to be an utterance boundary '#' (Allen & Christiansen, 1996; Aslin et al., 1996; Brent, 1999; Christiansen et al., 1998). We explore both of these approaches for bigrams and trigrams. In sum, we examine mutual information (**M**), unit-predicting bigrams (**B**), boundary-predicting bigrams (**#B**), unit-predicting trigrams (**T**), and boundary-predicting trigrams (**#T**).

In addition to these four basic levels and five types of statistics, we also define a parameter setting that is best understood as a combination of level and statistic type. The combination of consonant level and trigram statistics captures some nonadjacent dependencies between consonants, but these statistics only capture dependencies up to a fixed distance. In order to capture longer-distance dependencies within consonant clusters, we define a setting we refer to as cluster (**CL**). This setting allows some of the same consonant cluster dependencies as captured by syllables to be reflected in our set of cues. The intuition behind the cluster setting is that the probability of a boundary occurring at different positions within a consonant cluster may depend on the sequence of consonants preceding it and may reflect the appropriateness of treating that sequence as a syllable coda. Accordingly, we define one cluster statistic (**#CL**) as the bigram probability of an utterance boundary, given the entire sequence of consonants (up to a vowel or another utterance boundary) preceding it[3]. Like bigrams and trigrams, the cluster statistic measures the probability of one unit (here, an utterance boundary), given others. But unlike bigrams and trigrams, the length of the unit that predicts the utterance boundary is not fixed. It varies depending upon

[3] In our implementation, the cluster statistic distinguishes sequences of consonants following vowels from sequences following '#', and thus has access to whether or not a given cluster occurs at the beginning of an utterance.

the number of adjacent consonants at a given position. We define a second cluster statistic (**#|CL|**) as the bigram probability of '#' given the *length* of the preceding cluster, a (small) non-negative integer.

Table 2 shows cue examples for two different contexts. Since the cluster settings incorporate a level of representation and a type of statistic, they are listed under both parameters and are discussed with respect to both parameters throughout the analyses. However, with respect to cue parameter combinatorics, clusters do not vary along the level and statistic parameters – they are a fixed combination of level and statistic. Thus, there are 22 combinations of settings along the level and statistic parameters: two cluster settings plus twenty combinations of the four basic levels and five types of statistics.

The next parameter according to which we organize the set of cues refers to the **direction** in which the cues are calculated. Since most of the statistics described above are directional (with the exception of mutual information, which is symmetric[4]), they can be calculated such that the predicted element follows the conditioning element(s) in the speech stream (forward**: F**) or such that the predicted element precedes the conditioning element(s) in the speech stream (backward**: B**). Although the vast majority of previous experimental and computational work on segmentation examines forward transitional probabilities, many of the statistical dependencies used in the artificial languages are consistent with backward transitional probabilities (Pelucchi et al., 2009a; Pelucchi, Hay & Saffran, 2009b; Perruchet & Desaulty, 2008). We are aware of one prior corpus analysis that examined the usefulness of backward transitional probabilities for segmentation (Swingley, 1999). Recently there has been a surge of interest in backward transitional probabilities with studies showing that both adults (Perruchet & Desaulty, 2008) and infants (Pelucchi et al., 2009a) are able to segment artificial speech using only backward bigrams. Work on adult speech production has also found effects of backward

[4] While mutual information is symmetric, we calculate it in both directions because in our implementation Direction also controls what is calculated across utterance boundaries. In the forward direction, positions corresponding to utterance boundaries are associated with statistics that refer to the final units in the utterance followed by #, while in the backward direction utterance boundaries are associated with # followed by the initial units of the utterance.

transitional probabilities (see e.g. Bell et al., 2003). We therefore include Direction as a parameter in order to examine the relative information content of cues calculated in both directions across languages.

Finally, the last parameter we consider refers to the relationship between a cue's value and the values associated with neighboring positions (**relation**). In addition to examining the predictiveness of the value of each statistic at a given position (**Curr**), as explored in a number of computational studies (Cairns et al., 1997; Christiansen et al., 1998; Elman, 1990), we consider several relationships. We examine the possibility that the value at the subsequent position is indicative of the current position's likelihood of being a boundary (**Next**). The intuition behind this manipulation is that neighboring positions' low probability of being a boundary may be predictive of the current position's high probability of being a boundary and vice-versa since words tend to be longer than one phoneme. Following previous work, we also examine the predictiveness of 'peaks' and 'dips' in the statistics relative to the values of the surrounding units' statistics (Adriaans & Kager, 2010; Brent, 1999; Gambell & Yang, 2006; Saffran et al., 1996; Yang, 2004). We include one relative cue setting that calculates the difference between the current value and the previous value (**Diff**), and another that calculates a ranking of the current value relative to its two neighbors (**Rank**). The Rank setting permits precisely the kind of segmentation strategy suggested by Saffran et al. (1996) to be considered in the analyses below, while the Diff variant provides an alternative relational statistic, one that allows a more gradient relational measure.

Altogether, we examine 22 combinations of Level and Statistic, 4 settings along the Relation dimension, and 2 Directions. This results in 176 distributional cues, as summarized in Table 2.

Analyses

The following sections present four analyses examining the informativeness of the 176 distributional cues defined above individually and in combination, within and across languages. Analysis 1 examines the predictiveness of each of the 176 cues individually and compares the cues' performance across languages. Analysis 2 explores the degree to which information from multiple cues can be productively integrated by examining segmentation performance relying on successively larger sets of cues. In Analysis 3 we investigate whether the set of cue parameter settings defined above can be reduced

without affecting segmentation performance. As discussed earlier, these logistic regression analyses use word boundaries to find the best way to extract information from these cues. In Analysis 4, we present an initial exploration of how such a weighting could be accomplished without access to word boundaries.

Analysis 1

In the first analysis we examine the amount of information contained within individual cues in each of the languages, comparing cues across languages and to one another.

Method

For each cue in each language, we fit a regression model, use that regression model to predict boundaries in the corpus, and evaluate the f-score of that predicted segmentation. The fitted regression model generates a probability of a word boundary at each position in the corpus. We use this probability to categorically decide between boundary or non-boundary. Since our goal is to determine the amount of information captured by the cues, we choose the threshold of probability that maximizes the f-score for each cue. This method of thresholding captures the amount of information that could in principle be extracted from each cue. An alternative method, placing a boundary whenever the regression model predicts a boundary with at least 0.5 probability, would underestimate the information content of weak cues. This is because the probability of word boundaries is well below 0.5 in all three languages, and the curves of many weak cues that nevertheless capture information are very flat, never rising above 0.5.

In order to avoid over-fitting of the regression models, we perform two-fold cross-validation. We divide the corpus into two halves, use one half to fit the regression model and choose the threshold, and then use that regression model and threshold to predict boundaries on the other half. We do this separately for each half and then calculate the f-score over the combined segmentations. Furthermore, in order to check the robustness of the f-scores, we perform simple bootstrapping on a portion of the cues as described below (Efron & Tibshirani, 1993). This involves taking a random sample (the same size as the original data) with replacement from the original data, fitting models to data from the sample, and then evaluating the resulting models and thresholds on the original data.

Also, in order to determine whether the cues contain any information whatsoever, we determine a

baseline to which the performance of the cues for each language can be compared. Following Brent (1999), we implemented a baseline that randomly places *n* word boundaries in each corpus, where *n* is the actual number of word boundaries in that corpus. The baseline thus relies on language-specific knowledge, namely the actual proportion of word boundaries, but assigns the locations of these boundaries randomly. Due to the randomness, we repeated the random assignment of boundaries 1000 times for each language to determine the average f-scores of the baseline. The f-scores of this baseline calculated over word tokens are 12.2%, 8.5%, and 6.5% for English, Polish, and Turkish, respectively.

Results

Fig. 1 shows the cross-validated f-scores of all the cues in each language (in descending order of f-score) as well as the baselines for each language. This makes it easy to see the range of f-scores attained within each language and the proportion of cues performing above the threshold. Overall performance in English is highest, with the best cues reaching 68.9%. In Polish the f-scores reach 39.3%, and in Turkish they reach 39.5%. Thus, the languages show marked differences with respect to the performance of their best individual cues, with substantially higher performance in English. The figure also shows, however, that there are a substantial number of cues within each language whose cross-validated f-scores fall above baseline. Using simple bootstrapping to confirm the robustness of above-baseline performance[5], we found the number of cues performing above baseline to be 73 in English, 80 in Polish, and 74 in Turkish. Also, 99 of the cues were above baseline in one or more languages. Thus, a substantial number of cues capture important information about word boundaries within and across languages.

