Manuscript Running Header: CLASSIC-Utterance-Boundary Chunking-Based Model

Article Type: EMPIRICAL STUDY

Manuscript Title: CLASSIC-Utterance-Boundary: A Chunking-Based Model of Early Naturalistic Word Segmentation

Author(s): [Francesco Cabiddu (he/him) [a], Lewis Bott (he/him) [a], Gary Jones (he/him) [b], Chiara Gambi (she/her) [a, c]]

Author Affiliations: [a]Cardiff University [b]Nottingham Trent University [c]University of Warwick

Abstract: Word segmentation is a crucial step in children's vocabulary learning. While computational models of word segmentation can capture infants' performance in small-scale artificial tasks, the examination of early word segmentation in naturalistic settings has been limited by the lack of measures that can relate models' performance to developmental data. Here, we extended CLASSIC (Jones et al., 2021) - a corpus-trained chunking model that can simulate several memory, phonological and vocabulary learning phenomena - to allow it to perform word segmentation using utterance boundary information (henceforth CLASSIC-UB). Further, we compared our model to children on a wide range of new measures, capitalizing on the link between word segmentation and vocabulary learning abilities. We show that the combination of chunking and utterance-boundary information used by CLASSIC-UB allows a better prediction of English-learning children's output vocabulary than other models.

Keywords: Word segmentation; computational modeling; CLASSIC; chunking; transitional probability

Author Twitter information: *@FrancescoCabid1*

Author notes / acknowledgements

CRediT author statement– **Francesco Cabiddu**: conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; writing – original draft preparation; writing – review and editing. **Lewis Bott**: conceptualization; formal analysis; methodology; writing – review and editing; supervision. **Gary Jones**: conceptualization; formal analysis; methodology; writing – review and editing; supervision. **Chiara Gambi**: conceptualization; formal analysis; methodology; writing – review and editing; supervision.

We would like to thank Professor Padraic Monaghan and the anonymous reviewers for their valuable feedback on the manuscript.

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

The research reported in this article was supported by a Cardiff University School of Psychology PhD Studentship, and a British Academy Small Grant SRG1920\100600.

Correspondence concerning this article should be addressed to Francesco Cabiddu, School of Psychology, Cardiff University, 70 Park Place, Cardiff, CF10 3AT, CabidduF@cardiff.ac.uk

Introduction

Word segmentation is a fundamental process in infant language development. Phonological word forms are not given *a priori* but must be extracted from continuous speech input. While several computational models capture basic word segmentation phenomena displayed by infants in small-scale artificial tasks, assessing whether models can scale-up to naturalistic inputs has been hampered by limited sets of measures to compare performance against. Here, we present a new word segmentation model which extends CLASSIC (Jones & Rowland, 2017; Jones et al., 2021; Jones, 2016; Jones, Justice et al., 2020), a chunking model that uses naturalistic inputs to successfully simulate key developmental phenomena in memory and language. Our extended model, CLASSIC-Utterance-Boundary (CLASSIC-UB), performs unsupervised word segmentation using large-scale naturalistic inputs. Importantly, we assess our model against existing segmentation models using both standard evaluation metrics and novel developmental measures to provide a more comprehensive assessment of segmentation performance.

Chunking models successfully account for adult (e.g., Frank et al., 2010) and infant (e.g., French et al., 2011; Perruchet & Vinter, 1998) word segmentation in lab tasks by extracting and storing frequent input sequences (chunks) as candidate words that guide subsequent segmentation. This allows chunking models (e.g., Kurumada et al., 2013) to account for lexical effects in infant segmentation such as easier extraction of novel words when preceded by familiar words (e.g., Bortfeld et al., 2005). Lexical effects are not predicted by competing models that assume a dedicated mechanism that estimates the location of word boundaries in speech by tracking sublexical regularities (i.e., forward and backward sound transitional probabilities) (e.g., Cleeremans & McClelland, 1991; Saksida et al., 2016). Further, chunking also accounts for infants' sensitivity to sublexical regularities (e.g., Hay et al., 2011; Pelucchi et al., 2009; Saffran, Aslin et al., 1996; Saffran et al., 1997;

Saffran, Newport et al., 1996), because the component parts of a chunk are mutually linked, giving equal weight to forward and backward relations (e.g., French et al., 2011; Perruchet & Desaulty, 2008; Perruchet & Poulin-Charronnat, 2012; Perruchet & Vinter, 1998; although see McCauley & Christiansen, 2019 for a hybrid model of speech comprehension and production that forms chunks via backward transitional probability without the need to capture forward relations).

