COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 1988–2021

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12454

Using Predictability for Lexical Segmentation

Çağrı Çöltekin

Department of Linguistics, University of Tübingen

Received 24 June 2015; received in revised form 23 August 2016; accepted 25 August 2016

Abstract

This study investigates a strategy based on predictability of consecutive sub-lexical units in learning to segment a continuous speech stream into lexical units using computational modeling and simulations. Lexical segmentation is one of the early challenges during language acquisition, and it has been studied extensively through psycholinguistic experiments as well as computational methods. However, despite strong empirical evidence, the explicit use of predictability of basic sub-lexical units in models of segmentation is underexplored. This paper presents an incremental computational model of lexical segmentation for exploring the usefulness of predictability for lexical segmentation. We show that the predictability cue is a strong cue for segmentation. Contrary to earlier reports in the literature, the strategy yields state-of-the-art segmentation performance with an incremental computational model that uses only this particular cue in a cognitively plausible setting. The paper also reports an in-depth analysis of the model, investigating the conditions affecting the usefulness of the strategy.

Keywords: Language acquisition; Computational modeling; Word segmentation; Predictability entropy

1. Introduction

Predicting what comes next in time or space is an activity that the human cognitive system is constantly engaged in. Predicting the environment at multiple perceptual domains is at the very center of how human cognition works. Predictability plays such a prevalent role in many aspects of cognition that it is argued that "brains are essentially prediction machines" (Clark, 2013). Although we still know rather little about the way our brains work, researchers from different areas of cognitive sciences seem to agree that our brains are engaged in predicting the environment we live in at multiple levels.[1] The cognitive system does not only make use of successful predictions of the next percept in

Correspondence should be sent to Çağrı Çöltekin, Department of Linguistics, University of Töbingen, Wilhelmstr. 19, 72074 Tübingen, Germany. E-mail: cagri@coltekin.net

time or space. When predictions fail, it has further consequences on the cognitive system. We remember and learn most from surprising events.

Not surprisingly, predicting what comes next in the auditory input is also important in language processing and acquisition. Listeners predict and use what comes next at multiple levels, that is, using different linguistic units, during language comprehension. In this article, we focus on a particular form of predictability, where learners discover and make use of the statistical regularities between consecutive basic sub-lexical unit sequences (such as phonemes or syllables) for determining likely lexical units. Throughout this article, the term *predictability* is used in this restricted sense of predictability or uncertainty associated with a short sequence of basic units such as phonemes or syllables.

Segmenting continuous speech into lexical units is one of the early tasks an infant needs to tackle during language acquisition. The segmentation problem is more difficult than may be appreciated at first sight. Children need to find words in a continuous stream of speech, with no knowledge of words to start with. Fortunately, studies in psycholinguistics suggest that children are not helpless in this task. They are sensitive to, and make use of, some properties of naturally occurring speech very early in the acquisition process, which leads to relatively simple computational strategies for segmenting input utterances. Children are known to attend to a number of cues in the speech stream that are useful for discovering lexical units. These cues include, but are not limited to, *predictability statistics* (Saffran, Aslin, & Newport, 1996), *lexical stress* (Cutler & Butterfield, 1992; Jusczyk, Houston, & Newsome, 1999), *phonotactics* (Jusczyk, Cutler, & Redanz, 1993), *allophonic differences* (Jusczyk, Hohne, & Bauman, 1999), *vowel harmony* (van Kampen, Parmaksız, van de Vijver, & Höhle, 2008; Suomi, McQueen, & Cutler, 1997), and *coarticulation* (Johnson & Jusczyk, 2001). Furthermore, as the learner's (implicit) knowledge about the target language grows, extracting words from the input stream is aided by the learner's lexicon and, in general, her or his knowledge of the language.

While a large number of experimental studies, including the ones cited above, have shown that a certain cue or strategy is used during language acquisition, the stimuli used in these experimental studies are necessarily simplified. As a result, it is difficult to assess the extent to which the same strategy is useful under more realistic input distributions. Computational models and simulations offer useful ways to investigate viability of a particular cue or strategy and provide further insights by creating an explicitly specified model of the (partial) task the human learner has to perform. There have been an increasing number of models of segmentation, particularly within the last two decades (e.g., Aslin, 1993; Brent, 1999; Brent & Cartwright, 1996; Cairns, Shillcock, Chater, & Levy, 1994; Christiansen, Allen, & Seidenberg, 1998; Elman, 1990; Fleck, 2008; Goldwater, Griffiths, & Johnson, 2009; Johnson & Goldwater, 2009; Monaghan & Christiansen, 2010; Venkataraman, 2001; Xanthos, 2004).

One of the well-attested cues from the psycholinguistic experiments is the predictability or surprise associated with sequences of syllables (Saffran, Aslin, & Newport, 1996). However, computational models of segmentation that investigate the human performance in these experiments are rather scarce. Furthermore, the models that rely on predictability are often reported to have substantially worse performance results compared to the models that use other cues or strategies. However, some recent studies have shown that the models based on predictability alone can yield results competitive with the other models in the literature (Cohen, Adams, & Heeringa, 2007; Çöltekin, 2011; Çöltekin & Nerbonne, 2014, as well as the supervised model by Jarosz & Johnson, 2013). However, there appears to be a common impression (e.g., Gambell & Yang, 2006; Lignos, 2012; Pearl, Goldwater, & Steyvers, 2010) in the field that the predictability-based models do not perform as well as other models.

In this article, we argue that a model relying solely on predictability of consecutive basic units does not only perform similar to the state-of-the-art computational models, but it also does so under a strictly unsupervised and online learning regime. In addition, we present further experiments with the model, investigating some properties of the input and the representations that may affect the effectiveness of a predictability-based segmentation strategy. Although the primary contribution of this article, usefulness of predictability cue, fits into the computational level of Marr's (1982) influential classification, the modeling practice we follow here lends itself to studies seeking explanations at algorithmic level as well.

The remainder of this paper is organized as follows: the next section discusses the predictability in the context of lexical segmentation and clarifies the notion of predictability that we are concerned with in this study. Section 2 describes the model, and Section 3 presents the experiments and the results obtained with the model (results from additional experiments are presented in Appendix A and Appendix B). In Section 4, we provide a general discussion of the results. Section 5 points to some future directions after concluding remarks.

1.1. Predictability and segmentation

The aim of the present study is to provide an in-depth investigation of a single cue or strategy based on predictability of consecutive basic units. The aim here is not to develop a complete model of early lexical segmentation. We acknowledge that predictability is not the only source of information used by adults and children during lexical segmentation, but it is a particularly interesting one that warrants deeper investigation.

An important aspect of the predictability cue is that it does not require any language or domain-specific information to start with. Children do not even need to look for boundaries; the brain's habit of learning patterns that occur in the perceived environment and predicting the next input based on the known patterns is enough for a good starting point. In other words, predictability is not only useful when the learner does not know any words in the target language; it is also useful when the learner does not even assume that the input is formed by concatenation of lexical units.[2] On the other hand, most other cues that are useful in the lexical segmentation task are language specific and can be useful only after the learner has a lexicon populated with some of the words of the target language. For example, *lexical stress*, a well-attested cue for languages with a regular stress pattern such as English, can be learned only after one has a large enough lexicon from which a common stress pattern can be discovered.

Children's use of statistical cues that are useful for segmentation is demonstrated in an experimental study by Saffran, Aslin, & Newport (1996). This highly influential study showed that predictability of consecutive syllables is used by 8-month-old infants to learn word-like units in the input stream. Saffran, Aslin, & Newport (1996) familiarized infants with stimuli constructed from three-syllabic artificial words. In the stimuli used during the familiarization phase, the transition of syllables within the words was deterministic, while the *transition probability* between the words was lower (1/3). Crucially, the stimuli did not contain any other cues to artificial word boundaries. After 2 min of familiarization, infants were able to distinguish a novel sequence constructed from the artificial words in the familiarization phase from another sequence formed by part-words with the same frequency of occurrence of the syllables as in the training stimuli. The computational principle that explains this behavior is simple: Posit a boundary where it is difficult to predict the next syllable. A large number of experimental studies have confirmed that the same principle is used by adults and children for learning various aspects of language (e.g., Aslin, Saffran, & Newport, 1998; Graf Estes, Evans, Alibali, & Saffran, 2007; Newport & Aslin, 2004; Perruchet & Desaulty, 2008; Thiessen & Saffran, 2003; Thompson & Newport, 2007).

Being one of the first studies that (re)introduced usefulness of statistical learning in the study of language acquisition, the results from Saffran, Aslin, & Newport (1996) and subsequent studies investigating the same strategy are typically interpreted as the evidence of use of *statistics* by infants in the segmentation task. The *prediction* aspect of these findings is generally under-articulated.

In this paper, our main focus is the type of predictions that can explain the performance of the infants in the study by Saffran, Aslin, & Newport (1996). By using computational simulations on child-directed speech corpora (as many other computational studies do), we explore the usefulness of this strategy using input collected outside a laboratory setting. Compared to the stimuli used in artificial language experiments, the child-directed speech corpus used in this study has a larger variety of basic units (phonemes or syllables) and a larger number of lexical units. Furthermore, the co-occurrences of basic units are not controlled artificially. Only restrictions on the co-occurrence probabilities come from the naturally occurring speech directed to children.

The use of this form of predictability by adults and children is shown repeatedly, and its use in segmentation is detectable in the laboratory as early as 6–7 months of age (Thiessen & Saffran, 2003). Many studies, however, showed that when pitted against more precise language-specific cues at later periods of development, children rely on the language-specific cues, for example, lexical stress in English, rather than predictability (Thiessen & Saffran, 2007). As a result, predictability is often suggested as a cue that bootstraps the others (e.g., Swingley, 2005; Thiessen & Saffran, 2007). The present study shares this viewpoint: Predictability is particularly important during early stages of learning as a method to bootstrap the language-specific cues (potentially together with other language-general cues like utterance boundaries and isolated words). We also note that even though the weight assigned to predictability may diminish as the learner masters more precise cues, the strategy seems to prevail throughout life, and it is called for duty even by adults when other cues are not reliable or available (Saffran, Newport, & Aslin, 1996).

