


ARTICLE

Word segmentation from transcriptions of child-directed speech using lexical and sub-lexical cues

Zébulon GORIELY , Andrew CAINES and Paula BUTTERY

Department of Computer Science and Technology, University of Cambridge, Cambridge, UK.

Corresponding author: Zébulon Goriely; Email: zg258@cam.ac.uk

(Received 10 October 2022; revised 26 June 2023; accepted 16 July 2023)

Abstract

We compare two frameworks for the segmentation of words in child-directed speech, PHOCUS and MULTICUE. PHOCUS is driven by lexical recognition, whereas MULTICUE combines sub-lexical properties to make boundary decisions, representing differing views of speech processing. We replicate these frameworks, perform novel benchmarking and confirm that both achieve competitive results. We develop a new framework for segmentation, the DYnamic Programming MULTIPLE-cue framework (DYMULTI), which combines the strengths of PHOCUS and MULTICUE by considering both sub-lexical and lexical cues when making boundary decisions. DYMULTI achieves state-of-the-art results and outperforms PHOCUS and MULTICUE on 15 of 26 languages in a cross-lingual experiment. As a model built on psycholinguistic principles, this validates DYMULTI as a robust model for speech segmentation and a contribution to the understanding of language acquisition.

Keywords: word segmentation; CHILDES; statistical learning

Introduction

Unlike many written languages, where words are separated by spaces, spoken communication is delivered in continuous utterances with only occasional pauses and no clear demarcation of words (Cole & Jakimik, 1980). Yet, adults are usually able to segment speech with no problem, without even realising that there are no such markings. They are assisted in part by more developed lexicons, which they use to identify familiar words in the speech stream. Children, on the other hand, are born with no lexicon to consult, yet by the age of six months they are already capable of segmenting the speech stream into words and phrasal units (Jusczyk, 1999).

The question of how children are able to learn to segment speech and bootstrap their lexicons is the WORD SEGMENTATION PROBLEM. In the 1980s and 1990s there was a renewed interest in examining the statistical properties of language, and in particular how these may impact the understanding of language acquisition and comprehension (Christiansen et al., 1998). Psycholinguistic studies from this time found that children use statistical properties of language to help solve the word segmentation problem. Such

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

properties include lexical stress (Cutler & Carter, 1987; Cutler & Mehler, 1993; Jusczyk, Cutler, et al., 1993), phonotactics (Jusczyk, Friederici, et al., 1993; Mattys et al., 1999; Mattys & Jusczyk, 2001), predictability statistics (Saffran, Aslin, et al., 1996a; Saffran, Newport, et al., 1996; Thiessen & Saffran, 2003), allophonic differences (Jusczyk, Hohne, et al., 1999), coarticulation (Johnson & Jusczyk, 2001), vowel harmony (Suomi et al., 1997) and prosody (Cooper & Paccia-Cooper, 1980; Gleitman et al., 1988).

Interest in the segmentation problem, combined with the evidence provided by these psycholinguistic studies, has led to the design of a variety of computational models for an abstract version of the task. In the established paradigm, utterances are represented symbolically as strings of phones or phonemes without word boundaries, and models have the task of finding these boundaries without supervision. Besides offering insight into the segmentation problem, such models have also developed into successful algorithms for segmenting written text in languages where word boundaries are not marked (Feng et al., 2004; Sproat & Shih, 1990).

In this study, we compare two approaches taken by such computational models; the BOUNDARY-FINDING approach and the LANGUAGE MODELLING approach. The boundary-finding approach considers statistical information present around each inter-phoneme position to make local boundary decisions, often operating phoneme-by-phoneme. Language modelling methods operate utterance-by-utterance, calculating the most-likely segmentation of each, based on lexical recognition. We re-implement the top-performing models for these two approaches, a language modelling approach known as PHOCUS (Blanchard et al., 2010; Venkataraman, 2001) and a boundary-finding model known as MULTICUE (Çöltekin, 2017; Çöltekin & Nerbonne, 2014), both of which achieve similar scores on the child-directed utterances in the English BR corpus, the de-facto standard for evaluating segmentation models (Brent, 1999). Word segmentation models are typically trained on speech corpora directed at children aged less than two years because early indicators of infants' word segmentation abilities are well-attested in the first year (Bergelson & Swingley, 2012; Johnson & Jusczyk, 2001; Saffran, Aslin, et al., 1996a) and most children are regularly producing multi-word utterances by this stage. Therefore, child-directed utterances in the first two years are held to be both crucial and sufficient for children learning to segment words.

Through the comparison of these two approaches, we observe that the boundary-finding methods can combine information from multiple sub-lexical cues, but cannot make decisions based on the placement of other boundaries. We also find that the language modelling methods can make decisions based on the placement of other boundaries, but cannot combine information from multiple sub-lexical cues.

The aim of our research is to investigate whether giving a statistical model access to both lexical and sub-lexical cues improves its ability to segment words across a range of languages. By considering cues that are accessible to children and allowing the model to utilize whichever cues it deems valuable, an increased ability to segment words would suggest that both types of cue provide complementary information useful for word segmentation. This finding would contribute to our understanding of language acquisition by highlighting the extensive knowledge that statistical methods can acquire from the linguistic signal alone, leading to further inquiry into the additional linguistic knowledge that can be learned jointly with or subsequent to word segmentation.

In this study, we develop the DYNAMIC programming MULTIPLE-cue (DYMULTI) framework for modelling word segmentation. This framework combines the strengths of both the boundary-finding and language modelling approaches and allows for the consideration of sub-lexical and lexical cues, achieving higher F_1 -scores on the BR corpus

than any previous model that uses the same constraints. We also undertake novel cross-lingual evaluation of these models, finding that our model outperforms PHOCUS and MULTICUE on 15 of 26 languages, confirming its validity as a computational model for infant word segmentation. In doing so, we also find that there may be previous research bias towards performance on English corpora. The contributions of our paper are as follows:

- We give a thorough review of the word segmentation problem and the previous psycholinguistic and computational modelling studies that have investigated it. We introduce the DYMULTI framework for segmentation, which achieves the highest F_1 -scores to date.
- We perform a thorough and robust benchmarking of segmentation models, comparing the PHOCUS, MULTICUE and DYMULTI frameworks. This includes an investigation into the effect of utterance order, a comparison of learning rates across models, and cross-lingual evaluation across 26 languages.
- We release our implementations of PHOCUS, MULTICUE and DYMULTI as an open-sourced repository for reproducibility and future research¹.

Background

In this section, we give the psycholinguistic background to the word segmentation problem. We then discuss the computational models that have been designed to explore it, detailing the boundary-finding and language modelling approaches.

Cues for segmentation

Despite the lack of consistent acoustic gaps between spoken words, adults are able to segment the speech stream into linguistically significant units and therefore access their meaning, a process called segmentation. Early models of speech processing declared segmentation to be a by-product of lexical identification (Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978), later described as “serendipitous” or “interactionist” segmentation models (Cairns et al., 1997; Cutler, 1996). These models identify words in the speech stream by matching them against the listener’s lexicon, either processing the utterance in a strictly temporal order, as in the COHORT model of Marslen-Wilson and Welsh (1978) or by using the activation of competing lexical items to cut up the input, as in the TRACE model of McClelland and Elman (1986). These models can make use of sub-lexical cues, such as adults’ sensitivity to phonotactic information, to make judgements about possible words (Greenberg & Jenkins, 1966), but are primarily driven by the lexicon.

Another view of speech processing is that segmentation occurs purely on the basis of information in the speech signal without making use of any lexical influences. Cutler (1996) calls these “explicit” segmentation models and multiple studies have found that adults can segment using purely using low-level information. Saffran, Newport, et al. (1996), for example, found that within 20 minutes of exposure to an artificial language, adults are able to use phonotactic information to tell non-words apart from words in a speech stream. Such studies do not refute interactionist accounts, as these can still incorporate low-level information, but they do provide evidence that adult segmentation is not fully driven by lexical recognition.

¹<https://github.com/codebyzeb/DYMULTI-23>.

When it comes to infants, there is evidence that lexical recognition is used to solve the segmentation problem, supporting the interactionist view. Six-month-olds learn new words from utterances containing familiar names (Bortfeld et al., 2005). French eight-month-olds use function words such as *des* and *mes* for segmentation (Shi & Lepage, 2008) and infants at this stage can even make semantic associations with nouns (Bergelson & Swingley, 2012). It is clear that infants are able to recall familiar sound patterns and use them weeks later for segmentation (Jusczyk & Hohne, 1997).

The problem with a model of speech segmentation that only considers lexical recognition lies in explaining how these familiar words are acquired in the first place; infants cannot have any innate assumptions about rhythmic and phonological regularities as these vary between languages (Cutler & Carter, 1987). One hypothesis is that these proto-lexicons are initially populated with single words spoken in isolation (Suomi, 1993). Indeed, in English Parentese (the particular register and style used by caregivers when talking to children), about one-tenth of utterances consist of isolated words (Brent & Siskind, 2001). The issue with this hypothesis is that there is no universal heuristic for identifying single-word utterances and many words will never occur in isolation. Brent and Siskind (2001) claim that if entire multisyllabic utterances are initially added to the lexicon, lexical recognition alone could be sufficient for bootstrapping the lexicon. This claim is supported by a more recent study that found that the proto-lexicon of eleven-month-old French-learning infants contains both words and non-words (Ngon et al., 2013).

On the other hand, there is substantial empirical evidence to suggest that infants use a wide variety of SUB-LEXICAL cues to solve the initial segmentation problem. Many of these are based on the simple principle that predictability within lexical units is high, and predictability between lexical units is low (Harris, 1955). It did not become clear that infants are able to use this principle for segmentation until the influential studies of Saffran, Aslin, et al. (1996a, 1996b) and Saffran, Newport, et al. (1996). Following their study in adults, they found that infants as young as eight months calculate the TRANSITIONAL CONDITIONAL PROBABILITIES of adjacent syllables A and B, defined as

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

where $\Pr(AB)$ is the estimated probability of the syllable pair (calculated as the relative frequency) and $\Pr(A)$ is the estimated probability of the syllable *A*, and use these to place word boundaries when the transitional probability is low (Aslin et al., 1998; Saffran, Aslin, et al., 1996a, 1996b).

These probabilities are also gathered at lower levels. At the phoneme level, for instance, differences in probabilities between within-word and across-word consonant clusters are used to segment novel phrases such as *fang tine*, as the pair of phones [ŋt] does not occur within English words (Mattys & Jusczyk, 2001). At the lowest level, seven-and-a-half-month-old infants use their knowledge of allophonic variations to segment utterances, such as the variants of /t/ and /r/ that distinguish *nitrate* and *night rate* (Jusczyk, Hohne, et al., 1999).

Infants also seem to be sensitive to prosodic cues, those as young as 7.5 months learn to use the predictable strong-weak stress pattern in English (as in *BAbby*) for segmentation (Cutler & Mehler, 1993; Jusczyk, Cutler, et al., 1993; Jusczyk, Houston, et al., 1999). While statistical cues may precede stress cues in their use (Thiessen & Saffran, 2003), stress and

coarticulation cues are weighed more heavily by infants once adopted (Johnson & Jusczyk, 2001). Stress alone is unlikely to be a universal cue for segmentation, as it is unclear whether all languages even provide reliable prosodic cues (Saffran, Newport, et al., 1996). Indeed, it has generally been accepted that no single cue is solely responsible for solving the segmentation problem and that a complete model for explicit segmentation must consider information from multiple cues (Blanchard et al., 2010; Christiansen et al., 1998; Çöltekin & Nerbonne, 2014; Jusczyk, 1999).

Taking these accounts together, it is unclear whether initial segmentation in infants is purely explicit, or whether a combination of lexical and sub-lexical information is used. There are many overlapping and competing cues in these studies, so it is difficult to justify one view over the other. For example, segmentation around familiar words could be a result of phonotactic regularity rather than lexical recognition. This motivates the development of computational models in order to test hypotheses in isolation and therefore also solve the word segmentation problem. In particular, the DYMULTI framework developed in this study lets us test whether sub-lexical and lexical cues are alternative or complementary explanations for speech segmentation.

Segmentation models

Computational models for studying the segmentation problem are often designed to study one of two questions: (a) how statistical information can be used to segment speech, and (b) what computational problem is being solved.

These are often discussed using terminology from Marr's computational theory of vision (Marr, 1982): the first question operates at Marr's ALGORITHMIC LEVEL, focusing on the algorithm, and the second operates at Marr's COMPUTATIONAL LEVEL, focusing on the problem being solved.

Algorithmic-level studies are concerned with the implementation of algorithms that incorporate cognitively plausible mechanisms for the segmentation problem. These models propose efficient algorithms that follow three constraints:

1. They must start with no knowledge of the target language.
2. They must learn unsupervised.
3. They must operate incrementally.

The first constraint follows from the fact that all languages have different phonotactic constraints and vocabularies, yet children can learn any of them. The second is established because children are not always explicitly given the boundaries between words, so neither should computational models. The third follows from the fact that we process speech as it is heard, not in batches sometime later.

Numerous models have been proposed based on these constraints, taking a wide variety of approaches. Two broad categories stand out: boundary-finding methods and language modelling methods. These are somewhat related to interactionist and explicit views of speech processing, although top-performing models make use of both lexical and sub-lexical cues. Investigating these two approaches is the focus of this study. Note, however, that these models tend to operate on phonemic transcripts and so do assume some knowledge of the target language, not quite meeting the first constraint. However, this is still preferable to using orthographic transcriptions, as the phonemic forms are still representations of the sound signal.

By contrast, computational-level studies are concerned with defining the goal of segmentation and the logic of the strategy used to meet that goal. As the focus of these studies is not the algorithm, the models developed need not meet the three constraints. An example is the probabilistic model of Goldwater et al. (2009), who find that the assumption that words are statistically independent units leads to under-segmentation by an ideal learner. As a computational-level study, their algorithm does not have to be cognitively plausible. It operates in batches over the corpus, using a hierarchical Dirichlet Process bi-gram model estimated using a Gibbs sampling algorithm. Besides the batch processing violating the third constraint for an algorithmic-level model, the computation time is also over 2000 times longer than most algorithmic-level models when presented with the same amount of data (Fleck, 2008). In this study, we do not work directly on these computational-level models, although many algorithmic-level models often offer insight at the computational level.

