BabyLM's First Words: Word Segmentation as a Phonological Probing Task

Zébulon Goriely 🍊 **Paula Buttery** 🍊🍋
🍊 Department of Computer Science & Technology, University of Cambridge, U.K.
🍋 ALTA Institute, University of Cambridge, U.K.
🍊 firstname.secondname@cl.cam.ac.uk

Abstract

Language models provide a key framework for studying linguistic theories based on prediction, but phonological analysis using large language models (LLMs) is difficult; there are few phonological benchmarks beyond English and the standard input representation used in LLMs (subwords of graphemes) is not suitable for analyzing the representation of phonemes. In this work, we demonstrate how **word segmentation** can be used as a phonological probing task, allowing us to study the representations learned by phoneme-based language models trained on child-directed speech across 31 languages. Following computational models of word segmentation, we present unsupervised methods for extracting word boundaries from a trained model using the observation that prediction-error peaks at the start of words. We also use linear probes to identify that these models implicitly track word boundaries, even when they do not appear in training. This cross-lingual work corroborates statistical learning theories of acquisition and empirically motivates new methods for training subword tokenizers.

1 Introduction

Small models trained on developmentally plausible data have led to numerous advancements across pre-training strategies, architectures and tools for linguistic analysis (Hu et al., 2024). Yet most of this work involves training on English orthographic data with subword tokenization, restricting the ability to study phonological representations and word learning. A few recent studies have demonstrated that these so-called "BabyLMs" can be trained on individual phonemes (Goriely et al., 2024; Bunzeck et al., 2024), supporting phoneme-based phonological analysis. However, the majority of this work continues to center on English, in part due to the lack of phonological benchmarks for other languages.

Figure 1: Three strategies for unsupervised word segmentation using cues extracted from an auto-regressive language model trained to predict phonemes.

In this work, we explore the phonological capabilities of phoneme-based BabyLMs across 31 languages using the **word segmentation task**. Following computational models of word segmentation studies in the acquisition literature, we investigate models by assessing their ability to correctly place word boundaries in a sequence of phonemes when word boundaries are not provided during training. Successful segmentation indicates implicit phonological knowledge and when performed zero-shot on developmentally plausible data, contributes to statistical learning theories of language acquisition.

In some of the earliest sequential models, it was noted that *prediction-error* (the degree to which the model struggles to predict the next token) often corresponded with word boundaries (Elman, 1990). Using this observation, we identify four word boundary cues that can be extracted from trained models and three unsupervised strategies for placing boundaries using these cues, as illustrated in fig. 1. We additionally follow the supervised approach of Hahn and Baroni (2019), training linear probes on final layer embeddings to deter-

mine if word boundaries are implicitly tracked in order to improve phoneme prediction.

We train phoneme-based BabyLMs on the phonemic transcriptions of child-centered speech comprising the IPA CHILDES dataset (Goriely and Buttery, 2025). We find that these models implicitly encode word boundaries across all 31 languages and identify two factors that may provide useful priors depending on the language: the length of words and the distribution of phonemes at the end of words.

We discuss the validity of orthographic word boundaries as gold labels and note the similarities between our results and recent work that uses byte-level prediction entropy to improve the tokenization step in large language model (LLM) pretraining (Pagnoni et al., 2024). We conclude that this framework not only supports the study of distributional phonology and acquisition, but could also have implications for improving the efficiency and robustness of LLMs.

Finally, we release our code and pre-trained models to facilitate future work.

2 Related Work and Motivations

Since their inception, language models have been used to study the structures of language and explore mechanisms that humans may use to learn them.

Early "connectionist" language models were trained on sequences of letters or phonemes, often using developmentally plausible data in order to explore theories of word learning and phonology (Seidenberg and McClelland, 1989; Norris, 1994; Coltheart et al., 2001). Modern *large* language models (LLMs) are still probed for grammatical information, but standard benchmarks are generally based on higher-order structures: syntax and semantics rather than morphology and phonology. This is due to LLM design being optimized for downstream tasks, not linguistic analysis. For instance, LLMs are typically trained on graphemic text using subword tokens. While this representation is practical for large-scale training, these tokens are not very cognitively plausible (Beinborn and Pinter, 2023), are less effective than character-based tokens for learning word structure (Bunzeck and Zarrieß, 2025) and cannot be used to explore representations of phonological units. Additionally, modern LLMs are inappropriate for theories of acquisition, due to the scales of data they are trained on (Warstadt et al., 2023).

Here, we are interested in evaluating models that train directly on individual phonemes, without word boundaries. When trained on individual words, phoneme LMs have been used to study the acquisition of morphological rules (Kirov and Cotterell, 2018) and compare phonotactic complexity across languages (Pimentel et al., 2020). When trained on running text, phoneme LMs have been used for text-to-speech (Li et al., 2023) and lyric generation (Ding et al., 2024). When compared to grapheme-based models on standard linguistic benchmarks, phoneme models slightly underperform (Nguyen et al., 2022; Bunzeck et al., 2024) but this could be attributed to pre-processing, punctuation and the fact that LLM architectures and evaluation sets have been optimized for written text (Goriely et al., 2024). Despite the benefits of phoneme-based training, phonological evaluation is limited, and few phoneme LMs exist beyond English. Goriely and Buttery (2025) trained phoneme LMs on child-directed speech across 11 languages, but were only able to use an English benchmark for studying how phonological and syntactic knowledge scales in phoneme LMs.

In this work, we propose the word segmentation task as a language-independent method for probing the representations learned by phoneme LMs. Below, we summarize past approaches for investigating the phonological capabilities of language models. We then give historical background on the word segmentation task. Finally, we discuss past examples of word segmentation being used as a probing task.

2.1 Phonological Evaluation of LLMs

While many studies have explored the representations learned by phoneme LMs trained on individual words, there are very few benchmarks for phoneme LMs trained on running text.

One method for testing phonology is to use minimal pairs of words and pseudowords as a lexical decision task. One benchmark that uses this approach is BabySLM (Lavechin et al., 2023), which provides a lexical decision metric for phoneme LMs or speech LMs (which learn directly from audio) using a vocabulary based on child-directed speech. Bunzeck et al. (2025) use a similar approach in order to compare grapheme LMs to phoneme LMs. They also use two probing tasks to examine the representations of sentences; age prediction and rhyme prediction.

PhonologyBench (Suvarna et al., 2024) is a benchmark that uses prompts to test chat-based English LLMs. However, by using prompts, they treat phonology as an emergent ability tested through metalinguistic judgment, an evaluation strategy which Hu and Levy (2023) argues is inferior to using quantities directly derived from a model's representations.

These benchmarks also only test English models, in part due to the lack of phoneme LMs in other languages, but also due to a lack of resources for constructing phonological tasks. For example, pseudowords are typically generated using `wuggy` (Keuleers and Brysbaert, 2010), which only supports three languages for phonetic pseudoword generation. An example of language-independent evaluation of phoneme LMs is the phonetic feature probe used in Goriely and Buttery (2025), which only requires feature vectors for each IPA symbol. The word segmentation task requires no language-specific data, only utterances labeled with word boundaries.

2.2 Computational Models of Segmentation

Unlike in written text, where lexical units are separated by spaces and punctuation, spoken communication consists of continuous utterances with no clear demarcation of words (see, e.g. Cole and Jakimik, 1980). Somehow, without a lexicon to consult, children are able to segment speech into words and phrasal units by the age of six months (Jusczyk, 1999). How children learn to segment words and bootstrap their lexicon is known in psycholinguistics as the *word segmentation problem*, and statistical learning experiments have established a wide variety of statistical cues which children may use to segment speech (Cutler and Carter, 1987; Gleitman et al., 1988; Jusczyk et al., 1993; Saffran et al., 1996b; Jusczyk et al., 1999a; Suomi et al., 1997).