Various forms of predictability have been used in earlier segmentation models. However, the notions of predictability explored or exploited in these studies have some important differences from the form of predictions that we concentrate on here. One such strategy is based on predicting the boundaries, for example, utterance or phrase boundaries that are clearly marked in the input stream, and positing lexical unit boundaries where the probability of an utterance boundary is high. This strategy has been used in a number of computational models (notably, Aslin, Woodward, LaMendola, & Bever, 1996; Christiansen et al., 1998; Daland & Pierrehumbert, 2011; Fleck, 2008; Ma, Çöltekin, & Hinrichs, 2016), and it has been found to be successful in the segmentation task. However, this strategy cannot have been used by the participants in the study by Saffran, Aslin, & Newport (1996), where the infants listened to a continuous 2-min stream without any utterance or phrase boundaries. Furthermore, the strategy cannot be used without an expectation for boundaries, and a heuristic that relates phrase or utterance boundaries to the lexical unit boundaries.

Another use of predictability can be found in many state-of-the-art models that learn a lexicon (as opposed to the models guessing boundaries). The models that incorporate word-context information (Goldwater et al., 2009, and many other studies following the same strategy) predict the *words* based on a limited window of neighboring words. This form of predictions, which is shown to improve segmentation performance, is also different from what we investigate here.

The notion of predictability we investigate here is most similar to some of the early neural network models that are trained by predicting the next input. In these models, a boundary is posited when it is difficult to predict the next input (Cairns et al., 1994; Elman, 1990). Unlike the neural network models, however, our model lends itself to easier analysis and interpretation, and as we demonstrate in Section 3, it also performs closer to the current state-of-the-art models.

The strategy we use can simply be summarized as "predictability *within* the lexical units is high, predictability *between* the lexical units is low." The early uses of the strategy date back to Harris (1955), where Harris proposes a measure called the *successor variety* (SV) for determining the morpheme boundaries. The SV is simply the number of possible phonemes that can follow a given phoneme sequence. If the number of possible phonemes, the SV, is high after a given sequence, it is difficult to guess the next phoneme, and hence, the high SV is an indication of a lexical boundary.

The method suggested by Harris (1955) was operationalized in early natural language processing literature (Hafer & Weiss, 1974). However, it has not been used in modeling human language acquisition (see Çöltekin, 2010, for a summary and for some improvements to the original defintion). In recent literature, Brent (1999) compares his target model with two simple (baseline) models that use *transitional probabilities* (TP, as defined by Saffran, Aslin, & Newport, 1996) and *pointwise mutual information* (MI, an information theoretic measure of association). Although he notes that the single-phoneme context used in the study does not necessarily exploit the full utility of the strategy, Brent (1999, or any later study) did not investigate the method at any detail. Swingley (2005), in his investigation of how a language-specific segmentation strategy (the use of lexical stress pattern) may emerge, suggests a segmentation strategy based on pointwise mutual information between consecutive syllables. Gambell and Yang (2006) also use a simple model based on predictability to compare with the model they propose. The segmentation performances reported by all three papers are rather low in comparison to other segmentation models in the literature. Although not targeted directly for modeling human language acquisition, in another related study, Cohen et al. (2007) present a segmentation model based on *entropy* (an information theoretic measure of unpredictability or surprise) with promising segmentation performance.[3]

In the remainder of this paper, we present an in-depth analysis of a strategy for speech segmentation that depends only on predictability through computational simulations using real-world child-directed speech data.

2. The model

In modeling most cognitive phenomena, we know rather little about the target system, the human cognition. Learning segmentation is not an exception. Although we have learned a lot about how adults and infants segment continuous speech into lexical units within the last few decades, our knowledge is still limited. As a result, our modeling efforts are not yet suitable for making precise predictions that are useful or applicable (e.g., as in models of atmosphere used in meteorology), but they are useful for understanding the cognitive phenomenon in question better. The model we present in this section is intended for investigating the utility of predictability statistics in learning segmentation, aiming for the second, more modest, objective.

Similar to all segmentation models that do not make use of lexical knowledge as in the models of (adult) word recognition (see Dahan & Magnuson, 2006; Davis, 2006, for comprehensive reviews), the model defined in this section lends itself best in investigating early language acquisition.[4] However, we assume that the predictability cue remains in effect throughout one's lifetime, even though it may not be detectable in the presence of stronger cues. Furthermore, we also assume that the switch from relying on predictability to stronger language-specific cues is not instantaneous. Hence, the predictability cue interacts with other cues heavily during these early stages of the development. Although these assumptions affect our discussion of predictability cue later in this paper, our focus in this work is to only model the contribution of the predictability cue. In this section, we define the model in detail after a brief non-formal introduction.

2.1. An overview of the model

One of the crucial differences between the current model and the earlier models that rely on predictability is the use of basic units of varying size. Earlier models typically track and use statistics on a single basic unit (i.e., phoneme or syllable). In this work, we combine information from multiple overlapping sequences of basic units (one to four phonemes for the main experiments) for boundary decisions. For example, in the utterance /IzD&t6kIti/ "Is that a kitty?," while deciding whether there is a boundary after the first /t/, we make use of uncertainty or predictability after /zD&t/,

/D&t/, /&t/ as well as /t/ itself. The way the model combines these multiple sources of information is defined in Section 2.4.

As the example above indicates, the model assumes that the learner can make distinctions of segment- or phoneme-size units.[5] Since the model tracks statistics of combinations of consecutive basic units, however, this allows the model to also discover generalizations that are possible by making use of larger units, for example, syllables.

In a nutshell, the model quantifies the predictability or uncertainty using entropy and reverse entropy (Section 2.2) associated with sequences of one to four phonemes (see Section 2.4 for a discussion of choice of context length, and Appendix A for experiments varying context size) uses the local increase of uncertainty before a candidate boundary position and local decrease of uncertainty after a candidate boundary position as boundary indications (Section 2.3), and it combines the indicators using a weighted majority voting scheme (Section 2.4) for the final boundary decision. The model works in an online fashion, evaluating each boundary candidate from left to right and making an irreversible boundary decision based on the statistics collected from the input seen so far. The model parameters (statistics on phoneme n-gram sequences and weights of the indicators) are updated after each utterance is segmented.

2.2. *Quantifying predictability*

The notion of predictability has to be quantified before it can be studied through a computational model. There are a relatively large number of ways to quantify predictability (or surprise). In Section 1.1, we have already noted four different ways to measure predictability used in the earlier studies. These measures are *successor variety* (SV), *transition probability* (TP), *entropy* (*H*), and *pointwise mutual information* (MI). Although these measures are related, they are not identical. Also, it has been demonstrated that they even complement each other to some extent in finding lexical unit boundaries (Çöltekin, 2011). Noting that slightly better results can be obtained by combining multiple measures of predictability, we report results from simulations using only entropy for the sake of simplicity. Formal definitions of the other three measures alongside a comparison of their performances are provided in Appendix B.

Entropy is the information theoretic measure of unpredictability. In this study, we use a particular form of entropy that we call *boundary entropy*.[6] It is defined as

$$H(l) = -\sum_{r \in A} P(r|l) \log_2(P(r|l)) \tag{1}$$

where l is a sequence of phonemes to the left of the candidate boundary position, and r ranges over the members of the alphabet A which contains all observed phonemes. The conditional probabilities $P(r|l)$ are the empirical probabilities in the input seen so far.

The choice of entropy in this study is mainly motivated by the fact that it is a mathematically sound measure of uncertainty (MacKay, 2003). It is also compatible with the experimental evidence so far. Although most experimental studies (e.g., Saffran, Aslin, &

Newport., 1996) typically control and contrast the differences of TP within and between the words, the resulting stimuli has low (or in case of Saffran, Aslin, & Newport, 1996, no) entropy within the lexical units, but higher entropy between the lexical units. A learner relying on entropy for measuring unpredictability is expected to find boundaries after the lexical units in stimuli used in the experimental studies so far.

The entropy measure defined in Eq. 1 is related to transitional probability (TP) and pointwise mutual information (MI) used often in earlier studies. The main difference is that TP and MI are statements about events, while entropy (and SV) are statements about probability distributions. Appendix B provides comparisons between these alternative measures of predictability or surprise.

It is also worth noting that entropy is only based on expectation. That is, unlike TP and MI, it only depends on the sequence observed before the candidate boundary location. A learner acting based only on entropy places a boundary to the right of a *single* input sequence with high entropy, not between *two* sequences with low conditional probability (TP) or weak association (MI). In a way, entropy is compatible with a fully predictive mode of operation. This is not to claim that humans do not make use of the units or information that come after the candidate boundary location. In fact, we are about to define the reverse version of the boundary entropy which does exactly that. However, it is interesting to disentangle these two sources and to investigate their limits and relative contributions to segmentation.

Although predicting past events based on the current state may seem odd for a truly online and predictive system, we also incorporate the reverse version of the entropy defined above. The definition of *reverse boundary entropy*, H_r, can be obtained by simply reversing l and r in Eq. 1 above. The justification of using reverse predictability measures for segmentation comes from two sources. First, intuitively, it seems that what we hear at a particular moment changes our interpretation of past input, especially if the previous interpretation is uncertain in some way. At higher levels of linguistic processing, for example, it is not unusual that when reading some text or listening to someone, things we read or heard start making sense only after we hear or read more. The second, more concrete evidence is from developmental psycholinguistics. Pelucchi, Hay, and Saffran (2009) showed that 8-month-old infants (the same age as the infants in Saffran, Aslin, & Newport's 1996 study) were able to track statistical regularities that are only possible to detect if the subjects were sensitive to some reverse predictability measure between the successive syllables. The additional information the model gains from the reverse entropy also captures (in some way) the effect of the association between units before and after the boundary candidate as TP and MI also do. The individual contributions of forward and reverse entropy to the segmentation performance are presented in Section 3.6.

2.3. *From predictability measures to boundary indicators*

As stated in Section 2.2, uncertainty of what comes next after a sequence of phonemes, measured by high entropy, is an indication of a boundary after the sequence. The higher the entropy, the more likely it is that there is a lexical boundary after the sequence.

However, there is no natural cutoff value after which we should posit a lexical boundary. The optimal cutoff value depends on many factors, including the input language, the size of the input, and the choice of the context size. An easy way to determine a threshold would be estimating this from a corpus where the boundaries are already known. However, we cannot assume that the early language learners know the word boundaries. Hence, tuning a threshold value this way is not tenable for a model of early language acquisition. Another option is to find a natural threshold value, for example, assuming an entropy value over the mean of all entropy values indicates a boundary (a version of this idea is used by Cohen et al., 2007). Although this approach does not require direct supervision, it still requires computing the mean entropy or pointwise mutual information values in advance. Furthermore, intuitively, a level of uncertainty or surprisal just above the mean does not necessarily indicate a boundary. Even if we assume that the average values can be obtained from earlier exposure to the language, the decision of threshold would still not be solved. For example, a conservative learner, as children are generally assumed to be, would be expected to set a threshold higher than the mean entropy value.