Boundary-finding methods for segmentation

Boundary-finding methods for segmentation relate to the explicit view of speech processing, that segmentation is driven by local information at each inter-phoneme position rather than lexical recognition. Models that use these methods follow directly from experimental studies. For example, Saksida et al. (2017) follow the findings of Saffran, Aslin, et al. (1996a), showing that children segment utterances at low points of transitional probability. Their unsupervised algorithm places boundaries between a syllable pair when the transitional probability of the syllable pair is lower than the two neighbouring pairs. Using syllables as the basic unit of segmentation is widely debated (Coltekin, 2011) and also has a high-performing baseline since the vast majority of child-directed English words are monosyllabic (Gambell & Yang, 2006).

Earlier studies made use of connectionist models for infant segmentation, as was the trend for investigating many cognitive phenomena at the time (Christiansen et al., 1998; Elman, 1990). As cognitively-plausible models for segmentation need to be unsupervised, these models could not be trained to predict word boundaries directly. Instead, they were often trained on an alternative task. Elman (1990) trained a recurrent neural network to predict phonemes, finding that relatively high error in prediction could indicate word boundaries. Cairns et al. (1994) found that peaks in the error score could also indicate word boundaries. Finally, Christiansen et al. (1998) developed a recurrent neural network to predict utterance boundaries, phonemes and lexical stress information in an utterance, finding that the prediction of an utterance boundary was a good indicator of a word boundary. This model allowed them to test these different cues together and in isolation, finding that the best performance was achieved when all three cues were combined.

Inspired by this model, Çöltekin and Nerbonne (2014) developed an explicit model for segmentation, arguing that it is difficult to interpret what connectionist models learn. Their model uses statistical information at each inter-phoneme position, as with transitional probability models, but extends this by introducing a cue-combination method to combine statistical information from multiple sources, also achieving far better performance than the connectionist models. This is the boundary-finding approach that we re-implement in this study.

Language modelling methods for segmentation

Language modelling methods are based on the interactionist view of speech processing, that segmentation and lexical recognition occur serendipitously, driven by lexical

knowledge. These models typically build word n -gram models and use statistical criteria to define the best segmentation of an utterance, bootstrapping a lexicon that is then used for further segmentation.

Brent (1999) and Venkataraman (2001) both developed probabilistic language models and used dynamic programming to infer the best segmentation. In Venkataraman's model, the probability of a segmented utterance is given as the joint probability of all words in that utterance. The Viterbi algorithm is then used to find the segmentation that maximises an estimate of this probability. Word probabilities are approximated by n -grams, with a back-off procedure to lower-order n -grams. As the model processes more utterances, these probability estimates are refined and more words are added to the lexicon, further improving the model. These models produced state-of-the-art results unmatched by boundary-finding methods until later work (Çöltekin, 2017; Çöltekin & Nerbonne, 2014; Fleck, 2008).

It is worth noting that none of the successful language modelling methods rely only on lexical recognition. A model that only matches utterances with previously-seen utterances will fail, as “short, frequently occurring utterances are likely to be segmented within larger word-level chunks resulting in an over-segmentation of words into their segmental phonology” (Monaghan & Christiansen, 2010). For instance, if *no* has been added to the lexicon, then *note* could later be segmented as *no* and *te*, followed by increasingly smaller segmentations. As such, these models often gather sub-lexical statistics or make sub-lexical assumptions to prevent over-segmentation. The PUDDLE model, for example, uses word-initial and word-final phoneme clusters derived from its lexicon to restrict segmentation (Monaghan & Christiansen, 2010). The model of Venkataraman (2001) incorporates phoneme-level statistics to estimate the probability of unseen words. Finally, the model of Blanchard et al. (2010) extends Venkataraman's model with an additional constraint that all segmented words must contain a syllabic nucleus. It is these latter two models that we re-implement and compare in this study.

Segmenting from raw speech

In this article, we focus on the abstract version of the word segmentation task, with utterances consisting of sequences of discrete, symbolic phonemes. Although this has remained an established paradigm for the study of word segmentation, in recent years the speech research community has made great advances in the area of ZERO-RESOURCE SPEECH PROCESSING. These studies aim to develop unsupervised methods that learn from raw speech audio only, pioneered in recent years by the Zero Resource Speech Challenge (ZRC) series (Dunbar et al., 2022).

One of the four tasks presented by the ZRC series is Spoken Term Discovery, the text-less counterpart to word segmentation. The general approach proposed by ZRC is to first match speech fragments consisting of the same sequence of phonemes (the matching sub-task), then build a lexicon of word types (the lexicon discovery sub-task) and finally then use these to find word boundaries (the word segmentation sub-task). “Match-first” systems focus first on the matching problem, placing boundaries at the edges of discovered segments (Räsänen & Blandón, 2020). “Segmentation-first” systems prioritise the discovery of boundaries – for instance, by jointly optimising segmentation and building clustered word embeddings using Bayesian modelling (Kamper et al., 2017) or by using self-expressive autoencoders to build a segmentation from matched learned acoustic units (Bhati et al., 2020). The most recent approaches do not even attempt to build a lexicon of types, either using a Bayesian approach directly on learned tokens (Algayres et al., 2022)

or by using peaks in surprisal across sequences of learned units (Kamper, 2023), similarly to traditional text-based boundary-finding methods for segmentation.

In this study, we were motivated by the availability of phonemic transcripts for 26 languages provided by Caines et al. (2019) to carry out novel cross-lingual analysis of word segmentation models. As the original audio recordings have not been made available for the majority of these transcriptions, we were unable to consider the models developed for ZRC series. However, many of the models presented in this study could operate on raw audio; their only requirement is that the units are discrete, so the input could easily be replaced with features derived from speech frame units. We discuss these ideas further at the end of the article.

Summary

Experimental psycholinguistic studies provide evidence that infants use sub-lexical statistical and speech cues for solving the segmentation problem, supporting the explicit view of speech processing. Other studies find that infants make use of lexical knowledge, supporting the interactionist view. To study the problem in a controlled manner, computational models have been designed to solve an abstract version of the problem, where continuous speech is represented as a series of symbolic phonemes. These models either explore what problem is being solved or present cognitively plausible algorithms for solving the problem. To be cognitively plausible, these algorithms must segment incrementally, start with no knowledge of the target language and learn unsupervised. Boundary-finding algorithms correspond to the explicit view of speech processing; and language modelling algorithms correspond to the interactionist view.

Implementation of Segmentation Models

In this section, we present our re-implementation of the state-of-the-art models for the boundary-finding approach of Çöltekin and Nerbonne (2014) and the language modelling approach of Venkataraman (2001) and its extension presented by Blanchard et al. (2010). We discuss the benefits and drawbacks of these two approaches and produce a new model that combines their strengths.

Çöltekin and Nerbonne's multiple-cue boundary-finding model

The model presented by Çöltekin & Nerbonne (2014) iterates through utterances phoneme-by-phoneme, placing boundaries by combining votes from a set of indicators based on a variety of cues. It is explicit in nature, although it does use statistical cues derived from the lexicon. We refer to this model as MULTICUE.

Cue combination algorithm

The core strength of MULTICUE lies in its cue combination algorithm, which allows for the consideration of an arbitrary number of psychologically-motivated boundary indicators. All of the cues are language-independent and the task of the algorithm is to determine how to use them without any supervision. As no single cue is solely responsible for the placement of word boundaries, this allows for a more comprehensive model for explicit segmentation.

Each boundary indicator labels every inter-phoneme position as either ‘boundary’ or ‘word internal’. The model then makes a final decision based on a variation of the weighted majority voting algorithm (Littlestone & Warmuth, 1994). In Çöltekin (2017), the following condition for deciding on the ‘boundary’ label is given:

$$\sum_i^K w_i 1_i > \frac{K}{2},$$

where K is the number of boundary indicators, w_i is the weight and 1_i gives the boundary decision for indicator i , equal to 1 for ‘boundary’ and 0 for ‘word-internal’. This is a conservative threshold, relying heavily on informed indicators. It also requires that the weights are all over 0.5 on average (otherwise the model would never be able to place a boundary). For instance, if all K boundary indicators were assigned a weight of 0.5, the model would never be able to place a boundary. This is because even if all indicators voted for a boundary, the weighted majority vote would only equal $\sum_i^K 0.5 = \frac{K}{2}$.

Instead, our implementation places a boundary if the weighted vote for the ‘boundary’ label is greater than the weighted vote for the ‘word-internal’ label:

$$\sum_i^K w_i 1_i > \sum_i^K w_i (1 - 1_i). \quad (1)$$

By noticing that the two sides sum to $\sum_i^K w_i$, and so the boundary can simply be placed if the normalised weighted vote exceeds 0.5, this equation can also be rewritten as:

$$\frac{\sum_i^K w_i 1_i}{\sum_i^K w_i} > \frac{1}{2}.$$

We discussed this with Çöltekin and he agreed that this boundary decision formulation is better, representing a more general case where we make no assumptions about the weights. He also noted that this makes the model more robust to bad indicators, but may favour recall over precision — recall indicating the retrieval of true boundaries and precision indicating how accurate a model’s predicted boundaries are.

The majority-vote algorithm is a common and effective method for combining multiple classifiers (Narasimhamurthy, 2005). In this case, the WEIGHTED majority-vote is used so that votes from boundary indicators that make fewer errors have larger weights. As the model must be unsupervised, the ground-truth boundary locations cannot be used to update the weights. Instead, an error happens when an individual cue disagrees with the majority vote. At each inter-phoneme position, the incremental algorithm gathers votes from each indicator i , decides whether the position is a ‘boundary’ or is ‘word-internal’, and then increments the error count e_i for each indicator that disagreed with this decision. Finally, the weight w_i of each indicator is updated:

$$w_i = 1 - 2 \frac{e_i}{N},$$

where N is the total number of inter-phoneme positions seen, producing weights in $[-1, 1]$.

Çöltekin & Nerbonne (2014) state that this update rule “sets the weight of a vote that is half the time wrong to zero, eliminating incompetent voters” and that with this model, the success of boundary decisions depends on the precision of individual boundary indicators. In reality, this score is related to the ACCURACY of these indicators. As there are fewer true ‘boundary’ labels than ‘word-internal’ labels, an indicator that never places a boundary will achieve higher accuracy than an indicator that always places a boundary, so setting weights to zero when the accuracy is 0.5 is misleading. In our implementation, we use the following:

$$w_i = 1 - \frac{e_i}{N}. \quad (2)$$

These weights are in the range $[0, 1]$ and are exactly equal to the accuracies of each indicator with respect to the final votes.

Cues and boundary indicators

Using the majority-voting framework, any number of indicators can be considered. describes a series of indicators derived from four sets of cues; predictability statistics, utterance boundaries, lexical stress and the lexicon, all deriving from psycholinguistic studies.

All of the indicators calculate a certain measure based on these cues. To propose boundaries, they use a PARTIAL-PEAK strategy. This is based on the peak strategy of transitional probability models where a boundary would be suggested if the transitional probability at an inter-phoneme position was lower than the transitional probabilities on either side of that boundary. Each cue is split into two indicators, splitting this peak in half. An example is given in Figure 1, where the first indicator proposes a boundary after a decrease in transitional probability and the other proposes a boundary before an increase in transitional probability. The model can then learn weights associated with each indicator, using the weighted majority-vote algorithm.

Çöltekin & Nerbonne (2014) also include indicators that calculate statistics over a larger context of three phonemes, capturing higher-order regularities, as well as indicators that calculated reverse measures, following the study of Pelucchi et al. (2009) that found that children can also use REVERSE transitional probabilities for segmentation. Çöltekin, (2017) later used MULTICUE to explore various predictability cues in isolation. He found that the best performance was achieved when including indicators with a context size of one, two, three and four phonemes and also found that SUCCESSOR VARIETY was a better

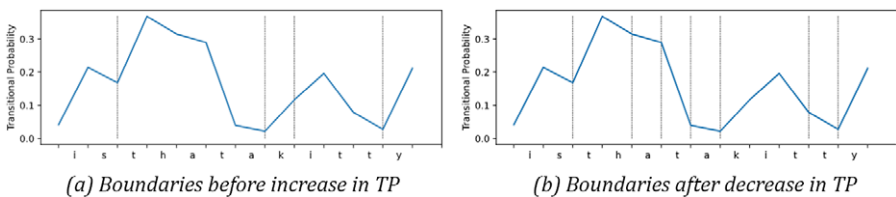


Figure 1. Two indicators following the partial-peak strategy, segmenting the phrase is that a kitty. One segments at an increase in transitional probability and the other segments at a decrease. Letters are used instead of phonemes for clarity.

predictability measure than transitional probability. Successor variety is the predictability measure originally described by Harris (1955) as being a good measure for predicting morpheme boundaries. For a set A of phonemes in the input language, the successor variety for a substring of phonemes l is given by:

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

where

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

We reimplemented all cues used by Çöltekin & Nerbonne (2014) as well as the predictability cues used by Çöltekin (2017). Using a simple command-line interface, we re-implemented their models, which we henceforth refer to as MULTICUE-14 and MULTICUE-17 respectively. These models only vary in which indicators are included; MULTICUE-14 has 44 stress, predictability, lexicon and utterance-boundary indicators and MULTICUE-17 has 16 predictability indicators. For our reported results of MULTICUE-17, we use the successor variety predictability cue, as this was the measure that Çöltekin (2017) found to give the best performance. See Çöltekin & Nerbonne (2014) and Çöltekin (2017) for detailed descriptions of these cues.

An updated set of cues

Based on the success of using a variety of cues (Çöltekin & Nerbonne, 2014) and the success of using higher-order n -grams and the successor variety cue (Çöltekin, 2017), we propose a new set of indicators that combine these ideas. This set consists of the successor variety cue of MULTICUE-17 and the lexicon and utterance boundary cues of MULTICUE-14. Indicators are created for n -gram values from 1 to 4. The stress cue is not included, following the finding of Çöltekin & Nerbonne (2014) that it decreases performance and also because the cross-lingual corpora that we use for evaluation do not provide stress alignment information.

We refer to the MULTICUE model using this new set of cues as MULTICUE-23. A summary of MULTICUE-14, MULTICUE-17 and the new MULTICUE-23 model is given in Table 1.

Venkataraman's language modelling algorithm

Venkataraman's model follows a language-modelling approach to segmentation. As a lexicon is developed and phonemic distributional statistics are learned, utterances are decoded using the Viterbi algorithm to find the maximum-likelihood segmentation. This is an interactionist approach as it is driven by lexical recognition rather than boundary placement.