Typically, computational investigations have used artificial language tasks to assess the plausibility of learning mechanisms involved in infant (e.g., French et al., 2011; Perruchet & Vinter, 1998) and adult word segmentation (e.g., Endress & Langus, 2017; Frank et al., 2010). Although modelers have also examined scale-up to naturalistic input (e.g., Daland & Pierrehumbert, 2011; Monaghan & Christiansen, 2010; Saksida et al., 2016), such investigations have suffered from one important limitation: The benchmark for models' segmentation accuracy have been the word boundaries present in adult vocabularies, but these are unlikely an accurate reflection of infants' and children's segmentation (e.g., Monaghan & Christiansen, 2010). In contrast, we introduce new measures based on developmental data, and specifically on the composition of children's early vocabularies. The key insight here is that children's vocabularies should be a reflection of early word segmentation processes: Word forms that are more easily discovered in the input should enter the child's vocabulary earlier on in development. We use these novel developmental measures alongside traditional evaluation measures to provide a much richer assessment of the developmental plausibility of word segmentation mechanisms. Specifically, we use this suite of measures to compare CLASSIC-UB to other models that have shown different strengths in modeling early naturalistic segmentation.

We first present CLASSIC's architecture and its extension CLASSIC-UB. We then discuss new measures of segmentation and show how models' developmental plausibility can

be assessed when trained on naturalistic speech corpora. Finally, we compare the results of CLASSIC-UB to competing segmentation models (transitional probability models, Saksida et al., 2016; and a different chunking-based model, PUDDLE, Monaghan & Christiansen, 2010) and conclude that chunking might play a significant role in infant word segmentation.

CLASSIC

CLASSIC uses a domain-general chunking mechanism (Gobet et al., 2001) to model linguistic knowledge acquisition via experience with the sequential structure of the language. It is not a model of auditory perception or production per se (as basic processes that transfer information to the learning mechanism are not modeled), but a learning model representing performance increases derived from perceptual learning and efficiency in production (Jones, Justice et al., 2020). The accumulation of language experience is essentially represented by the chunking of adjacent items, gradually shifting the model's representations from sublexical to lexical and multi-word units. A key assumption in CLASSIC is that the child already knows how to identify word boundaries. This has been implemented in CLASSIC because past simulations (presented in the next paragraphs) have investigated phenomena at an age where the child is likely to have already learned how to segment speech into words. Below we illustrate how CLASSIC works using a simplified example in which the model repeatedly processes the phonetically transcribed utterance [d, æ, d | ɪ, z | k, ʌ, m, ɪ, ŋ][1] (i.e., *dad is coming*; where | demarcates word boundaries which, as explained above, are given as input to the model). CLASSIC first chunks adjacent phonemes that do not cross a word boundary and forms biphone representations (dæ, æd | ɪz | kʌ, ʌm, mɪ, ɪŋ). Any learned chunks can subsequently be used to encode the input. For example, at the second iteration, the model would represent the utterance as [dæ, d | ɪz | kʌ, mɪ, ŋ] – i.e., proceeding from left to right, it uses the longest available chunks to encode each demarcated word. This way of

encoding preserves the input temporal structure and represents a proxy for the increased processing efficiency derived from acquired knowledge[2]. The model then continues to join adjacent chunks; for example, the third iteration would result in the representation [dæd | ɪz | kʌmɪ, ŋ], where CLASSIC has learned two of three words in the utterance. When two adjacent chunks are words themselves, CLASSIC crosses word boundaries and learns multi-word sequences (i.e., dæd|ɪz in the example); thus, at the fourth iteration, CLASSIC would encode the utterance as a two-word sequence followed by a word: [dæd|ɪz, kʌmɪŋ]. Finally, in a last iteration the model would represent the whole utterance as a single multi-word chunk: [dæd|ɪz|kʌmɪŋ].

CLASSIC accounts for the role of sublexical, lexical and multi-word sequences in language development. For example, in Jones (2016), incremental exposure to naturalistic speech supported CLASSIC's building up of chunks at different grain sizes, capturing 85% of variance in nonword repetition performance - a task closely related to vocabulary learning (e.g., Hoff et al., 2008) – from six studies involving 2- to 6-year-old children. CLASSIC has also simulated vocabulary learning more directly (Jones et al., 2021): Similarly to 2-3-year-old children learning to produce words, CLASSIC gradually learns longer, more infrequent words that have a smaller number of similar words in the language (i.e., lower neighborhood density) and higher internal predictability (i.e., higher average biphone probability, or "phonotactic probability"). Jones et al. (2021) also showed that novel words entering children's productive vocabularies are more likely to share large phonological chunks with words they already use, indicating a pivotal role for phonological knowledge in vocabulary learning. In sum, these studies have shown that sublexical knowledge can be used to learn and produce pseudowords and real words (see Baayen et al., 2019; Chuang et al., 2021 for similar conclusions using linear discriminative learning).

Finally, Jones, Justice et al. (2020) showed that phonological knowledge plays an important role in learning multi-word sequences. CLASSIC captured the faster increase in children's short-term memory for digit over word sequences, likely because chunks that span multiple digits are learned more quickly from random combinations of digits occurring in naturalistic speech. This study also showed how knowledge of multi-word sequences facilitates lexical processing (e.g., processing of the individual items *five* and *six* becomes more efficient when the two are presented within a familiar multi-word sequence *five-six*).

In sum, CLASSIC is a chunking-based model that has captured important developmental phenomena in word learning, but has not yet been applied to word segmentation. Here we show how CLASSIC can be extended to perform word segmentation, thus making the model more developmentally plausible: Infants must of course discover word forms before they can learn novel words and integrate them into their existing vocabulary (Newman et al., 2016).