In this study, we use a method similar to a number of earlier studies where peaks in uncertainty are considered indications for lexical unit boundaries (e.g., Hafer & Weiss, 1974). A peak is a location in the input where the entropy is higher than the entropy at the previous and the next locations. A particular shortcoming of this strategy is that, since there cannot be two peaks in a row, it can never find single-unit lexical items. This problem has also been noted by Gambell and Yang (2006), who used a similar strategy as an adversary to the heuristic segmentation strategy they propose. They use the local minima, or “valleys,” between consecutive TP values calculated on syllables as (hard) indicators of boundaries. As a result, their model can only find lexical units of two syllables or longer. Since their TP model tracks statistics between syllables, and the majority of words in their corpus are monosyllabic, the problem is particularly severe in their case. However, they do not offer any solution.

Although the “peak-based” segmentation is attractive because it does not require any threshold-tuning implausible for a model of human segmentation, the exact peak method does not have any psycholinguistic grounding. The intuition is that uncertainty or surprise is higher at the boundaries in comparison to the word-internal locations. In this study, instead of looking for prefect peaks of uncertainty, measured by H, we take an increase in H before the candidate boundary and a decrease in H after the candidate boundary as two separate indications.

Fig. 1 demonstrates the changes in H values in the example utterance “Is that a kitty?” Note that since we are taking both “increase before” and “decrease after” a candidate boundary location as separate indicators, there are two indicators for every potential boundary location (excluding the utterance boundaries). In the first boundary candidate in Fig. 1 (after /I/), none of the two indicators indicate a boundary: The entropy decreases before, and increases after this position. After /Iz/, on the other hand, both indicate a boundary: Entropy increases before, and decreases after this position. The indicators are not always correct, nor do they agree all the time. For example, after the penultimate phoneme /t/, both indicators agree for a boundary, where there is none according to our

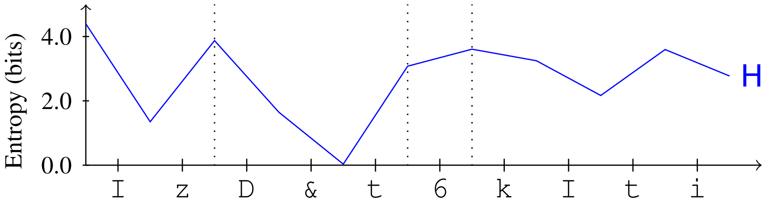

Fig. 1. The entropy values for all potential boundary locations in the example utterance /IzD&t6kIti/ “Is that a kitty?” The values are calculated using statistics over a phoneme-bigram left context on a child-directed speech corpus. Dashed vertical lines mark the boundaries in the gold-standard segmentation.

reference segmentation. Here, both indicators are wrong in suggesting a boundary. As expected, both boundaries of the single-phoneme lexical unit /6/ cannot be identified by proper peaks. Although the boundary to the right of /6/ is identified by a peak, the boundary to the left of /6/ is not marked by a peak. The uncertainty increases before the left boundary of /6/, but there is no proper peak here since it continues to increase afterwards. Hence, with the single measure plotted in Fig. 1, the boundary decision is inconclusive here. Fig. 1 demonstrates only two indicators from a single measure (H) using a single context size. Besides the use of both H and H_r, the varied context size leads to a large number of indicators for each boundary candidate. We will next discuss combining the decision of these multiple indicators to obtain better results.

2.4. *Combining multiple indicators and learning from the input*

With multiple measures of predictability and multiple context sizes, the steps described earlier in this section result in a potentially large set of boundary indicators. The model’s final decision is based on a variation of the weighted majority voting algorithm (Littlestone & Warmuth, 1994). The weighted majority voting algorithm calculates a weighted sum of each indicator. If the weighted sum exceeds half the number of indicators, it proposes a boundary. More precisely, we calculate the sum

$$\sum_{i}^{K} w_i 1_i$$

where w_i is the weight of the ith indicator, 1_i is the corresponding indicator function yielding 1 if the indicator posits a boundary and 0 otherwise. The index i ranges over all K indicators. If the sum above is larger than $K/2$, the combined decision is in favor of a boundary, otherwise for a word-internal position. Ties are broken randomly.

The success of the boundary decisions depends on two factors: the precision of the individual boundary indicators and the weights assigned to each indicator during the majority voting (Boland, 1989). The model presented here improves both the quality of individual indicators and the combination of their decision by learning from the input in an incremental manner.

As well as its inherent usefulness for the task, an individual indicator's precision depends on the relevant phoneme n-gram statistics collected from the earlier input. The learner simply counts the number of phoneme n-grams (up to 4-grams in this study) in the input as it processes each utterance. Hence, as the learning progresses, the learner obtains a better estimate of n-gram frequencies in the input language, the entropy values become more representative, and the model achieves a better segmentation performance.

The learner also adjusts the weights w_i after every boundary decision. The weight adjustment is a common practice in computational systems that employ weighted majority voting. Typically, weights of all voters are set to one at the beginning, and the weight of an indicator is updated based on the number of errors made by the voter. In the case of supervised learning, one can count errors directly by comparing the indicator's decision to the expected outcome. In our case, on the other hand, the model does not know the correct decision. Instead, we adopt a mechanism where weighted majority vote is considered the correct decision. More precisely, the number of errors (disagreements with the weighted majority vote) made by each indicator is counted after every boundary decision. Then, the weight w_i of the indicator i after the Nth boundary decision is calculated as

$$w_i \leftarrow 2\left(0.5 - \frac{e_i}{N}\right)$$

where e_i is the number of times the indicator i disagreed with the weighted majority vote within the N decisions made by the learner so far.

This update rule sets the weight of an indicator that is half the time wrong (a voter that votes randomly) to zero, eliminating the incompetent voters. If the votes of a voter are in accordance with the rest of the voters almost all the time, the weight stays close to one.

3. Simulations and results

This section presents the results of simulations conducted, using the model described in Section 2. The experiments are conducted using child-directed input utterances from the CHILDES database (MacWhinney & Snow, 1985). The software used for running the simulations and the processed version of the child-directed corpora used in this study are available at http://dx.doi.org/10.5281/zenodo.14537. In the experiments reported below, we have used a combination of indicators, using boundary entropy (H) and reverse boundary entropy (H_{r}). The context sizes (l or r) are varied between one and four phonemes, resulting in a set of 16 voters at every possible boundary location. We test the system on phonemically transcribed child-directed speech corpora and present the results in this section. Before presenting the results, we briefly introduce the data and the evaluation metrics we use.

3.1. Data

The results presented in the rest of this section are based on two child-directed speech corpora. First, to allow comparison with the earlier studies, we use a well-known corpus of English child-directed speech. This corpus was collected by Bernstein Ratner (1987) and processed by Brent and Cartwright (1996). It has become the de facto standard for evaluating segmentation models. Following the convention in the literature the corpus will be called the *BR corpus*. The original corpus is a part of the CHILDES database.

The original orthographic transcription of the corpus was converted to a phonemic transcription by Brent and Cartwright (1996). All words are transcribed the same at every occurrence, and onomatopoeia and interjections are removed. The BR corpus consists of 95,809 phonemes, 33,387 words, and 9,790 utterances. A complete description of the corpus can be found in Brent (1999).

The BR corpus has been used by many other computational studies of segmentation. The corpus is also distributed with the implementation of the models presented by Venkataraman (2001) and Goldwater et al. (2009). The copies of the corpus in these sources are identical, and the same copy was used in this study, except 12 boundary mismatches between segmentation of two words in the text version and phonemic transcriptions were corrected. The phonemic transcriptions of 10 instances of the word /ebisi/ "ABC" and two instances of the word /Enim%/ "anymore" have been modified to match the text version. In all cases, this resulted in removing boundaries in the instances of /e bi si/ and /Eni m%/.

Second, to verify the results on a larger, more diverse dataset, we also report results using a corpus obtained by combining child-directed utterances from all American English transcripts in the CHILDES database as of April 2011. To limit the age range to a similar range with the BR corpus, only the recording sessions where target children were younger than 1 year were used. The resulting corpus was a partial combination of the following sections of the CHILDES: Brent (Brent & Siskind, 2001), Higginson (Higginson, 1985), Providence (Demuth, Culbertson, & Alter, 2006), Rollins (Rollins, Pan, Conti-Ramsden, & Snow, 1994; the section of the corpora for normally developing children), and Sodesrstrom (Soderstrom, Blossom, Foygel, & Morgan, 2008). All child-directed utterances in these sessions are processed and converted to phonemic transcriptions following Brent (1999). The resulting corpus contains 53,770 child-directed utterances for 24 different children recorded in 171 sessions. The ages of children were between 0;6 and 0;11.29 (M = 9;11, SD = 48 days). The order of the utterances in each session was kept intact, and the sessions were combined according to the age of the child from younger to older.

3.2. Evaluation metrics

Ideally, we want to evaluate the models of language acquisition, and the models of cognition in general, not based on how well they perform, but how well they match the humans performing the same task, for example, including the age-appropriate "mistakes" in the period of development modeled. The lack of knowledge and appropriate corpora, however, limits our evaluation methods. The issues of evaluation have been discussed in the field, and evaluation methods that solve some of the issues are suggested in the recent literature (Phillips & Pearl, 2015). We report well-established evaluation measures that compare the model's output to the gold standard segmentations (corresponding to typical adult segmentation) in a well-known corpus of child-directed speech. This also allows easy comparison of the results with the results reported in earlier literature.

Two quantitative measures, *precision* and *recall*, originate in the information retrieval literature and have become the standard measures of evaluation of computational simulations. Precision (P) is defined as follows:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP is the number of true positives (items identified correctly by the model) and FP is the number of false positives (items identified by the model that are wrong according to the gold standard). Precision can be seen as a measure of exactness, and it is sometimes called *accuracy* in the cognitive science literature.[7]

Recall (R) is a measure of *completeness*, and sometimes called so in cognitive science literature. Recall is defined as follows:

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where FN denotes false negatives (the number of items missed by the model).