Language model

A standard language model is used to calculate the likelihood of a segmentation. Given a segmentation $\mathbf{W} = w_1, \dots, w_n$ composed of n individual words $w_i \in \mathbf{L}$ in a lexicon \mathbf{L} , the most likely segmentation \hat{W} is

Table 1. Summary of Models Implemented in this Study

Model	Sub-Lexical Cues	Lexical Cues	Segmentation Algorithm
MULTICUE-14	Predictability, Stress, Lexicon Boundaries, Utterance Boundaries	None	Weighted Majority Voting
MULTICUE-17	Predictability	None	Weighted Majority Voting
MULTICUE-23	Predictability, Stress, Lexicon Boundaries, Utterance Boundaries	None	Weighted Majority Voting
PHOCUS-1	None	1-gram Language Model	Viterbi Decoding
PHOCUS-1S	None	1-gram Language Model, Syllabic Nucleus Constraint	Viterbi Decoding
DYMULTI-14	Predictability, Stress, Lexicon Boundaries, Utterance Boundaries	Syllabic Nucleus Constraint, Lexical Recognition ($\alpha > 0$)	Viterbi Decoding with Weighted Majority Voting
DYMULTI-17	Predictability	Syllabic Nucleus Constraint, Lexical Recognition ($\alpha > 0$)	Viterbi Decoding with Weighted Majority Voting
DYMULTI-23	Predictability, Stress, Lexicon Boundaries, Utterance Boundaries	Syllabic Nucleus Constraint, Lexical Recognition ($\alpha > 0$)	Viterbi Decoding with Weighted Majority Voting

$$\begin{aligned}\widehat{\mathbf{W}} &= \arg \max_{\mathbf{W}} P(\mathbf{W}), \\ &= \arg \max_{\mathbf{W}} \prod_{i=1}^n P(w_i | w_1, \dots, w_{i-1}).\end{aligned}$$

To prevent underflow errors in computation, an equivalent calculation is made using log-likelihoods:

$$\widehat{\mathbf{W}} = \arg \min_{\mathbf{W}} \sum_{i=1}^n -\log P(w_i | w_1, \dots, w_{i-1}).$$

A common approximation when implementing language models is the n -gram approximation, collapsing these conditional probabilities to consider at most $n-1$ words. Venkataraman (2001) makes a three-gram approximation, estimating $P(w_i | w_{i-2}, w_{i-1})$ with relative frequencies and using a back-off procedure to estimate the probability of unseen n -grams with lower order n -grams (Katz, 1987). He uses a back-off technique given by Witten & Bell (1991), using phoneme 1-grams to estimate unseen words.

Venkataraman (2001) implemented 1-gram, 2-gram and 3-gram models, finding a trade-off between precision and recall, with 1-grams giving the best performance overall. This is surprising, as n -gram contexts typically improve performance in such systems. Venkataraman claimed that this is because the 2-gram and 3-gram models are more conservative, as longer n -grams are more infrequent, leading to whole utterances being often inserted into their lexicons. As such, we only implement the 1-gram model.

Viterbi search

The language model defines the likelihood of a segmentation, but a search procedure is required to find the most likely segmentation. Exhaustive search is computationally intractable as there are 2^{n-2} possible segmentations for an utterance of n phonemes, so this would be an unreasonable model for human segmentation. Instead, Venkataraman uses Viterbi search (Viterbi, 1967) to decode each utterance, a dynamic programming algorithm that only explores $(n-2)^2$ segmentations.

The algorithm begins with an empty lexicon and no knowledge of phoneme frequencies, building these incrementally as each utterance is processed. The process is unsupervised, as no word boundaries are ever provided to the model. As such, all three constraints for algorithmic-level segmentation are satisfied.

Blanchard's extended algorithm

Blanchard et al. (2010) extend Venkataraman's 1-gram model to produce PHOCUS, for PHonotactic CUe Segmenter. They introduce two phonotactic cues: language-specific and language-universal. The first extends the unseen word estimate to use conditional probabilities of phoneme n -grams, rather than the phoneme 1-grams of Venkataraman. This cue is language-specific as phonotactic constraints (permissible phoneme combinations) vary between languages, so the phoneme n -gram probabilities must be learned. The models that keep track of phoneme n -grams are referred to as PHOCUS- n , with PHOCUS-1 being equivalent to Venkataraman's model. For simplicity, we do not consider higher-order phoneme n -grams here.

The second cue is the universal constraint that words must have at least one SYLLABIC NUCLEUS. Syllabic nuclei in English consist of all vowels and some consonant sounds (such as the [l,m,r] sounds in *awful* [ɔfl], *rhythm* [ɹɪðm], *butter* [bʌtɹ] and *even* [ivn]). There is much debate about the validity of syllables as a perceptual unit (Mehler et al., 1981; Räsänen et al., 2018; Ziegler & Goswami, 2005), but Blanchard et al. claim that this constraint is plausibly a prior that does not need to be learned, as it can be explained without making assumptions about the perceptual status of syllables: instead, this assumption only depends on sonority (for vowels) or manner of articulation (nasals, liquids) and the fact that every word requires at least one of these. To implement this constraint, probabilities of words that do not have a syllabic nucleus are set to 0. Adding this constraint to PHOCUS-1 gives PHOCUS-1S.

Full algorithm

PHOCUS-1S iteratively processes each utterance using the Viterbi algorithm to find the segmentation that maximises the product of estimated word probabilities. After segmenting each utterance, the proto-lexicon and phoneme counts are updated, improving the language model.

An example of this loop is given in Figure 2, where possible segmentations of the utterance *andadoggy* are considered. For the first possible segmentation, *and adoggy*, the probability of the word *and* is given by its relative frequency in the proto-lexicon. The word *adoggy* has not been seen before, so its probability is calculated using the relative frequencies of each of its symbols (graphemes in this example, phonemes in our experiments). For the second possible segmentation given in the example, *andado gg y*, the word *gg* contains no syllabic nucleus, so has a probability of 0, resulting in a probability of 0 for the whole utterance. Assuming 0.02 is the highest score out of all segmentations considered by the Viterbi algorithm, *and adoggy* would be selected as the best segmentation for this utterance.

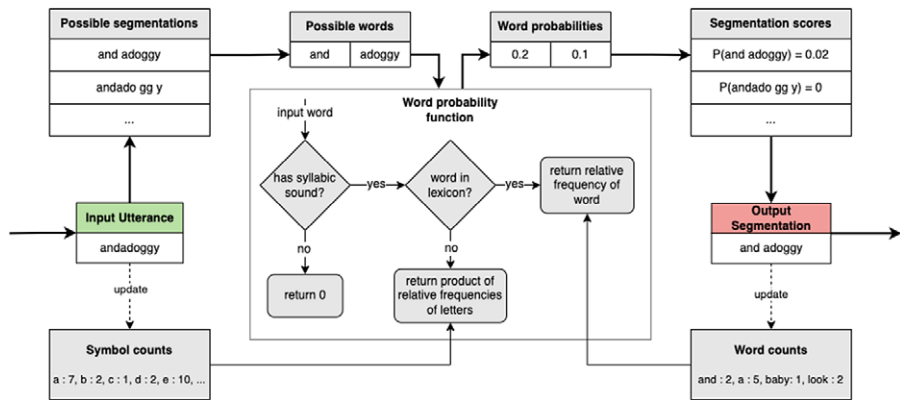


Figure 2. An example of PHOCUS-1S segmenting the utterance *andadoggy*, using letters as our discrete symbolic unit instead of phonemes for clarity.

Model summary

The PHOCUS-1 model of Venkataraman (2001), extended by Blanchard et al. (2010) to produce PHOCUS-1S, uses a language model to define the probability of a segmentation based on seen-word frequency and phoneme frequency for unseen words. In PHOCUS-1S, probabilities of words not containing a syllabic nucleus are set to 0. A Viterbi algorithm finds the most-likely segmentation of an utterance. We implemented both PHOCUS-1 and PHOCUS-1S for comparison with the MULTICUE models. A summary of both models is given in Table 1.

DYMULTI: A combined segmentation model

MULTICUE is in principle a boundary-finding model for segmentation, but MULTICUE-14 does use indicators based on the lexicon, so it could be considered an interactionist model. PHOCUS is also an interactionist model, using a language model for calculating the probability of segmenting an utterance, but it does use sub-lexical information for estimating the probability of unseen words. Therefore, both models involve a complicated interaction of lexical and sub-lexical information, which is consistent with studies showing that infants use both sources of information for solving the segmentation problem. There are, however, drawbacks to both approaches.

One of the key benefits of MULTICUE is that it can combine an arbitrary number of sub-lexical boundary indicators. This is a good model for explicit segmentation, as experimental studies have shown that infants are sensitive to a wide variety of cues. PHOCUS, on the other hand, cannot consider an arbitrary number of sub-lexical indicators. This is not just a drawback of PHOCUS, but of any language modelling approach to segmentation. To add a new indicator, the entire language model would need to be redefined and this would be very difficult to do without making prior assumptions about the cues.

The strength of PHOCUS lies in the Viterbi search process. The segmentation of an utterance is decided at the lexical level, based on the scores assigned to each word in the segmentation. This means that it is easy to incorporate lexical-level constraints, such as the syllabic nucleus constraint of Blanchard et al. (2010). Such a constraint cannot be

easily incorporated into the MULTICUE model, or indeed into any boundary-finding approach to segmentation, as boundary-finding models place boundaries independently of each other using only the local context around that boundary. Hence, the decision cannot depend on the placement of previous or future boundaries.

In this section, we present a new framework for segmentation models that combines the two approaches. It collects scores for each inter-phoneme position from multiple indicators using the weighted majority-vote algorithm of MULTICUE, then uses a modification of the Viterbi algorithm from PHOCUS to choose the best segmentation, rather than just placing boundaries greedily. This combined model allows for the consideration of multiple sub-lexical and lexical cues, addressing the drawbacks of both the boundary-finding and language-modelling approaches to segmentation. We name this framework DYMULTI for DYNAMIC programming MULTIPLE-cue model.

Using weighted boundary votes with the Viterbi algorithm

In DYMULTI, the Viterbi algorithm finds the best segmentation according to boundary scores rather than word scores. These boundary scores are adapted from the weighted majority-vote algorithm of the MULTICUE model, adjusting equation (1) to give a real-valued score instead of a binary decision:

$$\text{score}(j) = \frac{\sum_i^k w_i 1_{ij}}{\sum_i^k w_i} - \frac{\sum_i^k w_i (1 - 1_{ij})}{\sum_i^k w_i}, \quad (3)$$

where w_i are the weights for indicator i , as given by equation (2). These scores lie between -1 and 1 with scores over 0 indicating a boundary and scores close to 1 or -1 suggesting strong agreement between indicators.

The function returns the score at position j in the utterance where 1_{ij} is the vote of indicator i at this inter-phoneme position.

We then adapt the Viterbi algorithm to maximise the sum of these boundary scores, rather than minimise the sum of negative log word probabilities. The word score function now simply returns $\text{score}(j)$, the score given by the weighted majority-vote algorithm between phoneme $j - 1$ and j . At the utterance boundaries, $\text{score}(j)$ always returns 1.

Without any other changes, this algorithm simply places boundaries at every position where the score is greater than 0, as this maximises the sum over the utterance. As scores over 0 indicate where MULTICUE would have placed a boundary according to equation (2), this means that DYMULTI will act exactly like MULTICUE if the same indicators are provided. The difference with this new framework is that lexical-level processes can be introduced by adjusting the word score function, as described in the next two sections.

Introducing the require-syllabic-sound lexical constraint

The first lexical-level process we introduce to DYMULTI is the syllabic nucleus constraint of Blanchard et al. (2010). We adjust the word score function so that if the word has no syllabic nucleus, the function returns -100. This number is chosen to be far smaller than any positive sum could account for, similarly to the large negative log probability used to simulate a probability of 0 in our implementation of PHOCUS-1S.

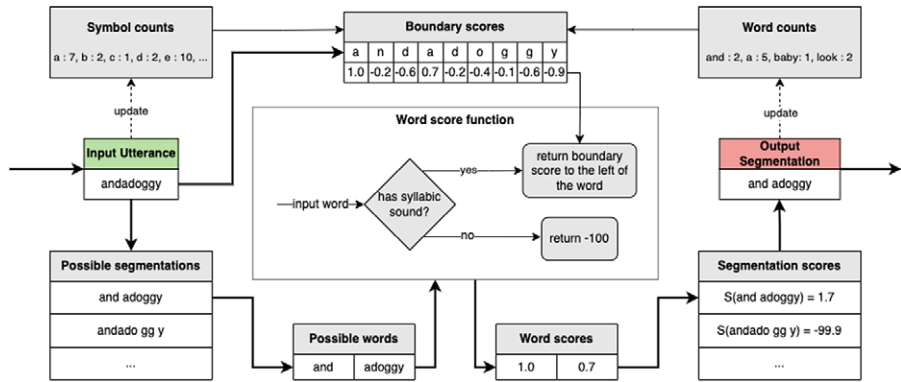


Figure 3. An example of DYMULTI segmenting the utterance *andadoggy*, using letters as our discrete symbolic unit instead of phonemes for clarity.

An example of the DYMULTI framework segmenting the utterance *andadoggy* is given in Figure 3. First, the boundary scores for the utterance are calculated using equation (3). These scores are then used to calculate the score of each segmentation, using the Viterbi algorithm. For the segmentation *and adoggy*, the boundary scores to the left of the two words are 1.0 and 0.7, so the score for the segmentation is 1.7. As with PHOCUS-1S, the syllabic nucleus constraint prevents *andado gg y* from being a valid segmentation, giving a score of -100 to the word *gg*.

Introducing a lexical recognition model

Using the Viterbi algorithm, other lexical processes can also be introduced to DYMULTI. Here, we propose one such process: a rudimentary lexical recognition process to favour previously-seen words. This mirrors the lexical recognition driving many of the language modelling methods for segmentation (Blanchard et al., 2010; Brent, 1999; Monaghan & Christiansen, 2010; Venkataraman, 2001).

This lexical recognition introduces a single parameter, α , to DYMULTI. To favour previously-seen words, this process simply adds α to the score of a word w if $w \in L$, where L is the proto-lexicon populated with words in previous segmentations. Reasonable values of α lie in $[0, 1]$, where $\alpha = 0$ is equivalent to not using the lexical recognition process. Setting $\alpha = 1$ has the effect of always trying to place boundaries around previously-seen words, as it will always return scores above 0 since $\text{score}(s) \in [-1, 1]$. Note that due to the syllabic nucleus constraint, these boundaries will not necessarily be placed, but there will still be a very strong bias towards them. Intermediate values of α result in a balance between the lexical recognition process and the boundary-finding process.