We also examined cue performance to determine whether cue performance is generally consistent across languages, that is, whether the same cues tend to perform well across languages. In general, cue performance is consistent across languages. The correlation in f-scores between English and Polish is 89.3, 78.1 between English and Turkish, and 83.2 between Turkish and Polish. Despite this overall

[5] To determine consistency we ran 100 bootstrap samples for all cues within 10% of their baselines (counting only those that fell above their baseline on all runs). Cues 10% or more above baseline were also counted as consistent.

consistency, there are significant differences in cue performance across languages. For example, the top-performing cue in English (the rank of the boundary-predicting backwards phoneme-level trigram probability) achieves f-scores of only 28.6% in Polish and 20.1% in Turkish.

Table 3 provides a more systematic comparison of how the languages vary with respect to which cues capture the most information. Table 3 shows f-scores of the highest performing cues for each of the parameter settings for each language, making it possible to compare performance of the best individual cues across settings and across languages[6]. This also makes it possible to determine whether there are informative cues at each of the parameter settings. Table 3 also shows the results of simple bootstrapping for each of these cues, which reveals that the f-scores are robust, with most f-scores varying little across the 100 bootstrap samples (standard deviations less than 0.1%). With a few exceptions, the bootstrapping thus indicates that f-score differences between the best cues for various parameter settings are meaningful. With the exception of vowel-level cues in English, the top cues for each parameter setting score above their respective baselines in each language, showing that some information about word boundaries is available for each parameter setting across languages. In other words, all the parameter settings we consider contain reliable information about word boundaries in at least one language.

In addition, Table 3 shows there are notable differences in relative performance across the languages. The top cues at the Level parameter show the greatest range of f-scores, indicating that the level of representation is crucial for choosing the most informative cues. Furthermore, the languages differ with respect to the relative informativeness of cues at the various levels in a way that reflects their phonological structures. For example, while vowel level cues are least informative and close to or at baseline in English and Polish, they provide reliable information in Turkish. In fact, the 45th best individual cue in Turkish is a vowel-level cue in spite of the fact that these cues are at an inherent disadvantage due to their inability to distinguish among positions between consecutive vowels. The

[6] Our focus is on best-performing cues, but see Table 5 (Appendix) for average f-scores within parameter settings. See first author's website (http://pantheon.yale.edu/~gjs42) for complete results of Analysis 1 & 2.

informativeness of vowel level cues in Turkish likely reflects the regularities created by the system of vowel harmony in Turkish. Stress level cues, which share the same disadvantage, fare better relative to cues in Polish than in the other languages, likely reflecting the regularity of lexical stress in the Polish corpus. Also, it is noteworthy that in the language with the largest consonant cluster inventory, Polish, cluster level cues are the most informative kinds of cues.

Discussion

These analyses have established several key facts. First, the best performing cues have cross-validated f-scores of 68.9%, 39.3%, and 39.5% in English, Polish, and Turkish, respectively. These figures estimate how much information could in principle be extracted by distributional learners relying on individual cues. The results for English confirm the supervised performance discussed above based on diphone probabilities and also demonstrate that this level of performance can be achieved with minimal reliance on words boundaries. In particular, whereas the supervised diphone approaches fit one parameter for each phoneme pair (depending on the number of phonemes, this can be up to 6241 parameters) based on statistics calculated from a segmented corpus, in our approach word boundaries are used only to set one parameter, the weight of the cue in the regression model - the cues themselves are estimated from an unsegmented corpus. The analysis also establishes significantly lower maximal f-scores of around 40% for child-directed speech in Polish and Turkish, which to our knowledge have not been examined previously. Although our set of 176 cues is by no means exhaustive, these results strongly suggest that successful segmentation must involve more than singleton distributional cues.

One of the most interesting empirical findings of this analysis is that the top cues in all three languages are calculated in the backwards direction. This is notable considering that most previous work, both experimental and computational, relies on forward calculation of sequential statistics. As discussed earlier, our investigation of backwards statistics parallels recent experimental results showing that humans can use backwards bigrams to segment speech (Pelucchi et al., 2009a; Perruchet & Desaulty, 2008). There is also related work showing an important role of backwards predictability effects in adult production (Bell et al., 2003; Jaeger & Kidd, 2008). Together with these results, our finding that

backwards statistics are in fact a highly informative cue to word boundaries across languages motivates further examination of infants' abilities to segment speech based on backwards cues defined at various levels of representation.

A major novel contribution of this analysis, however, is in establishing that there is reliable distributional information present across a vast and diverse set of distributional cues across languages. A substantial number of cues (99) perform above the baselines in one or more languages, and these high-performing cues are varied, with reliable distributional cues present at all the parameter settings. These results complement experimental findings showing infants' sensitivities to a wide array of distributional regularities. However, our results also highlight the variation in reliability of cues. While the sources of information are rich and varied in all languages, the relative informativeness of cues depends on the language. These results also indicate that the precise way in which distributional cues are formulated is crucial and has major consequences for the amount of information that can be extracted from distributional sources in any particular language. They imply that if learners are to make the most of distributional sources of information, they will need to be sensitive to the relative reliabilities of different cues in order to identify the most reliable cues in the ambient language.

Analysis 2

Analysis 1 showed that information about word boundaries is available in a rich set of distributional cues cross-linguistically. However, it is not clear to what extent the information captured by distinct cues overlaps. In particular, do different cues contain mutually redundant information or is the information content of different cues largely complementary? Analysis 2 investigates this question by examining the cumulative information content of multiple cues.

Method

In this analysis, we perform a step-wise multiple logistic regression for each language. This is a cumulative procedure that adds cues one-by-one to a multiple regression model until all cues are included in the full model. During each iteration the procedure considers each of the unused cues by fitting multiple logistic regression models that include each of them plus the accumulated cues that were selected

on earlier iterations. It selects and adds the cue that improves likelihood the most to the current model. This procedure is not optimal (in the sense that each iteration is not guaranteed to contain the set of cues that captures maximal information); however, it provides a good approximation of how the incorporation of additional cues affects the cumulative information content.

After the addition of each cue, the f-score of the resulting regression model is evaluated using two-fold cross validation, as in Analysis 1. That is, a regression model with the current set of cues is fitted for each half of the data and used to predict boundaries on the other half of the data. For these boundary predictions we simply use a fixed threshold of .5 for all cues[7]. The f-score is calculated on the basis of the predicted segmentations for both halves. Also as in Analysis 1, we confirm the reliability of the f-scores by performing simple bootstrapping on the full models incorporating all cues.

Results

The results are summarized in Fig. 2. As cues are added, the f-scores increase dramatically at first, then more slowly, and finally plateau. Overall, multiple regressions in each of the languages dramatically improve upon the performance of the best individual cues. In English the f-scores increase from 68.9% to 90.8%, in Polish from 39.3% to 81.2%, and in Turkish from 39.5% to 78.7%. Validated performance based on 100 bootstrap samples indicates that the f-scores of the complete models are robust, with average f-scores of 91.6%, 81.7%, 80.9% and standard deviations of 0.12%, 0.11%, and 0.23% for English, Polish, and Turkish, respectively[8]. Furthermore, a large number of cues are needed before the f-scores of the full models are reached. In English, the plateau is reached after roughly 65 cues, in Polish after roughly 115 cues, and in Turkish after roughly 85 cues.

Additionally, although the complete results of these multiple regressions are too cumbersome to

[7] The multiple regression models quickly yield strong enough predictions to obviate oracle selection of thresholds.

[8] The bootstrap f-scores for multiple regression models (see also results of Analysis 3) are often slightly above the cross-validated f-scores. The cross-validation, by splitting the data in half for train and test, provides a stricter evaluation because the two sets of data represent speech of different speakers. Slight differences can be expected since previous results show variation in performance for different speakers (Monaghan & Christiansen, 2010).

present in full, we would like to make several observations. In Analysis 1 we saw that the best vowel level cue in Turkish was ranked 45th. In the step-wise multiple regression analysis, however, a vowel level cue is the third cue to be added to the model. Thus, after just two cues had been included in the model, a vowel level cue was found to contain the most non-redundant additional information. Similarly, in Polish, the top stress level cue was ranked 44th individually, but a stress level cue was the second cue to be added to the multiple regression model. These results underscore the relative importance of vowel level and stress level cues in the Turkish and Polish data, respectively. Perhaps more importantly they illustrate that the information content of cues depends crucially on what other information is also available: cues that appear weak individually may nonetheless capture substantial non-redundant information.

In sum, a large portion of the distributional cues examined here capture distinct, non-redundant information. This is evidenced by the dramatic increase in f-score of the full models as compared to top performing individual cues and the large number of cues needed to reach the f-score of the full models. Further, these analyses show that information content of cues is conditional on availability of information from other cues: consideration of the utility of a cue in isolation is a misleading measure of its relative contribution in a richer context because strong cues can be largely redundant with other strong cues.