High precision is possible at the expense of recall, when the model is not able to find many relevant items. Similarly, high recall is possible at the expense of precision, when the model finds many incorrect items along with the relevant ones. To have a balanced single indication, a derived measure, F_1-score, is used, which is the harmonic mean of precision and recall.[8]

$$F_1\text{-score} = 2 \times \frac{P \times R}{P + R}$$

As in recent studies of computational segmentation, we report three different types of precision and recall values.

- *Boundary precision* (BP) and *boundary recall* (BR) calculations use TP, FP, and FN values calculated for the boundaries. Here, TP is the number of correctly identified boundaries, FP is the number of boundaries suggested by the model in word-internal positions (according to the gold standard), and FN is the number of boundaries the model fails to identify. Since utterance boundaries are clearly marked in the input, they are not included in the calculation of the boundary scores. The F-measure calculated using BP and BR will be denoted BF.

- *Token*, or *word*, *precision* (WP) and *word recall* (WR) scores require both boundaries of a word to be found to be counted as a TP. Hence, discovering only one of the boundaries of a word does not indicate success for these measures. The token scores are naturally lower than the boundary scores. The *F*-measure calculated from WP and WR will be denoted WF.
- *Type*, or lexicon, precision (LP), lexicon recall (LR), and lexicon *F*-measure (LF) are similar to token scores; however, the comparisons are done over the word types (unique words) the model proposes and word types in the gold standard. These scores are typically lower than the word scores. If a model does a good job only at segmenting high-frequency words (e.g., function words), the lexical scores are lower than the word scores, but if the model is good at segmenting low-frequency words as well, the lexical scores would be closer to the word scores. In case the model is particularly bad at segmenting high-frequency words, but good at segmenting low-frequency words, the lexical scores can be higher than the word scores.

Precision, recall and *F*-measure are the standard measures that are well understood and widely used in the literature. However, it is often insightful to study where a system fails. For this reason, we describe two error measures relevant to segmentation, and, when possible, report these measures alongside the precision, recall and *F*-measure.

A segmentation error can be due to one of two reasons. First, the model may fail to detect a boundary, causing *undersegmentation*. Second, the model may insert a boundary where there is none, causing *oversegmentation*. The simple counts of oversegmentation and undersegmentation errors change depending on the size of the corpus. Hence, they are not comparable across the simulations run on different corpora. Furthermore, in a typical corpus, there are more word-internal positions than boundaries. As a result, there are more chances of making an oversegmentation error compared to an undersegmentation error. To normalize for the size of the input and differences in the ratio of boundaries, we use the following error measures for oversegmentation (E_o) and undersegmentation (E_u), respectively:

$$E_\mathrm{o} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

$$E_\mathrm{u} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

where, as before, TP, FP, and FN are the same quantities used for calculating BP and BR, and TN indicates true negatives (the number of correctly predicted word-internal positions).

In plain words, E_o is the ratio of the false boundaries inserted by the model divided by the total number of word internal positions in the corpus. Similarly, E_u is the ratio of boundaries missed to the total number of boundaries.

The two error measures described above are related to precision and recall. The relationship is straightforward between the E_u and boundary recall ($E_u = 1 - \text{BR}$). As E_u is related to BR, the E_o is related to BP. In general, high E_o leads to low precision. However, the relationship between E_o and BP is not straightforward. While BP quantifies the ratio of correct boundary choices made among the model's choices, E_o quantifies the ratio of the wrong boundary decisions to the number of potential mistakes that can be made based on the gold standard segmentation. Hence, the error measures provide a direct indication of the room left for improvement.

All segmentation models we are interested in use unsupervised learning methods in the sense that the algorithms do not have access to information regarding real boundary locations. As a result, it is common practice to present the results on a single dataset without training–test data separation. Although it is fair not to split the input of an unsupervised learner into training and test sets, having the systems tested on the entire input results in an unfair comparison between batch and incremental learners (see next subsection for more detail). Hence, we also provide an incremental evaluation of the model to present the model's performance at the end of the learning as well as the learning progress over time.

3.3. Segmentation performance of the predictability model

Our aim in this study is not to achieve the best segmentation performance. We know that children are sensitive to and make use of other cues. Hence, a model relying on the predictability cue alone is necessarily impoverished. However, since predictability is a well-attested cognitive mechanism, and a good candidate for bootstrapping other cues, it is instructive to see the segmentation performance one can get from the predictability cue alone.

Table 1 presents the evaluation metrics for our predictability model along with a random baseline model and a selection of state-of-the art models from the literature which are evaluated on the same corpus. Note that the random segmentation used here follows the common, slightly informed, random baseline used often in the segmentation literature (since Brent & Cartwright, 1996). It inserts boundaries randomly; however, the number of boundaries inserted is the same as the number of boundaries in the gold-standard segmentation.

The results in Table 1 show that the predictability model as defined here performs similar to the other models in the literature with respect to boundary and word (token) *F*-measure. However, although it performs clearly above the random baseline, it shows a rather low lexical (type) *F*-measure compared to other models. In all measures, the model has higher recall, but lower precision. This behavior is also reflected by the under- and oversegmentation errors, which are 10.60% and 7.60%, respectively.

The low lexicon performance is expected, as the model has no notion of lexical items. However, it also is related to the fact that the model is incremental. The performance scores are adversely affected by the initial mistakes made before the model was able to learn from the data. Hence, using performance scores calculated over all decisions during

Table 1
Performance scores for our predictability model in comparison to other models in the literature (see discussion in the main text for caution needed for comparing these numbers directly).

	Boundary			Word			Lexicon		
Model	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Brent (1999)	80.3	84.3	82.3	67.0	69.4	68.2	53.6	51.3	52.4
Venkataraman (2001)	81.7	82.5	82.1	68.1	68.6	68.3	54.5	57.0	55.7
Goldwater et al. (2009)	90.3	80.8	85.2	75.2	69.6	72.3	63.5	55.2	59.1
Blanchard et al. (2010)	81.4	82.5	81.9	65.8	66.4	66.1	57.2	55.4	56.3
Predictability	81.6	89.4	85.4	70.6	75.3	72.9	37.4	65.0	47.5
Random	27.9	27.5	27.7	12.9	12.8	12.8	6.1	44.8	10.8

The evaluation metrics, precision (*P*), recall (*R*), and *F*-measure (*F*), are defined in Section 3.2. The performance scores for the other models are listed as reported in the related publications. If there were multiple models reported in a study, the model with the highest lexicon *F*-measure is presented. All scores are obtained on the BR corpus. The scores are presented as percentages.

a simulation leads to an unfair advantage in favor of the batch models. The segmentation decisions of an incremental model include early, “naive,” state of the learner, while all decisions of a batch model only reflect the final, “learned,” state of the learner.

To demonstrate the performance with increasing input, we present the performance and error scores of the predictability model as a function of amount of input utterances in Fig. 2. As expected, during the initial stages of the learning, the model makes more mistakes. Particularly, the amount of undersegmentation errors is very high because of the

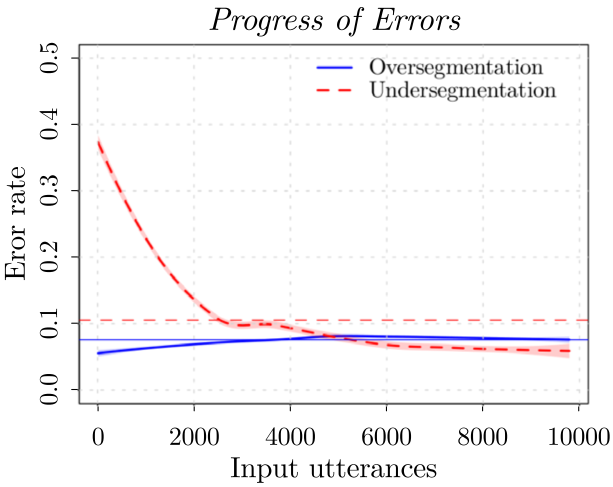

Fig. 2. The performance and error scores calculated progressively for every 10 input utterances on the BR corpus. The curves are obtained fitting a locally weighted regression model using loess function in R statistical computing environment (R; R Core Team, 2014). The shaded areas represent ±1 *SE* around the curves. The thin horizontal lines represent oversegmentation and undersegmentation errors calculated over the complete corpus.

fact that the learner did not yet collect enough information to posit any boundaries. With increasing input, the under-segmentation errors are reduced drastically. The oversegmentation errors, on the other hand, are relatively stable except for a slow increase at the beginning.

Since undersegmentation does not increase while oversegmentation decreases, *F*-measures also improve substantially if measured later in the learning process. For the last 790 utterances (last block in a 1,000-utterance block division of the BR corpus) in the simulation presented in Fig. 2, the BF, WF, LF values are 87.70%, 76.20%, and 65.10%, respectively. These figures are not only better than scores calculated over the complete corpus (85.40%, 72.90%, 47.50%, respectively, repeated here for convenience), but also higher than many of the other scores that are obtained by more sophisticated models in Table 1.[9] The improvement is particularly large for lexical *F*-measure, since the number of early undersegmentation errors constitutes an important part of all of the lexical units proposed by the system.

3.4. Qualitative error analysis

The numbers presented in the previous subsection show that the performance of the predictability model is similar to the other models presented in the literature. The quantitative evaluation also reveals some of the general characteristics of the model. For example, we observe that the model as defined here tends to oversegment. However, to gain more insight into the model's behavior, we present and discuss some of the actual segmentation results. Fig. 3 presents the first and last 10 utterances from the model's output.

The utterances in Fig. 3a are mostly undersegmented. Even though this behavior does not map directly to the long-term segmentation behavior of human learners, the conservative behavior at the beginning is in line with the general human learning pattern, which is also observed during learning segmentation (Dahan & Brent, 1999). The model also seems to start segmenting rather quickly. Despite the small amount of information at that point, the majority of the indicators operating on different basic unit lengths and predictability measures indicate a boundary before and after the string /lUk/ "look" at the fourth input utterance. Hence, the utterance is segmented even in that very early stage. It is also worth noting that the successive discovery of the sequence /lUk/ as a lexical unit is not directly related to the earlier discovery of this particular sequence. All decisions are based on phoneme n-gram statistics; no information is used from the boundaries discovered earlier.

As expected, the results in Fig. 3b are a lot better, although the model's eager segmentation is visible here. Most of the errors are oversegmentation errors. The large number of single-phoneme items in the output of the model indicates that it would benefit from a phonotactics component, or a higher level mechanism to constrain the sensible sequences of words.[10] The few undersegmentation errors, for example, /kAmQt/ "come out" and / D6d%/ "the door," involve common words or word sequences found in the input corpus. This behavior is in line with other segmentation models reported in the literature.