Particularly influential were the experiments of Saffran et al. (1996a), who established that 8-month-old children use distributional information to segment speech, specifically noting that low conditional probability between two adjacent syllables often indicated a word boundary. These experiments inspired the development of computational models proposing cognitively plausible learning mechanisms for word segmentation, most of which are based on the principle that units within words are far more predictable than units across word boundaries (Harris, 1955). Many models draw on Brent (1999), who use unigram statistics to segment speech, with later models using higher-order n-grams (Venkataraman, 2001), incorporating phonological constraints (Blanchard et al., 2010) or leveraging prior distributions over word frequencies and phonological shapes (Goldwater et al., 2009). Other models explicitly calculate several statistical cues at each potential word boundary and combine cues using a majority voting framework (Çöltekin and Nerbonne, 2014; Çöltekin, 2017; Goriely et al., 2023). Each cue provides a signal over the utterance (as illustrated in fig. 1) with peaks in each cue indicating a potential boundary.

Peaks in predictability can also be observed in neural language models. In the foundational work of Elman (1990), a simple recurrent network (SRN) is trained to predict letters in an unsegmented sequence (one of the first examples of auto-regressive language modeling). Elman observes that the prediction-error increases at the onset of each new word, concluding that "there is information in the signal that could serve as a cue to the boundaries of linguistic units which must be learned".

Christiansen et al. (1998) later used an SRN to segment speech by using the probability of an *utterance* boundary, rather than prediction-error, to place word boundaries. This followed previous work suggesting that children could use utterance boundaries to bootstrap their lexicon (Aslin et al., 1996) and is a cue used in the models of Çöltekin and Nerbonne (2014); Goriely et al. (2023).

In this study, we combine ideas from past computational models for word segmentation. Rather than explicitly calculate n-gram statistics, our cues are based on prediction-error and utterance boundary probability extracted from LLMs trained on the next-phoneme prediction task. As these cues are based on the language model's prediction of phonemes, successful segmentation indicates that implicit phonological knowledge of word-like units in these models.

While our experimental setup draws on previous computational work in word segmentation, we do not claim that our phoneme-level language models simulate child language acquisition (see section 6). Rather, we use the segmentation task — with phoneme-level input — as a diagnostic tool that allows us to characterize the cross-linguistic distributional structure of speech sounds and test whether

language models naturally cluster sequences into units that coincide with our notion of word-hood. Although our findings may support aspects of statistical learning theories, we acknowledge the limitations of using phoneme-based representations in appendix A.

2.3 Probing for Word Boundaries

Previous work has explored the representations of word boundaries in LLMs. Sanabria et al. (2021) explored methods for extracting word boundaries from attention weights in an LSTM, finding that attention had limited value for segmentation. Hahn and Baroni (2019) trained character-level RNNs and LSTMs without word boundaries, finding that individual activations correlated with word boundaries and that a linear probe trained on all activations also identified boundaries. They claimed that removing word boundaries resulted in a 'near tabula rasa' training paradigm but trained on billions of graphemic words Wikipedia, which is not developmentally plausible. Here, we use this probe on the final layer of phoneme LMs trained on developmentally plausible data, a more 'tabula rasa' paradigm.

Other studies have verified Elman's observations that prediction-error corresponds with word boundaries. For instance, Al-Rfou et al. (2019) train a 64-layer character-level transformer and in qualitative analysis note that three measures of prediction-error sharply increase at the start of words. However, their model is trained on graphemic text from Wikipedia without removing the word boundaries and they do not explicitly use these measures to evaluate word segmentation performance. Here, we use their three measures to propose an unsupervised word segmentation algorithm using phoneme LMs trained without word boundaries.

3 Word Segmentation Task

We use the *word segmentation task* as a zero-shot method for studying the phonological properties of language models trained on phoneme sequences. Given a list of utterances, each of which consists of a non-delimited phoneme sequence, the task is to produce a *segmentation* of each utterance by using an unsupervised method for placing word boundaries. For instance, given the utterance "what do you see", represented phonemically as wʌtduːyuːsiː, successful segmentation would return wʌt duː yuː siː, as demonstrated in fig. 1. Note that phonemes are individual tokens (e.g. uː is a single token, not two) and, crucially, word boundaries are removed during training, although utterance boundaries are present.

Our method for unsupervised word segmentation is based on the observation made by Elman (1990), that cues for word boundaries can be extracted from a sequence prediction model. Given a language model that at each position i provides the probability of a phoneme x given a context $x_1 \dots x_{i-1}$, we extract the following four cues at each potential boundary position:

- **Entropy:** The entropy (in bits) across the probabilities for all items in the vocabulary.
- **Loss:** The cross-entropy loss (bits) calculated as the negative log probability of the subsequent phoneme p_i.
- **Rank:** The rank of x_i in the list of possible tokens at position i sorted by likelihood.
- **Utterance Boundary Probability (UBP):** The probability assigned to the utterance boundary token.

The first three cues are put forward by Al-Rfou et al. (2019), where they are used to qualitatively examine the error rate of their character-based language model. Our use of these cues for word segmentation is novel. The fourth cue, UBP, relates to the model of Christiansen et al. (1998), who found that the prediction of the utterance boundary marker in a SRN increased at word boundaries. All four cues are utilized in the segmentation models of Çöltekin and Nerbonne (2014); Goriely et al. (2023) but rather than being explicitly calculated using n-gram frequencies, we calculate them using the probability distribution produced by a language model.

For each of these cues, we have three methods for placing word boundaries. The first is to identify peaks in each cue: placing word boundaries whenever the cue's value is higher at position i than at position $i-1$ or $i+1$ in the sequence. The second is to learn a single threshold value, placing word boundaries when the cue exceeds it. The third combines both strategies, placing word boundaries when the relative increase of the cue's value from position $i-1$ to i exceeds a learned threshold. We call these the **peak**, **threshold** and **relative** strategies, respectively, as illustrated in fig. 1. We acknowledge that the threshold and relative strategies

Suite Size	Model Parameters	Tokens (words)	Languages
Tiny	400k	100k (~20k)	31
Small	600k	700k (~180k)	17
Medium	5M	1.8M (~500k)	11
Large	19M	18M (~5M)	1

Table 1: The model size in number of (non-embedding) parameters and data size used for each suite of models. Languages are sub-sampled according to the token count for consistency, as word length varies across languages.

are not fully unsupervised, using a single learned parameter.

Finally, in order to explore whether word boundary information is present in the model's representations, we follow Hahn and Baroni (2019) and train a linear probe to predict word boundaries from the final layer embeddings. We implement their 'balanced' probe, training on embeddings taken from an equal number of word-final and word-internal positions, and ensure that no words in the training set are contained in the test set.

4 Experimental Setup

We train a suite of GPT-2 models on each of the 31 languages in the IPA CHILDES corpus. As the size of each subset varies considerably,[1] for a fair comparison we must subsample our training data to the size of the smallest subset and use a very small model to prevent over-fitting. In order to explore the use of larger models and more training data, we train four suites of models, each using a different sample size and model size, setting model parameters according to the scaling experiments of Goriely and Buttery (2025). These suites are detailed in table 1 with parameter configurations and training parameters given in appendix B. The smallest model (only 2 layers) is trained on 100k tokens from all 31 languages, and the largest model (6 layers) is trained on 18M tokens of English.

For the linear probes, we follow Hahn and Baroni (2019) and report accuracy. They claim that chance performance is 50% due to the balanced training data, but our results suggest otherwise. In order to evaluate our unsupervised strategies, we follow past work (see section 2.2) compute the F1 score of boundary placement, excluding boundaries placed at the start and end of utterances (as these are 'free' from the utterance boundaries).