The full word score function with both lexical processes is given in Figure 4.

Summary

The new DYMULTI framework addresses the drawbacks of the language modelling and boundary-finding approaches to segmentation. The model uses the weighted majority-vote algorithm of MULTICUE to produce scores that are then used to select the best segmentation using the Viterbi algorithm of PHOCUS. Using this dynamic programming

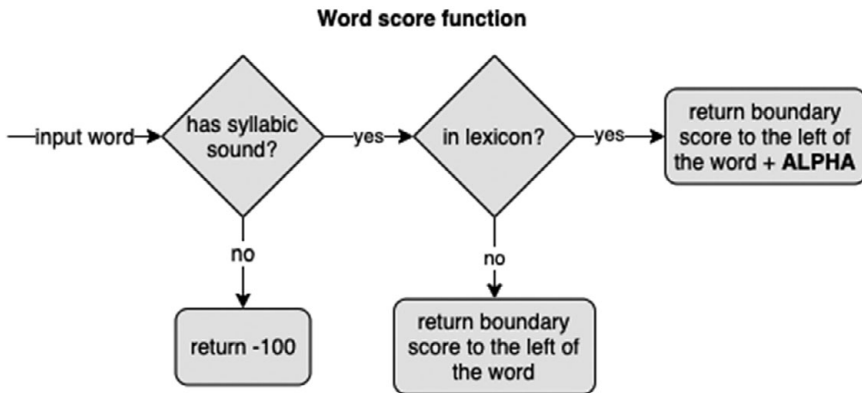


Figure 4. The expanded word score function for DYMULTI with the lexical recognition process.

algorithm, the model is able to incorporate lexical processes. We describe two such processes: the syllabic nucleus constraint from PHOCUS-1S (Blanchard et al., 2010) and a rudimentary lexical recognition process that takes a single parameter to adjust the weighting given to previously-seen words. The full model efficiently combines multiple sub-lexical and lexical cues for segmentation, without the drawbacks of previous models.

A summary of the DYMULTI models used in this study is given in Table 1, highlighting how the concepts behind MULTICUE and PHOCUS have been combined. We implement DYMULTI with the cues from MULTICUE-14, MULTICUE-17 and our new MULTICUE-23 model.

Data and Evaluation

In this section, we discuss the procedure used to evaluate PHOCUS, MULTICUE and DYMULTI. This includes the data used, the baseline segmentation model and the evaluation metrics.

Corpora

To evaluate computational models for speech segmentation, it is customary to use transcriptions of real child-directed speech as input data. The first corpus we use in this study is the BR CORPUS, the de-facto standard for evaluating computational models for segmentation. It was originally collected by Bernstein Ratner et al. (1987) by recording the conversations between nine mothers and their children. It makes up part of the English section of CHILDES, a large database that contains orthographic transcriptions of speech between carers and children of a variety of ages in a multitude of languages (MacWhinney & Snow, 1985).

The BR corpus was later hand-processed by Brent and Cartwright (1996) to produce a phonemic transcription, keeping only child-directed utterances and removing onomatopoeia and interjections. They removed all word boundaries, keeping only utterance boundaries, for a total of 95,809 phonemes, 33,387 words and 9,790 utterances. The transcription system used is not standard, often combining diphthongs, r-colored vowels and syllabic consonants into a single symbol. As there are only 50 symbols used, there is

Table 2. First Five Utterances in the BR Corpus

Input utterance	Correct segmentation	Orthographic equivalent
yuwanttusiD6bUk	yu want tu si D6 bUk	you want to see the book
lUKD*z6b7wIThIzh&t	lUK D*z 6 b 7 wIT hIz h&t	look there's a boy with his hat
&nd6dOgi	&nd 6 dOgi	and a doggie
yuwanttuUk&tDIs	yu want tu lUK &t DIs	you want to look at this
lUK&tDIs	lUK &t DIs	look at this

an average of 2.9 phonemes per word. The lexical stress was later added by Çöltekin & Nerbonne (2014) semi-automatically according to stress patterns in the MRC psycholinguistic database. Examples of utterances from the corpus can be seen in Table 2. Segmentation models have the task of correctly placing word boundaries given these input utterances, without any supervision. For example, if the input is *yuwanttusiD6bUk* then the correct output is *yu want tu si D6 bUk* (you want to see the book). We also note that the corpus represent a tiny fraction of the total input available to children, which has been estimated to be between 2M and 7M words per year (Gilkerson et al., 2017).

In this study, we also evaluate models cross-lingually, using phonemic transcriptions of child-directed speech from 26 different languages created by Caines et al. (2019). Their dataset consists of 132 monolingual CHILDES corpora, each containing 10,000 child-directed utterances aimed at children two years or younger. They processed these corpora using the eSpeak Next Generation (NG) speech synthesizer text analysis module² or segments grapheme-to-phoneme transformer³ to produce phonemic transcriptions. We direct readers to their study for a full description of the corpora. Of these 132 transcripts, we select one for each language⁴, for a total of 26. This was done to facilitate comparison: as otherwise, the corpora for some languages would be much larger than others – there are 28,000 total utterances for North American English but only 10,000 utterances for Basque.

These transcriptions use the International Phonetic Alphabet which contains more symbolic phonemes than the alphabet used for the BR corpus. In these transcriptions, there is an average of 3.7 ± 0.7 phonemes per word due to the more fine-grained phonetic detail. This also varies between languages. For instance, the Turkish transcript has an average of 5.4 phonemes per word but the Cantonese transcript only has an average of 2.6 phonemes per word, reminding us that the notion of a “word” is not equal across languages.

It also must be stated that the BR corpus and these cross-lingual corpora represent an idealisation of the natural scenario. There is likely to have been a degree of error in the transcription stage. Additionally, the phonemic representations of the transcriptions are idealised productions based on dictionary pronunciation; orthographic words are phonemically transcribed in the same way each time they occur, regardless of context. Finally, the largest simplifying assumption made by working with phonemic transcripts is that

²<https://github.com/espeak-ng/espeak-ng>.

³<https://github.com/cldf/segments>.

⁴Their study had 28 languages but counted English and Portuguese twice each, as these are separated by region (North America and UK for English, Brazil and Portugal for Portuguese) in CHILDES. We selected North American English and Brazilian Portuguese for this study.

infants at this age are able to segment speech into phonemes, requiring them to both group phone realisations as phonemes and to have access to phone boundaries, neither of which is a simple task. We discuss the validity of this assumption and the implications for future work in child language acquisition at the end of the paper.

Evaluation metrics

We report each model's performance standard measures; precision, recall and F_1 -score as:

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

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

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

TP is the number of true positives identified by the model, FP is the number of false positives (items identified by the model that are incorrect with respect to the gold standard) and FN is the number of false negatives (items missed by the model). The F_1 -score is calculated as the harmonic mean of precision and recall, providing a single balanced measure. As is conventional, we report F_1 -scores as percentages.

Studies of computational segmentation report these measures in three different ways (Brent, 1999; Çöltekin & Nerbonne, 2014), BOUNDARY, TOKEN and TYPE:

Boundary scores

TP, FP and FN are calculated according to the boundaries placed. For instance, TP is the number of correctly identified boundaries. This gives BP, BR and BF for the BOUNDARY PRECISION, BOUNDARY RECALL and BOUNDARY F_1 -SCORE. Note that utterance boundaries are not included in these calculations, as these are assumed to be trivial to place.

Token/word scores

These are stricter measures that indicate how well word tokens have been identified in the speech stream. As such, true positives are counted only if both boundaries of a word are found without an intervening boundary

between them. These scores are necessarily lower than the boundary scores. This gives WP, WR and WF for the WORD PRECISION, WORD RECALL and WORD F_1 -SCORE. Note that these include utterance-initial and utterance-final words.

Type/lexicon scores

These are similar to word scores, but true positives are marked over word types rather than word tokens, so are not skewed by the frequency of each type. This is done by comparing the final lexicon learned by the model to the expected lexicon, the true set of word types in the corpus. If the model is better at segmenting high-frequency words, the

lexicon scores will be lower than the word scores. The lexicon scores are LP, LR and LF for the LEXICON PRECISION, LEXICON RECALL and LEXICON F₁-SCORE.

Finally, we report two error measures to give insight into how the models may fail, related to two of the error types a model may make. First, a model may miss a boundary, causing UNDER-SEGMENTATION. Second, the model may place a boundary where there should not be one, causing OVER-SEGMENTATION. As the simple error counts will change depending on the size of the corpus and as there are many more word-internal positions than boundaries, normalised measures are used for under-segmentation (E_u) and over-segmentation (E_o):

$$E_u = \frac{FN}{FP + TP},$$

$$E_o = \frac{FP}{FP + TN},$$

where TP, FP and FN are the quantities used for the boundary measures and TN gives the true negatives (the total count of correctly placed word-internal positions). Intuitively, E_u gives the fraction of boundaries marked as word internal and E_o gives the fraction of word internal positions incorrectly marked as boundaries.

Evaluation procedure

In the machine learning literature, models are typically evaluated by training them on one section of the corpus then testing them on another. Computational models for segmentation, however, are unsupervised. Thus, we follow the procedure outlined by previous studies, training our models on a single run of the whole corpus and reporting the average scores across this training period. Although models improve as they learn, they cannot correct past mistakes, resulting in lower average scores than if a test-train split were used.

As a baseline model, we implemented BASELINE, which assigns boundaries randomly but with the correct probability (the true proportion of word boundaries). BASELINE is therefore more informed than a truly random classifier, as this probability is difficult to estimate. This has been the customary baseline for evaluating segmentation models since Brent & Cartwright (1996).

Most previous studies only report their results on one run of the unshuffled corpus, in order to best represent the input that children receive, but some studies report results averaged over multiple shuffles, as is standard practice when performing empirical evaluation of machine learning systems. In order for our study to be as thorough and consistent with previous work as possible, we report both types of results. All models reported here are deterministic, so only one run is needed per shuffle. Comparing unshuffled to shuffled results has the additional benefit of isolating the effect of utterance order, allowing us to identify if parents unknowingly bias the ordering of utterances spoken to their children to increase learnability.

Results

In this section, we first evaluate our re-implementations of MULTICUE and PHOCUS. We compare older sets of cues to the new set of cues used in MULTICUE-23. We then

evaluate DYMULTI against PHOCUS and MULTICUE by comparing the average performance, learning rates across the BR corpus and different values of the lexical recognition constant α . We then compare DYMULTI to previous studies to place its performance in context. Finally, we perform cross-lingual evaluation, comparing DYMULTI, PHOCUS and MULTICUE on 26 different languages. This marks the first time that state-of-the-art segmentation models have been compared on so many languages (Caines et al. (2019) compared baseline models with one state-of-the-art model).

In the analysis below, significance is tested using a pairwise t-test, $\alpha = 0.001$ using samples collected from running each model over 10 shuffles of the input data. When not otherwise stated, we set the DYMULTI's lexical recognition parameter α to 0.

Reimplementation of PHOCUS and MULTICUE models

MULTICUE models

Table 3 gives the results for MULTICUE, run with different sets of indicators. Our implementations of MULTICUE-14 and MULTICUE-17 achieve similar scores to Çöltekin & Nerbonne (2014) and Çöltekin (2017), with slightly higher error rates than the reported results but far exceeding the baseline. These differing error rates are likely due to fine-grained implementation differences, such as how probability estimates are calculated and how utterances are internally represented.

Table 3 also compares running MULTICUE-14 without the stress cue, as Çöltekin & Nerbonne (2014) found the stress cue to decrease performance. The model achieves slightly lower scores than the published results in this case, with a higher under-segmentation error rate. As such, we can confirm the finding of Çöltekin & Nerbonne (2014), that including the stress cue leads to worse overall performance. Also included are the results of MULTICUE-23, whose set of indicators combines the strengths of the MULTICUE-14 and MULTICUE-17 models. This set of indicators clearly leads to substantial improvements, with MULTICUE-23 achieving better F_1 -scores than MULTICUE-14 and MULTICUE-17.

Table 3. Comparison of Reimplemented MULTICUE Models on the BR Corpus

Model	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_u	E_o
MULTICUE-14	93.4	78.0	85.0	80.3	70.9	75.3	27.4	60.9	37.8	21.9	2.1
<i>Reference</i> ^a	92.8	75.7	83.4	78.3	68.1	72.9	26.8	62.7	37.5	24.3	2.2
MULTICUE-14/S	84.3	88.3	86.2	73.6	76.1	74.8	38.7	64.8	48.5	11.7	6.2
<i>Reference</i> ^a	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	8.8	6.7
MULTICUE-17	84.0	87.7	85.8	73.1	75.3	74.2	35.6	66.6	46.4	12.3	6.3
<i>Reference</i> ^b	84.9	88.5	86.7	74.3	76.5	75.4	38.0	67.0	48.5	11.5	5.9
MULTICUE-23	89.7	87.1	88.4	80.2	78.5	79.3	41.2	69.6	51.7	12.9	3.8
BASELINE	27.6	29.5	28.5	12.5	13.1	12.8	6.0	43.4	10.6	70.5	29.3

Note. BP, BR and BF stand for boundary precision, recall and F_1 -score. W and L scores are similar for the word and lexicon measures. E_u and E_o give over-segmentation and under-segmentation. The highest scores and lowest error rates (not including referenced results) are given in **bold**. Italicised lines give the scores reported for each model in their corresponding studies. MULTICUE-14/S is the MULTICUE-14 model without the stress cue. The models are only run once on the unshuffled corpus to facilitate direct comparison with the corresponding reported scores.

^aAs reported by Çöltekin & Nerbonne (2014).

^bAs reported by Çöltekin (2017).

PHOCUS models

The performance of the two PHOCUS models is given in Table 4. Reference rows are not available because Venkataraman (2001) and Blanchard et al. (2010) use different evaluation schema. Venkataraman reports only WP, WR and LP for PHOCUS-1, averaged over 100 shuffles of the corpus, giving 67.7, 70.2 and 52.9 respectively. Averaging over 10 shuffles, we get 69.2 ± 2.9 , 67.2 ± 2.5 and 47.4 ± 1.2 respectively. Deriving a WF score from his WP and WR scores gives 68.9, which is close to our WF score of 68.2 ± 2.6 .