`yuwanttusiD6bUk`	*you want to see the book*
`lUkD*z6b7wIThIzh&t`	*look there's a boy with his hat*
`&nd6dOgi`	*and a doggie*
`yuwanttu lUk &tDIs`	*you want to look at this*
`lUk &tDIs`	*look at this*
`h&v6drINk`	*have a drink*
`okenQ`	*okay now*
`WAtsDIs`	*what's this*
`WAtsD &t`	*what's that*
`WAt IzIt`	*what is it*

(a)

`DIs op~ z h(`	*this opens here*
`huz In s9d D6 hQs`	*who's inside the house*
`DIs &n 6mL`	*this animal*
`k&n kAmQt If yu pU l h#d In Af`	*can come out if you pull hard enough*
`go 6hEd`	*go ahead*
`yu wan t mami tu tek hI m Qt`	*you want mommy to take him out*
`kloz D6d%`	*close the door*
`nQ D6 d% Iz op~`	*now the door is open*
`yu k&n pUt hIm In h(`	*you can put him in here*
`no 9dId ~t TIN k It wUd f It iDR`	*no i didn't think it would fit either*

(b)

Fig. 3. The model's output for the first (a) and the last (b) 10 utterances. Single-word utterances without oversegmentation errors are excluded.

3.5. Stability of the results over large input

The results presented so far are results obtained on a small, well-known reference corpus (BR corpus). This choice was motivated by the fact that this corpus has been used by a large number of studies. Hence, results can be compared with other models. A reasonable objection about the oversegmentation tendency shown by the model is that this trend may continue until all input utterances are segmented into the basic unit (the phonemes in this study). This is especially worrying since the amount of input (9,790 utterances) used in this study is only a very small fraction of what children hear. To assure that the model is stable in the long run, and does not evolve into a model that only produces single phonemes as lexical items, we run the same simulation on a larger child-directed corpus.

Fig. 4 presents the under- and oversegmentation errors on the larger child-directed corpus introduced in Section 3.1. It is clear that the oversegmentation errors increase at the beginning, but they stabilize slightly over 0.10, and do not show a clear trend of increase. The undersegmentation errors, on the other hand, seem to be reduced with more data. As a result, at the end of this particular corpus the model achieves boundary, word, and lexical *F*-measure of 88.10%, 75.60%, and 61.10%, respectively.

3.6. Contributions of forward and reverse entropy

All experiments presented in the earlier sections are performed with the full model combining forward and reverse entropy. In this section we present experiments that use only forward or reverse entropy to show the relative contributions of each to the full

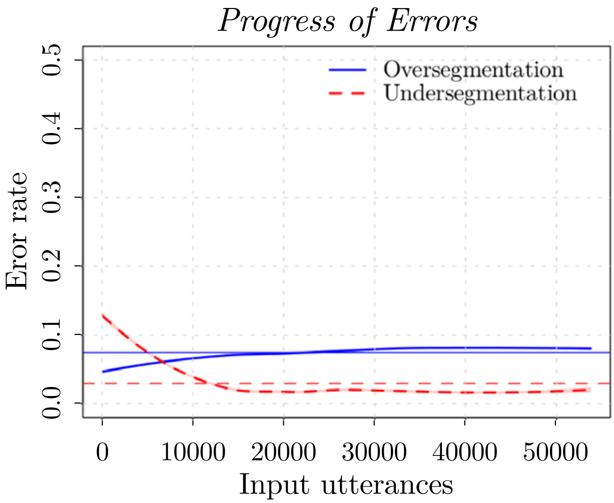

Fig. 4. The error scores calculated progressively for every 50 input utterances on the larger child-directed speech corpus. The curves are obtained by fitting a locally weighted regression model. The shaded areas represent ±1 *SE* around the curves (barely visible here because of the small standard error). The thin horizontal lines represent the corresponding error scores calculated over the complete corpus.

Table 2
Performance scores for models that only use forward and reverse entropy. (The row labeled “full model” is the same as the predictability model in Table 1, repeated here for convenience)

	Boundary			Word			Lexicon			Error	
Model	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	E_o	E_u
Forward	70.1	82.1	75.6	53.9	60.5	57.0	25.7	55.7	35.2	13.2	17.9
Reverse	58.0	82.1	68.0	36.6	47.3	41.2	21.9	47.4	30.0	22.4	17.9
Full model	81.6	89.4	85.4	70.6	75.3	72.9	37.4	65.0	47.5	7.6	10.6

model. The combination of both is motivated by the fact that forward entropy does not fully capture the effect of surprise, since it is only based on the sequence seen so far. The use of reverse entropy compensates for this to some extent. Furthermore, the effect of reverse entropy can also be interpreted as providing a short window of opportunity to affect the earlier decisions based only on forward entropy.

Table 2 presents the performance scores of both reduced models in comparison to the full model. The individual models still combine indicators with different context sizes in both cases (see Appendix A for the effect of context size). The forward entropy alone seems to perform quite well. While the reverse entropy performs worse than its forward counterpart, the combination of both provides substantial improvements over all performance scores.

4. Discussion

This paper presents experiments with a segmentation model based purely on predictability of consecutive sub-lexical speech units in the input stream. Unlike most of the recent models in the literature, the model does not try to find the optimum segmentation of a given input utterance or corpus. Instead, it guesses boundaries using local statistics over a limited neighborhood of the boundary candidate. A large number of studies in developmental psycholinguistics demonstrated children’s sensitivity to (or use of) a number of local cues for segmentation. The approach is not new for computational models of segmentation, either. It has been used by many earlier studies either implicitly or explicitly. However, it has “fallen out of fashion” probably due to the impression that it does not yield good segmentation performance. The results presented in this paper bring up a number of interesting observations and questions that we discuss in this section.

4.1. The performance of the model

As stated earlier, the aim of this study is not to produce a complete segmentation model. Here, we focus on a single cue or strategy alone. One of the motivations for this study has been to establish the extent to which predictability can be useful for segmentation. The results presented in Section 3 show that the model performs similarly to the other state-of-the art models in the literature. Furthermore, towards the end of training on the BR corpus, the model presented here achieves boundary, word, and lexicon *F*-measure of 88%, 76%, and 65%, respectively. Although the model uses only the predictability cue, these scores are even higher than some of the state-of-the-art models trained on the same corpus (presented in Table 1). Nonetheless, we note again that our aim in this study is not to show that the present model outperforms others. The competitive score underlines the fact that a model that relies only on predictability can perform similarly to the state-of-the-art models, in contrary to earlier reports. Hence, it is a reliable cue for bootstrapping and complementing other cues for segmentation.

Because of the use of different input and/or evaluation methods, the scores presented in this paper are not directly comparable to the earlier studies that used predictability strategy. However, results reported earlier in the literature using only the predictability cue have been much lower than the results reported in this paper. Christiansen et al. (1998) report 37% WP and 40% WR with an SRN using phonotactics and utterance boundary cues on another child-directed speech corpus (Korman, 1984). Graphs presented by Brent (1999) indicate about 50%–60% WP and WR and 20%–30% LP for his baseline model utilizing pointwise mutual information on the BR corpus. Cohen et al. (2007) report 76% BP, and 75% BR on George Orwell's *1984*. Gambell and Yang (2006) report 41.60% BR, and 23.30% BP. Despite the differences in the evaluation methods and the corpora used, the performance scores presented here are clearly better than these figures from the earlier studies.

The model presented in this paper, as the majority of segmentation models in the literature, is an *analytic* segmentation model. That is, it breaks longer sequences (utterances) into smaller ones (words), as opposed to *synthetic* models that build words from sub-word units (Batchelder, 2002; De Marcken, 1996; Olivier, 1968). The analytic models initially do not segment the utterances. They increase their segmentation accuracy by admitting more and more boundaries as the input provides more information. As a result, a potential problem for such models is segmenting all utterances to the basic units in the long run. The results presented in Section 3.5 show that our model is not susceptible to this behavior. Note that there is no ad hoc mechanism that prevents the model from segmenting the input to single phonemes. For some input, for example, constructed artificially, the model may learn to segment the input into single basic units. The only constraint that keeps the model in check is the statistical distributions in the input language.

As it is implemented in our model, the predictability cue tends to oversegment (but without a snowball effect of oversegmentation to basic units on the long run). If one takes the output of the model as the only source for building a lexicon, this may cause problems for a learner. However, the model presented here is not offering a complete solution for lexical acquisition. Although predictability is a good candidate for one of the first strategies at work in lexical acquisition (see Section 1.1), it is clearly not the only source of information for lexical acquisition. Human learners combine many cues in the input, but they are also likely to make use of their linguistic and world knowledge at many levels for acquiring an adult-like lexicon.

4.2. *Basic units for segmentation*

A common discussion in the segmentation literature is about the basic perceptual unit in the speech signal. Most computational studies in the literature, including the present one, use a phonemically transcribed input. On the other hand, it is often argued that syllable is more appropriate as the basic unit, probably due to its perceptual salience (Lignos, 2012; Phillips & Pearl, 2012, 2014). The linguistic evidence for or against both units seems to be inconclusive (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Steriade, 1999).

Although we use phonemes as the indivisible unit in this study, an interesting aspect of the current model is that it does not commit to a single basic unit for making generalizations. We collect statistics over larger phoneme n-grams to capture the regularities that can be captured by multi-phoneme units such as syllables. In combination with the weighted majority voting algorithm, this strategy learns, in a naive way, the units that are useful for segmentation. To some extent, our model discovers units useful for the purpose from the data. The use of different units in potentially differing roles is also in line with findings of Newport and Aslin (2004). In the current model, the weights are associated with all phoneme n-grams of the same size rather than individual phoneme n-grams. This reduces the number of parameters to learn, hence the computational cost. However, it also makes it difficult to relate the weights learned by model with linguistically more meaningful units. A possible direction for the future research is to lift this limitation by using a better representation of the input and better models of learning at multiple levels (e.g., an incremental version of the models with a notion of learning at multiple levels similar to Johnson & Goldwater, 2009).