Discussion

The main result of this analysis is that the cumulative information content of the set of distributional cues is substantial and well beyond the performance of individual cues. Distributional cues are not entirely redundant: they can be combined to yield significant gains in information about word boundaries. The word token f-score of around 90% for English is above previously reported f-scores based on supervised analyses of distributional cues, and shows that in principle there is enough rich distributional information for segmentation performance comparable to that of state-of-the-art lexicon-building approaches. The extent to which this rich source of information can be harnessed without relying on an oracle to set the cue weights is an open question, but this result provides reason to be optimistic about this possibility. The word token f-scores of around 80% for Polish and Turkish provide even stronger evidence for the cumulative information content of distributional cues since these represent an

increase from around 40% for individual cues. This dramatic increase is significant because it illustrates weaker cues can productively combine to help narrow the gap in performance between languages. This suggests that part of the key to explaining how successful segmentation occurs across languages may lie in learning strategies that incorporate information from a wide array of distributional cues. It should be reiterated that the regression models assume information from multiple cues is combined via weighted sum and therefore estimate an upper-bound for learners under this assumption. It is possible that even more distributional information could be extracted using more powerful models.

Analysis 3

The results of Analysis 2 also suggest that a large number of cues are needed in order to extract all the available distributional information. Analysis 3 addresses this question from another perspective by asking whether it is possible to reduce the number of parameter settings without sacrificing performance. Analysis 3 also explores how the conditional information content of distributional cues varies across languages and how it differs from the unconditional performance examined in Analysis 1.

Method

The analysis presented in this section is a systematic examination of the contribution of various parameter settings to overall segmentation performance. In this analysis we use step-wise multiple regression performed over the sets of cues associated with each parameter setting. For each parameter, we iteratively add in the set of cues (corresponding to a parameter setting) that improves model fit most until all parameter settings have been included. For example, for the Level parameter, we consider a series of five increasingly richer models to which at each iteration all the cues calculated at a particular level of representation are added, until cues from all levels have been included. Every iteration we evaluate the performance of the models using two-fold cross validation and perform simple bootstrapping to determine the robustness of the resulting f-scores.

In addition to examining the necessity of parameter settings, this analysis provides a systematic evaluation of how information content of cues is conditional on the presence of other cues in the model. It investigates how the relative information content of cues in a cumulative model can differ from the

relative information content of cues in isolation. It also examines how the conditional information content of cues at various parameter settings varies across the three languages. In this analysis we collapse the two cluster statistics and refer to then as CL in the tables below. These parameter settings contain fewer cues than other settings at the Level and Statistic parameters so this move reduces the inherent disadvantage these parameter settings face in these multiple regression analyses.

Results

Table 4 presents the results of the step-wise multiple regressions over sets of cues within parameter settings[9]. F-scores of the full models are repeated from Analysis 2 for convenience, and the table also shows the results of 100 bootstrap samples for each of the models[10]. As with the previous analyses, the bootstrapping results indicate that f-scores are robust and vary little across the 100 bootstrap samples (all standard deviations are below 0.32%)[11]. As a result, most of the f-score differences in the table can be interpreted as meaningful, with a few exceptions. The models whose f-score ranges overlap with the f-score ranges of the simpler models of the previous step are shaded in grey. Specifically, the vowel-level cues in English and Polish make little improvement to f-scores, with bootstrap f-scores overlapping the f-scores of simpler models without the vowel-level cues. This indicates that vowel-level cues do not provide significant improvement in English and Polish. The other case of insubstantial improvement is for the Statistic parameter in Polish, for which the f-score ranges for the last two steps overlap with the previous steps. The unit-predicting bigrams and mutual information statistics do contribute some information in Polish, but the model without these statistics performs nearly as well as the full model. In all other cases, however, these results indicate that all remaining parameter settings contribute non-redundant information in at least one of the languages. To confirm that the step-wise

[9] In doing this analysis, we also evaluated the cumulative information content for the set of cues at each of the parameter settings individually. These results are summarized in Table 6 in the Appendix.

[10] Table 7 in the Appendix shows the BIC (Bayes Information Criterion) for each of these models.

[11] For why bootstrap f-scores are sometimes slightly higher than the cross-validated f-scores see fn. 8.

regression did not select a sub-optimal set of parameters in the penultimate step, we also considered for each parameter setting the model without that parameter setting. These results (not reported) corroborate the results in Table 4; none of these models outperformed the penultimate models in the table. Thus, while several parameter settings contribute little information in some languages, our results indicate all the parameter settings we consider are needed for models in all languages to reach their full potential.

Several observations regarding performance of particular settings can be made. Although stress level cues were not very informative individually, they are the second group to be added (after phoneme level cues) in all languages. This indicates that stress level cues capture the most additional information about word boundaries given the information captured at the phoneme level. This result complements experimental findings showing that stress cues are important in early word segmentation (Jusczyk et al., 1999; Mattys et al., 1999; Thiessen & Saffran, 2004) and corroborates prior modeling results showing improved segmentation performance after incorporation of stress information (Christiansen et al., 1998; Hockema, 2006). Conversely, cluster level cues were among the most informative cues individually, yet they are added last or second-to-last in the step-wise regressions. Their addition substantially improves performance in Polish, but their late addition nonetheless suggests that the information they capture is largely redundant with the information captured by phoneme level cues. Additionally, the first two statistics settings in all languages include boundary-predicting transitional probabilities and unit-predicting transitional probabilities, suggesting that these two kinds of statistics capture a good deal of complementary information. Likewise, the first two settings along the Relation dimension added in all languages include one absolute (Curr or Next) and one relative cue (Rank), again suggesting that these kinds of cues capture different types of information. As in Analysis 2, this step-wise regression shows that information content of cues is relative – performance within individual parameter settings is not a good predictor of the cumulative information each setting contributes to the full model once other settings are factored in.

Discussion

This analysis examined the richness of distributional information by investigating whether the set

of parameters could be reduced without sacrificing performance. Although Analysis 2 showed that there is some redundant information contained within the set of cues as a whole, the results of this section show that there is no systematic way to reduce the set of cue parameters without affecting performance in one or more languages. This does not mean that some smaller set of cues could not be found to perform as well or nearly as well as this set of 176. What it does illustrate, however, is that the interaction of information captured by multiple cues is complex and varied across languages. Cues that perform poorly in isolation may provide vital information once other sources of information are incorporated. Conversely, cues that perform well in isolation may be largely redundant with other highly performing cues. Furthermore, the relative importance of particular distributional cues is language-dependent, with some cues that capture crucial information in one language providing nearly no improvement in another. This suggests that in order to make full use of distributional information, learners must be able to consider a diverse set of distributional cues to identify the most reliable combination in the ambient language.

Analysis 4

The findings of Analyses 1-3 show that distributional cues are a rich source of information about word boundaries across languages when multiple cues are used simultaneously and weighted appropriately. However, as Analyses 1-3 illustrate, the interactions among distributional cues are complex and varied across languages. If learners are to make full use of this rich information, they must be capable of adapting their segmentation strategies in response to the language input in order to determine the relative reliability of different cues. In order to evaluate the potential information content of cues, the preceding analyses relied on an oracle to set the relative weights of the cues in such a way as to extract maximal information. Given these findings, it is not clear, however, to what extent this rich source of information can be harnessed by unsupervised learners without access to such an oracle. Although a complete answer to this question is beyond the scope of this paper, the analysis described in this section is an initial investigation into the utility of rich distributional cues in a fully unsupervised setting.

Method

The method we employ in this analysis is an extension of the methods employed in Analyses 1 –

3. Logistic regression is not inherently suited to unsupervised learning, and we emphasize that this analysis is a preliminary investigation of the usefulness of these cues in an unsupervised setting and likely underestimates their potential. The basic insight is to use distributional cues evaluated at and around utterance boundaries to predict word boundaries, an idea that has been explored in a number previous studies (Allen & Christiansen, 1996; Aslin et al., 1996; Brent, 1999; Christiansen et al., 1998; Daland & Pierrehumbert, 2011; Fleck, 2008). We use all positions corresponding to utterance boundaries as positive examples of boundaries for fitting purposes. In order to provide the fitting process with non-boundaries as well, we make the simplifying assumption that the positions adjacent to utterance boundaries are non-boundaries. This assumption is mostly harmless as the proportions of false negatives it assumes are just 0.021, 0.082, and 0.015, in English, Polish, and Turkish, respectively. Thus, we create new training sets for each language that consist of just utterance boundaries and the positions adjacent to them on the left and right, labeling these adjacent positions as non-boundaries.