[1]The North American English section contains 10M words but Farsi only contains 40k.

5 Results

We present the results of the word boundary probe in fig. 2 and the maximum boundary F1 scores of our unsupervised segmentation strategies in fig. 3. The individual scores for each combination of language, suite size, boundary cue and segmentation strategy are provided in appendix C.

Overall, both the word boundary probe and the unsupervised strategies successfully identify word boundaries — all probes achieve accuracies significantly higher than the untrained baseline, as do the unsupervised strategies (see appendix D for details on significance tests). The probe accuracies show that models implicitly track word boundaries in their contextual embeddings, suggesting that they are learning phonological rules to aid in next-phoneme prediction. The unsupervised segmentation results indicate that word boundaries can be extracted through prediction across many languages, corroborating previous statistical learning results about the role of distributional cues in language acquisition.

Below, we analyze these results in more detail.

180k words are sufficient for learning word boundaries. We note that across all languages, the accuracy of the word boundary probes increases from the Tiny suite to the Small suite (where models are trained on about 180k words, as seen in table 1), but improvements are minimal for models in the larger suites. This also occurs with the unsupervised approach, despite receiving several orders of magnitude more training data and training with many more parameters. We conclude that 180k words is sufficient for a model to learn word-like units in our framework, but other models may require more or less data.

Utterance boundaries are better predictors of word boundaries than prediction-error. Figure 3 provides the maximum boundary F1 score achieved for each model in each suite across the four boundary cues and three segmentation strategies, for a total of 12 combinations. In table 2 we summarize the cue and strategy combinations that achieved these scores. The UBP cue is the most effective in each suite, out-performing the three cues based on prediction-error, and the relative strategy out-performs the other two strategies. For reference, we give the best combinations for each language in appendix C. Generally, the best cue stays consistent across suites for a particular

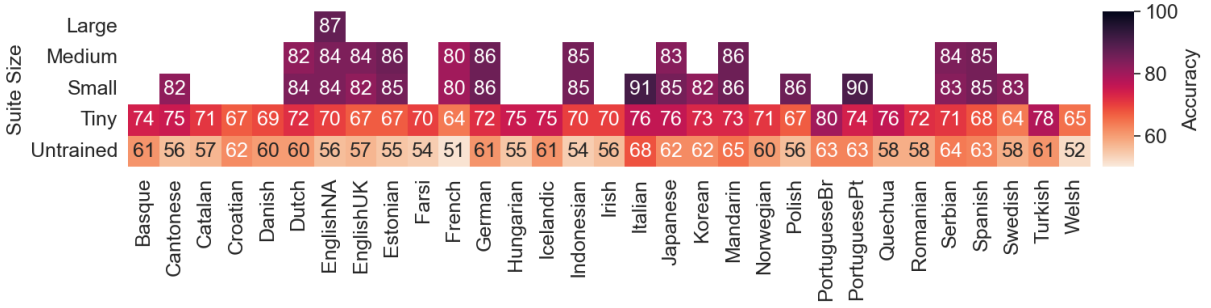

Figure 2: Accuracy scores for the word boundary probe trained on the contextual embeddings of phonemes across models in each suite. Training and test instances are balanced and each word used for training embeddings is removed from the test set. Probe results for each untrained model in the Tiny suite are included as a baseline.

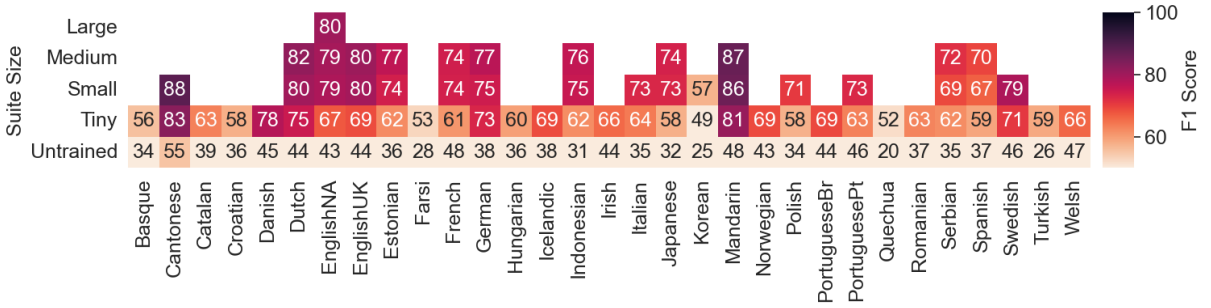

Figure 3: Boundary placement F1 scores achieved using the unsupervised segmentation strategies across models in each suite. For each score, we report the maximum across the 4 cues and 3 segmentation strategies. The Untrained row give the maximum scores achieved by each model in the Tiny suite before training.

Cue & Strategy	Tiny	Small	Medium	Large
UBP (threshold)	3	2	1	-
UBP (relative)	3	6	4	-
UBP (peak)	11	4	3	1
Entropy (threshold)	1	-	1	-
Entropy (relative)	-	4	2	-
Entropy (peak)	-	1	-	-
Loss (relative)	9	-	-	-
Rank (relative)	3	-	-	-
Rank (peak)	1	-	-	-

Table 2: Counts of the word boundary cues and segmentation strategies that achieved the highest F1 scores in each suite.

language (e.g. Entropy is the best cue for Italian), but this is not always the case, and the best strategy also varies.

The peak segmentation strategy fails to capture subsequent boundaries. We compare the four segmentation cues using the peak strategy segment utterances from the EnglishNA section of IPA CHILDES in fig. 4. We identify two failure modes for this strategy. The first is that since two peaks cannot directly follow one another, subsequent boundaries cannot both be successfully placed. In this example, the h in "help" is incorrectly placed by all four cues. A second failure case is that the relative size of peaks is not considered; three cues incorrectly place a boundary within the word "fingers" due to a very small peak at ə. The threshold and relative segmentation strategies address both of these issues but for English the peak strategy is still best overall.

Italian has a strong prior for learning word boundaries. Hahn and Baroni (2019) claim that since the probes are trained on balanced examples, chance accuracy should be 50%. However, we find that the probes trained on completely untrained models (see fig. 2) achieve accuracies ranging from 51% for French up to 68% for Italian. This is because the balancing procedure does not account for the fact that phonemes have different probability distributions depending on their position within words. For example, in fig. 5 we find that at the end of Italian words, a small number of phonemes have particularly high frequencies (the vowels ɐ, o, e and i end 84% of words) whereas the distribution of French word-final phonemes is not as skewed. This skewed distribution provides a useful prior for the Italian probe, which can achieve high accuracies