Another issue related to the choice of basic unit is that the *phone*, rather than the phoneme, is a better representation of input for a model of early language acquisition. Although we acknowledge this as a weakness of the simulations, we note that the use of phone versus phoneme does not necessarily affect the performance of the model in a positive way. Using a corpus without phonetic variation decreases the number of units to consider, hence, decreasing the computational cost and amount of data needed for generalizations. However, it also removes some useful information from the input. For example, having the distinction between aspirated and unaspirated allophones of the voiceless stop consonants in English is a very valuable cue to word boundaries (they are aspirated when they are word-initial). This information would be available to a phone-based input representation, while the phonemic representation does not allow the learner to make use of this information. Furthermore, any hand-annotated corpora will include theoretical or practical idealizations that will affect the representations of units in potentially arbitrary ways. A potential solution to this problem is to learn (continuous) input representations together with learning segmentation (e.g., Ma et al., 2016).

From the modeling perspective, the use of one unit or the other does not change the underlying computation. The choice of the basic unit, however, may affect the performance of the model. The effect of using different units in computational models of segmentation is discussed in Çöltekin (2015).

4.3. Predictability as a domain general bootstrapping method

Predictability is not only one of the many cues that are shown by experimental studies to be useful in the segmentation task, but it also has additional properties that make it interesting for computational modeling of segmentation. First of all, predicting events or percepts from the environment seems to be what brains are busy with at multiple levels and modalities, at all times. The machinery in the brain is good at, and used to use predictability as a cue for many tasks. Hence, it seems to be a good domain-general method for finding patterns and learning. Furthermore, most other cues that are attested to be used for segmentation are language specific. For example, another well-known cue for segmentation is lexical stress which was also shown to outweigh predictability when available (Cutler & Butterfield, 1992; Jusczyk, Hohne, et al., 1999; Jusczyk, Houston, et al., 1999; Thiessen & Saffran, 2004). However, lexical stress, like many others, requires language-specific knowledge. Not all languages exhibit a stress pattern, and not all of the languages that exhibit stress have the same lexical stress pattern. The predictability cue, on the other hand, does not require any prior knowledge of the target language.

Logically, predictability is a good candidate for bootstrapping other, language-specific, cues. Of course, it can only be a viable cue for bootstrapping, if it is able to identify lexical units successfully. Although the level of success required for this purpose is unclear, a necessary condition for a cue to bootstrap others is to perform well. The results presented in this paper show that the strategy performs similar to other, more elaborate strategies. Hence, supporting the position of (predictability) statistics as a cue for bootstrapping other cues (Swingley, 2005; Thiessen & Saffran, 2007).

4.4. Measuring predictability

As discussed in earlier sections, we have substantial evidence for use of predictability in many cognitive tasks, including segmentation. However, we do not (yet) have access to how exactly the human brain computes predictability. For a computational model, as well as in an experimental design, we need to quantify the notion of predictability. In this study we measure (un)predictability using boundary entropy. The choice of this particular measure, along with a few other choices made during the implementation of the model, is an example of choices for convenience or idealization that are found in any modeling practice. Nevertheless, there are some interesting aspects of these "modeling artifacts" that warrant some discussion.

Entropy is a principled measure of uncertainty. Hence, it is a natural choice for measuring (un)predictability in the context of segmentation. However, as we noted already, there are other ways to quantify (un)predictability. Appendix B presents further data on three other measures, successor variety (SV), transitional probability (TP), and pointwise mutual information (MI), that are used in earlier segmentation literature. The data presented in Table B1 suggest that all measures perform comparably with an indication that SV and entropy perform better than TP and MI in the setting presented in Appendix B.

An interesting question arises because of the use of both forward and reverse entropy. The forward entropy measures the uncertainty given only the preceding sequence. It does not make use of the sequence following the boundary position. For example, while deciding whether there is a boundary after /Iz/ in our earlier example /IzD&t6kIti/ "Is that a kitty?," forward entropy can be calculated only based on /Iz/, while TP or MI requires the unit(s) after the boundary candidate. The results we present in Section 3.6 also indicate that, although combining forward and reverse entropy is beneficial, a good part of the performance comes only from the forward entropy. This is also intuitively appealing. A human listener would have a good idea of whether a word boundary follows or not without hearing the beginning of the next word. However, the experimental evidence so far does not show whether early learners use this forward-only mechanism (as measured by the forward entropy) or they need the following context (more compatible with our combined model, or models using TP or MI). The measures overlap to a large extent, and the stimuli used in all experimental studies so far are compatible with all measures listed above. The typical experimental settings, such as head-turn preference, only measure the preference towards the overall stimuli, rather than employing real-time measurements at potential boundary locations. Specially crafted stimuli or methods that measure (brain) response during the processing of the stimuli may be able to tap into this rather subtle but interesting difference.

The choice of the context size, and as a consequence the choice of the number of indicators, is a free parameter of the model presented in this paper. The results in Appendix A indicate that the model performs similarly around phoneme 4-grams that we used in the main text. However, the performance drops as the context size is increased or decreased. The reason for lower performance of the models using small context size is clear. Using only a small context size does not allow the model to utilize the valuable information longer context sizes provide. The explanation for lower performance of the models using larger context size is more involved. From the computational perspective, the lower performance as context size increases has to do with the requirements of the weighted majority algorithm used by the model. The success of weighted majority algorithm depends on performance of the individual voters (boundary indicators). In particular, each voter is expected to be accurate, that is, performing better than a random voter, and diverse, that is, voting based on (partially) independent information (Hansen & Salamon, 1990). As we increase the context size, the individual indicators degrade for two reasons. First, the frequencies of larger n-grams cannot be estimated reliably due to data sparseness, causing them to act closer to a random voter. Second, the larger n-grams do not provide any new information over the information provided by their shorter subsequences. The second reason may also be tied to the fact that the information processing capacity of the cognitive system is limited. These type of limits, the "magical numbers," are known to affect human processing of language (Miller, 1956). In turn, the product of this system, the language input to the learners, is affected by these limits in such a way that only a limited context is useful.

4.5. Incremental and predictive models of segmentation

The focus of this study is to show that predictability associated with sequences of basic units in the input to the learners is useful for segmentation. This question clearly seeks explanations at Marr's computational level (Marr, 1982). However, we note that the model presented here has some properties that may allow it to be beneficial in investigating questions at Marr's algorithmic level.

Unlike the majority of the recent models in the segmentation literature, the model presented here processes the input in an online fashion, without waiting for the end of the utterance or the complete corpus. Furthermore, at any point during processing of the input, the model bases its decisions on what *may* come next. Hence, exhibiting the sort of predictive behavior we observe with human processing of language input.

Another crucial aspect of the present model is that it uses a cue that is known to be used by humans. Combined with the fact that processing and learning in the model is incremental, the present model takes a step towards models of segmentation at Marr's algorithmic level. This is not true only because of the fact that the current model is incremental. Incremental algorithms that are (approximately) equivalent to batch algorithms can be devised (e.g., Pearl et al., 2010). However, the approximation or the equivalence is typically established on the grounds of formal mathematical properties of these algorithms. The crucial aspect of the current model is not only that it is incremental, but it also uses a strategy shown to be used by humans in the same task.

As pointed out by Goldwater et al. (2009), similar to the other models that may provide explanations at Marr's algorithmic level, some of the behaviors of the model are due to the (sometimes arbitrary) choices made during design and implementation of the model. However, this is not necessarily a weakness. Diverging from ideal computational models by adding cognitively plausible mechanisms or constraints may allow us to develop predictive models on the long run. As we learn more about the underlying human cognitive processes, we can replace the arbitrary choices in our models with the ones that stem from our knowledge of human cognition.

5. Conclusion and outlook

This paper presented a computational model of lexical segmentation based only on predictability. The aim of this modeling effort has been to investigate the performance of this particular strategy as a strategy for early segmentation. The concept of predictability is known to be prevalent in all areas of cognition (Clark, 2013) and also shown to be used by infants during early stages of language acquisition, including lexical segmentation (Saffran, Aslin, & Newport, 1996). Our results show that, contrary to earlier reports, the predictability strategy leads to a segmentation performance that is not far from the state-of-the-art models presented in the literature. Combined with its language- and domain-independent nature, this finding indicates that the predictability cue is a good candidate for "bootstrapping" lexical segmentation.

Besides good segmentation performance, we also report on some aspects of the model that allow us to understand the predictability cue better. Particularly, we show that combining the information from different basic unit sizes, the phoneme-n-gram lengths used in this study, is helpful. We also observe that the model presented here tends to oversegment, but it does so at a consistent level that does not increase with more input.

Contrary to the recent trend of segmentation models that aim to learn a compact lexicon, the model presented here uses low-level local cues in the input stream for lexical segmentation. These cues have been studied extensively in experimental psycholinguistic studies. The present study also shows that there still are aspects of these cues and their use in lexical segmentation that we can investigate and learn through computational modeling. Further research in combining these local cues with modern machine learning methods is likely to inform us more about the nature and interaction of these cues, hence contributing to our understanding of language acquisition. Testing these models on languages other than English and using more realistic input, for example, including variability in the speech signal, are also further directions for future research.

Notes

1. See, for example, Clark (2013). Clark puts forward a specific version of the claim that brains are prediction machines in this highly debated article. The debate, carried out in 30 commentaries published with the article, is around the specifics of the proposed prediction machine, none disputing that the brains are prediction machines.
2. Since concatenating (lexical) units is one of the fundamental ways human language processing works, expecting early learners to look for boundaries is not necessarily a far-fetched assumption. Nevertheless, relying on fewer assumptions, even those that seem to be well warranted by the data at hand, is a desirable property of a language- and domain-general learning strategy.
3. Although the model in Cohen et al. (2007) has been improved in subsequent work (Hewlett & Cohen, 2009, 2011a,b), the ideas presented in Cohen et al. (2007) are most relevant to the present study.
4. It is even claimed that segmentation is not necessary at all to extract words from a speech stream given a comprehensive lexicon (Baayen, Shaoul, Willits, & Ramscar, 2016).
5. Even though phones would have been a better choice of input unit for modeling early language acquisition, since we use a phonemically transcribed corpus for simulations, our results are based on phonemes. See Section 4.2 for a detailed discussion of the use of alternative units.
6. Boundary entropy defined here is similar to but different from a well-known entropy measure, conditional entropy, which is defined as $-\Sigma_{r \in A} P(r, l) \log 2\ (P(r|l))$. In the preliminary experiments that were conducted, the results obtained for both measures in the segmentation task were similar. The boundary entropy is adopted here since it was used in previous research for segmentation (e.g., Hafer & Weiss, 1974).