These new training sets are much smaller and simpler than the original data, and few cues are needed to model these data perfectly. Recall that the cues themselves have access to utterance boundaries so the task of predicting them is much simpler than that of predicting the word boundaries. Since we are not interested in fitting this data perfectly, but rather want to be able to generalize from these training sets to the rest of the data, we select a subset of cues to use for each language using these training sets. Specifically, for each language we perform step-wise logistic regression on these new training sets until the f-scores reach 100%[12]. This procedure is not guaranteed to identify the best cues for segmentation, but it identifies viable cues capable of distinguishing utterance boundaries from the positions adjacent to them, assuming those are non-boundaries. We then examine each of the models considered by the step-wise regression procedures for their generalization capacities, evaluating them on the original data.

[12] Once f-scores reach 100%, there is no basis on which to prefer some cues over others. Furthermore, if additional cues were added, there would be no controlling for how the fitting process weighted the various cues, as many solutions would be possible, so the contribution of any particular cue could not be guaranteed or controlled.

Once fitted on the new training set, each of the models can be used to generate the probability of a boundary for each position in the original data. Since we did not want to assume any language-specific information for this analysis, we evaluate and present the predictions of the models for the same range of thresholds in all languages. Specifically, for each model we perform separate f-score evaluations at rates of boundary prediction ranging between .15 and .40. For example, for a rate of .35 a threshold is selected so that the proportion of boundaries predicted equals 35%. We chose this method of thresholding because the actual probability thresholds corresponding to these rates are highly arbitrary.

Results

The generalization results for all models considered during the step-wise regressions are shown in Fig. 3. For readability, we present the f-scores at thresholds .25, .3, and .35 only – performance at higher and lower thresholds was poorer, while performance at intermediate values was similar. Perfect f-scores on the new training data were reached within 11 cues for English, 14 cues for Polish, and 3 cues for Turkish. The ability of the models to generalize from utterance boundaries is highly dependent on the cues used and their weights, and our procedure for automatically choosing cues is not capable of gauging generalization ability in any way (it only measures how well the cues distinguish utterance boundaries from non-boundaries). Therefore, as cues are added to the models, and the weights for the regressions are calculated from scratch, performance on the original data sometimes goes down and then up.

The best performance for English (Fig. 3a) is at a threshold of .35 and reaches an f-score of 77.3% after two cues. The best performance for Polish is 41.2%, which is reached after six cues at a threshold of .25 (Fig. 3b). Finally, the best performance for Turkish is 34.4% and is reached after two cues at threshold .25 (Fig. 3c). In sum, performance reaches f-scores of 77.3%, 41.2%, and 34.4% in English, Polish, and Turkish, respectively. Notably, in each of the languages the highest f-score is reached after multiple cues are used, again showing a cumulative effect of multiple distributional cues, this time in an unsupervised setting. Performance on English is once again much higher than for the other two languages. Since the method employed here relies on generalizing from utterance boundaries to all word boundaries, the differences in performance may reflect differences in how representative of word

boundaries utterance boundaries are in these languages. We emphasize that this analysis is a preliminary exploration; further work employing methods better suited to unsupervised learning is needed to uncover the true potential of unsupervised learners to exploit the richness of distributional information.

Discussion

Despite the preliminary nature of this approach, the performance of the resulting models provides reason to be optimistic about the prospects of unsupervised learning with rich distributional information. As discussed earlier, prior unsupervised segmentation models relying on distributional information alone achieved word token f-scores in the range of 45-55% for English child-directed speech. Our unsupervised results for English, reaching word token f-scores of 77.3%, are substantially higher than previous models relying on distributional cues alone. Indeed, the performance on English rivals that of recent lexicon-building approaches, which, until Johnson and Goldwater's 2009 result, achieved word token f-scores in the range of 70-80% on English child-directed speech (Batchelder, 2002; Blanchard et al., 2010; Brent, 1999; Fleck, 2008; Goldwater et al., 2009; Johnson, 2008b; Venkataraman, 2001). In Polish and Turkish, the unsupervised results are less impressive, with less of the potential distributional information in these languages (around 80% f-score) captured by these models. As discussed above, the explorations in this section are preliminary, and we believe better-suited techniques will be able to extract much more of this information. Nonetheless, these results demonstrate that rich distributional cues hold great potential for the task of unsupervised word segmentation, a task we hope future work will explore more fully.

General Discussion

This study investigated the reliability and richness of distributional cues to word boundaries in spontaneous child-directed speech in English, Polish, and Turkish. Analysis 1 showed that information about word boundaries is available in a large and diverse set of distributional cues across languages. It also showed for two previously unexplored languages that the reliability of individual cues is not constant across languages, with the best singleton cues providing much less information in Turkish and Polish than in English. Analyses 2 and 3 focused on the combined information content of multiple cues, showing that the cumulative information content from distributional sources is substantial and helps to narrow the

performance differences between languages. These analyses also illustrate that cues that appear weak individually may provide substantial complementary information in combination with other cues. All three analyses highlight the richness of distributional information and the language differences in the relative reliability of cues, showing that the most informative sets of cues vary depending on the language. Finally, Analysis 4 provided initial investigations using this rich information in an unsupervised setting, providing reason to be optimistic that this rich source of information can be exploited without supervision. Together these findings suggest that successful segmentation across languages may depend on learners' abilities to consider a wide range of distributional regularities and to integrate information from many distributional cues in a way that reflects their relative reliabilities in the ambient language.

Our findings highlighting the role of interacting cues across various levels of representation parallel recent developments in theoretical and computational phonology. Optimality Theory, the dominant theoretical framework in phonology, formalizes phonological well-formedness in terms of the interaction of constraints stated over cross-cutting levels of representation (Prince & Smolensky, 1993/2004). Probabilistic extensions of Optimality Theory have been used successfully to model gradient phonotactic knowledge relying on an integration of soft constraints referencing various aspects of phonological representation (Boersma, 1997; Coetzee & Pater, 2008; Hayes & Londe, 2006; Hayes & Wilson, 2008). Some of these formal proposals rely on a model of constraint interaction that is closely related to the kind of cue interactions assumed in our logistic regression analyses (Hayes & Wilson, 2008; Pater, 2009). Thus, there are close connections between the kind of cue integration explored here and a large body of literature on the modeling of gradient phonotactics. Most of the work on gradient phonotactics has focused on modeling well-formedness of isolated words; however, there is also exciting new work approaching segmentation from this perspective (Adriaans & Kager, 2010). We hope future experimental and computational work will further pursue these connections by developing and testing models of infant segmentation that build on the joint findings in the segmentation literature and the literature on modeling gradient phonotactics via the interaction of multiple cues.

An open question and one of much recent debate in the phonological literature is how much of

the formal machinery should be ascribed to innate endowments and how much can be acquired from the available language data. The traditional assumption in generative linguistics is that the constraints and representations over which they are calculated are innately available to the learner. In an influential paper, Hayes & Wilson (2008) explore the feasibility of learning the constraints and their weights from the language input. They are able to achieve successful modeling of gradient well-formedness across input data from several languages only when the models are provided with access to phonological features, metrical structure, and representations allowing direct computation over phoneme sequences within natural classes (tiers), similar to our vowel and consonant levels. The ways in which we've calculated cues across different levels of representation relies on access to certain aspects of phonological representation. For example, in order to calculate the vowel and consonant level cues, the model must have the ability to differentiate vowels from consonants and to construct representations that reference their sequences separately. Thus, the regression models are certainly not entirely linguistically ignorant. At the same time, as Hayes and Wilson discuss, there is a distinction to be made between a genetic endowment that can access certain representations and perform certain calculations over these representations and an endowment that includes a set of prespecified constraints or principles. To give a concrete example, our simulations assume learners can identify vowels and make calculations over them in order to discover regularities; however, this is qualitatively different from providing learners with specific vowel constraints, such as the constraint that each word must contain a vowel (Brent & Cartwright, 1996). Our results certainly do not provide a definitive answer to the question of genetic endowment, but they do parallel the findings of Hayes and Wilson in showing that access to appropriate representations can go a long way. There is still much work to be done in developing cognitively motivated models relying on rich distributional information, but our findings indicate there is a substantial amount of information available in the signal. Therefore, our results do not support the conclusion (contra Yang, 2004) that distributional cues provide insufficient information for successful segmentation.