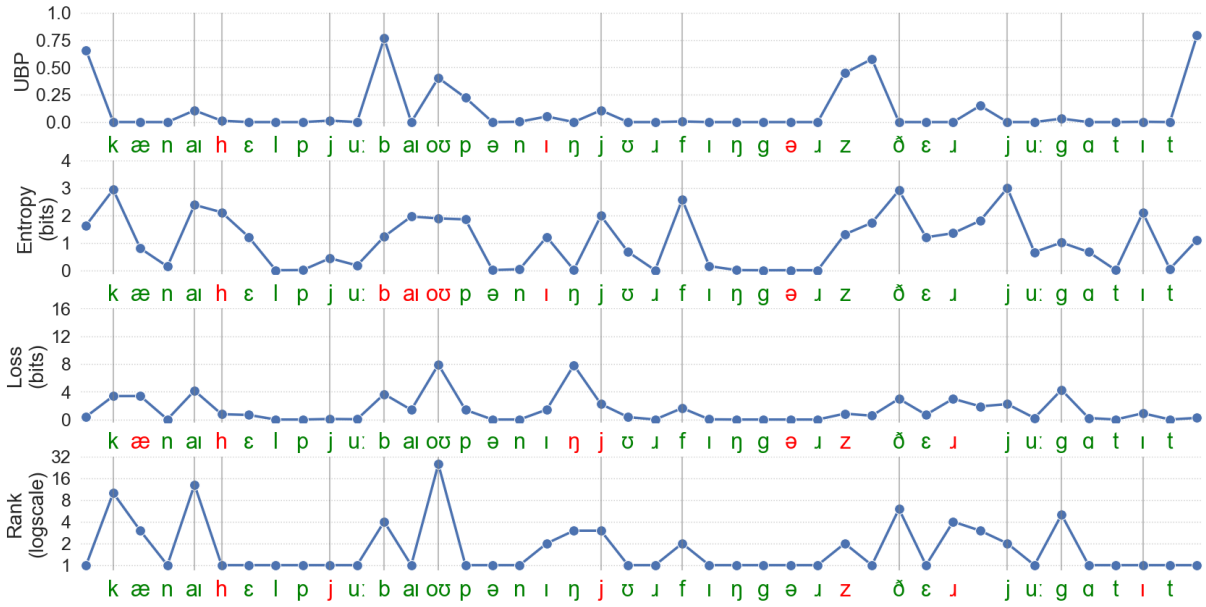

Figure 4: Per-phoneme boundary probability, entropy, loss and rank assigned by the Medium English model for the sequence of utterances "can I help you by opening your fingers", "there", "you got it". Spaces indicate utterance boundaries, vertical lines indicate gold word boundaries and phonemes are marked as green if they are correctly identified as word boundaries using the **peak** strategy or if they follow an utterance boundary (red otherwise).

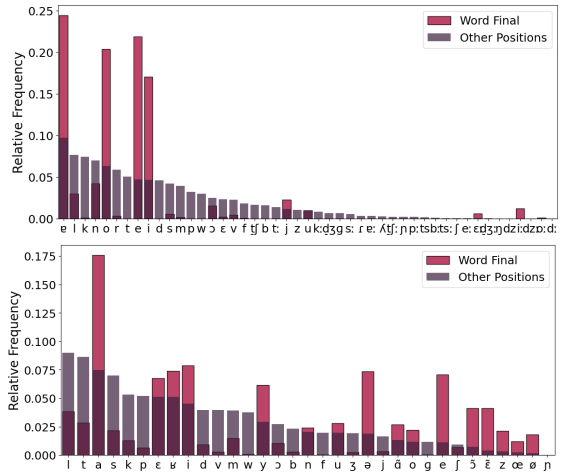

Figure 5: Relative frequencies of phonemes appearing in word-final positions and all other positions for Italian (top) and French (bottom).

by relying on these phoneme frequencies (the only signal available when using embeddings from an unsupervised model). To measure the relative benefit of each prior, we can compute the **normalized entropy** of the word-final phoneme distributions in each language,

$$H_{\text{norm}} = \frac{H(P)}{H_{\text{max}}} = \frac{\sum_{i=1}^{n} p_i \log_i p_i}{log_2(n)},$$

which ranges from 0 (deterministic distribution) to 1 (uniform distribution). We find that not only do Italian and French have the lowest and highest normalized entropies with 0.51 and 0.84, respectively, but in general, this normalized entropy has a high negative correlation with probe accuracy for the untrained models (Pearson $\rho = -0.69$). This correlation is still present for the Tiny suite (Pearson $\rho = -0.52$) but is not significant for the Small and Medium suites, indicating that although the word-final phoneme distribution prior is useful, the embeddings do still encode information about word boundaries that the probes can detect.

Word length is a confounding factor for unsupervised segmentation. Just as with the probes, using our unsupervised methods on untrained models can reveal confounding factors, as shown in fig. 3. The F1 scores for the untrained models range from 20 for Quechua up to 55 for Cantonese. For 25 of the 31 languages, this score comes from the UBP cue with the relative strategy; since the probability of an utterance boundary from an untrained model will randomly vary over the phoneme sequence, boundary placement using the relative strategy essentially places boundaries randomly, which can still yield relatively high F1 scores if words are short. This seems to be the case here; Quechua has the highest average word length in IPA CHILDES and Cantonese has the lowest, with

6.2 and 2.4 phonemes per word, respectively. Generally, we find that word length has a high negative correlation with the F1 scores with Pearson $\rho = -0.94, -0.71, -0.79, -0.42$ for the Untrained, Tiny, Small and Medium suites, respectively (although the final correlation is not significant).

This confounding factor means that we cannot easily compare word segmentation scores between languages, only scores for each language across suite sizes. Compared to the untrained models, the unsupervised word segmentation strategy still achieves significantly higher F1 scores for every language, demonstrating that distributional information is a useful cue for bootstrapping a lexicon.

6 Discussion

In this work, we train BabyLMs on phonemic transcriptions of 31 languages in IPA CHILDES and explore the word segmentation task as a method for probing these models for phonological knowledge. Our results indicate that prediction-error and utterance boundary probability can be used as cues for unsupervised word segmentation. Our study is the first to use prediction-error extracted from LLMs for unsupervised word segmentation, extending previous work that explicitly calculated these cues using n-gram models (Çöltekin and Nerbonne, 2014; Çöltekin, 2017; Goriely et al., 2023). We also update previous neural models of word recognition (Elman, 1990; Christiansen et al., 1998) by using modern architectures and evaluating cross-lingually. We now turn to the broader implications of our findings.

Statistical learning. Viewing our models as statistical learners, we find that no single cue or strategy consistently yields the best segmentation performance across different model sizes and languages. This is perhaps unsurprising, as many of the cues are highly interrelated (for example, entropy and surprisal often correlate) and all segmentation strategies are grounded in the same underlying principle: identifying boundaries at points of high prediction uncertainty. It is this general principle, rather than any specific cue or strategy, that proves sufficient for segmenting utterances into word-like units. Nevertheless, most cues and strategies perform reasonably well on their own. Previous segmentation models have explored combining multiple distributional cues through unsupervised majority voting (Çöltekin, 2017; Goriely et al., 2023), an approach that could be fruitfully applied to the cues investigated here in future work.

Cross-lingual comparison. Comparing models across languages is a challenge. Our study is the first cross-lingual study using the word segmentation task to compare 31 languages, but we identify two confounding factors that inhibit cross-lingual comparison. Firstly, we find that the distribution of phonemes in word-final slots provides a prior not previously accounted for in studies that probed contextual embeddings for word boundary information. Secondly, we find that word length provides a prior for the unsupervised strategies, since randomly placing boundaries yields a higher F1 score when words are shorter, which has not previously been accounted for in cross-lingual word segmentation studies. Nevertheless, both the probes and the unsupervised strategies achieve significant scores for all 31 languages, indicating the importance of the distributional cue for learning to segment speech in any language. These findings also highlight the importance of accounting for frequency information as a prior when training probes or comparing models trained on different datasets.

Simulating acquisition. Our results focus on the performance of our models at the end of training, whereas past work has compared the learning dynamics of phoneme-based models to developmental patterns observed in human acquisition (Kirov and Cotterell, 2018). Although our findings indicate the utility of the distributional cue for identifying word-like units, we do not claim that our models simulate language acquisition. In particular, given recent advances in models that operate directly on raw audio, the use of phoneme-level representations may be insufficient for capturing the full complexity of language learning, as discussed in appendix A.