Blanchard et al. only give the WF score of PHOCUS-1S, reporting WF = 80, but do not include the first 1000 utterances in this calculation. Replicating this, we achieve a very close WF score of 80.8. Overall, our implementation of these models seems to perform similarly to the original studies.

Comparing the scores achieved by our implementations of the two models, it is clear that the syllabic nucleus constraint introduced in PHOCUS-1S leads to a significant increase in all F_1 -scores, showing the benefit of a boundary-finding model that can use this constraint. Significance scores are given in Table 5. The over-segmentation error rate is halved from 6.1 to 3.0, a significant reduction ($t = -12.4, p = 5.8 \times 10^{-07}$), indicating that this constraint is preventing boundaries from being placed at word-internal positions that would otherwise lead to producing words without syllabic nuclei.

Table 4 also reveals that F_1 -scores decrease when the input corpus is shuffled. This indicates that the specific ordering of utterances in the BR corpus is useful for segmentation. As the ordering of utterances comes from real child-directed speech, this suggests that parents may positively bias the ordering of utterances spoken to their children to assist with segmentation, such as pairing new word types with previously-uttered word types.

Performance of DYMULTI

Comparing DYMULTI to MULTICUE when using the same set of indicators

Table 6 gives the full scores comparing MULTICUE to DYMULTI, considering just the syllabic nucleus constraint. Every F_1 -score is significantly improved using the DYMULTI model, with scores given in Table 5. DYMULTI-23 achieved the best F_1 -scores; 89.5, 81.8 and 51.9 for BF, WF and LF respectively. Generally, using the DYMULTI model significantly decreases the over-segmentation error rate of the MULTICUE models. This confirms that the syllabic nucleus constraint alone is a useful addition, correctly preventing the model from placing erroneous boundaries. The increase in WF and LF scores

Table 4. Comparison of Reimplemented PHOCUS Models on the BR Corpus

Model	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_u	E_o
PHOCUS-1	82.3	84.4	83.3	70.0	71.3	70.7	53.8	55.7	54.7	15.6	6.9
PHOCUS-1 (avg) ^a	83.2	79.9	81.5	69.2	67.2	68.2	47.6	54.0	50.6	20.1	6.1
PHOCUS-1S	91.3	84.4	87.7	81.6	77.2	79.3	57.9	67.8	62.5	15.6	3.0
PHOCUS-1S (avg) ^a	90.9	79.9	85.1	78.8	72.1	75.3	51.8	66.5	58.2	20.1	3.0
BASELINE	27.6	29.5	28.5	12.5	13.1	12.8	6.0	43.4	10.6	70.5	29.3

Note. BP, BR and BF stand for boundary precision, recall and F_1 -score. W and L scores are similar for the word and lexicon measures. E_u and E_o give over-segmentation and under-segmentation. The highest scores and lowest error rates are given in bold.

^aScores averaged over ten shuffles of the BR corpus.

Table 5. Result of a Pairwise Student's t-test Comparing F_1 -scores of Segmentation Models

Models	BF		WF		LF	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
PHOCUS-1 vs PHOCUS-1S	8.6	1.2×10^{-05}	10.6	2.2×10^{-06}	19.2	1.3×10^{-08}
MULTICUE-14 vs DYMULTI-14, $\alpha = 0$	13.0	4×10^{-07}	11.7	9.5×10^{-07}	4.9	0.00081
MULTICUE-14 vs DYMULTI-14, $\alpha = 0.5$	23.8	1.9×10^{-09}	16.1	6.1×10^{-08}	43.2	9.5×10^{-12}
MULTICUE-14 vs DYMULTI-14, $\alpha = 1$	17.0	3.8×10^{-08}	8.0	2.2×10^{-05}	30.5	2.2×10^{-10}
MULTICUE-17 vs DYMULTI-17, $\alpha = 0$	32.9	1.1×10^{-10}	36.0	4.8×10^{-11}	39.0	2.4×10^{-11}
MULTICUE-17 vs DYMULTI-17, $\alpha = 0.5$	14.8	1.3×10^{-07}	13.9	2.2×10^{-07}	39.8	2×10^{-11}
MULTICUE-17 vs DYMULTI-17, $\alpha = 1$	9.6	5×10^{-06}	7.7	3.1×10^{-05}	23.5	2.2×10^{-09}
MULTICUE-23 vs DYMULTI-23, $\alpha = 0$	10.2	3.1×10^{-06}	10.9	1.8×10^{-06}	6.9	7.4×10^{-05}
MULTICUE-23 vs DYMULTI-23, $\alpha = 0.5$	8.1	1.9×10^{-05}	6.1	0.00017	33.9	8.4×10^{-11}
MULTICUE-23 vs DYMULTI-23, $\alpha = 1$	-0.2	0.86	-1.5	0.17	14.9	1.2×10^{-07}

Note. BF, WF and LF stand for boundary, word and lexicon F_1 -scores. Each pairwise comparison considers the same set of indicators (e.g., DYMULTI-14 uses the same set of indicators as MULTICUE-14). We set $\alpha = 0$ for the DYMULTI models. Each model is run on ten shuffles of the BR corpus and scores are paired for each shuffle. All scores are significant at the $p < 0.001$ level except for the final test.

Table 6. Comparison of MULTICUE and DYMULTI Models on the BR Corpus

Model	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_u	E_o
MULTICUE-14	92.8	78.3	84.9	80.0	71.1	75.3	27.6	61.9	38.2	21.7	2.3
DYMULTI-14	93.9	80.0	86.4	82.2	73.6	77.6	28.8	62.1	39.4	20.0	2.0
MULTICUE-17	83.5	88.1	85.7	72.6	75.4	74.0	34.9	65.6	45.6	11.9	6.6
DYMULTI-17	90.1	88.4	89.3	82.0	80.9	81.4	38.2	70.2	49.5	11.6	3.7
MULTICUE-23	88.4	87.4	87.8	78.5	77.9	78.2	40.3	69.8	51.1	12.6	4.4
DYMULTI-23	92.0	87.0	89.5	83.4	80.2	81.8	40.8	71.2	51.9	13.0	2.8
BASELINE	27.6	29.5	28.5	12.5	13.1	12.8	6.0	43.4	10.6	70.5	29.3

Note. BP, BR and BF stand for boundary precision, recall and F_1 -score. W and L scores are similar for the word and lexicon measures. E_u and E_o give over-segmentation and under-segmentation. Each pairwise comparison considers the same set of indicators (e.g., DYMULTI-14 uses the same set of indicators as MULTICUE-14). We set $\alpha = 0$ for the DYMULTI models. Each model is run on ten shuffles of the BR corpus, averaging scores, with the highest scores and lowest error rates in **bold**.

shows that the Viterbi algorithm, considering this constraint, is finding the correct words to segment, leading to a more accurate lexicon.

Learning rates of segmentation models

The scores in Table 6 were calculated by taking the average performance of each model over the whole BR corpus, including the many initial mistakes made as the models gather statistical information. While this gives an indication of how well each model learns to

segment after beginning with no knowledge of the target language, it can be more informative to see how the performance of each model progresses across the corpus.

Figure 5 gives the WF and LF learning rates for a selection of the models described in this study. These models initially perform very poorly, but quickly improve over the first 1000–2000 utterances, after which scores do not increase or decrease by more than 10 points. This is expected, as the models begin with very poor representations of the target language and so make poor boundary decisions. As such, the average scores over the whole corpus are not representative of the final performance of each model. For example, over the whole corpus MULTICUE-14 has an average WF score of 75.4 ± 0.5 and MULTICUE-23 has an average WF score of 78.2 ± 1.0 but there is no significant difference between their WF scores on the final block of 200 utterances. It seems that MULTICUE-14 initially learns very slowly, likely due to a large number and a high variety of indicators that need to be learned.

From these learning rates, we also see the consistent benefit of the syllabic nucleus constraint. At every stage in the learning process, PHOCUS-1S achieves higher F_1 -scores than PHOCUS-1 and DYMULTI-23 achieves higher F_1 -scores than MULTICUE-23. Indeed, DYMULTI-23 achieves the highest F_1 -scores out of any model at almost every stage of the learning process, achieving WF and LF scores of 86.3 ± 1.7 and 82.9 ± 2.0 respectively on the final 200 utterances of the BR corpus. This confirms the validity of this model and the benefits of combining the boundary-finding and language modelling approaches to segmentation.

Lexical recognition process

Figure 6 compares MULTICUE to DYMULTI, considering three values for the lexical recognition parameter α . In all cases but one, the DYMULTI model performs significantly better than the MULTICUE model when both models use the same set of indicators (see Table 5 for the significance test results). It seems the LF scores are more sensitive to this change in model than the WF scores. The LF score for MULTICUE-14, for instance, jumps by more than 20 points when DYMULTI is used with $\alpha = 0.5$. This suggests that the lexical processes of DYMULTI help capture infrequent words.

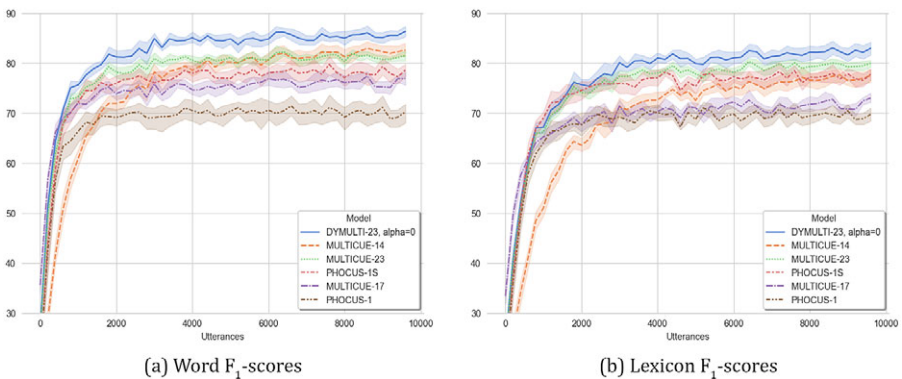


Figure 5. Word and Lexicon F_1 -scores (WF, LF) for a selection of the models implemented in this study, calculated over blocks of 200 utterances. Scores are calculated by running each model separately on ten shuffles of the BR corpus and averaging results.

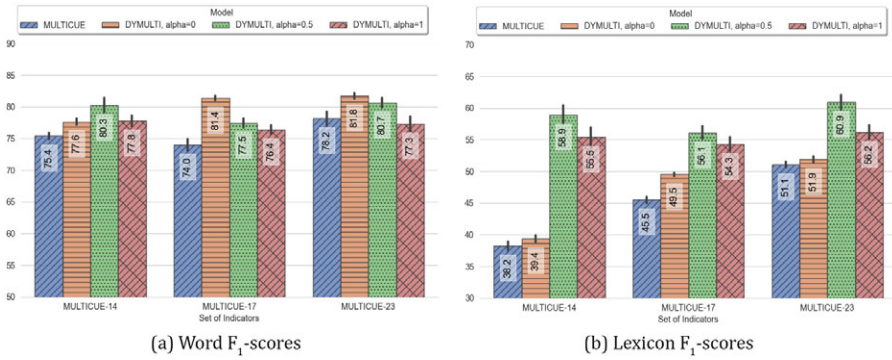


Figure 6. Word and Lexicon F₁-scores (WF, LF) for four models using three sets of indicators. MULTICUE-14, MULTICUE-17 and MULTICUE-23 are compared to three DYMULTI models using the same sets of indicators, setting $\alpha = 0, 0.5, 1$. Scores are calculated by running each model separately on ten shuffles of the BR corpus and averaging results.

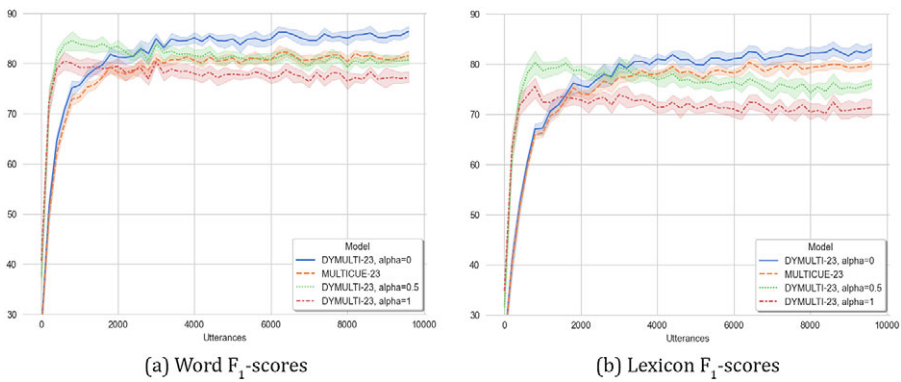


Figure 7. Word and Lexicon F₁-scores (WF, LF) for MULTICUE-23 and DYMULTI-23, setting $\alpha = 0, 0.5, 1$. Scores are calculated by running each model separately on ten different shuffles of the BR corpus and averaging results.

The learning rates for these models, however, tell a different story. **Figure 7** compares the learning rates of these models, revealing that the relatively higher AVERAGE WF and LF scores over the corpus for DYMULTI with $\alpha = 0.5$ and $\alpha = 1$ given in **Figure 6** are actually due to a very steep initial learning rate. Indeed, the final WF and LF scores for these two values of α are actually lower than DYMULTI-23 with $\alpha = 0$ and are in fact lower than MULTICUE-23. It seems that segmenting on the basis of previously-seen words is a useful strategy at the very start of learning to segment while the boundary cues are still gathering statistical information. As the boundary-based process improves, this lexical recognition procedure actually harms the model, leading to a decrease in WF and LF over time. We also experimented with smaller values of α but never reached significantly higher F₁-scores by the end of learning.

We believe that this is because the lexical recognition process described in this study is very rudimentary, simply adding a fixed value to the score when a word is recognised. While the model is gathering statistical information, the boundary-based votes are likely

to be inaccurate and extreme, so the lexical recognition process will help prevent incorrect segmentation. As the boundary-based votes become more fine-grained, however, the fixed score of the lexical recognition process will dominate. This will prevent the boundary-based votes from discovering new words, explaining the decrease in LF over time. A more nuanced lexical recognition process should account for this, perhaps by including a decay parameter that decreases α gradually, relying less on the lexical recognition process as the boundary-based votes become more stable predictors. Using DYMULTI, it is easy to explore the inclusion of such a lexical process, without needing to define or run a new model.