Our findings motivate further exploration of the roles of interacting distributional cues in infant segmentation as well as computational modeling. One consistent finding of the analyses above is the

complexity of the interactions of distributional cues. This study investigates how cues should combine in order to provide maximal information about word boundaries, but it does not address how infants actually integrate different distributional cues during language learning. The analyses discussed above show that the information content of distributional cues is highly variable both within and across languages. This raises the question of how the information content of distributional cues affects infants' sensitivity to them: does infants' weighting of distributional cues reflect their relative information content in the ambient language? Behavioral results from speech perception, visual perception, and sentence processing indicate that language users do weight cues according to their reliability (Bejjanki, Clayards, Knill & Aslin, 2011; Clayards, Tanenhaus, Aslin & Jacobs, 2008; Ernst & Banks, 2002; Fine & Jaeger, 2011; Kleinschmidt & Jaeger, 2011). For example, Ernst & Banks (2002) found that adults combine information from visual and haptic cues according to the variability associated with these information sources: cues that provide more reliable estimates are weighted more heavily by the subjects. In the segmentation domain, there is already important work examining the relative weighting among broad classes of cues (Johnson & Jusczyk, 2001; Mattys, 2004; Mattys et al., 1999; Mattys et al., 2005; Morgan & Saffran, 1995; Thiessen & Saffran, 2003; Weiss et al., 2010). An important avenue for future experimental work is determining the relationship between estimates of segmentation cue reliability and cue weighting by infants to determine whether infants weight cues according to their information content in the ambient language. On the computational side, our results motivate development of segmentation models that integrate multiple distributional cues. There is much work in the visual domain on modeling the integration of multiple cues with varying degrees of reliability (Jacobs, 2002; Kersten, Mamassian & Yuille, 2004). There are also models of phonetic category learning (Bejjanki et al., 2011; Feldman, Griffiths & Morgan, 2009; Kleinschmidt & Jaeger, 2011), sentence processing (Fine & Jaeger, 2011), and word segmentation (Adriaans & Kager, 2010; Cairns et al., 1997; Christiansen et al., 1998; Goldwater et al., 2009; Johnson, 2008b; Norris & McQueen, 2008) that incorporate multiple statistical information sources. Particularly relevant is a recent model by Toscano & McMurray (2010) that learns to combine and weight multiple cues to phonetic categories from the distributional information in the input. The

approaches to cue integration and weighting formalized in these models are promising avenues for the development of an unsupervised model of infant segmentation based on multiple distributional cues.

Although the focus of this work has been on the information content of rich distributional information, we do not mean to suggest that segmentation models or human learners must rely only on distributional cues. Indeed, recent work suggests that human learners do not rely exclusively on sequential statistics, with experimental evidence suggesting that additional biases associated with lexicon-building strategies are at work (Frank, Tily, Arnon & Goldwater, 2010; Giroux & Rey, 2009). However, see Daland & Pierrehumbert (2011) for arguments that early segmentation and word-learning are separate processes. As discussed earlier, lexicon-building models that make limited use of distributional cues do not perform consistently well on all languages. The above experimental results on learner biases together with our finding that differences in performance across languages can be narrowed by combining multiple weak distributional cues suggest there is much potential for combining the advantages of lexicon-building approaches with reliance on rich distributional information. Also, it is important to keep in mind that experimental work has shown that infants are capable of using finer-grained acoustic cues not explored in the present study, such as coarticulation and allophony (Johnson & Jusczyk, 2001; Jusczyk et al., 1999; Thiessen & Saffran, 2004; Weiss et al., 2010). An important question for future work is how the information content of these lower level cues varies across languages and whether they can be used to further reduce the performance differences between languages.

Finally, there are several issues and simplifying assumptions that warrant further investigation. Together with most previous modeling work on segmentation, we have assumed that the language input is represented as a sequence of discrete phonetic symbols. This is a simplification since speech is actually a continuous stream with phonetic variation. Indeed, infants seem to solve the phonetic categorization and segmentation problems during roughly the same period, suggesting that by the time infants begin segmenting they have not yet fully learned to parse the speech stream into discrete phonemes (Feldman et al., 2009). In recent work, Rytting, Brew & Fosler-Lussier (2010) showed that segmentation performance is markedly lower on phonetically variable speech as compared to speech transcribed using

dictionary methods. Future work must investigate the impact of phonetic variability not only on the informativeness of distributional cues investigated in the present work but also on the performance of various computational models of word segmentation. In addition, although we have shown that combining many weak distributional cues can narrow performance differences between languages, performance in English is still substantially higher in our analyses and tends to be higher in English in previously reported modeling studies. In addition to exploring lower-level phonetic cues as discussed above, it is important first to rule out the possibility that some of the language differences are a result of properties of the particular corpora or the transcription conventions. The Brent corpus has become the standard corpus for evaluation of segmentation models in English, but a number of the transcription choices it assumes are controversial and appear to raise segmentation performance to some degree (Blanchard & Heinz, 2008). The effect of using orthographic word boundaries to represent phonological word boundaries should also be investigated further as this choice may affect performance differently in different languages (Blanchard & Heinz, 2008). Finally, it is not clear what effect the age of the children addressed in the corpora has; the non-English corpora have tended to involve older children than the English corpora, which may be affecting the segmentation performance.

The investigations in this study were motivated by a desire to connect experimental findings on infant speech segmentation with computational models of segmentation via statistical analysis of the learner's input. We hope the findings of this work contribute to the development of a more complete picture of the process by which infants extract words from fluent speech.

References

Adriaans, F. & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*(3), 311 - 331.

Allen, J. & Christiansen, M. H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the 18th annual conference of the cognitive science society.*

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (pp. 117-34).

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). The celex lexical database (cd-rom). *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167 - 206.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *Plos ONE*, *6(5)*.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, *113*, 1001.

Bernstein-Ratner, N. (1987). The phonology of parent-child speech. *Children's Language*, *6*, 159-174.

Blanchard, D. & Heinz, J. (2008). Improving word segmentation by simultaneously learning phonotactics. In *Conll '08: Proceedings of the 12th conference on computational natural language learning.* Association for Computational Linguistics.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, *First View*, 1-25.

Boersma, P. (1997). How we learn variation, optionality, and probability. *IFA Proceedings*, *21*, 43-58.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1), 71-105.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93 - 125.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*(2), 111 - 153.

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

Christiansen, M. H., Onnis, L., & Hockema, S. A. (2009). The secret is in the sound: From unsegmented speech to lexical categories. *Developmental Science*, *12*(3), 388-395.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Speech perception reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804-809.

Clements, G. N. & Sezer, E. (1982). Vowel and consonant disharmony in turkish. *The Structure of Phonological Representations*, *2*, 213-255.

Coetzee, A. & Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in muna and arabic. *Natural Language & Linguistic Theory*, *26*(2), 289-337.

Daland, R. & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*, 119–155.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211.

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the cognitive science society*.

Fine, A. B. & Jaeger, T. F. (2011). Language comprehension is sensitive to changes in the reliability of lexical cues. In *The 33rd annual meeting of the cognitive science society*.

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In *Proceedings of the 46th annual meeting of the association for computational linguistics.*

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107-125.

Frank, M. C., Tily, H., Arnon, I., & Goldwater, S. (2010). Beyond transitional probabilities: Human learners apply a parsimony bias in statistical word segmentation. In *Proceedings of the 32nd annual meeting of the cognitive science society.*

Gambell, T. & Yang, C. (2006). *Word segmentation: Quick but not dirty.* In *Word segmentation: Quick but not dirty.* Yale University, New Haven, CT.

Giroux, I. & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*(2), 260-272.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21 - 54.

Gussmann, E. (1992). Resyllabification and delinking: The case of polish voicing. *Linguistic Inquiry*, *23*(1), 29-56.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. 2009.* Springer, New York.

Hayes, B. & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of hungarian vowel harmony. *Phonology*, *23*(01), 59-104.

Hayes, B. & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379-440.

Hockema, S. A. (2006). Finding words in speech: An investigation of american english. *Language Learning and Development*, *2*(2), 119-146.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345-350.

Jaeger, T. F. & Kidd, C. (2008). Toward a unified model of redundancy avoidance and strategic lengthening. In *21St annual CUNY conference on human sentence processing, chapel hill, NC.*

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

Johnson, M. (2008a). Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the 10th meeting of ACL SIGMORPHON.*

Johnson, M. (2008b). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th annual meeting of the association for computational linguistics.*

Johnson, M. & Goldwater, S. (2009). Improving nonparameteric bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *NAACL '09: Proceedings of the annual conference of the north american chapter of the association for computational linguistics.* Association for Computational Linguistics.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, *61*(8), 1465-1476.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, *39*(3-4), 159 - 207.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, *55*, 271-304.

Kleinschmidt, D. & Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. *ACL HLT 2011*, 10.

Liang, P. & Klein, D. (2009). Online EM for unsupervised models. In *NAACL '09: Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics.* (pp. 611-9).

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271-196.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 397 - 408.

Mattys, S. L. & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants.