CLASSIC-UB

To extend CLASSIC to perform word segmentation, we retained CLASSIC's architecture but removed word boundary information from the model input (i.e., the model is not constrained to chunk items within demarcated words). We also added utterance boundary information using positional markers (↲), which signal utterance start or end. Such positional markers were coded by transcribers of the input corpora used in this study, based on various syntactic (e.g., utterances are centered around a main clause) and prosodic cues (e.g., pauses, intonation patters distinguishing declarative, interrogative, or other clauses); note that only written transcriptions were available for most of the input, not the original speech recordings, so it was not possible to automatically assign positional markers based on – for example - changes in phonetic features. Positional markers have been used in previous computational

work (e.g., Aslin et al., 1996; Christiansen et al., 1998; Saksida et al., 2016) as a proxy for the increased saliency that phonological units at utterance boundaries gain in child-directed speech (e.g., Fernald & Mazzie, 1991). This has been modeled via conjunctive use of utterance-boundary markers and phonological units to perform distributional learning (e.g., *utterance-boundary+syllable* constitutes a pair of units for which transitional probabilities can be obtained; Saksida et al., 2016). In a similar way, CLASSIC-UB treats utterance-boundary markers as additional units that can be used to form chunks (i.e., a chunk becomes longer when an utterance-boundary marker is attached to a phonological sequence).

We present a version of CLASSIC-UB that uses utterance-final markers, and one that uses both initial and final markers. Infants may privilege utterance-final words (e.g., Aslin et al., 1996; Christiansen et al., 1998), as these gain perceptual prominence from syllable lengthening (Wightman et al., 1992) and sentential accent in English (Cinque, 1993). However, some studies suggest that infants may use both initial and final markers in segmentation (Seidl & Johnson, 2006; 2008). In fact, different cues could facilitate segmentation of utterance-initial words (e.g., exaggerated amplitude, duration, pitch, and formant structure; Cruttenden, 1986). Therefore, presence of initial markers should provide additional facilitation over utterance-final cues. We are not aware of any computational studies assessing the relative contribution of initial and final boundaries, thus comparing CLASSIC-UB with final markers to CLASSIC-UB with both initial and final markers could shed light on the factors that facilitate word segmentation at utterance edges.

Figure 1 illustrates how CLASSIC-UB segments input, after this has been transcribed using the CMU pronouncing dictionary (Lenzo, 2007), which contains over 134,000 phonetic transcriptions of English words and provides an automatic way to convert large orthographic input into phonetic form (using alphabetic codes for phonemes rather than IPA e.g., AE rather than æ). When encoding the utterance-final biphone *AED* in the first utterance, the

model learns the chunk with an associated utterance-final marker (i.e., *AED*↲). If the chunk *AED* appears in later utterances, even in word-medial positions, the model will recognize that it can be used in word-final position assuming a word boundary at this location (see the third utterance *dad is coming*). This also shows how the following phone (*IH*) is marked as 'can begin a word' based on the model flagging *AED* as ending the preceding word *DAED* (bold chunk of Figure 1). The same logic applies to utterance-initial markers. In essence, the function of the ↲ markers within chunks is akin to "this chunk can appear at the [beginning/end] of a word".