Rather, we use this framework to investigate the distributional patterns of phonemes across languages and whether language models trained to predict upcoming phonemes implicitly track meaningful sub-sequences that align with words. While many computational models of word segmentation treat segmentation as a necessary precursor for language understanding, this assumption has been questioned. For example, Baayen et al. (2016) show that a tri-phone model, operating on unsegmented utterances can make predictions consistent with infants' sensitivity to linguistic structure. Likewise, recent phoneme-level language models

perform well on both linguistic benchmarks and downstream tasks without explicit segmentation (Goriely et al., 2024) — although our results suggest that some degree of implicit segmentation may be occurring to enhance these models' predictive performance.

Word boundaries as gold labels. Throughout this work, we have used word boundaries from orthographic text as the gold labels for evaluation, but these boundaries may not correspond with lexical units in speech. In early stages of acquisition, children may treat both predictable multi-word phrases as single lexical units (MacWhinney, 1978) and unsupervised word segmentation strategies may be segmenting morphemes, rather than words (Fleck, 2008). From an information-theoretic angle, word boundaries may only exist to optimize the trade-off between syntax and morphology across languages (Koplenig et al., 2017; Mosteiro and Blasi, 2025) and in general, what exactly defines a 'word' is still up for debate (Dixon and Aikhenvald, 2003; Haspelmath, 2023).

Unsupervised segmentation for tokenization. Instead of evaluating against word boundaries, we can treat our cues as *graded* measures of co-occurrence statistics, as noted by Elman (1990). This idea can be leveraged to improve the tokenization step in modern LLM pre-training. Instead of forming subwords by merging frequently occurring byte pairs, token sequences that are highly predictable can be combined. Pagnoni et al. (2024) apply this concept to a "token-free" model, where bytes are joined into 'patches' according to the entropy of the probability distribution for each byte (probabilities are computed using a byte-level LLM). They use two constraints for merging bytes which exactly correspond to our threshold and relative segmentation strategies, but only use entropy as a cue. In our experiments, entropy was less effective than utterance boundary probability (UBP) for unsupervised word segmentation and in an initial investigation (see appendix E) we found that creating a subword tokenizer using both cues improves the linguistic abilities of models trained on phonemes compared to regular BPE and that the UBP cue is more effective than entropy. This creates a parallel between word segmentation research and practical applications for tokenization in NLP and we encourage further work in this area.

7 Conclusion

Phoneme-level language models trained on developmentally plausible corpora are a valuable tool for studying cross-lingual phonology and theories of acquisition. In this study, we demonstrate how the **word segmentation task** can be used to probe these models for phonological knowledge and introduce novel unsupervised methods leveraging prediction-error and utterance boundary probability to identify words. Our findings show that models trained on 31 languages can all detect word boundaries; however, cross-linguistic comparisons are influenced by confounding factors such as word length and word-final phoneme distribution. These factors, while positing challenges, also offer new avenues for understanding the role of distributional cues in language processing cross-lingually. Finally, we explore the connection between word segmentation and information-driven tokenization schemes, highlighting how this research can inform and improve practical applications in natural language processing.

Acknowledgments

We are grateful to Pietro Lesci and Julius Cheng for their careful reading of this article and their insightful feedback, which greatly contributed to its improvement.

Our experiments were performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council. Zébulon Goriely is supported by an EPSRC DTP Studentship.

References

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166.

Richard N Aslin, Julide Z Woodward, Nicholas P LaMendola, and Thomas G Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In *Signal to syntax*, pages 117–134. Psychology Press.

R. Harald Baayen, Cyrus Shaoul, Jon Willits, and Michael Ramscar and. 2016. Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, 31(1):106–128.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Daniel Blanchard, Jeffrey Heinz, and Roberta Golinkoff. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, 37(3):487–511.

Michael R. Brent. 1999. Efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1):71–105.

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2024. Graphemes vs. phonemes: Battling it out in character-based language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 54–64.

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2025. Small language models also work with small vocabularies: Probing the linguistic abilities of grapheme- and phoneme-based baby llamas. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6039–6048, Abu Dhabi, UAE. Association for Computational Linguistics.

Bastian Bunzeck and Sina Zarrieß. 2025. Subword models struggle with word learning, but surprisal hides it.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. Call for papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

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

Ronald A Cole and Jola Jakimik. 1980. A model of speech perception. *Perception and production of fluent speech*, 133(64):133–42.

Çağrı Çöltekin and John Nerbonne. 2014. An explicit statistical model of learning lexical segmentation using multiple cues. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pages 19–28.

Çağrı Çöltekin. 2017. Using Predictability for Lexical Segmentation. *Cognitive Science*, 41(7):1988–2021.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Anne Cutler and David M. Carter. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2(3-4):133–142.

Maureen de Seyssel, Marvin Lavechin, and Emmanuel Dupoux. 2023. Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language*, 50(6):1294–1317.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

R. M. W. Dixon and Alexandra Y. Aikhenvald. 2003. *Word: a typological framework*, page 1–41. Cambridge University Press.

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The zero resource speech challenge 2021: Spoken language modelling. In *Proc. Interspeech 2021*, pages 1574–1578.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Naomi H. Feldman, Sharon Goldwater, Emmanuel Dupoux, and Thomas Schatz. 2021. Do Infants Really Learn Phonetic Categories? *Open Mind*, 5:113–131.

Margaret M Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138.

Lila R Gleitman, Henry Gleitman, Barbara Landau, and Eric Wanner. 1988. Where learning begins: Initial representations for language learning. *Linguistics: The Cambridge Survey: Volume 3, Language: Psychological and Biological Aspects*, pages 150–193.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Zébulon Goriely, Andrew Caines, and Paula Buttery. 2023. Word segmentation from transcriptions of child-directed speech using lexical and sub-lexical cues. *Journal of Child Language*, pages 1–41.

Zébulon Goriely, Richard Diehl Martinez, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 37–53, Miami, FL, USA. Association for Computational Linguistics.

Zébulon Goriely and Paula Buttery. 2025. Ipa-childes & g2p+: Feature-rich resources for cross-lingual phonology and phonemic language modeling.

Michael Hahn and Marco Baroni. 2019. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics*, 7:467–484.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Martin Haspelmath. 2023. Defining the word. *Word*, 69(3):283–297.

Jennifer Hu and Roger P Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Peter W. Jusczyk. 1999. How infants begin to extract words from speech.

Peter W. Jusczyk, Anne Cutler, and Nancy J. Redanz. 1993. Infants' Preference for the Predominant Stress Patterns of English Words. *Child Development*, 64(3):675–687.

Peter W. Jusczyk, Elizabeth A. Hohne, and Angela Bauman. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61(8):1465–1476.

Peter W Jusczyk, Derek M Houston, and Mary Newsome. 1999b. The beginnings of word segmentation in english-learning infants. *Cognitive psychology*, 39(3-4):159–207.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42:627–633.

Christo Kirov and Ryan Cotterell. 2018. Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, 12(3):e0173614.

Marvin Lavechin, Maureen de Seyssel, Marianne Métais, Florian Metze, Abdelrahman Mohamed, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2024. Modeling early phonetic acquisition from child-centered audio data. *Cognition*, 245:105734.

Marvin Lavechin, Maureen De Seyssel, Hadrien Titeux, Hervé Bredin, Guillaume Wisniewski, Alejandrina Cristia, and Emmanuel Dupoux. 2022. Can statistical learning bootstrap early language acquisition? a modeling investigation.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592, Dublin, Ireland. ISCA.

Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023. Phoneme-level BERT for enhanced prosody of text-to-speech with grapheme predictions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Brian MacWhinney. 1978. The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, 43(1/2):1–123.

Bob McMurray. 2022. The myth of categorical perception. *The Journal of the Acoustical Society of America*, 152(6):3819–3842. Publisher: Acoustical Society of America.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Pablo Mosteiro and Damián Blasi. 2025. Word boundaries and the morphology-syntax trade-off. In *Proceedings of the New Horizons in Computational Linguistics for Religious Texts*, pages 86–93.