Cognition, *78*(2), 91 - 121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants,. *Cognitive Psychology*, *38*(4), 465 - 494.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477 - 500.

Monaghan, P. & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *First View*, 1-20.

Morgan, J. L. & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*(4), 911-936.

Myers, J., Jusczyk, P. W., Nelson, D. G. K., Charles-Luce, J., Woodward, A. L., & Hirsh-Pasek, K. (1996). Infants' sensitivity to word boundaries in fluent speech. *Journal of Child Language*, *23*(01), 1-30.

Newport, E. L. & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127 - 162.

Norris, D. & McQueen, J. M. (2008). Shortlist B: A bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

Oflazer, K. & Inkelas, S. (2006). The architecture and the implementation of a finite state pronunciation lexicon for turkish. *Computer Speech & Language*, *20*(1), 80 - 106.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, *33*, 999-1035.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-Month-Old infants track backward transitional probabilities. *Cognition*, *113*(2), 244 - 247.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674-685.

Perruchet, P. & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Mem Cognit*, *36*(7), 1299-305.

Prince, A. & Smolensky, P. (2004). *Optimality theory : Constraint interaction in generative grammar* .

Malden, MA : Blackwell Pub. (Original work published 1993)

R Development Core Team (2008). R: A language and environment for statistical computing. [Computer Software] Vienna, Austria: R Foundation for Statistical Computing.

Rubach, J. & Booij, G. E. (1985). A grid theory of stress in polish. *Lingua*, *66*(4), 281 - 320.

Rytting, C. A., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: Subsegmental variation in segmental cues. *Journal of Child Language*, *37(3), Special Issue on Computational Models of Child Language Learning*, 513-543.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606 - 621.

Slobin, D. I. (1982). Universal and particular in the acquisition of language. *Language Acquisition: The State of the Art*, *57*.

Swingley, D. (1999). Conditional probability and word discovery: A corpus analysis of speech to infants. In *21St proceedings of the annual meeting of the cognitive science society*.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86 - 132.

Thiessen, E. D. & 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*(4), 706 - 716.

Thiessen, E. D. & Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception & Psychophysics*, *66*(5), 779.

Toscano, J. C. & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434-464.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Comput. Linguist.*, *27*(3), 352-372.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2010). Colliding cues in word segmentation: The role of cue strength and general cognitive processes. *Language and Cognitive Processes*, *25*(3), 402-422.

Weist, R. M., Wysocka, H., Witkowska-Stadnik, K., Buczowska, E., & Konieczna, E. (1984). The defective tense hypothesis: On the emergence of tense and aspect in child polish. *Journal of Child Language*, *11*(02), 347-374.

Xanthos, A. (2004). Combining utterance-boundary and predictability approaches to speech segmentation. In W. G. Sakas (Ed.), *Proceedings of the first workshop on psycho-computational models of language acquisition at COLING 2004* (W. G. Sakas, Ed.). (pp. 93-100). Geneva, Switzerland.

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451 - 456.

Table 1 - Characteristics of Phonemically Transcribed Child-Directed Speech

	English	**Polish**	**Turkish**
Tokens			
Utterances	9,498	9,361	10,160
Words	32,106	34,125	33,492
Clusters	46,585	66,969	82,899
Phonemes	94,730	140,138	168,839
Types			
Words	1,198	5,040	2,516
Clusters	916	1618	425
Phonemes	47	38	38
Average Lengths			
Words per utterance	3.38	3.65	3.30
Phonemes per utterance	9.97	14.97	16.62
Phonemes per word	2.95	4.11	5.04
Phonemes per cluster	1.24	1.23	1.16

Table 2 – Cue Parameters & Settings with Examples for two contexts

Level	**Example Combination (Lev × Stat × Dir × Rel)**	**[jusi_mi] "you see me" stress: [2 2 1]**	**[hiæsks_mi] "he asks me" stress: [1 2 1]**
P (phoneme)	P × T × F × Curr	Pr([m] | [si])	Pr([m] | [ks])
V (vowel)	V × T × F × Curr	Pr([i] | [ui])	Pr([i] | [iæ])
C (consonant)	C × T × F × Curr	Pr([m] | [])	Pr([m] | [ks])
S (stress)	S × T × F × Curr	Pr(1 | 22)	Pr(1 | 12)
CL (cluster)	(CL ×) #CL × F × Curr	Pr(# | [])	Pr(# | [sks])
Statistic			
B (bigram)	P × B × F × Curr	Pr([m] | [i])	Pr([m] | [s])
T (trigram)	P × T × F × Curr	Pr([m] | [si])	Pr([m] | [ks])
M (mutual information)	P × M × F × Curr	MI([m], [i])	MI([m], [s])
#B (# bigram)	P × #B × F × Curr	Pr(# | [i])	Pr(# | [s])
#T (# trigram)	P × #T × F × Curr	Pr(# | [si])	Pr(# | [ks])
#CL (cluster)	(CL ×) #CL × F × Curr	Pr(# | [])	Pr(# | [sks])
#|CL| (cluster length)	(CL ×) #|CL| × F × Curr	Pr(# | 0)	Pr(# | 3)
Direction			
F (forward)	P × T × F × Curr	Pr([m] | [si])	Pr([m] | [ks])
B (backward)	P × T × B × Curr	Pr([i] | [mi])	Pr([s] | [mi])
Relation			
Curr	P × B × F × Curr	Pr([m] | [i])	Pr([m] | [s])
Next	P × B × F × Next	Pr([i] | [m])	Pr([i] | [m])
Diff	P × B × F × Diff	Pr([i] | [s]) - Pr([m] | [i])	Pr([s] | [k]) - Pr([m] | [s])
Rank	P × B × F × Rank	rank of Pr([m] | [i]) vs. Pr([i] | [m]) & Pr([i] | [s])	rank of Pr([m] | [s]) vs. Pr([s] | [k]) & Pr([i] | [m])

Table 3 – Max Word Token F-scores (Bootstrap μ, σ) of Individual Cues by Parameter Setting

English		**Polish**		**Turkish**	
Level					
P	68.9 (68.9, 0)	CL	39.3 (39.5, 0.02)	P	39.5 (39.1, 0.09)
CL	61.7 (61.5, .35)	P	35.7 (35.7, 0.03)	CL	25.1 (25.1, 0)
C	30.8 (30.8, 0)	S	16.5 (16.5, 0)	C	14.4 (14.4, 0.13)
S	17.3 (17.3, 0)	C	14.2 (14.2, 0)	V	11.6 (11.9, 0.03)
V	14.5 (9.5, .70)	V	9.7 (9.4, 0.23)	S	9.4 (9.4, 0)
Direction					
B	68.9 (68.9, 0)	B	39.3 (39.5, 0.02)	B	39.5 (39.1, 0.09)
F	62.8 (62.8, 0)	F	27.9 (27.6, 0.48)	F	33.6 (34.1, 0.93)
Relation					
Rank	68.9 (68.9, 0)	Diff	39.3 (39.5, 0.02)	Diff	39.5 (39.1, 0.09)
Diff	67.4 (67.3, 0.13)	Curr	37.8 (37.8, 0.009)	Curr	35.6 (35.6, 0.005)
Curr	61.6 (61.6, 0.11)	Rank	28.6 (28.6, 0)	Rank	21.9 (21.9, 0)
Next	49.3 (49.3, .08)	Next	23.9 (24.1, 0.56)	Next	21.6 (21.9, 0.89)
Statistic					
#T	68.9 (68.9, 0)	#CL	39.3 (39.5, 0.02)	#T	39.5 (39.1, 0.09)
#CL	61.7 (61.5, .35)	#T	35.7 (35.7, 0.03)	B	33.6 (34.1, 0.93)
M	56.8 (56.8, 0)	T	27.9 (27.6, 0.48)	T	33.4 (33.6, 0.42)
#|CL|	54.0 (54.0, 0)	#|CL|	26.9 (26.0, 0.97)	M	29.8 (30.1, 0.34)
#B	48.3 (48.3, 0)	#B	25.8 (25.8, 0)	#CL	25.1 (25.1, 0)
T	47.6 (47.6, 0)	M	25.5 (25.5, 0)	#|CL|	21.0 (21.0, 0)
B	43.5 (43.5, 0)	B	24.2 (24.2, 0)	#B	17.8 (17.6, 0.49)

Grey shading indicates parameter settings whose bootstrap f-score range overlaps with the f-score range of neighboring cues (shaded light grey).