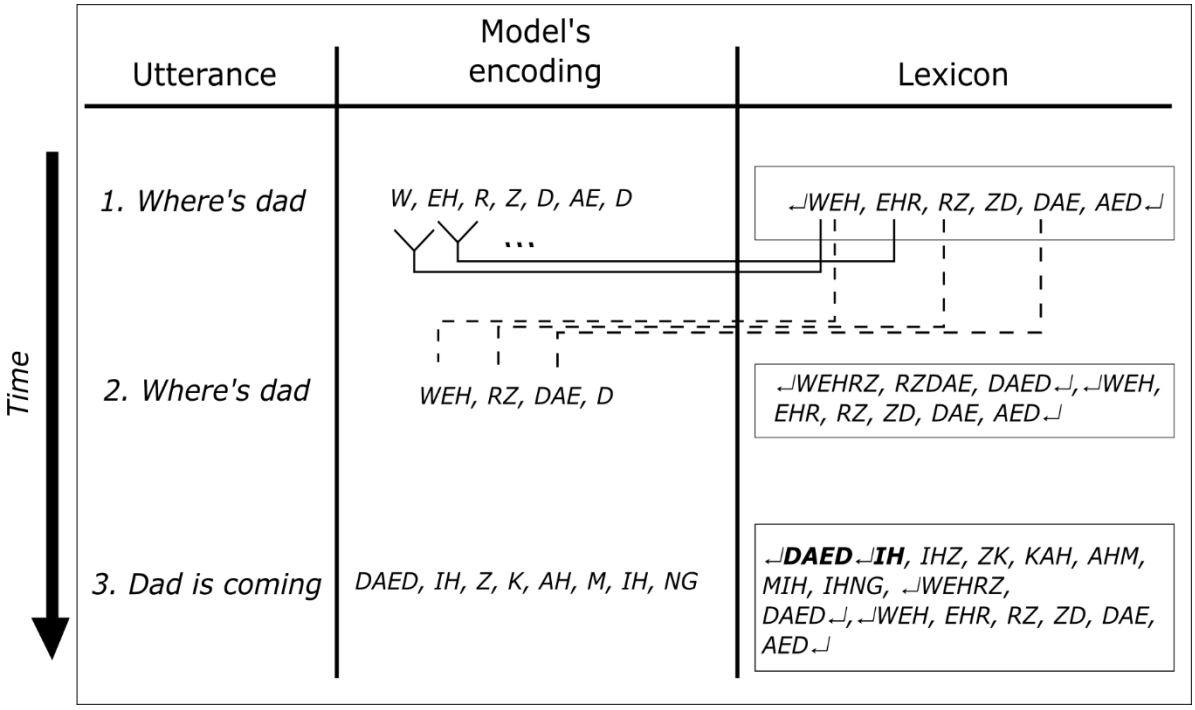


Figure 1. CLASSIC-UB generalization of utterance-boundary markers to utterance-medial position. Solid lines indicate grouping of adjacent items into single chunks and storage into the lexicon. Dashed lines indicate use of stored chunks to segment speech. Lines are only shown for the first utterance. Time indicates independent presentations of new child-directed utterances. All English phonemes are present in the lexicon but are not shown for reason of space. The transcription used is based on the CMU pronouncing dictionary (Lenzo, 2007).

Like CLASSIC, CLASSIC-UB processes phonemic input. As such, it assumes that children already know phoneme categories, in line with an early phonetic category learning approach (e.g., Werker, 2018) and previous computational studies in word segmentation (e.g., Batchelder, 2002; Daland & Pierrehumbert, 2011; Goldwater et al., 2009; but there are alternative approaches, which we briefly refer to in the General Discussion). Knowledge of sound categories and co-occurrences of sounds might begin to develop at the same time or soon after infants start segmenting speech into words at around 6 months of age (Bortfeld et al., 2005). For example, between 3 and 9 months infants discriminate between and learn new phonetic categories using distributional cues (e.g., Cristià, McGuire et al., 2011; Maye et al., 2008; Mersad et al., 2021; Yeung et al., 2014), they can use this information in word segmentation (e.g., Jusczyk & Aslin, 1995), and soon after in word recognition (12 months; Mani & Plunkett, 2010) and word learning tasks (14 months; Fais et al., 2012). Similarly, between 4 and 9 months infants attune to native phonotactic patterns (Cristià, Seidl et al., 2011; Jusczyk et al., 1994), and can use this knowledge in word segmentation (e.g., Mattys & Jusczyk, 2001). Nevertheless, we also run all of our simulations on syllabified input (see Method), because infants may initially perceive syllables as basic linguistic units (e.g., Bertoncini & Mehler, 1981).

As with CLASSIC, items that co-occur often will have more opportunities to be chunked together by CLASSIC-UB. This facilitates subsequent segmentation in two ways. First, when a word is frequent in the input, its sublexical components will have more opportunities to be chunked together, reaching a whole-word representation faster. This makes the model frequency-sensitive, even though frequency is not explicitly tracked (unlike in other chunking models, such as PUDDLE; Monaghan & Christiansen, 2010; see Appendix S1 for a detailed description of this model). Second, learning words that share phonological

material with other words will be facilitated by the reuse of existing chunks (e.g., learning *just* can make the sequence *ust* available to subsequently learn *crust*). Other models, such as PUDDLE, do not include this mechanism and rely on frequency information alone.

The number and size of chunks changes as more input is processed. CLASSIC-UB processes input incrementally (i.e., one utterance at a time), as other segmentation models (e.g., French et al., 2011; Monaghan & Christiansen, 2010; Perruchet & Vinter, 1998). As shown in Figure 1, each utterance is encoded, from left to right, by using existing chunks present in the model lexicon. Consistent with previous chunking models (e.g., Batchelder, 2002; French et al., 2011; Perruchet & Vinter, 1998), preference is given to encoding larger chunks over shorter ones. For example, the chunk *AED*↵ which contains a boundary marker is preferred over the shorter chunk *AED*, which does not contain a boundary marker. At the same time, new/larger chunks are stored in the model lexicon by joining adjacent encoded items together, facilitating subsequent segmentation. This makes the learning process plausible, as children's learning happens incrementally, as a function of their accumulating knowledge of the language (e.g., Jones et al., 2021).

Crucially, selecting larger chunks over shorter ones means that chunks formed by sublexical sequences and utterance-boundary markers are dispreferred to words, thus avoiding oversegmentation. At the same time, the presence of utterance-boundary markers prevents the model from building large undersegmented chunks. Together, these two mechanisms favor segmentation at the (intermediate) word-level. However, there is no explicit rule defining when the model should stop building chunks of increasing size. In fact, at later stages, the model stores multi-word chunks (consistent with representation of multi-word sequences from 11 months; e.g., Jones, Cabiddu et al., 2020; Skarabela et al., 2021). Notably, such longer chunks can include multiple boundary markers, which means the model can represent multi-word sequences while also retaining knowledge of the individual words

composing the sequence. For example, an utterance such as *I'll do it later* could be encoded using the two chunks ↲*I'll*↲*do*↲*it*↲ and *later*↲.