Tu Anh Nguyen, Maureen De Seyssel, Robin Algayres, Patricia Roze, Ewan Dunbar, and Emmanuel Dupoux. 2022. Are word boundaries useful for unsupervised language learning? *arXiv preprint arXiv:2210.02956*.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The zero resource speech benchmark 2021:

Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Dennis Norris. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, et al. 2024. Byte latent transformer: Patches scale better than tokens. *arXiv preprint arXiv:2412.09871*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996a. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996b. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4):606–621.

Ramon Sanabria, Hao Tang, and Sharon Goldwater. 2021. On the difficulty of segmenting words with attention. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 67–73.

Thomas Schatz, Naomi H Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 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.

Mark S Seidenberg and James L McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kari Suomi, James M. McQueen, and Anne Cutler. 1997. Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language*, 36(3):422–444.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Anand Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):350–372.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Limitations

We acknowledge the following limitations of our work.

Limitations of phonemic data: Using phonemic data for the word segmentation task is the typical framework for exploring relevant acquisition theories. However, the phonemic transcriptions in IPA CHILDES do have limitations. Having been generated using grapheme-to-phoneme (G2P) conversion, they may have been subject to conversion error, and the original transcriptions may also contain errors. The G2P process also removes natural variation in speech, such as accents and allophonic variation. The symbolic nature of phonemes may also be an unrealistic starting point for acquisition; it is unclear if infants have access to phonetic categories at this stage of acquisition (Feldman et al., 2021; McMurray, 2022). Researchers who advocate for using language models as cognitive models argue that the training data should be as developmentally plausible as possible (Dupoux, 2018; Warstadt and Bowman, 2022), and that phonemes may be as implausible as text for simulating early acquisition (Lavechin et al., 2023).

From this perspective, a more appropriate framework is to learn segmentation directly from raw audio, as pursued in the Zero Resource Speech Challenge (Nguyen et al., 2020; Dunbar et al., 2021). Audio-based models naturally incorporate prosodic cues, which play a key role in language acquisition (Cutler and Carter, 1987; Jusczyk et al., 1993, 1999b). Unsupervised models have demonstrated the ability to perform statistical learning directly from raw speech (Lavechin et al., 2022; de Seyssel et al., 2023), and have found that the resulting units tend to be shorter than phonemes, consistent with early perceptual categories (Schatz et al., 2021). While such models show promising signs of early phonetic learning and perform well on word-level tasks, they currently require significantly more data to match the performance of text-based models (Lavechin et al., 2023). Moreover, training on curated audiobook datasets gives these models a considerable advantage over learning from noisier, long-form audio that better resembles real-world input—but ongoing work is making such realistic simulations increasingly viable (Lavechin et al., 2024).

Distribution of languages: When training models cross-lingually, we were limited by the scale of each language partition of the IPA CHILDES dataset. The dataset has a very skewed distribution: the EnglishNA section contains 18M words but the Farsi section only contains 43k words. We addressed this skew by training four suites of models in order to provide a cross-lingual comparison while also exploring how segmentation performance increased in scale for the languages with more data available.

Language coverage: To the best of our knowledge, our work is the most cross-lingual exploration word segmentation to date, but is still limited in language coverage: the languages we compare are predominantly European and Asian, with no languages indigenous to the Americas, Australia or Africa. Word segmentation of languages that are more globally distributed should be explored in future work.

B Implementation Details

We conduct our experiments using the `PyTorch` framework (Paszke et al., 2019) and the `Transformers` library (Wolf et al., 2020).

B.1 Hardware Details

We use a server with one NVIDIA A100 80GB PCIe GPU, 32 CPUs, and 32 GB of RAM for all experiments. Below, we report a subset of the output of the *lscpu* command:

```
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Address sizes:         46 bits physical,
                       48 bits virtual
Byte Order:            Little Endian
CPU(s):                32
On-line CPU(s) list:   0-31
Vendor ID:             GenuineIntel
Model name:            Intel(R) Xeon(R)
                       Silver 4210R CPU
                       @ 2.40GHz
CPU family:            6
Model:                 85
Thread(s) per core:    1
Core(s) per socket:    1
Socket(s):             8
Stepping:              7
BogoMIPS:              4800.11
```

B.2 Model Parameters and Training Procedure

We describe the model and training parameters in table 3. The model parameters were chosen according to the scaling experiments of Goriely and Buttery (2025), who trained a suite of GPT-2 models

Parameter	Tiny	Small	Medium	Large
Layers	2	3	6	6
Heads	4	4	8	8
Dropout	0.3	0.3	0.3	0.1
Embedding Size	128	128	256	512
Inner Size	512	512	1024	2048
Max Example Length	128			
Learning Rate	0.001			
Optimizer	AdamW			
Scheduler Type	Linear			
Max Steps	200k			
Warm-up Steps	60k			
Per Device Batch Size	32			

Table 3: Hyperparameter settings for training the GPT-2 architecture in each suite. Vocabulary size varies according to the language, but all other parameters are constant across experiments. Where values are not reported, they may be assumed to be default values.

for different subsets of the English section of IPA CHILDES and used the lexical score in BabySLM (Lavechin et al., 2023) to determine the best parameters. We note that since these parameters were optimised for English, there may be better parameters for the other languages, but differences in perplexity between languages were generally larger than the differences in perplexity between models in the scaling experiments we reference.

Data is prepared into batches by first tokenizing the entire dataset, combining all tokens into one long vector, and then splitting the vector into chunks of 128 tokens. Only the very last example is padded, if required. At each step during training, random chunks are selected and combined into batches.

Checkpoints are taken every 20,000 steps during training. At each checkpoint, the perplexity is evaluated on the held-back evaluation set, and at the end of training the checkpoint with the lowest perplexity is returned as the best model. For the Tiny suite, many of the best models were from the very first checkpoint, since due to the small training dataset and small model, the model had already fit the data by this point.

C Full Word Segmentation Results

All boundary placement F1 scores for the Tiny, Small, Medium and Large suites are given in fig. 6, fig. 7, fig. 8 and fig. 9, respectively. The best combination of cue and segmentation strategy for each language is given in table 4.

D Significance Tests

All word boundary probes for a particular language are trained and tested on the same evaluation set. We compute significance between two probes using McNemar's Test (McNemar, 1947) over the predicted word boundaries for the evaluation set, with a significance threshold of $p < 0.05$. The same procedure is used when comparing the unsupervised methods.

E Using Word Segmentation Cues for Subword Tokenization

We briefly explore the use of our unsupervised word boundary cues to create a subword tokenizer. Typically, the vocabularies for these tokenizers are generated using methods like Byte-Pair Encoding (Sennrich et al., 2016), where the vocabulary initially consists of each individual byte, and pairs of bytes that frequently co-occur in a training dataset are 'merged' into a new token, with this process repeated until a fixed vocabulary size is reached. We use the same principle, but base merges on the word boundary cues from a language model trained on the dataset.

Our method is as follows:

1. We take a trained phoneme-level LM and compute either the UBP cue or the entropy cue at every position in the a given dataset.

2. We initialize our vocabulary V to match the vocabulary of the phoneme LM (so it contains every phoneme plus the utterance boundary token).

3. For every pair of tokens $x_i, x_j \in V$ that co-occur in the dataset, we compute the score for that pair by finding the average value of the word boundary cue at the position of the second token in the pair (e.g. for the pair ð,ɛ, we find the value of the cue at every position where ɛ appears after ð and return the average).