Table 4 – Word Token F-scores of Step-wise Multiple Regressions by Parameter Setting

English		Polish		Turkish	
Level					
P	84.6 (84.8, 0.11)	P	54.1 (55.2, 0.12)	P	59.2 (60.1, 0.32)
…+S	87.6 (87.8, 0.09)	…+S	74.5 (74.9, 0.13)	…+S	70.3 (71.1, 0.32)
…+C	89.9 (90.3, 0.10)	…+C	77.9 (78.7, 0.09)	…+V	73.8 (75.9, 0.28)
…+CL	90.7 (91.5, 0.10)	…+CL	80.7 (81.5, 0.10)	…+C	77.9 (79.4, 0.25)
…+V		…+V		…+CL	
(FULL)	90.8 (91.6, 0.12)	(FULL)	81.2 (81.7, 0.11)	(FULL)	78.7 (80.9, 0.23)
Direction					
B	83.8 (84.4, 0.12)	B	75.0 (75.4, 0.18)	F	65.2 (66.1, 0.26)
…+F		…+F		…+B	
(FULL)	90.8 (91.6, 0.12)	(FULL)	81.2 (81.7, 0.11)	(FULL)	78.7 (80.9, 0.23)
Relation					
Rank	84.5 (84.9, 0.09)	Curr	70.4 (71.4, 0.13)	Rank	60.6 (61.0, 0.25)
…+Next	88.6 (88.9, 0.10)	…+Rank	77.0 (77.3, 0.11)	…+Curr	72.5 (73.2, 0.23)
…+Curr	89.9 (90.5 0.10)	…+Diff	79.8 (80.1, 0.09)	…+Next	75.2 (77.1, 0.24)
…+Diff		…+Next		…+Diff	
(FULL)	90.8 (91.6, 0.12)	(FULL)	81.2 (81.7, 0.11)	(FULL)	78.7 (80.9, 0.23)
Statistic					
#T	78.9 (79.0, 0.09)	#T	60.8 (61.4, 0.15)	#T	58.9 (58.8, 0.22)
…+T	85.5 (85.8, 0.09)	…+T	71.2 (72.5, 0.13)	…+B	69.6 (70.1, 0.23)
…+#B	87.7 (88.1, 0.11)	…+CL	78.6 (79.1, 0.10)	…+#B	72.9 (74.0, 0.26)
…+CL	89.0 (89.4, 0.09)	…+#B	80.0 (80.4, 0.12)	…+T	75.7 (77.3, 0.23)
…+B	89.9 (90.5, 0.12)	…+B	80.3 (81.1, 0.10)	…+M	77.9 (79.4, 0.28)
…+M		…+M		…+CL	
(FULL)	90.8 (91.6, 0.12)	(FULL)	81.2 (81.7, 0.11)	(FULL)	78.7 (80.9, 0.23)

Grey shading indicates parameter settings whose bootstrap f-score range overlaps with the f-score range of the preceding parameter setting. All settings at the Level parameter correspond to 40 cues, except CL, which corresponds to 16. Backward and Forward each correspond to 88 cues. Each of the Relation settings corresponds to 44 cues, and each of the Statistic settings corresponds to 32 cues, except CL, which corresponds to 16.