In sum, CLASSIC-UB learns chunks including both phonological and utterance-boundary information. Chunks gradually increase in size facilitating subsequent segmentation. Below, we discuss the problem of evaluating models of early naturalistic segmentation and describe new measures we used to assess CLASSIC-UB's developmental plausibility.

Evaluation of Naturalistic Speech Segmentation

Corpus-based evaluations of segmentation models usually compare models' output to segmented transcriptions of child-directed speech (e.g., Monaghan & Christiansen, 2010). Precision and Recall are two widely used measures. Precision is the number of words segmented by a model divided by the total number of items segmented (including segmentation errors) (i.e., how many of the items found are words). Recall is the number of words segmented by a model divided by the total number of words in the input (i.e., how many words present in the input are found). In these two measures, chunking models perform above models that segment speech randomly (e.g., Bernard et al., 2020; Monaghan & Christiansen, 2010), in line with results from computational studies capturing artificial language learning (e.g., French et al., 2011). For example, in Larsen et al. (2017) the chunking model PUDDLE showed the highest performance, reaching 82% Precision and 80% Recall. In contrast, another class of models that track sound transitional probabilities (see Appendix S1 for a detailed description) perform better than the random baselines (e.g., Bernard et al., 2020) but less well than chunking models (e.g., 43% Precision and 51% Recall in Larsen et al., 2017).

Although these measures capture how accurately models segment the input, they do not capture their developmental plausibility. The use of segmented input to evaluate model performance makes the implicit assumption that infants segment speech in an adult-like way but, as discussed by Larsen et al. (2017), this assumption is likely to be wrong, given evidence that infants' initial protolexicons are comprised of words and frequent phonotactically legal nonword sequences (e.g., Ngon et al., 2013).

Addressing this problem is not straightforward because we do not know how infants segment speech in naturalistic settings. Larsen et al.'s (2017) solution is to link model accuracy to word age of acquisition. For example, *dog* is understood by a higher proportion of children at 13 months than *deer*, and this should be reflected by a more accurate segmentation of *dog* than *deer* (i.e., *dog* is correctly segmented on more occasions). Theoretically, the reasoning behind using word learning as a proxy for segmentation performance is that vocabulary knowledge (word-meaning mapping) is facilitated by word segmentation (e.g., Estes et al., 2007; Hay et al., 2011). For example, in Estes et al. (2007), infants were able to extract, store and recognize word forms previously presented in fluent speech to successfully perform a label-object association task. In sum, words that are acquired early must also be accurately segmented at earlier ages.

Interestingly, when Larsen et al.' (2017) measure is used, transitional probability models perform better than chunking models despite discovering fewer words in the input as mentioned above. For example, a transitional probability model explained 19% of variance in age of acquisition (the highest performance in the study), while the chunking model PUDDLE explained only 7% (Larsen et al., 2017).

Here we also capitalize on the link between vocabulary knowledge and segmentation as suggested by Larsen et al. (2017), but instead of age of acquisition derived from parental report, we use age of first production derived from child speech (Grimm et al., 2017).

Looking at production rather than comprehension has drawbacks, but also important advantages. The words children produce are of course not a direct reflection of their segmentation abilities. Production involves additional factors related to recalling stored instances from the lexicon and to articulation, and of course what children spontaneously produced at the time of recording does not reflect the entirety of their comprehension vocabularies. Further, there are limitations inherent in estimating children's knowledge from a small number of relatively short samples of speech, filtered through adult transcribers' potentially biased judgement (e.g., leading to the omission of non-lexical productions). Nevertheless, using production vocabularies has two key advantages. First, it dramatically increases the number of words examined: the British CDI (Alcock, 2020), a parent-report measure of age of acquisition, only comprises 330 words[3], lacking sufficient statistical sensitivity. Second, we found that the CDI word sample has a word frequency distribution shifted toward high-frequency words, not reflecting the Zipfian input infants hear (i.e., many low frequency and few high-frequency word types; Hendrickson & Perfors, 2019)[4]. Using such sample might bias results because transitional probability models might perform well only because the distribution considered is less skewed toward low frequency words (Kurumada et al., 2013).

We additionally propose a new measure examining whether a model can capture word-level characteristics of child vocabularies. Previous measures do not examine whether a model capitalizes on sublexical/lexical regularities (similarly to how learning is evaluated in lab settings). Traditional measures are focused on finding a mechanism that minimizes segmentation errors, while the age of acquisition/production measure is focused on the time course of acquisition. In contrast, with our final set of analyses we assess whether the characteristics of the vocabulary learnt by a model reflect what children have produced in the language corpora. In other words, we assess whether models and children are sensitive to

input characteristics in a similar way. We focus on three lexical measures (word frequency, word length, neighborhood density) and one sublexical measure (phonotactic probability). These characteristics can explain approximately 50% of variance in word learning (Stokes, 2010, 2014; Storkel, 2009).