4. We find the pair with the lowest score, create a new token $V_i + V_j$, add it to the vocabulary and apply the merge to every token in the dataset. The cue's value at the newly merged token is set to be the sum of the cue's value of the two tokens before the merging occurs. For the entropy cue this follows from the chain rule and for the UBP cue this results in the

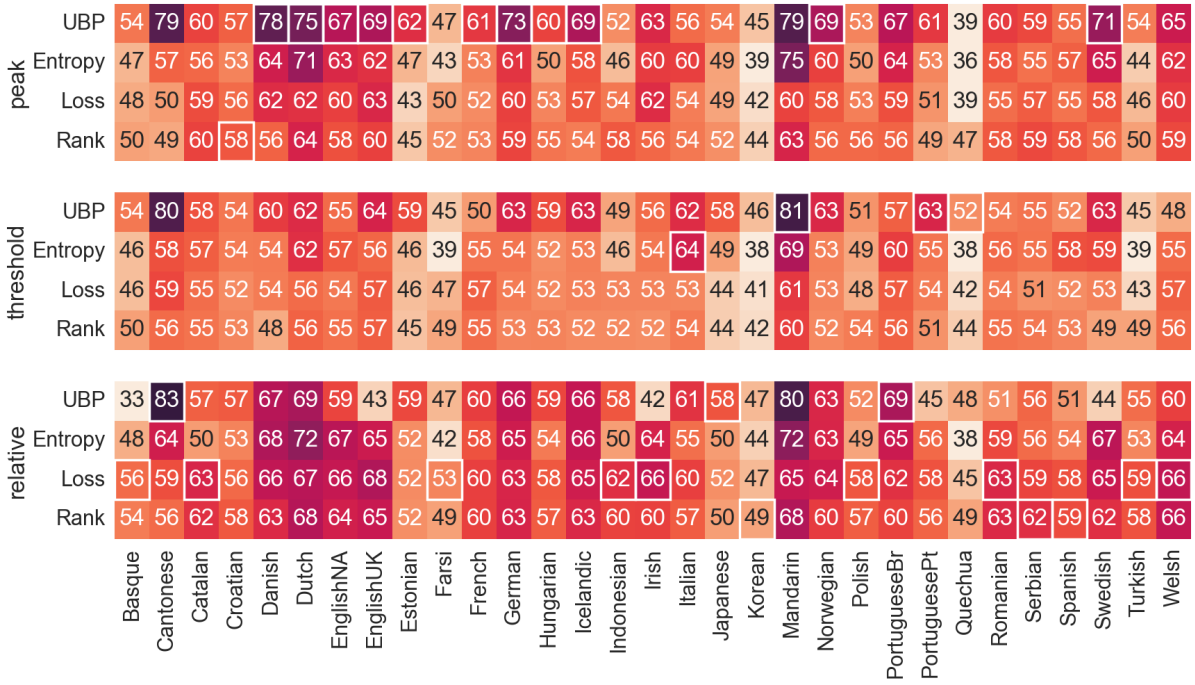

Figure 6: Boundary placement F1 scores achieved by the models in the **Tiny** suite for each cue and segmentation strategy, with the highest score for each language highlighted.

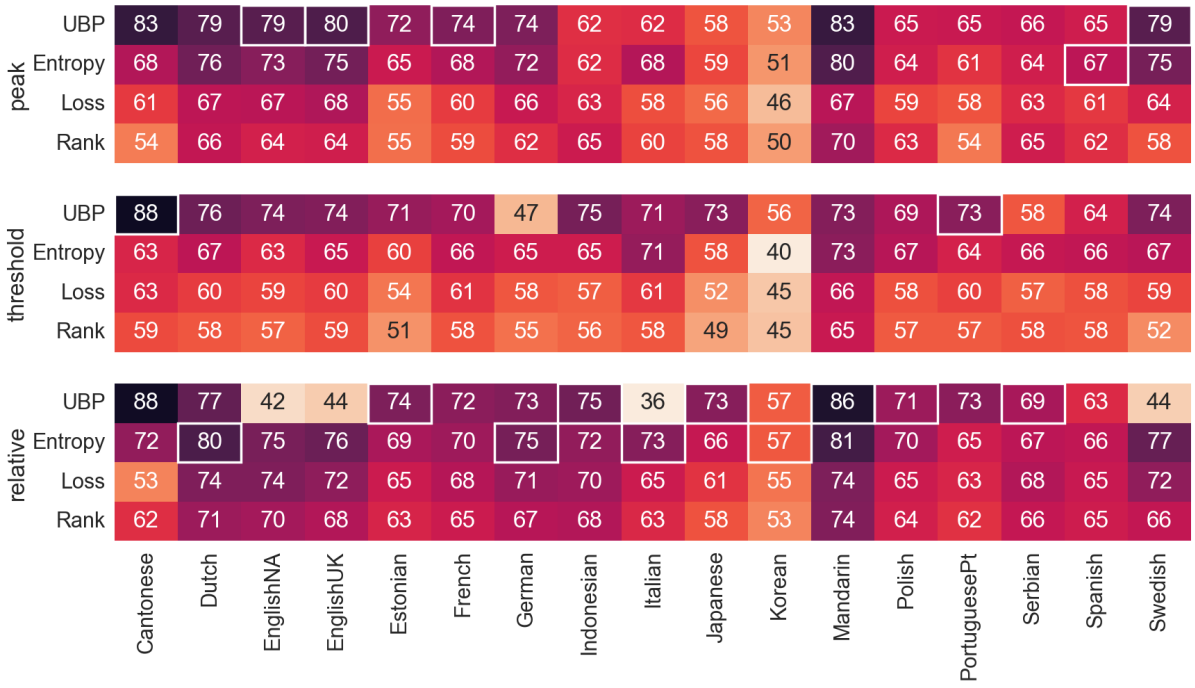

Figure 7: Boundary placement F1 scores achieved by the models in the **Small** suite for each cue and segmentation strategy, with the highest score for each language highlighted.

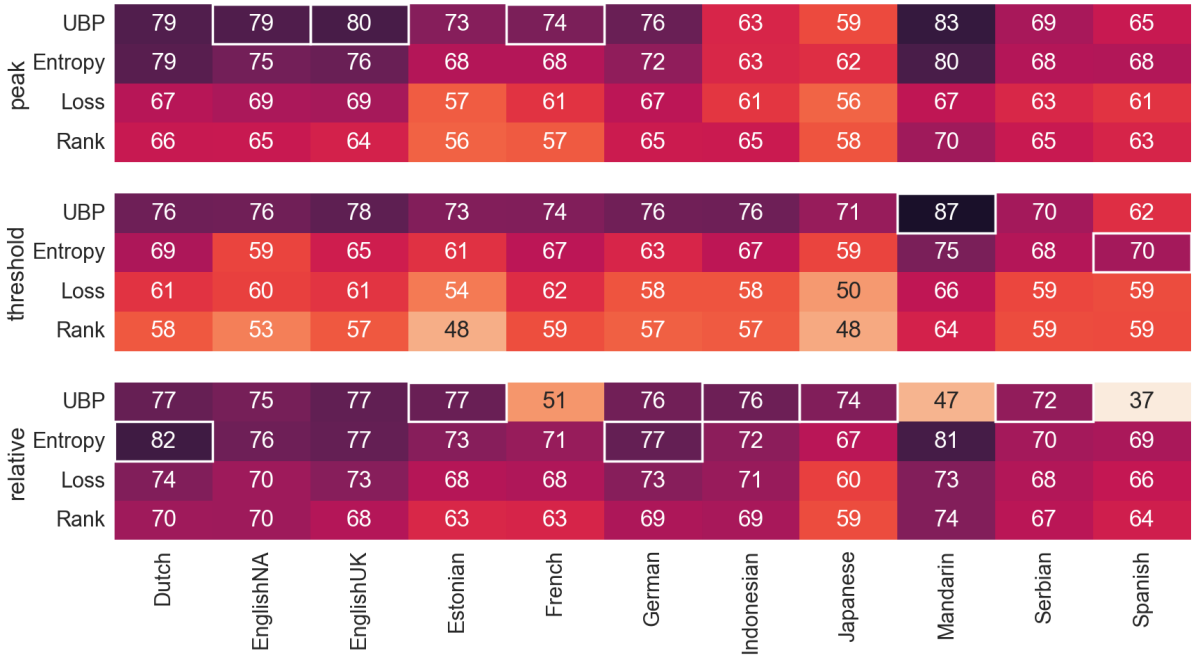

Figure 8: Boundary placement F1 scores achieved by the models in the **Medium** suite for each cue and segmentation strategy, with the highest score for each language highlighted.