Comparison to previous studies

Table 7 compares DYMULTI-23 and MULTICUE-23 to other models in the word segmentation literature. Note that these models differ in terms of the evaluation procedure. The first four models are incremental, so scores are calculated over a single run of the BR corpus (averaging over 100 independent runs over shuffles of the corpus in the case of Venkataraman (2001)). The next two models are incremental, but run over the corpus multiple times and only report the results for the final run. The following two models are batch-based, so scores are calculated after many iterations of training over the corpus (ranging from two to several thousand). DYMULTI-23 with $\alpha = 0$, achieves higher BF and WF scores than all of these, with a comparable LF score. Using $\alpha = 0.5$ results in the highest LF score, but this is potentially misleading, as discussed in the previous section. It is also interesting that DYMULTI-23 outperforms the several-run models of Fleck (2008); Ma et al. (2016) and the batch-based models of Elsner & Shain (2017); Goldwater et al. (2009) as these do not suffer from the lower performance in the initial learning phase. It is also worth noting that our implementation of PHOCUS-1S already outperforms most previous studies.

Many previous studies only explore one or two cues for segmentation, stating that they expect a model that considers more cues to perform better (Blanchard et al., 2010; Çöltekin, 2017; Ma et al., 2016). The set of cues chosen here seems particularly effective, with MULTICUE-23 already achieving higher BF and WF scores than in previous studies. The LF score for DYMULTI-23 with $\alpha = 0$ is still lower than many of the other models in Table 7, only outperforming the models of Çöltekin (2017); Fleck (2008); Ma et al. (2016) that do not store a lexicon at all, but setting $\alpha = 0.5$ remedies this, leading to the highest LF score.

This comparison confirms both the strength of the boundary cues included and the strength of the syllabic nucleus constraint at the lexical level.

Cross-lingual evaluation

Comparison of PHOCUS, MULTICUE and DYMULTI across 26 languages

The majority of studies that present models for child word segmentation only report results on English transcripts, typically only using the BR corpus. Exceptions include Blanchard et al. (2010), who also report results on a Sesotho corpus, Fleck (2008), who also reports results on Spanish and Arabic corpora, and Caines et al. (2019), who compared three different segmentation models on 26 languages. As these models are designed to represent the ability of a child to acquire any language, proper evaluation is incomplete if the models are not run on a wide variety of languages. Otherwise, the models

Table 7. Comparison of Computational Models for Word Segmentation

Model	BP	BR	BF	WP	WR	WF	LP	LR	LF	E_u	E_o
Brent (1999)	80.3	84.3	82.3	67.0	69.4	68.2	53.6	51.3	52.4	25.7	–
Venkataraman (2001)	–	–	–	67.7	70.2	68.9	52.9	–	–	–	–
Çöltekin & Nerbonne (2014)	83.7	91.2	87.3	74.1	78.8	76.4	43.9	67.7	53.3	8.8	6.7
Çöltekin (2017)	84.9	88.5	86.7	74.3	76.5	75.4	38.0	67.0	48.5	11.5	5.9
Fleck (2008)	94.6	73.7	82.9	–	–	70.7	–	–	36.6	26.3	–
Ma et al. (2016)	–	–	82.9	–	–	68.7	–	–	42.6	17.3	6.4
Goldwater et al. (2009)	90.3	80.8	85.2	75.2	69.6	72.3	63.5	55.2	59.1	19.2	–
Elsner & Shain (2017)	81	85	83	–	–	72	–	–	–	15	–
PHOCUS-1S ^a	91.3	84.4	87.7	81.6	77.2	79.3	57.9	67.8	62.5	15.6	3.0
MULTICUE-23	89.7	87.1	88.4	80.2	78.5	79.3	41.2	69.6	51.7	12.9	3.8
DYMULTI-23, $\alpha=0$	92.8	86.4	89.5	84.4	80.2	82.2	41.4	71.4	52.4	13.6	2.5
DYMULTI-23, $\alpha=0.5$	85.9	95.7	90.6	79.7	86.2	82.8	63.6	65.3	64.5	4.3	5.9
BASELINE	27.6	29.5	28.5	12.5	13.1	12.8	6.0	43.4	10.6	70.5	29.3

Note. MULTICUE-23 and DYMULTI-23 are compared to a variety of the top-performing models from the child word segmentation literature. BP, BR and BF stand for boundary precision, recall and F_1 -score. W and L scores are similar for the word and lexicon measures. E_u and E_o give over-segmentation and under-segmentation. Scores are obtained on the BR corpus, with the highest scores and lowest error rates in **bold**. If there were multiple models reported in a study, the model with the highest LF score is given. The scores across models are not always directly comparable, as some are calculated differently from others.

^aOur implementation of PHOCUS-1S is used as a stand-in for the model of Blanchard et al. (2010) as they only report WF.

could be biased towards learning English, and therefore not represent a truly universal language acquisition procedure.

Figure 8 compares the LF scores of PHOCUS-1S, MULTICUE-17 and DYMULTI-23 across 26 languages. These transcripts come from the study of Caines et al. (2019). The models they evaluate in their study either perform significantly worse than those presented here or process the transcripts several times, so are not comparable to these three single-run and incremental state-of-the-art models. To account for the initial learning curve which may differ between models and languages, we only report the scores achieved over the last 5000 utterances of each transcript, after the models have stabilised.

For 15 of the 26 languages, DYMULTI-23 achieves a significantly higher LF score than MULTICUE-17 and PHOCUS-1S (according to a paired t-test, with $p < 0.001$). This is significantly above the chance level of $\frac{26}{3}$ for this experiment⁵. If we order the languages by LF score, English comes second for MULTICUE-17 and PHOCUS-1S and fourth for DYMULTI-23. As DYMULTI builds on MULTICUE and PHOCUS, which themselves build on previous work, this suggests that the research into child segmentation has been biased towards performance on English corpora. Also note that PHOCUS-1S achieves a higher LF score than DYMULTI-23 for English, which was not the case for the BR corpus (as seen in Figure 5). This highlights the importance of testing on multiple corpora, as the transcription system seems to have an effect on the performance of the models.

Correlating DYMULTI performance with language features

As a preliminary investigation into why DYMULTI-23 performs better on some languages than others, we grouped the languages by their sub-families using Glottolog

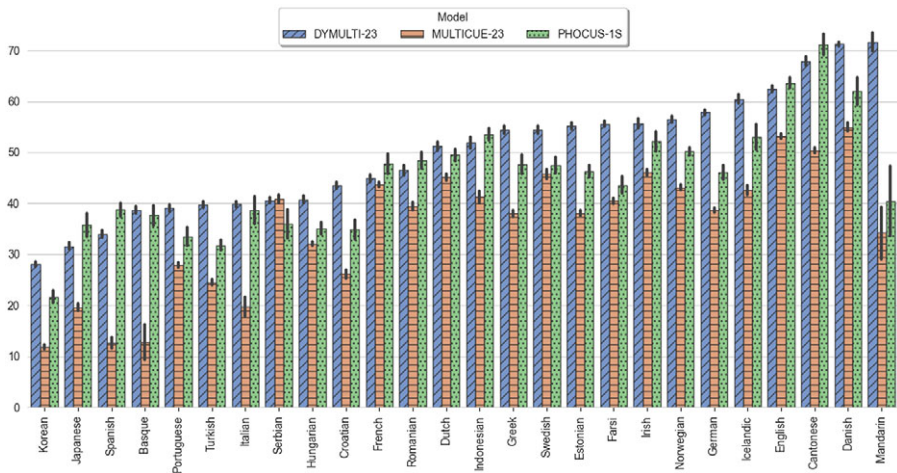


Figure 8. Lexicon F_1 -scores (LF) for PHOCUS-1S, MULTICUE-17 and DYMULTI-23 with $\alpha=0$ compared across 26 languages, sorted by LF scores for DYMULTI-23. Scores are calculated by running each model separately on ten shuffles of each transcript and averaging results over the last 5000 utterances of each run, accounting for differing initial learning rates.

⁵The chance that DYMULTI-23 would randomly outperform the other two systems on at least 15 languages can be calculated with a binomial distribution with $p = \frac{1}{3}$, giving $P(X \geq 15) < 0.01$

(Hammarström et al., 2022). Any languages that did not share a sub-family with any other languages were classed as ‘Other’. The resulting groups are:

- Sinitic (Cantonese and Mandarin)
- Germanic (Danish, Dutch, English, German, Icelandic, Norwegian and Swedish)
- Balto-Slavic (Serbian and Croatian)
- Italic (French, Italian, Portuguese, Romanian and Spanish)
- Other (Basque, Estonian, Farsi, Greek, Hungarian, Indonesian, Irish, Japanese, Korean and Turkish)

All languages in the Italic group are also Romance languages (a modern subgroup of Italic languages). These languages have a number of shared features; they are moderately inflecting, have a primarily subject-verb-object word order and accent with stress. Germanic languages also accent with stress but vary when it comes to inflection; German and Icelandic have complex inflectional morphology whereas English and Swedish are largely analytical. Germanic languages also have verb-second word order (English is an exception), unlike other families. Serbian and Croatian are mutually intelligible varieties of Serbo-Croatian, a highly inflectional language that has a flexible word order, often defaulting to subject-verb-object as with Italic languages. Serbo-Croatian also has a simple tone system. The Germanic, Balto-Slavic and Italic groups are all divisions of the Indo-European family. The Sinitic group, on the other hand, is a sub-division of the Sino-Tibetan family. These languages have relatively simple morphology, with no inflections or conjugations. The basic word order is subject-verb-object and modifiers usually precede the words they modify. One distinguishing feature of Sinitic languages is the use of tones to distinguish words. Of the languages in the ‘Other’ group, Greek, Farsi and Irish are also Indo-European, members of the Hellenic, Indo-Iranian and Celtic sub-families, respectively. The remaining languages are all members of different top-level families; Koreanic (Korean), Uralic (Hungarian, Estonian), Japonic (Japanese), Turkic (Turkish) and Austronesian (Indonesian). Basque is a language isolate, not classified into any family (King, 2018). Figure 9 presents DYMULTI-23 LF scores with languages grouped using this classification. We see that DYMULTI-23 performs better on all Sinitic and Germanic languages than any Balto-Slavic or Italic language. This suggests that the features of a language family can predict how easy it is to learn the segmentation of the languages within that family.

To investigate this further, we extracted the structural properties of each language using the World Atlas of Language Structures (WALS) database (Dryer & Haspelmath, 2013). WALS uses 192 grammatical, phonological and lexical features to describe the cross-linguistic diversity, each feature taking between 2 and 28 values. We then calculated the feature-value pairs that best predicted the DYMULTI-23 LF score achieved for each language⁶.

Figure 10 presents the best ten of these feature-value pairs. We see that out of languages in which the adjective precedes the noun, languages in which the object precedes the verb (line 6) tend to be harder to segment than languages in which the verb precedes the object (line 7). Six of the most predictive features are to do with word order, which is significantly higher than expected, given that out of the 510 feature-value pairs considered, only

⁶This was done using the SelectKBest method in the scikit-learn library (Pedregosa et al., 2011), selecting the feature-value pairs that maximised F_1 -score on a regression task

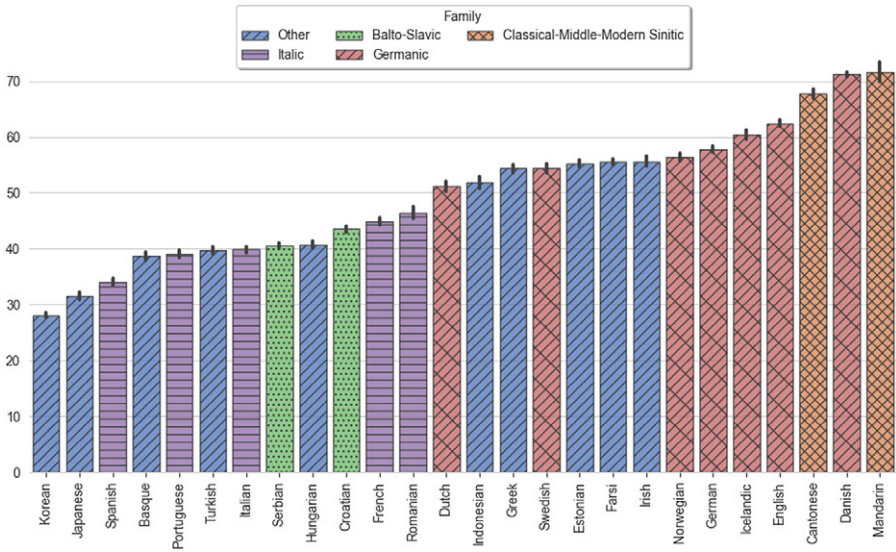


Figure 9. The Lexicon F₁-scores (LF) that DYMULTI-23 achieves for each language. Languages are grouped by family and LF scores are calculated using 10 runs over different shuffles of each transcript.

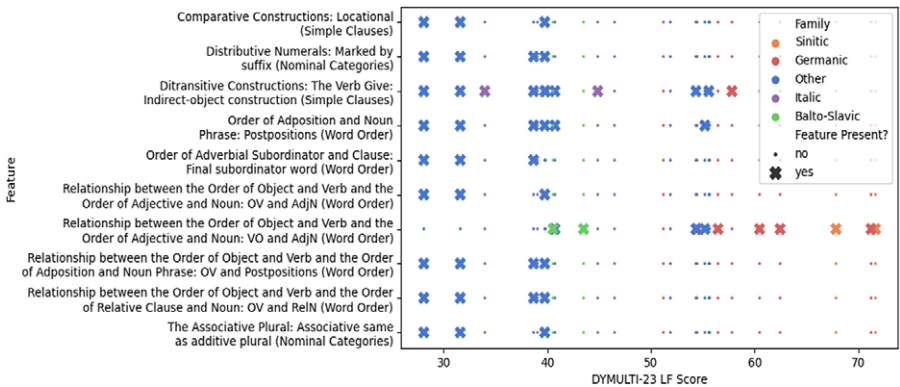


Figure 10. The 10 structural language features that best predict DYMULTI-23 LF score. Each point in a row is a language, with the x-value giving the average DYMULTI-23 LF score achieved for that language across 10 runs. Each point is marked with a cross if the language contains the corresponding feature. Languages are grouped by family.

123 are to do with word order⁷. Korean and Japanese share five of the word order features and DYMULTI-23 achieves the lowest performance on these two. As many of the cues used in DYMULTI-23 identify infrequent phoneme combinations across word boundaries, it is understandable that languages with particular word order features would be easier to segment than others with this model.