7. Unfortunately, accuracy is ambiguous in the cognitive science literature. Accuracy $\left(\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}\right)$, as it is commonly used in many branches of science, is not equal to precision.
8. The subscript "1" indicates that the measure gives equal weights for precision and recall. In its more generic original formulation, F_α-score gives higher weight to recall for higher values of α, and lower values give higher weight to precision (van Rijsbergen, 1979). In this paper, we only report F_1-scores, and simply refer to it as F-measure. It is also customary to present F-measure as percentages (100 × F-measure), which we also follow in this paper.
9. If the initial mistakes are excluded from the evaluation, substantial improvements in the performance scores are also expected for the two incremental models listed in Table 1, particularly for Venkataraman (2001). This is less clear for the model of Blanchard, Heinz, and Golinkoff (2010), since the scores reported already exclude first 1,000 input utterances.
10. Indeed, combining this strategy with other strategies or cues leads to better segmentation performance (see Çöltekin & Nerbonne, 2014, for a demonstration).

References

Aslin, R. N. (1993). Segmentation of fluent speech into words: Learning models and the role of maternal input. In B. D. Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 305–315). Dordrecth: Kluwer Academic Publishers.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. doi:10.1111/1467-9280.00063

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (Chap. 8, pp. 117–134). Mahwah, NJ: Lawrence Erlbaum Associates.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128. doi:10.1080/23273798.2015.1065336

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*, 167–206. doi:10.1016/S0010-0277(02)00002-1

Bernstein Ratner, N. (1987). The phonology of parent-child speech. In K. Nelson & A. van Kleeck (Eds.), *Children's language* (vol. 6, pp. 159–174). Hillsdale, NJ: Erlbaum.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, *37*(Special Issue 03), 487–511. doi:10.1017/S030500090999050X

Boland, P. J. (1989). Majority systems and the condorcet jury theorem. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *38*(3), 181–189. doi:10.2307/2348873

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1–3), 71–105. doi:10.1023/A:1007541817488

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125. doi:10.1016/S0010-0277(96)00719-6

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33–B44. doi:10.1016/S0010-0277(01)00122-6

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Modelling the acquisition of lexical segmentation. In Eve V. Clark (Ed.), *Proceedings of the 26th child language research forum* (pp. 32–41). Chicago, IL: University of Chicago Press.

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), 221–268. doi:10.1080/016909698386528

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. doi:10.1017/S0140525X12000477

Cohen, P., Adams, N., & Heeringa, B. (2007). Voting experts: An unsupervised algorithm for segmenting sequences. *Intelligent Data Analysis*, *11*(6), 607–625.

Çöltekin, Ç. (2010). Improving successor variety for morphological segmentation. In E. Westerhout, T. Markus, & P. Monachesi (Eds.), *Proceedings of the 20th meeting of computational linguistics in the Netherlands* (pp. 13–28). Utrecht, The Netherlands: LOT, Netherlands Graduate School of Linguistics.

Çöltekin, Ç. (2011). Catching words in a stream of speech: Computational simulations of segmenting transcribed child-directed speech. Doctoral dissertation, University of Groningen.

Çöltekin, Ç. (2015). Units in segmentation: A computational investigation. In R. Berwick, A. Korhonen, A. Lenci, T. Poibeau, & A. Villavicencio (Eds.), *Proceedings of EMNLP 2015 workshop on cognitive aspects of computational language learning* (pp. 55–64). Lisbon, Portugal: Association for Computational Linguistics.

Çöltekin, Ç., & Nerbonne, J. (2014). An explicit statistical model of learning lexical segmentation using multiple cues. In A. Lenci, M. Padró, T. Poibeau, & A. Villavicencio (Eds.), *Proceedings of EACL 2014 workshop on cognitive aspects of computational language learning* (pp. 19–28). Gothenburg, Sweden: The Association for Computational Linguistics.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*(2), 218–236. doi:10.1016/0749-596X(92)90012-M

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, *128*(2), 165–185. doi:10.1037/0096-3445.128.2.165

Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition. In M. J. Trexler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (Chap. 8, 2nd ed., pp. 249–283). London: Elsevier.

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155. doi:10.1111/j.1551-6709.2010.01160.x

Davis, M. H. (2006). Lexical segmentation in spoken word recognition. Doctoral dissertation, Birkbeck College, University of London.

De Marcken, C. G. (1996). Unsupervised language acquisition. Doctoral dissertation, Massachusetts Institute of Technology.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, *49*(2), 137–173. doi:10.1177/00238309060490020201

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568–1578. doi:10.1037/0096-1523.25.6.1568

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211. doi:10.1016/0364-0213(90)90002-E

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In J. D. Moore, S. Teufel, J. Allan & S. Furui (Eds.), *Proceedings of the annual meeting of the Association for Computational Linguistics (ACL-08)* (pp. 130–138). Columbus, Ohio: Association for Computational Linguistics.

Gambell, T., & Yang, C. (2006). Word segmentation: Quick but not dirty. Unpublished manuscript. Available at: http://www.ling.upenn.edu/ycharles/papers/quick.pdf. Accessed Nov 17, 2014.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54. doi:10.1016/j.cognition.2009.03.008

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254–260. doi:10.1111/j.1467-9280.2007.01885.x

Hafer, M. A., & Weiss, S. F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, *10*(11–12), 371–385. doi:10.1016/0020-0271(74)90044-8

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. doi:10.1109/34.58871

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*(2), 190–222. doi:10.2307/411036

Hewlett, D., & Cohen, P. (2009). Bootstrap voting experts. In Q. Yang & M. Wooldridge (Eds.), *Proceedings of the 21st international joint conference on artifical intelligence* (pp. 1071–1076). Palo Alto, California: AAAI Press.

Hewlett, D., & Cohen, P. (2011a). Fully unsupervised word segmentation with BVE and MDL. In Y. Matsumoto & R. Mihalcea (Eds.), *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies* (pp. 540–545). Portland, OR: Association for Computational Linguistics.

Hewlett, D., & Cohen, P. (2011b). Word segmentation as general chunking. In S. Goldwater & C. Manning (Eds.), *Proceedings of the fifteenth conference on computational natural language learning* (pp. 39–47). Portland, OR: Association for Computational Linguistics.

Higginson, R. P. (1985). Fixing-assimilation in language acquisition. Doctoral dissertation, Washington State University.

Jarosz, G., & Johnson, J. A. (2013). The richness of distributional cues to word boundaries in speech to young children. *Language Learning and Development*, *9*(2), 175–210. doi:10.1080/15475441.2011.641904

Johnson, M., & Goldwater, S. (2009). Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In M. Ostendorf et al. (Eds.), *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 317–325). Boulder, CO: The Association for Computational Linguistics.

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. doi:10.1006/jmla.2000.2755

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, *64*(3), 675–687. doi:10.1111/j.1467-8624.1993.tb02935.x

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*, 159–207. doi:10.1006/cogp.1999.0716

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, *5*, 44–45.

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In N. Arnett & R. Bennett (Eds.), *Proceedings of the 30th West Coast Conference on Formal Linguistics* (pp. 237–247). Somerville, MA: Cascadilla Proceedings Project.

Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation*, *108*(2), 212–261. doi:10.1006/inco.1994.1009

Ma, J., Çöltekin, Ç., & Hinrichs, E. (2016). Learning phone embeddings for word segmentation of child-directed speech. *Proceedings of ACL 2016 workshop on cognitive aspects of computational language learning* (pp. 53–63). Berlin, Germany: The Association for Computational Linguistics.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–296. doi:10.1017/S0305000900006449

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. doi:10.1037/h0043158

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(Special Issue 03), 545–564. doi:10.1017/S0305000909990511

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162. doi:10.1016/S0010-0285(03)00128-2

Olivier, D. C. (1968). Stochastic grammars and language acquisition mechanisms. Doctoral dissertation, Harvard University.

Pearl, L., Goldwater, S., & Steyvers, M. (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*(2–3), 107–132. doi:10.1007/s11168-011-9074-5

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–247. doi:10.1016/j.cognition.2009.07.011

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*(7), 1299–1305. doi:10.3758/MC.36.7.1299

Phillips, L., & Pearl, L. (2012). “Less is more” in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 863–868). Austin TX: Cognitive Science Society.

Phillips, L., & Pearl, L. (2014). Bayesian inference as a viable cross-linguistic word segmentation strategy: It’s all about what’s useful. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2775–2780). Quebec City, CA: Cognitive Science Society.

Phillips, L., & Pearl, L. (2015). Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. In M. van Schijndel & T. O’Donnell (Eds.), *Proceedings of the 6th workshop on cognitive modeling and computational linguistics* (pp. 68–78). Quebec City, Canada: Cognitive Science Society.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rollins, P. R., Pan, B. A., Conti-Ramsden, G., & Snow, C. E. (1994). Communicative skills in children with specific language impairments: A comparison with their language-matched siblings. *Journal of Communication Disorders*, *27*(2), 189–206. doi:10.1016/0021-9924(94)90040-X

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*(5294), 1926–1928. doi:10.1126/science.274.5294.1926

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. doi:10.1006/jmla.1996.0032

Soderstrom, M., Blossom, M., Foygel, R., & Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, *35*, 869–902. doi:10.1017/S0305000908008763

Steriade, D. (1999). Alternatives to the syllabic interpretation of consonantal phonotactics. In O. Fujimura, B. Joseph, & B. Palek (Eds.), *Proceedings of the 1998 linguistics and phonetics conference* (pp. 205–242). Prague: The Karolinum Press.

Suomi, K., McQueen, J. M., & Cutler, A. (1997). Vowel harmony and speech segmentation in finnish. *Journal of Memory and Language*, *36*(3), 422–444. doi:10.1006/jmla.1996.2495

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86–132. doi:10.1016/j.cogpsych.2004.06.001

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. doi:10.1037/0012-1649.39.4.706

Thiessen, E. D., & Saffran, J. R. (2004). Infants’ acquisition of stress-based word segmentation strategies. In A. Brugos, L. Micciulla & C. E. Smith (Eds.), *BUCLD 28: Proceedings of the 28th annual Boston University conference on language development* (pp. 608–619). Somerville, MA: Cascadilla Press.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*(1), 73–100. doi:10.1080/15475440709337001

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. doi:10.1207

van Kampen, A., Parmaksız, G., van de Vijver, R., & Höhle, B. (2008). Metrical and statistical cues for word segmentation: The use of vowel harmony and word stress as cues to word boundaries by 6- and 9-month-old Turkish learners. In A. Gavarro & M. J. Freitas (Eds.), *Language acquisition and development: Proceedings of GALA 2007* (pp. 313–324). Newcastle, UK: Cambridge Scholars Publishing.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Oxford, UK: Butterworth-Heinemann.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27*(3), 351–372. doi:10.1162/089120101317066113

Xanthos, A. (2004). An incremental implementation of the utterance-boundary approach to speech segmentation. In B. Decadt, V. Hoste, & G. De Pauw (Eds.), *Proceedings of computational linguistics in the Netherlands (CLIN) 2003* (pp. 171–180). Antwerpen: University of Antwerp.