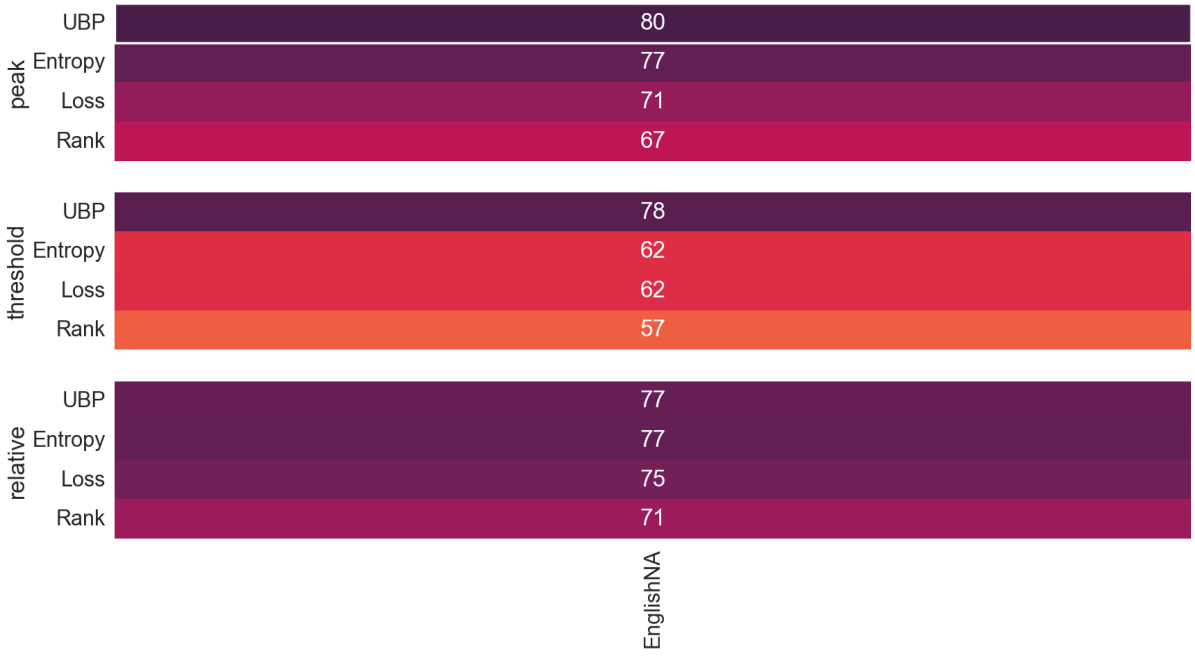

Figure 9: Boundary placement F1 scores achieved by the models in the **Large** suite for each cue and segmentation strategy, with the highest score for each language highlighted.

Language	100k	700k	2M	18M
Basque	Loss (relative)			
Cantonese	UBP (relative)	UBP (threshold)		
Catalan	Loss (relative)			
Croatian	Rank (peak)			
Danish	UBP (peak)			
Dutch	UBP (peak)	Entropy (relative)	Entropy (relative)	
EnglishNA	UBP (peak)	UBP (peak)	UBP (peak)	UBP (peak)
EnglishUK	UBP (peak)	UBP (peak)	UBP (peak)	
Estonian	UBP (peak)	UBP (relative)	UBP (relative)	
Farsi	Loss (relative)			
French	UBP (peak)	UBP (peak)	UBP (peak)	
German	UBP (peak)	Entropy (relative)	Entropy (relative)	
Hungarian	UBP (peak)			
Icelandic	UBP (peak)			
Indonesian	Loss (relative)	UBP (relative)	UBP (relative)	
Irish	Loss (relative)			
Italian	Entropy (threshold)	Entropy (relative)		
Japanese	UBP (relative)	UBP (relative)	UBP (relative)	
Korean	Rank (relative)	Entropy (relative)		
Mandarin	UBP (threshold)	UBP (relative)	UBP (threshold)	
Norwegian	UBP (peak)			
Polish	Loss (relative)	UBP (relative)		
PortugueseBr	UBP (relative)			
PortuguesePt	UBP (threshold)	UBP (threshold)		
Quechua	UBP (threshold)			
Romanian	Loss (relative)			
Serbian	Rank (relative)	UBP (relative)	UBP (relative)	
Spanish	Rank (relative)	Entropy (peak)	Entropy (threshold)	
Swedish	UBP (peak)	UBP (peak)		
Turkish	Loss (relative)			
Welsh	Loss (relative)			

Table 4: Best combination of boundary cue and segmentation strategy for each language and each suite.

probability that *either* original token was an utterance boundary.

5. We repeat (2)-(3), adding new tokens and applying merges until a fixed vocabulary size is reached.

Conceptually, creating merges using minimum average entropy will join highly predictable tokens together and result in tokens with comparable information and a uniformly dense signal that the model can learn from. Creating merges using the minimum average probability of an utterance boundary is similar, but instead tokens are joined according to the model's certainty that they do not cross an utterance boundary.

In order to test this method, we use the phoneme-level LM trained by Goriely et al. (2024) on a phonemized version of the 100-million word BabyLM dataset (Choshen et al., 2024) and train subword tokenizers using a phonemized version of the 10-million word BabyLM dataset. We create two tokenizers with a vocabulary size of 16k using the UBP cue and the entropy cue. We compare these to the BPE tokenizer trained by Goriely et al. (2024) on the same dataset, which also has a vocabulary size of 16k. Note that all three tokenizers are trained on a dataset without word boundaries, so it is possible for tokens to span word boundaries.

Goriely et al. (2024) trained a large model using their BPE tokenizer on the 100-million word BabyLM dataset and evaluated their results on two linguistic benchmarks, BLIMP (Warstadt et al., 2020) and BabySLM (Lavechin et al., 2023). We train and evaluate a model using the same procedure but replace their tokenizer for ours.

The results of this experiment are provided in table 5. We find that our two tokenizers improve all three scores compared to the BPE tbut instead okenizer with the UBP cue leading to a particularly large improvement for the BabySLM syntactic score.

Our method is similar to Pagnoni et al. (2024), who calculate the entropy cue over bytes using a small byte-level LLM, and use either a *global constraint* (corresponding to our threshold segmentation strategy) or a *monotonic constraint* (corresponding to our relative segmentation strategy) in order to group bytes into latent 'patches'. These

Tokenizer	BLIMP	BabySLM Syntactic	BabySLM Lexical
BPE	71.7	74.7	71.2
Entropy	72.7	77.6	81.3
UBP	72.6	85.6	84.4

Table 5: BLIMP and BabySLM scores achieved by a GPT-2 model trained on the BabyLM dataset. We compare BPE to our subword method, where merges are assigned using either entropy or UBP as a cue. BPE results are taken from Goriely et al. (2024).

patches are then fed into the main model, a large transformer, and the encoded patches are 'unpatched' and fed back into the byte-level LLM to predict the next byte. Future work should investigate whether their method is improved by using the cues explored in this study. When training with word boundaries, the prediction of the space character (or other word boundary characters) could also be used to group bytes.