Finally, although word comprehension as a marker of vocabulary growth is predominant (e.g., Fernald & Marchman, 2012), the use of evaluation measures based on early production is reasonable given both the relation between early vocalizations and vocabulary growth (McGillion et al., 2017), and the relation between early segmentation abilities and later expressive vocabularies (Newman et al., 2006; 2016).

In summary, we ask whether a novel chunking account of word segmentation can scale-up to naturalistic speech in a developmentally plausible way, by comparing CLASSIC-UB to a model that has shown a high performance in traditional measures of naturalistic segmentation (PUDDLE), and to backward and forward transitional probability models that might account for a high proportion of variance in child word knowledge (Larsen et al., 2017). We also ask whether utterance-initial edges play a role in segmentation beyond final edges, by comparing two different implementations of CLASSIC-UB. Finally, we ask whether transitional probability models can capture developmental data better than chunking accounts, by comparing PUDDLE to transitional probability models to test whether we replicate previous results (Larsen et al., 2017) but using different corpora and performance measures.

Method

Computational Models

We compared CLASSIC-UB to forward and backward transitional probability (Saksida et al., 2016), PUDDLE (Monaghan & Christiansen, 2010), and a random baseline relying on a coin

toss to place a boundary after each input unit (Lignos, 2012). A full description of these models can be found in Appendix S1. We implemented them to process syllables or phonemes as basic units (see Appendix S2 for details). Python and R scripts for preparing the input, running the models and analyzing the output are available at the OSF page https://doi.org/10.17605/OSF.IO/KBNEP.

Corpora

Seven CHILDES English corpora were used following Grimm et al. (2017) (see Table S2A and Appendix S2 for input preprocessing and characteristics). We only considered transcripts of target child age 2 as input for the models. While infants start segmenting speech much earlier than age 2, our choice to focus on this age group was motivated by the much smaller size of corpora of speech directed at children of younger ages (e.g., 54,274 utterances at age 1 vs. 604,000 utterances at age 2). As we show in Appendix S2, this limits the representativeness of input directed at children of younger ages. In total, the input to models contained 604,000 utterances (*MLU* = 4.39) from 332 different speakers, directed to 53 target children. Note that such input is 3 to 60 times larger than input used in previous studies (Christiansen et al., 1998; Daland & Pierrehumbert, 2011; Larsen et al., 2017; Monaghan & Christiansen, 2010; Saksida et al., 2016).

Measures of Model Performance

Precision and Recall. We compared models' performance by looking at the pairwise differences in mean Precision and Recall (e.g., Monaghan & Christiansen, 2010). We tested the last 10,000 utterances of output because models' performance was stable (see Figure 2) and because testing the entire output (i.e., 604,000) would have led to significant results even for trivial differences. We used a Welch's t-test for unequal variances, with p values and

bootstrap 95% confidence intervals corrected for multiple comparisons using Holm's correction.

Word Age of First Production. Transcripts' MLU was used as a proxy of word age of first production following Grimm et al. (2017) (see Appendix S3 for details). MLU is a useful estimator of child gross linguistic skills (i.e., developmental stage), controlling for the fact that children with similar age might be far apart in terms of language development. The sample comprised 5,480 words. We fitted linear regression models predicting word age of first production, as a function of the log10 number of times a target word was correctly segmented by each algorithm (Larsen et al., 2017). The latter was weighted by dividing it by input word frequency before fitting the regression models, as the two variables correlate highly with each other (e.g., for a random baseline, the Pearson correlation is .92). Word frequency correlates highly with the age of word acquisition (e.g., Morrison et al., 1997), therefore failing to control for its effect might lead to results that are an artifact of frequency. Indeed, input frequency tends to strongly affect models' performance (e.g., for the random model, the correlation between the number of correct segmentations and age of first production drops from .58 to .20 after controlling for frequency). Therefore, controlling for input frequency allows us to assess the performance of each segmentation algorithm over and above the fact that words that appear more often are acquired earlier.

Since previous studies have not used weighting by word frequency, we also include analyses for the unweighted measure in Appendix S6 to facilitate comparison. To foreshadow our findings, differences between models are consistent when using either the weighted or unweighted measure, with only one exception (pertaining to transitional probability models) which is discussed in the General Discussion.

Comparisons between models are based on pairwise differences in Adjusted R^2 from the regression models; we bootstrapped the 95% confidence interval of the difference between coefficients and corrected the interval using Holm's correction (Grimm et al., 2017). We conclude that two coefficients differ significantly from one another if the corrected confidence interval does not include zero.

Word-level Measures. The distributions of unique words discovered by each model were compared to children's actual vocabulary (i.e., the words produced by children in the corpus) in terms of phonemic length, word frequency, neighborhood density and phonotactic probability. Following Jones et al. (2021), the distribution of words relative to sublexical and lexical characteristics should be similar between children and model if the model's learning mechanism is developmentally plausible.

As in previous studies (e.g., Storkel, 2009; Swingley & Humphrey, 2018; Vitevitch & Luce, 1998), word length refers to the number of phonemes in a word; word frequency is the log10 frequency of a word across the input; phonotactic probability is the average probability of a phoneme pair to appear in a word; neighborhood density is the raw count of phonemic words that differ from a target word by one phoneme (i.e., by deletion, insertion or substitution). Note that phonotactic probability and neighborhood density were stress-unmarked to be consistent with previous work (e.g., Storkel, 2009; Swingley & Humphrey, 2018).