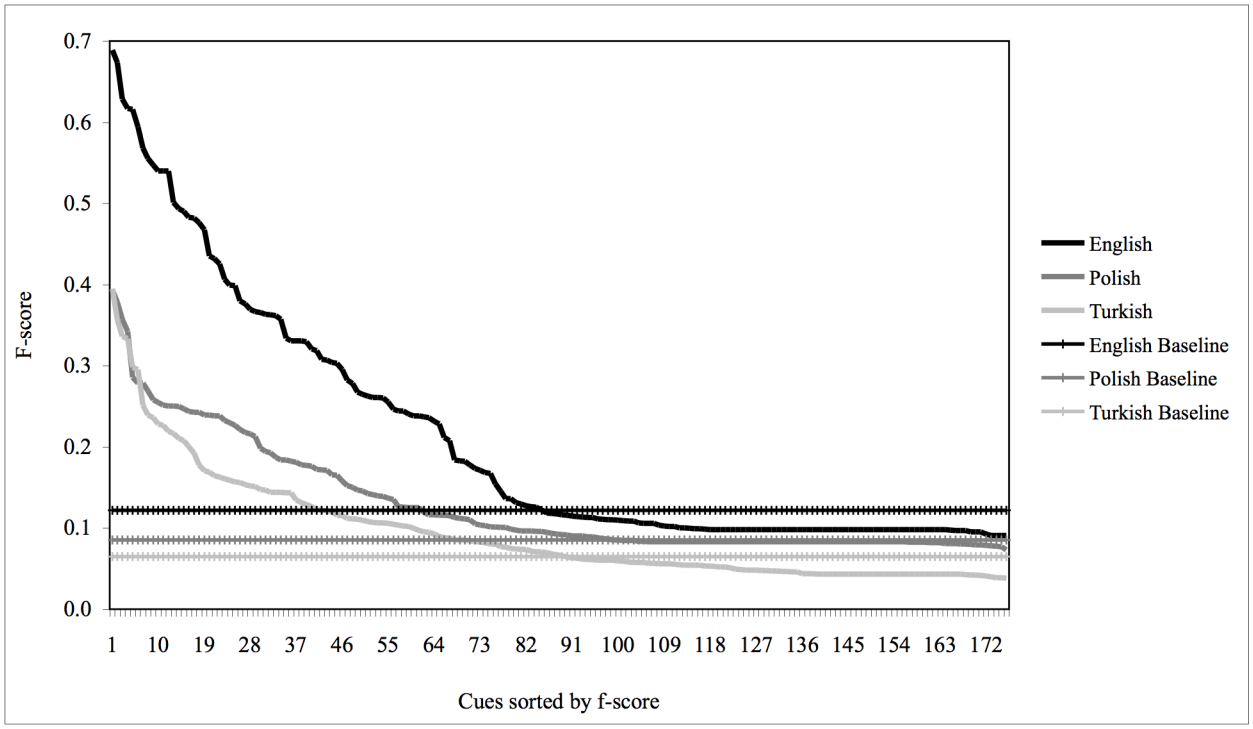

Fig. 1 – Sorted Word Token F-scores of Individual Cues vs. Baselines

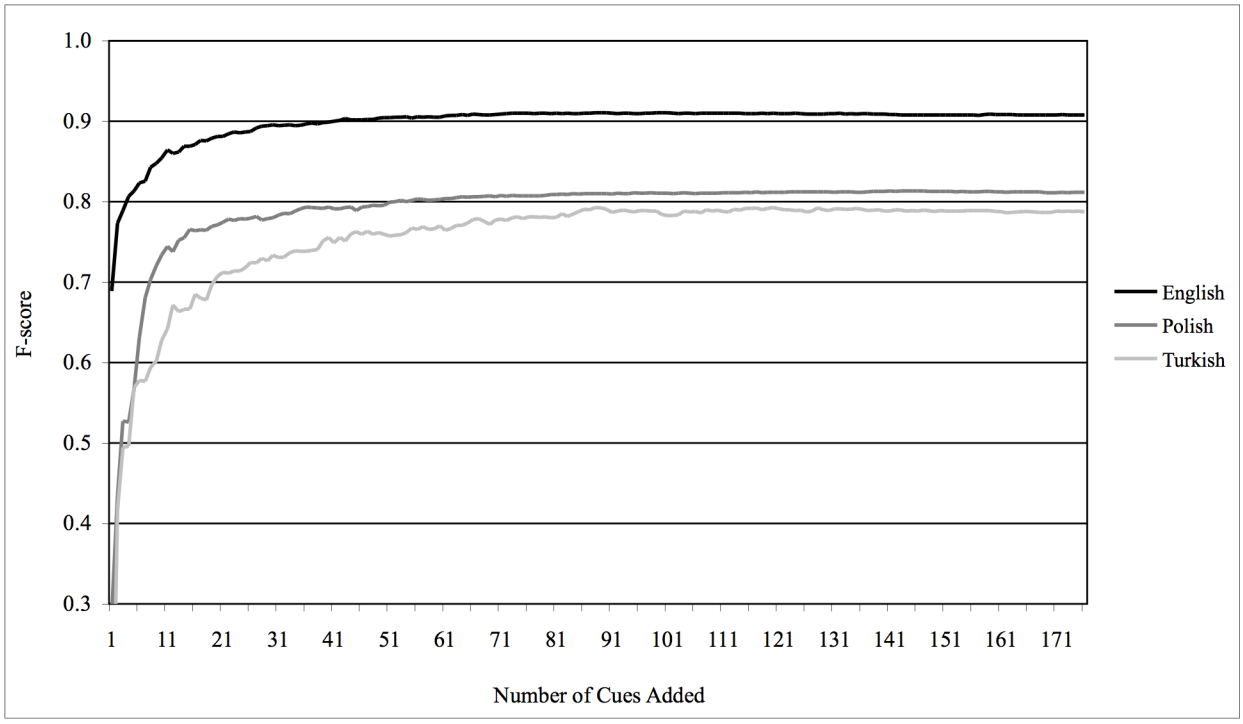

Fig. 2 – Word Token F-score Change During Step-Wise Multiple Regression

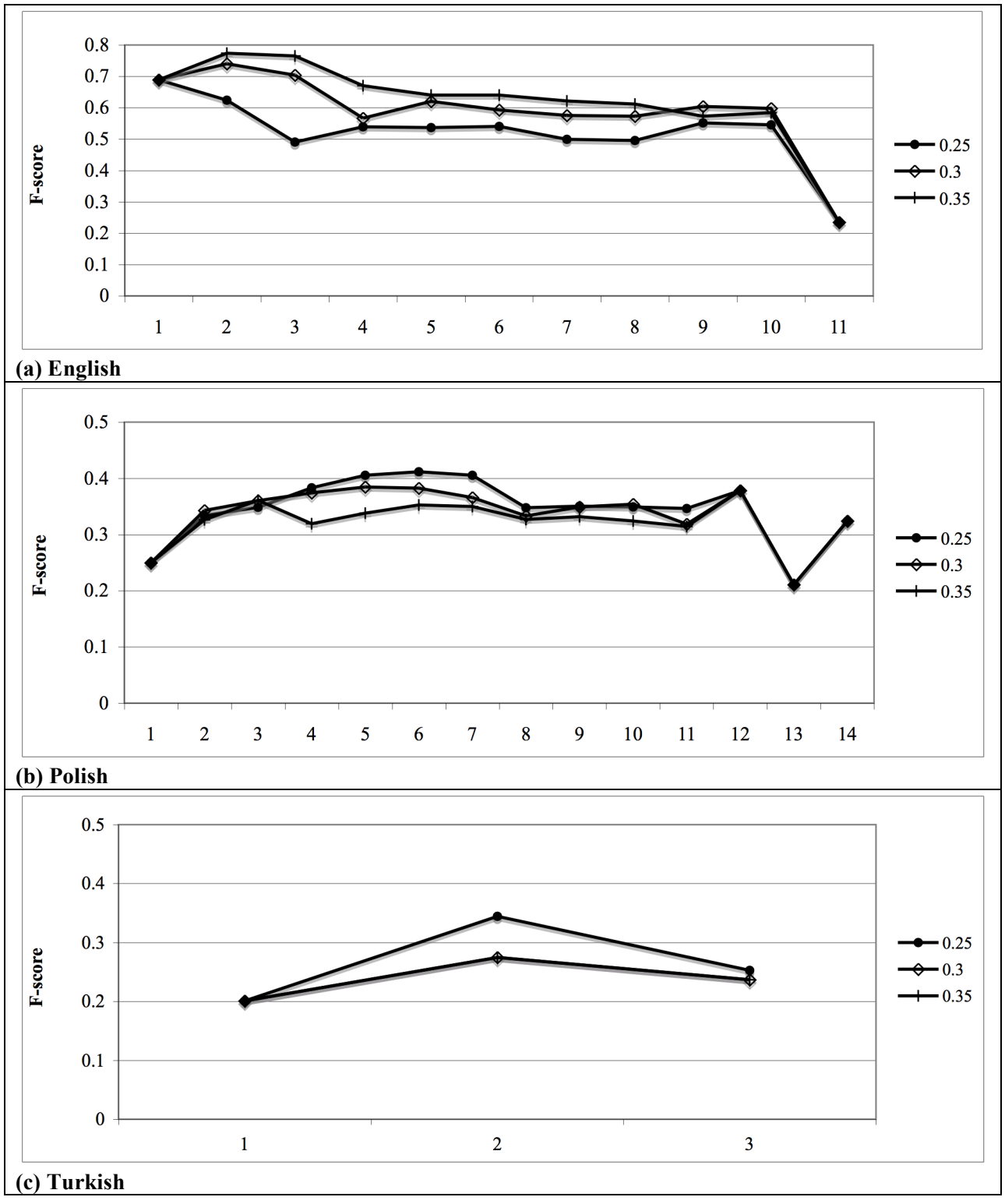

Fig. 3 - Unsupervised Word Token F-scores

APPENDIX

Table 5 – Mean Word Token F-scores of Individual Cues Within Parameter Settings

English		**Polish**		**Turkish**	
Level					
P	38.2	CL	22.5	P	16.9
CL	41.8	P	20.3	CL	14.2
C	17.0	S	9.3	C	8.0
S	10.7	C	10.1	V	6.1
V	10.4	V	8.4	S	5.1
Direction					
B	20.4	B	13.0	B	8.5
F	22.0	F	13.0	F	10.4
Relation					
Rank	24.8	Diff	13.7	Diff	10.5
Diff	23.0	Curr	14.0	Curr	11.4
Curr	21.2	Rank	13.2	Rank	8.5
Next	15.7	Next	11.0	Next	7.6
Statistic					
#T	21.9	#CL	24.1	#T	10.0
#CL	44.5	#T	13.7	B	8.6
M	19.9	T	11.3	T	9.3
#|CL|	39.1	#|CL|	21.0	M	9.8
#B	19.5	#B	11.7	#CL	16.2
T	17.3	M	12.7	#|CL|	12.2
B	17.0	B	10.6	#B	7.4

Table 6 – Word Token F-scores of Multiple Regressions within each Parameter Setting (# cues)

English		**Polish**		**Turkish**	
FULL MODEL	90.8	FULL MODEL	81.2	FULL MODEL	78.7
Level					
P (40)	84.6	P (40)	54.1	P (40)	59.2
CL (16)	63.1	CL (16)	38.7	CL (16)	29.0
C (40)	44.4	C (40)	34.3	C (40)	26.8
S (40)	22.7	S (40)	30.3	S (40)	10.8
V (40)	13.4	V (40)	7.1	V (40)	9.2
Direction					
B (88)	83.8	B (88)	75.0	F (88)	65.2
F (88)	83.0	F (88)	68.8	B (88)	61.0
Relation					
Rank (44)	84.5	Curr (44)	70.4	Rank (44)	60.6
Curr (44)	82.5	Diff (44)	61.2	Curr (44)	58.4
Diff (44)	77.3	Rank (44)	60.8	Diff (44)	57.2
Next (44)	53.7	Next (44)	47.6	Next (44)	39.0
Statistic					
#T (32)	78.9	#T (32)	60.8	#T (32)	58.9
#B (32)	65.8	M (32)	46.9	T (32)	50.2
T (32)	64.5	T (32)	45.4	M (32)	45.7
CL (16)	63.1	#B (32)	43.1	B (32)	43.5
M (32)	62.9	B (32)	40.1	#B (32)	31.6
B (32)	54.3	CL (16)	38.7	CL (16)	29.0

Table 7 – BIC of Step-wise Multiple Regressions by Parameter Setting

English		**Polish**		**Turkish**	
Level					
P	22115.38	P	66412.76	P	59634.58
…+S	18185.71	…+S	38975.8	…+S	41799.94
…+C	15167.96	…+C	33177.18	…+V	36330.03
…+CL	13411.23	…+CL	28624.88	…+C	32648.56
…+V		…+V		…+CL	
(FULL)	13533.28	(FULL)	28556.92	(FULL)	30345.13
Direction					
B	22491.38	B	42290.25	F	49201.49
…+F		…+F		…+B	
(FULL)	13533.28	(FULL)	28556.92	(FULL)	30345.13
Relation					
Rank	23294.7	Curr	45560.94	Rank	56038.09
…+Next	16893.51	…+Rank	35851.81	…+Curr	39273.54
…+Curr	14524.8	…+Diff	30997.91	…+Next	34977.52
…+Diff		…+Next		…+Diff	
(FULL)	13533.28	(FULL)	28556.92	(FULL)	30345.13
Statistic					
#T	30032.44	#T	57958.33	#T	63674.5
…+T	20543.38	…+T	41084.6	…+B	45759.34
…+#B	17840.77	…+CL	32791.56	…+#B	40300.46
…+CL	15960.65	…+#B	30726.16	…+T	34990.15
…+B	15053.84	…+B	29315.51	…+M	32648.56
…+M		…+M		…+CL	
(FULL)	13533.28	(FULL)	28556.92	(FULL)	30345.13

The BIC (Bayes Information Criterion) for each of the models in the step-wise multiple regressions in Analysis 3. BIC provides a measure of fit with a penalty for complexity. The differences between languages are not meaningful, but lower BIC within the same language corresponds to better models for that language. These figures confirm the results in Table 4, indicating that, with the exception of the final step at the Level parameter in English, the addition of all parameter settings improves performance