⁷The probability of 6 or more word order feature-value pairs being randomly selected is $p=0.0165$, calculated using the binomial distribution

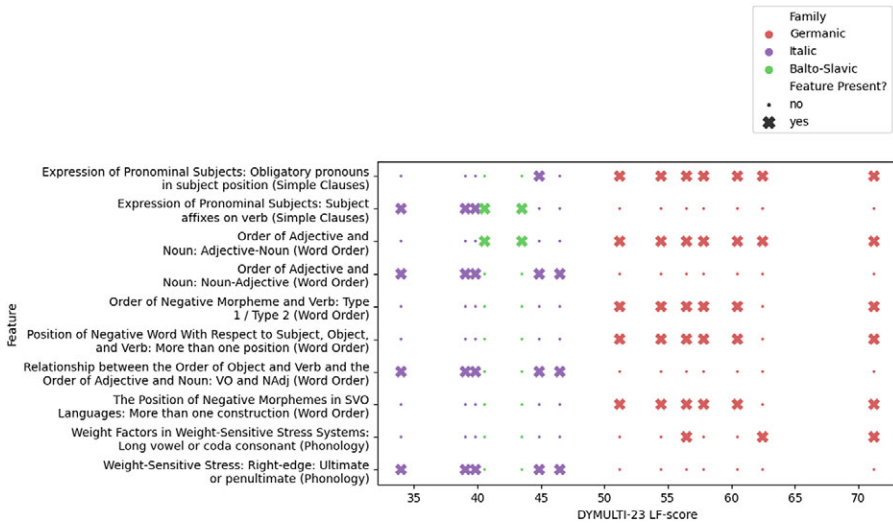


Figure 11. The 10 structural language features that best predict DYMULTI-23 LF score. Each point in a row is a language, with the x-value giving the average DYMULTI-23 LF score achieved for that language across 10 runs. Each point is marked with a cross if the language contains the corresponding feature. Languages are grouped by family, considering only Germanic, Balto-Slavic and Italic languages.

For more fine-grained analysis, we repeated this experiment considering just languages in the Balto-Slavic, Germanic and Italic sub-families. The results are presented in Figure 11. Once again, six of the ten features are to do with word order. The Italic languages have Noun-Adjective word order, whereas the Balto-Slavic and Germanic languages have Adjective-Noun word order. There are also two phonological features selected. The Italic languages all have ultimate or penultimate stress. Given that DYMULTI-23 does not use a stress cue, this could explain the relatively low performance on Italic languages; perhaps for these languages, with their relatively freer word order, stress is a crucial cue for segmentation.

Correlating DYMULTI performance with information-theoretic measures

As the cues used by DYMULTI-23 are largely statistical, we decided to compare the information-theoretic properties of each language. We calculated the following measures:

- the unique number of phonemes in each transcript,
- the entropy of phoneme *n*-grams, and
- the conditional entropy of phoneme *n*-grams.

The first measure was chosen as one of the most basic metrics for phonological complexity (Nettle, 1995). Historically, the number of vowels or the number of consonants have also been used as count-based measures (Moran & Blasi, 2014). Count-based measures are simple to calculate but do not consider the phonotactics of a language. Entropy measures capture the phonological complexity of each language according to the predictability of the phonemes in that language, inherently capturing the nuanced

interactions that the count-based measures do not (Piantadosi et al., 2011; Pimentel et al., 2020). When calculating these measures we use the target segmentation transcripts, which include spaces, so across-word and within-word probabilities are both considered.

Figure 12 presents DYMULTI-23 LF scores plotted against the 3-gram conditional entropy of each language transcript used. There is a strong negative correlation of -0.65 . This suggests that the more predictable a phoneme is within context, the more capable DYMULTI-23 is of identifying boundaries.

Figure 13 gives the correlations between each information-theoretic measure and each of the three F_1 -scores that DYMULTI achieves. It is clear that the information-theoretic statistics of each language are strongly linked to the performance of segmentation models. For Italic languages, whose phonemes are not as predictable within context compared to Germanic languages, additional cues such as stress may be required for successful segmentation.

Summary

In this section, we evaluated PHOCUS, MULTICUE and the new DYMULTI framework. Our re-implementations of PHOCUS and MULTICUE achieved similar performance to their original studies, successfully replicating their findings. We also found that our new set of cues for MULTICUE significantly improves its performance.

Comparing DYMULTI to PHOCUS and MULTICUE, we found that the syllabic nucleus significantly improved performance over MULTICUE models using the same set of cues. The lexical recognition process also improved performance, but this was due to a very fast initial learning period; performance actually decreased over time when α was set to 0.5 or 1. DYMULTI also outperformed prior work, including models that used batch training or other weaker constraints.

Finally, we performed cross-lingual evaluation, which has not been done at this scale for state-of-the-art segmentation models. This validated the performance of

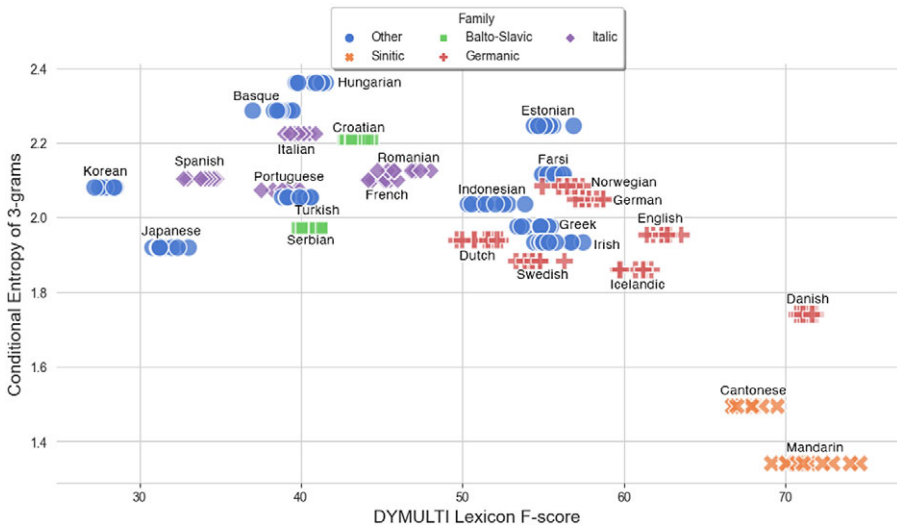


Figure 12. 3-gram conditional entropies of each language transcript used in the study. Languages are grouped by family and are plotted against the Lexicon F_1 -scores (LF) that DYMULTI-23 achieves in 10 runs over different shuffles of the transcript for that language.



Figure 13. Correlation scores between information-theoretic measures and the average F_1 -scores that DYMULTI-23 achieves for each language across 10 runs.

DYMULTI-23, which outperformed PHOCUS-1S and MULTICUE-17 on 15 of 26 languages. We analysed the structural properties of each language and found that word order features are predictive of DYMULTI-23 performance. We then performed information-theoretic analysis of the language transcripts, finding several correlations between these and the performance of DYMULTI-23 on these languages. Taken together, the above two findings mean that for less predictable languages (which tend to have freer word orders), additional cues such as stress may be needed for successful segmentation.

Discussion and Summary

In this study, we explored both boundary-finding and language modelling methods for word segmentation, producing a new segmentation framework, DYMULTI, that combines the powerful boundary decisions from the MULTICUE framework of Çöltekin & Nerbonne (2014) with the lexical constraints of the PHOCUS-1S model of Blanchard et al. (2010). In this section, we first consider the performance of DYMULTI with respect to the different views of speech processing. We then discuss our re-implementations of PHOCUS and MULTICUE and the novel benchmarking that we have carried out in this study. Finally, we describe the limitations of this study and future directions, concluding with the wider implications of this research.

Performance of DYMULTI

We presented two views of speech processing from the cognitive science literature. The interactionist view states that speech segmentation is driven by lexical recognition.

The explicit view states that segmentation is purely a result of placing boundaries using sub-lexical information, without making use of any lexical influences.

Our goal was to compare the language modelling and boundary-finding approaches to solving the speech segmentation problem (which relate to these views of speech processing) and to establish if combining these approaches would lead to improvements in performance on transcriptions of child-directed speech. To achieve this, we first reimplemented the PHOCUS models of Venkataraman (2001) and Blanchard et al. (2010) and the MULTICUE models of Çöltekin & Nerbonne (2014) and Çöltekin (2017). Both models can make use of sub-lexical and lexical information, but PHOCUS primarily uses lexical cues and MULTICUE primarily uses sub-lexical cues. As such, these models corroborate the studies in the cognitive science literature that find that children make use of sub-lexical and lexical cues for solving the word segmentation problem.

Until this study, it was not possible to conclude whether sub-lexical and lexical cues were complementary or alternative explanations for segmentation, as no model had been designed that was able to efficiently combine lexical and sub-lexical cues without constraint. Presenting the DYMULTI framework, we confirmed an improvement over PHOCUS and MULTICUE models. This implies that sub-lexical and lexical cues are indeed complementary and that both can be helpful for solving the word segmentation problem.

We also found that DYMULTI-23 outperforms previous state-of-the-art systems for segmentation on the BR corpus, including those that run over the corpus several times or learn in batch. As DYMULTI makes use of cues found to be used by infants for speech segmentation and builds on established, state-of-the-art models, this means that DYMULTI is a good representation of how children may learn to segment speech and begin to build their lexicon.

Novel benchmarking of state-of-the-art models

Besides presenting a new state-of-the-art model for segmentation, a major contribution of this study was the thorough benchmarking and replication of the PHOCUS and MULTICUE frameworks. Replication is an important scientific discipline and few state-of-the-art models have been re-implemented in prior work. Re-implementing these frameworks, we achieved comparable results to their respective studies. Running MULTICUE-14 without the stress cue, we confirmed the result of Çöltekin & Nerbonne (2014) that it increased performance. We also validated the finding of Blanchard et al. (2010) that the syllabic nucleus constraint improves performance, using this as the core motivation for the design of DYMULTI.

Despite most studies in this field using the same corpus for evaluation, they all evaluate their models differently, making cross-comparison difficult. This is also the case for the PHOCUS and MULTICUE frameworks. For example, the PHOCUS studies do not provide boundary scores and the MULTICUE studies do not provide the learning rates of their models. In this study, we compared these models with a wide range of experiments, including calculating average scores over the whole corpus, plotting learning rates over time and performing novel cross-lingual evaluation. This is the first time these models have been directly compared, producing a useful survey of the field. The cross-lingual evaluation is particularly noteworthy, as few state-of-the-art models have previously been compared on more than two languages. This needs to become a regular practice if the goal of these models is truly to understand how any language is acquired, not just English.

Limitations of child-directed corpora

One of the strengths of the DYMULTI framework is that it is much more flexible than previous models as it can easily consider multiple boundary-based cues and lexical processes. This allowed for the combination of phonotactic, utterance-boundary, lexicon and stress cues derived from phonetic transcriptions of child-directed speech.

The BR corpus is the de-facto standard for evaluating such computational models, but it has certain limitations. Containing only 9,790 utterances spoken by only nine speakers of U.S. English from the east coast in 1987, it is not very representative of child-directed speech in English and all its varieties, yet alone other languages. To validate our results, we ran PHOCUS, MULTICUE and DYMULTI on corpora from 26 different languages.

Through this cross-lingual evaluation, we found that the models perform consistently better on English than on most other languages. This is another limitation of using a single corpus for cross-comparison, as it suggests that previous work may have been biased towards performance on the BR corpus, producing models that perform well on English at the expense of other languages. For example, Çöltekin & Nerbonne (2014) found that including the stress cue decreased the performance of MULTICUE, but this may only be the case for English. In our cross-lingual analysis we found that DYMULTI-23 performs worse on Italic languages than Germanic languages. It may be that with the freer word order and lower phonemic predictability of Italic languages, stress cues play a more important role in segmentation. Using the DYMULTI framework, this could be investigated in future work by experimenting with different cues and implementing different lexical processes to see how these choices alter the performance across languages.

Another limitation of these corpora is the fact that they represent speech as symbolic phonemes. Not only could these transcripts be subject to translation error and bias, due to the way they were automatically produced, but it is also unclear if infants have access to phonetic categories at this stage in acquisition. There is increased awareness that infants may not be forming phonetic categories until later in life (Feldman et al., 2021) and there is mounting evidence that we do not fully collapse speech to discrete phonemic categories for higher-level processing, as assumed by the “myth” of categorical perception (see McMurray (2022) for a review). If infants do form categories, they may even be more fine-grained and numerous than phonemes (Schatz et al., 2021).

In this study we used phonemes as our base unit, focusing on lexical and sub-lexical cues, but it is clear that future work should consider incorporating sub-phoneme cues as well. If a corpus containing continuous audio were used, these cues could be extracted, providing allophonic variation and realistic stress information to the model. This was attempted by Rytting et al. (2010), who used the raw audio associated with the Brent corpus (which is not aligned with the phonemes in the BR corpus) to represent the input stream as phone probability vectors, thus preserving phonetic variation. Unfortunately, they do not make these vectors available, nor will their vectors necessarily align well with the phonemes of the BR corpus (which were derived from the orthographic transcription of the Brent corpus).

One child-directed speech corpus that does contain phonemic transcriptions aligned with raw audio is the CAREGIVER corpus (Altosaar et al., 2010). However, the utterances are scripted and the type-token ratio is only 0.002, much smaller than the 0.036 for the BR corpus. In initial experiments we found that segmentation models that rely on seeing a variety of phoneme combinations at boundaries struggle, resulting in very different results than when run on actual child-directed utterances.

Future work

Segmentation is a developmental process. English-learning infants display some ability to segment words at 7.5 months of age and do not achieve adult-level performance until close to 24 months (Johnson & Jusczyk, 2001; Jusczyk, 1999). During this time, perceptual tuning also occurs and infants' sensitivity to the universal set of distributional cues narrows to a native inventory (Liu & Kager, 2017). When we evaluate computational models of word segmentation, it is important to note that we are not equating model performance with infant performance. Instead, we treat such computational models as idealised learners of the distributional phonemic signal, with high performance indicating the utility of cues and learning methods for segmentation, rather than directly predicting infant ability. If we had proper corpora documenting the development of segmentation in infants over time, we could use DYMULTI to create empirical predictions to test against this data.