We carried out a Chi-Square Goodness of Fit Test to compare (observed) probabilities of a word being of a certain length (in the output of a segmentation model) to the (expected) probabilities in children's utterances (we focus on lengths 2-8 phonemes due to the low number of words at other phonemic lengths). Probabilities are defined as the proportion of types at each length. We then look at the pairwise differences in Chi-Square test statistics,

using bootstrap confidence intervals as described in the previous section. In other words, this analysis first looks at how close each model is to children, and then uses the estimates of such distance to compare models to one another.

For word frequency, neighborhood density and phonotactic probability (which are continuous measures) we followed a similar procedure to the above, but used a Kolmogorov–Smirnov test statistic. Following Piantadosi et al. (2012), we divided each of these measures by word length. Word length tends to be anti-correlated with word frequency (e.g., Zipf, 1936) and neighborhood density (Storkel, 2004), and positively correlated with phonotactic probability (Storkel, 2004). In our dataset, the correlations varied from moderate to strong (*Spearman's rho* [length, frequency] = -.37; *Spearman's rho* [length, neighborhood density] = -.86; *Spearman's rho* [length, phonotactic probability] = .42).

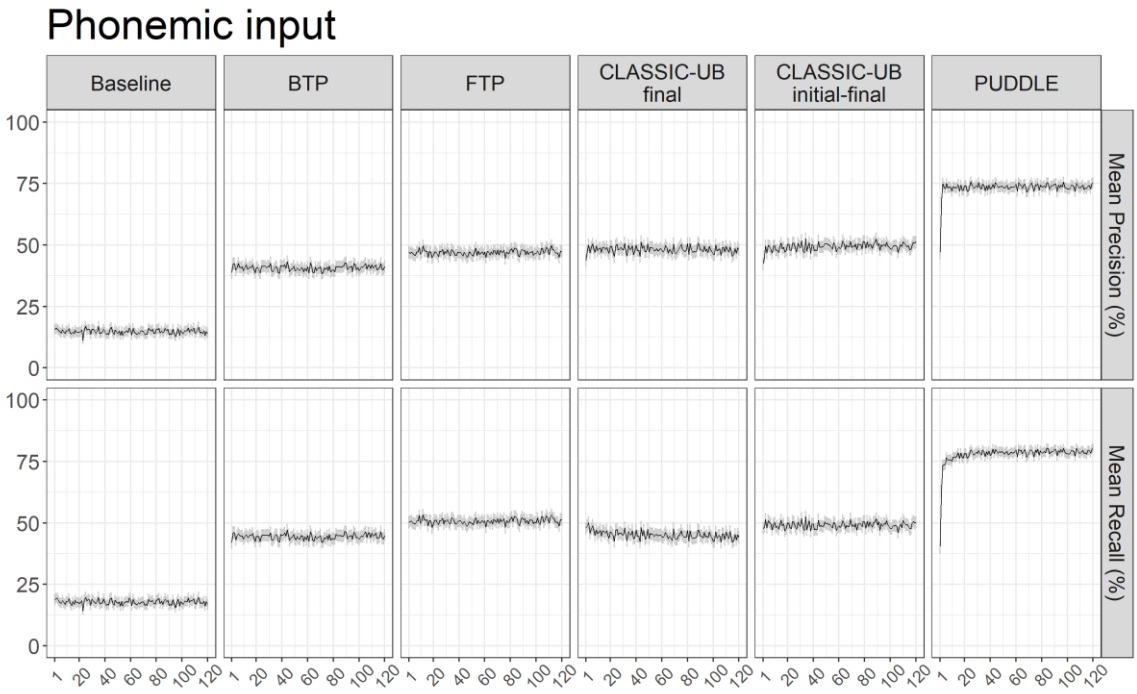
Results and Discussion

We first report results for Precision/Recall and age of first production, and finally for word-level measures. For ease of readability, in each subsection we only give a discursive presentation of key results and point to statistical results in the Appendix. Both CLASSIC-UB initial and CLASSIC-UB initial-final are included in this section, but for reasons of space a discursive comparison between the two can be found in Appendix S11.