Appendix A: The effect of context size

The size of the context that the predictions are based on affects the performance of the system. The results presented in the earlier sections are based on a combination of phoneme–n-grams of size 1–4. This section provides additional data on the effect of different context sizes. Table A1 presents the token, word, and lexicon *F*-measure and the over- and undersegmentation rates for the model with varying context sizes. The error scores are presented graphically in Fig. A1. Each row on Table A1 includes the indicators from the earlier row, only adding indicators based on one larger context size.

Both oversegmentation and undersegmentation errors decrease until around the context size of 3. If we continue adding higher n-gram contexts, the oversegmentation errors continue decreasing, but undersegmentation errors start increasing. Depending on the error or performance measure, the maximum context size between 3 and 6 seems to produce good results.

Table A1
Performance and error scores for the model using H and H_r with varying context size between 1 and 10. (The model reported in each row includes all the indicators from the previous rows and the n-grams with the size indicated at the context column)

	Boundary			Word			Lexicon			Error	
Context	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	E_o	E_u
1	43.5	75.4	55.2	22.7	34.5	27.4	16.0	27.6	20.3	37.0	24.6
2	60.9	89.9	72.6	46.2	61.8	52.9	27.9	41.5	33.4	21.8	10.1
3	72.6	92.5	81.3	60.7	72.4	66.0	36.4	55.4	43.9	13.2	7.5
4	81.6	89.4	85.4	70.6	75.3	72.9	37.4	65.0	47.5	7.6	10.6
5	85.3	84.4	84.8	73.1	72.5	72.8	33.5	68.0	44.9	5.5	15.6
6	86.8	78.9	82.7	72.9	68.3	70.5	28.8	66.5	40.2	4.5	21.1
7	87.5	74.1	80.2	72.1	64.3	68.0	25.4	64.1	36.4	4.0	25.9
8	87.6	69.9	77.8	70.9	60.7	65.4	23.4	62.6	34.1	3.7	30.1
9	87.5	66.5	75.6	69.8	57.9	63.3	22.1	61.3	32.5	3.6	33.5
10	87.4	64.0	73.9	69.0	55.9	61.8	21.3	60.5	31.5	3.5	36.0

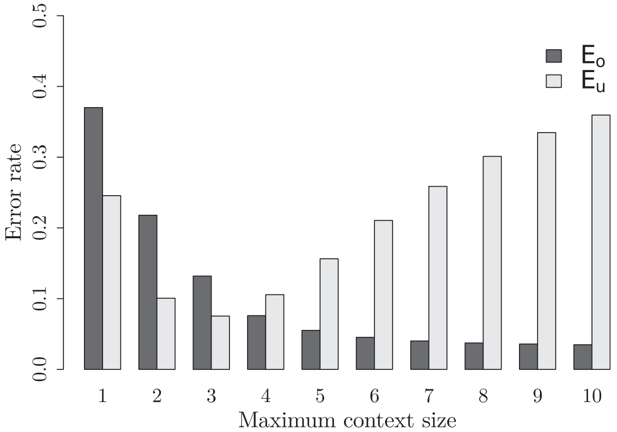

Fig. A1. A graphical display of error scores reported in Table A1 with varying maximum context size on the BR corpus.

The reason for increase in undersegmentation with increased context size has to do with the combination of two factors. First, since larger n-grams are sparse in the data, the indicators using larger context do not have enough number of observations, and they rarely vote for boundary decisions. Second, training method we use for weighted majority voting algorithm takes majority decision as the correct decision from the start. A large number of indicators that do not vote for any segmentation cause the learner not to segment at all, or in milder cases it causes the learner to start segmenting late. The learners that pay attention to larger contexts learn slowly, making more (undersegmentation) mistakes at the beginning. As a result, the discrepancy between the evaluation scores calculated over the whole corpus and the scores calculated at the end of the learning period discussed in Section 3.2 is expected to be even larger for the learners that use longer sequences. The learning curves presented in Fig. A2 demonstrate these differences. Note that phoneme 5-gram context is much more useful than phoneme 3-gram context toward the end of the larger child-directed corpus. It keeps the low undersegmentation rate as shown in the left panel of the Fig. A2, but also drops the oversegmentation rate to the same level as the learner, using up to 3-gram contexts. In fact, it provides a better average error rate even at the end of the BR corpus in comparison to the 4-grams we reported in the main discussion of this paper. Nevertheless, we note that increasing the context size indefinitely is not useful (neither it is cognitively plausible). The learner using up to 7-gram context size cannot catch up with the learner using up to 5-gram context even after about 50,000 utterances. In general, larger context sizes would not always be useful. The statistics related to larger n-gram suffer from data sparseness problem due to the Zipfian-like distribution of sequences found in human languages. Although it is difficult to determine the optimal context size with the data at hand, it is clear that neither very large nor very small contexts are useful.

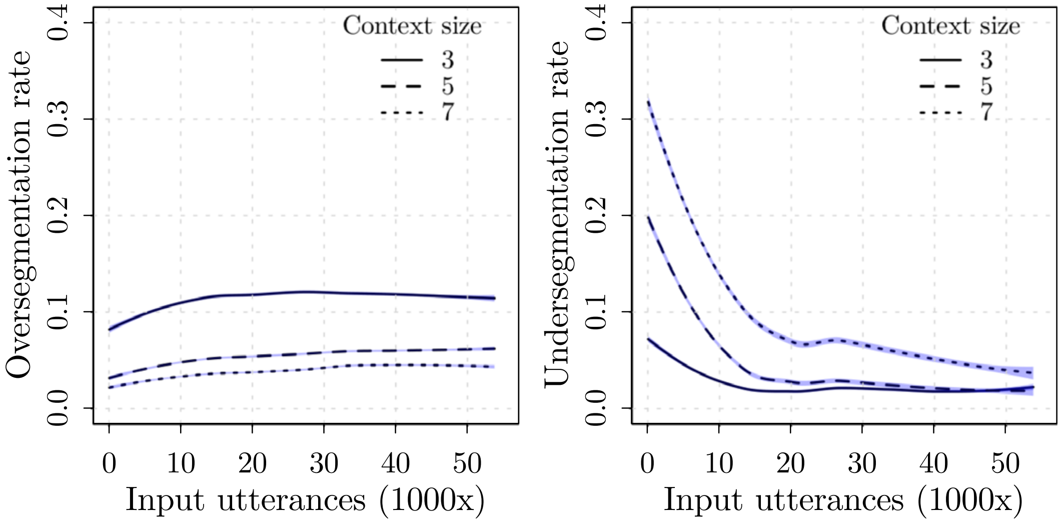

Fig. A2. Learning curves with varying context sizes. The error rates are calculated on larger child-directed speech corpus described in Section 3.1.

Appendix B: Other measures of predictability

In this study, we opted for measuring the (un)predictability using entropy as described in Section 2.2. We also noted that there are other measures that are used in the earlier literature. Three such measures, *transitional probability* (TP), *pointwise mutual information*, (MI) and *successor variety* (SV), are often used in the segmentation literature. In this section, we compare the segmentation performance of the combined entropy model defined in Section 2, with these three measures of (un)predictability. In all definitions, l refers to the left context, a sequence of phonemes preceding the candidate boundary location, and r refers to right context.

- We define *successor variety* (SV) as

$$\mathrm{SV}(l) = \sum_{r \in A} c(l, r)$$

where,

$$c(l, r) = \begin{cases} 1 & \text{if substring } lr \text{ occurs in the corpus} \\ 0 & \text{otherwise} \end{cases}$$

and A is the set of phonemes in the input language (the alphabet). Like entropy, the SV is only a function of l, and it may be complemented by a reverse version, “predecessor variety.” The SV is an unnormalized version of the entropy measure defined in Section 2.2. The entropy measure is sensitive to the probabilities of individual r values (high-probability events result in low entropy), while the SV is not affected by the probabilities of individual r values (all possible values are treated equally).

- *Transitional probability* (TP) is simply the conditional probability of observing r after l.

$$\mathrm{TP}(l,r) = \frac{P(l,r)}{P(l)}$$

where lr is the joint probability observing l and r in this configuration.

- *Pointwise mutual information* (MI) is a well-known measure of association, and defined as follows:

$$\mathrm{MI}(l,r) = \log_2 \frac{P(l,r)}{P(l)P(r)}$$

Table B1 presents the performance and error scores for these predictability measures, along with the boundary entropy that we discuss in the main text. For all measures we only report results using a fixed context size. The performances of the measures may differ based on the context size, as well as other aspects of the learning algorithm (see Çöltekin, 2011, chapter 6, for a more detailed comparison). All measures have similar undersegmentation rates. In this setting, TP and MI seem to make more oversegmentation errors and perform worse than SV and H. Nevertheless, all measures perform well above the baseline, and some perform close to the state-of-the-art models as discussed in Section 3.3.

Table B1
Performance and error scores using different measures of predictability

	Boundary			Word			Lexicon			Error	
Measure	P	R	F	P	R	F	P	R	F	E_o	E_u
SV	84.9	88.5	86.7	74.3	76.5	75.4	38.0	67.0	48.5	5.9	11.5
TP	69.4	88.5	77.8	54.5	65.1	59.3	30.8	54.1	39.3	14.8	11.5
MI	63.4	89.6	74.3	45.5	58.8	51.3	30.0	49.6	37.4	19.5	10.4
H	81.6	89.4	85.4	70.6	75.3	72.9	37.4	65.0	47.5	7.6	10.6
Random	27.9	27.5	27.7	12.9	12.8	12.8	6.1	44.8	10.8	26.8	72.5

[*] For all measures, the context size of up to four phoneme n-grams is used. For asymmetric measures SV, TP, and H, we also included reverse versions in the weighted majority voting. For MI, we altered both left and right context between one and four. The row random presents scores of the baseline segmentation discussed in Section 3.3.