There are many components of DYMULTI that could be expanded upon in future work. One such component is the representation of utterances. Instead of individual symbols, phonemes could be represented by features, such as the 11 phonetic features used in the model of Christiansen et al. (1998). Language-specific, distributed representations could also replace phonemes, such as the learned acoustic embedding vectors used in the models of Ma et al. (2016) and Kamper et al. (2016) or the probabilistic phone vectors of Rytting et al. (2010).

Furthermore, as suitable corpora become available – such as the release of new 'day long' corpora in Homebank (VanDam et al., 2016), the number and quality of input cues can be improved. Future work could also investigate semantic and multi-modal information that parents may provide their children, such as deictic gestures towards images, joint attention on entities in the environment or iconic gestures to demonstrate object shapes. It is likely that with more cues available, performance would increase, improving our understanding of language acquisition. With sufficient data, DYMULTI could even be used to explore how statistical learning varies between individual children, rather than assuming learned probabilities are similar across an entire population (Siegelman & Frost, 2015).

Future work should also seek to bridge the gap between models operating on phonemic transcripts and those operating on raw speech signals. The latter continue to perform significantly worse than the former — the top-performing model for the Zero Speech Challenge (ZRC) series segmentation task achieves a token F_1 -score of only 19.2 on the English portion of the TDE-17 test corpus, compared to 64.5 for the text-based topline system provided by the task. In an effort to explain this gap in performance, Dunbar et al. (2022) discuss how the higher granularity of analysis, the lack of invariant quantised acoustic representations and the variability of speech rate all contribute. DYMULTI could be run at a higher granularity of analysis, with features extracted directly from the speech stream, to help bridge this gap. Many of the ZRC series models operate on 10ms frames, which are much shorter than the average duration of a phoneme (about 70ms).

We also note that the ZRC series tasks evaluate their models on adult-directed speech corpora, whereas traditional segmentation models have strictly required that the input data be child-directed. Relaxing this requirement would help with the lack of suitable corpora, help to bridge the gap between these two approaches and could lead to new revelations about acquisition. Using her model, Fleck (2008) explored how infant speech segmentation could be upgraded to adult speech segmentation. She did this by introducing a simple syntactic process to prevent affixes from being segmented away from their stems, achieving WF and LF scores of 80.3 and 41.5 respectively on the Buckeye corpus (Pitt et al., 2005). In initial experiments, DYMULTI performs worse than Fleck's model, with WF and LF scores

of 69.9 and 40.6 respectively. Fleck hypothesised that models for infant speech segmentation may be segmenting morphemes, rather than words. Further work into DYMULTI could implement a syntactic process such as Fleck's to investigate this claim.

Conclusion

In this study, we presented the word segmentation problem and compared two state-of-the-art models for speech segmentation; the PHOCUS models Venkataraman (2001) and Blanchard et al. (2010), a language modelling approach to speech segmentation, and the MULTICUE models of Çöltekin & Nerbonne (2014); Çöltekin (2017), a boundary-finding approach. By re-implementing both models, we found that MULTICUE lacked the ability to consider lexical-level processes and PHOCUS lacked the ability to combine information from several cues. We created a new model, DYMULTI, which overcomes both drawbacks by using the boundary decisions of MULTICUE and converting them to scores that can be passed to the Viterbi algorithm of PHOCUS. In doing so, we achieved state-of-the-art performance on the BR corpus. Evaluating the model cross-lingually, we found that DYMULTI outperformed MULTICUE and PHOCUS on 15 of 26 child-directed speech corpora from different languages, but also that all three models achieved close to their best performance on English, suggesting possible research bias.

DYMULTI represents a flexible framework for exploring hypotheses related to the word segmentation problem, efficiently combining lexical and sub-lexical cues. In this study we have explored predictability, utterance and stress as sub-lexical cues, and a syllabic nucleus constraint and lexical recognition as lexical cues. Such cues could be enhanced with updated knowledge about infant speech cognition to produce a more comprehensive model for speech segmentation. The framework can also be upgraded by considering more nuanced representations of utterances, alternative cue-combination algorithms and other cues for segmentation, once suitable corpora are available.

The impacts of DYMULTI and future research into child speech segmentation are plentiful. Using DYMULTI's cue combination system, we can better our understanding of which cues are relevant to segmentation, aiding segmentation in speech recognition models. Adult speech segmentation could also be researched using DYMULTI, by examining which cues are relevant by testing on adult-directed speech corpora and adapting DYMULTI with new grammatical processes accordingly. Finally, this research has contributed to the ever-growing understanding of language acquisition. By designing segmentation models that perform well on child-directed speech, we can learn how children first solve this task, thereby improving how we teach language to children in the first place and how language disorders can be mitigated.

Acknowledgements. We thank Çağrı Çöltekin for clarifying the workings of his model. The first author was supported by the Cambridge Trust and the second and third authors were supported by Cambridge Assessment English, University of Cambridge. We also thank the editors and anonymous reviewers for their feedback which helped to greatly improve this article.

Competing interest. The authors declare none.

References

- Algayres, R., Ricoul, T., Karadayi, J., Laurençon, H., Zaiem, S., Mohamed, A., Sagot, B., & Dupoux, E. (2022). DP-Parser: Finding Word Boundaries from Raw Speech with an Instance Lexicon. *Transactions of the Association for Computational Linguistics*, 10, 1051–1065. https://doi.org/10.1162/tacl_a_00505

- Altoosar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & Heuvel, H. (2010). *A speech corpus for modeling language acquisition: CAREGIVER*.
- 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.
- Bergelson, E., & Swingle, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.
- Bernstein Ratner, N. (1987). The phonology of parent child speech. In K. Nelson & A. Van Kleeck (Eds.), *Children's Language*, 6.
- Bhati, S., Villalba, J., Želasko, P., & Dehak, N. (2020). Self-expressing autoencoders for unsupervised spoken term discovery. *ArXiv Preprint ArXiv:2007.13033*.
- Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, *37*(3), 487–511. <https://doi.org/10.1017/S030500090999050X>
- Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304.
- Brent, M. R. (1999). Efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1), 71–105. <https://doi.org/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.
- Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.
- Caines, A., Altmann-Richer, E., & Buttery, P. (2019). The cross-linguistic performance of word segmentation models over time. *Journal of Child Language*, *46*(6), 1169–1201. <https://doi.org/10.1017/S0305000919000485>
- Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Lexical segmentation: The role of sequential statistics in supervised and un-supervised models. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society: Atlanta, Georgia 1994*, 136–141.
- Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*(2), 111–153. <https://doi.org/10.1006/cogp.1997.0649>
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268.
- Cole, R. A., & Jakimik, J. (1980). A model of speech perception. *Perception and Production of Fluent Speech*, *133*(64), 133–142.
- Coltekin, C. (2011). *Catching words in a stream of speech: Computational simulations of segmenting transcribed child-directed speech* [PhD Thesis]. s.n.
- Çöltekin, Ç. (2017). Using Predictability for Lexical Segmentation. *Cognitive Science*, *41*(7), 1988–2021. <https://doi.org/10.1111/cogs.12454>
- Çöltekin, Ç., & Nerbonne, J. (2014). An explicit statistical model of learning lexical segmentation using multiple cues. *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLl)*, 19–28.
- Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and Speech*. Harvard University Press.
- Cutler, A. (1996). *Prosody and the word boundary problem*.
- Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3–4), 133–142. [https://doi.org/10.1016/0885-2308\(87\)90004-0](https://doi.org/10.1016/0885-2308(87)90004-0)
- Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics*, *21*(1–2), 103–108. [https://doi.org/10.1016/s0095-4470\(19\)31323-3](https://doi.org/10.1016/s0095-4470(19)31323-3)
- Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online (v2020.3)*. Zenodo. <https://doi.org/10.5281/zenodo.7385533>
- Dunbar, E., Hamilakis, N., & Dupoux, E. (2022). Self-Supervised Language Learning From Raw Audio: Lessons From the Zero Resource Speech Challenge. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1211–1226. <https://doi.org/10.1109/JSTSP.2022.3206084>
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

- Elsner, M., & Shain, C. (2017). Speech segmentation with a neural encoder model of working memory. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1070–1080. <https://doi.org/10.18653/v1/d17-1112>
- Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do Infants Really Learn Phonetic Categories? *Open Mind*, 5, 113–131. https://doi.org/10.1162/opmi_a_00046
- Feng, H., Chen, K., Deng, X., & Zheng, W. (2004). Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1), 75–93.
- Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. *Proceedings of ACL-08: HLT*, 130–138.
- Gambell, T., & Yang, C. (2006). Word segmentation: Quick but not dirty. *Unpublished Manuscript*.
- Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2), 248–265.
- Gleitman, L. R., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representations for language learning. *Linguistics: The Cambridge Survey: Volume 3, Language: Psychological and Biological Aspects*, 150–193.
- Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1), 21–54. <https://doi.org/10.1016/j.cognition.2009.03.008>
- Greenberg, J. H., & Jenkins, J. J. (1966). Studies in the psychological correlates of the sound system of American English. *Word*, 22(1–3), 207–242.
- Hammarström, H., Forkel, R., Hasepalm, M., & Bank, S. (2022). *Glottolog database 4.6*. <https://doi.org/10.5281/ZENODO.6578297>
- Harris, Z. S. (1955). From Phoneme to Morpheme. *Language*, 31(2), 190–222.
- 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. <https://doi.org/10.1006/jmla.2000.2755>
- Jusczyk, P. W. (1999). How infants begin to extract words from speech. In *Trends in Cognitive Sciences* (Vol. 3, Issue 9, pp. 323–328). Elsevier Current Trends. [https://doi.org/10.1016/S1364-6613\(99\)01363-7](https://doi.org/10.1016/S1364-6613(99)01363-7)
- 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. <https://doi.org/10.1111/j.1467-8624.1993.tb02935.x>
- Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' Sensitivity to the Sound Patterns of Native Language Words. *Journal of Memory and Language*, 32(3), 402–420. <https://doi.org/10.1006/jmla.1993.1022>
- Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, 277(5334), 1984–1986.
- Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61(8), 1465–1476. <https://doi.org/10.3758/BF03213111>
- Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The Beginnings of Word Segmentation in English-Learning Infants. *Cognitive Psychology*, 39(3–4), 159–207. <https://doi.org/10.1006/cogp.1999.0716>
- Kamper, H. (2023). Word Segmentation on Discovered Phone Units With Dynamic Programming and Self-Supervised Scoring. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 684–694. <https://doi.org/10.1109/TASLP.2022.3229264>
- Kamper, H., Jansen, A., & Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(4), 669–679. <https://doi.org/10.1109/TASLP.2016.2517567>
- Kamper, H., Jansen, A., & Goldwater, S. (2017). A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46, 154–174. <https://doi.org/10.1016/j.csl.2017.04.008>
- Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3), 400–401.
- King, G. (2018). *The Routledge Concise Compendium of the World's Languages*.
- Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation*, 108(2), 212–261.
- Liu, L., & Kager, R. (2017). Statistical learning of speech sounds is most robust during the period of perceptual attunement. *Journal of Experimental Child Psychology*, 164, 192–208. <https://doi.org/10.1016/j.jecp.2017.05.013>

- Ma, J., Çöltekin, Ç., & Hinrichs, E. (2016). Learning phone embeddings for word segmentation of child-directed speech. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, 53–63.
- MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*.
- Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**(1), 29–63. [https://doi.org/10.1016/0010-0285\(78\)90018-X](https://doi.org/10.1016/0010-0285(78)90018-X)
- Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, **78**(2), 91–121. [https://doi.org/10.1016/S0010-0277\(00\)00109-8](https://doi.org/10.1016/S0010-0277(00)00109-8)
- Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and Prosodic Effects on Word Segmentation in Infants. *Cognitive Psychology*, **38**(4), 465–494. <https://doi.org/10.1006/cogp.1999.0721>
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**(1), 1–86.
- McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, **152**(6), 3819–3842. <https://doi.org/10.1121/10.0016614>
- Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, **20**(3), 298–305.
- Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, **37**(3), 545.
- Moran, S., & Blasi, D. (2014). Cross-linguistic comparison of complexity measures in phonological systems. *Measuring Grammatical Complexity*, 217–240.
- Narasimhamurthy, A. (2005). Theoretical bounds of majority voting performance for a binary classification problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(12), 1988–1995.
- Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics*, **33**(2), 359–367.
- Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non) words,(non) words,(non) words: Evidence for a protollexicon during the first year of life. *Developmental Science*, **16**(1), 24–34.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.
- 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. <https://doi.org/10.1016/j.cognition.2009.07.011>
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, **108**(9), 3526–3529.
- Pimentel, T., Roark, B., & Cotterell, R. (2020). Phonotactic Complexity and Its Trade-offs. *Transactions of the Association for Computational Linguistics*, **8**, 1–18. https://doi.org/10.1162/tacl_a_00296
- Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, **45**(1), 89–95.
- Räsänen, O., & Blandón, M. A. C. (2020). Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics. *ArXiv Preprint ArXiv:2008.00731*.
- Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, **171**, 130–150.
- Rytting, C., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: Subsegmental variation in segmental cues. *Journal of Child Language*, **37**, 513–543. <https://doi.org/10.1017/S0305000910000085>
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical cues in language acquisition: Word segmentation by infants. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, 376–380.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996b). Statistical learning by 8-month-old infants. *Science*, **274**(5294), 1926–1928. <https://doi.org/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. <https://doi.org/10.1006/jmla.1996.0032>
- Saksida, A., Langus, A., & Nespore, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, 20(3), e12390.
- Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7), e2001844118. <https://doi.org/10.1073/pnas.2001844118>
- Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, 11(3), 407–413. <https://doi.org/10.1111/j.1467-7687.2008.00685.x>
- Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81, 105–120. <https://doi.org/10.1016/j.jml.2015.02.001>
- Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4), 336–351.
- Suomi, K. (1993). An outline of a developmental model of adult phonological organization and behaviour. *Journal of Phonetics*, 21(1–2), 29–60.
- Suomi, K., McQueen, J. M., & Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language*, 36(3), 422–444. <https://doi.org/10.1006/jmla.1996.2495>
- 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.
- VanDam, M., Warlaumont, A., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*, 37(2), 128–142. <https://doi.org/10.1055/s-0036-1580745>
- Venkataraman, A. (2001). A Statistical Model for Word Discovery in Transcribed Speech. *Computational Linguistics*, 27(3), 351–372. <https://doi.org/10.1162/089120101317066113>
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2), 260–269. <https://doi.org/10.1109/TIT.1967.1054010>
- Witten, I. H., & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4), 1085–1094.
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131(1), 3.