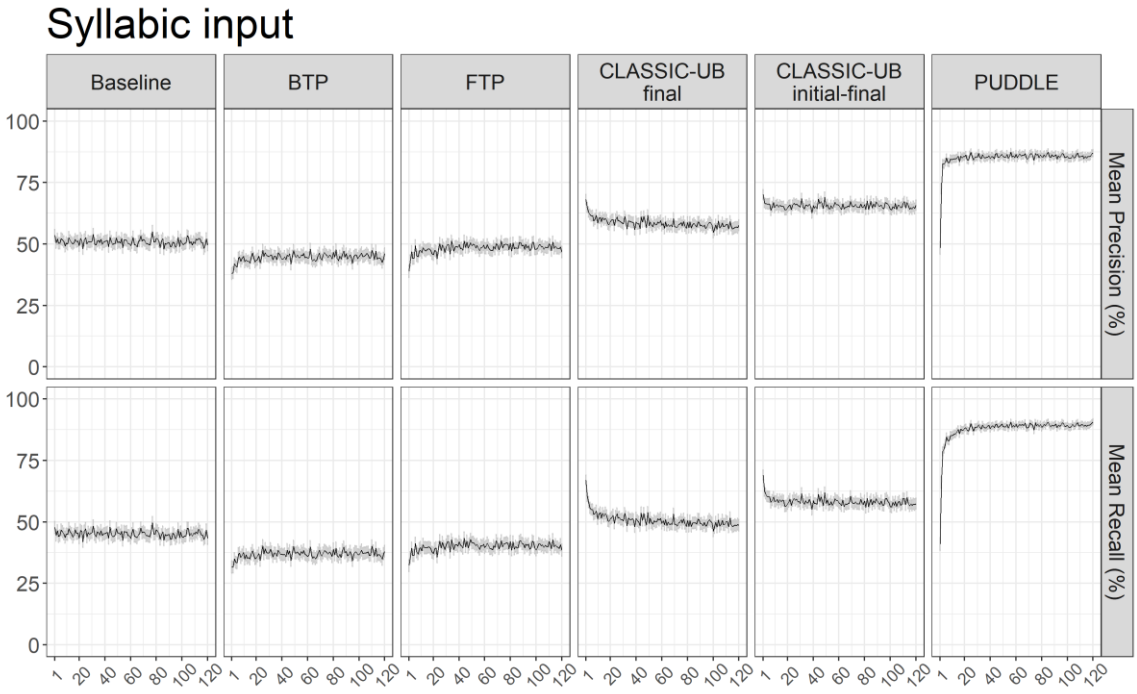
Precision and Recall

All models showed rapid learning (see Figure 2), reaching a ceiling in performance after approximately 40,000 utterances and indicating that the quantity of the input does not affect models' performance (consistent with Daland & Pierrehumbert, 2011). Pairwise statistical comparisons can be found in Appendix S4. All models segmented the input above chance

(baseline), except for transitional probability models when the input was syllabified (see Figure 2b and Table S4B).



2a.



2b.

Figure 2. Mean Precision and Recall performance with phonemic (2a.) and syllabic (2b.) input. The figure shows the random baseline, backward (BTP) and forward transitional probability (FTP), CLASSIC-UB with utterance-final and initial-final markers, PUDDLE. Performance is averaged every 1000 utterances (Stage). Only the first 120 stages are shown to better appreciate changes in performance and because the performance of the models is stable. Grey confidence bands indicate the 95% confidence interval around the mean.

In line with Larsen et al. (2017), PUDDLE shows the best performance, outperforming baseline, transitional probability and CLASSIC-UB models. When using phonemic input, 73% of items found by PUDDLE are words (Precision) and the model discovers 79% of input words (Recall). This model's accuracy is higher when segmenting syllabified input, reaching 85% Precision and 89% Recall. CLASSIC-UB's performance lies between PUDDLE and transitional probability models, with CLASSIC-UB initial-final reaching 50% Precision and Recall with phonemic input, and 66% Precision and 58% Recall with syllabified input.

Overall, the models segmented naturalistic speech above chance. However, while traditional measures examine models' accuracy, they tell us nothing regarding whether the model's segmentations reflect how infants segment speech, and we are not able to make any claim regarding the plausibility of a model compared to another. To address this issue, we turn to the next set of measures, which relate model performance to child data.

Word Age of First Production

Only CLASSIC-UB initial-final (*Adj* R^2 = .084 [.066, .103]), CLASSIC-UB final (*Adj* R^2 = .079 [.062, .100]) and PUDDLE (*Adj* R^2 = .078 [.060, .097]) - and only when run on phonemic input - outperformed the baseline (*Adj* R^2 = .036 [.023, .052]) at predicting word

age of first production (see Table 1 and all pairwise comparisons in Table S5A). Note that although the size of Adjusted R^2 is small, it is in line with Larsen et al. (2017), who for example showed that PUDDLE explained .067 of variance in child age of acquisition[5]. Surprisingly, when the models were run on syllabic input, none of them passed the baseline test (see Table S5A). We discuss this unexpected finding in Appendix S13. Also, note that the results reported above are based on weighting the predictor measure by frequency as explained in the Method section. Results for the unweighted measure are reported in Appendix S6.

Crucially, while CLASSIC-UB has lower Precision and lower Recall compared to PUDDLE (see Figure 2), the two models explain the same proportion of variance in child word age of first production (~8%), suggesting that achieving lower segmentation accuracy might not necessarily lead to lower developmental plausibility. Nevertheless, age of first production does not take into account the characteristics of the model's vocabulary, nor does it answer questions about whether model and children are sensitive to similar sublexical and lexical characteristics. The following fine-grained word-level measures address these questions.

Table 1

Adjusted R^2 for linear regression models predicting word age of first production as a function of weighted Log10 number of times a word was correctly segmented by each model. Heteroskedasticity-robust standard errors are computed using a HC2 estimator. Lower Bci and Upper Bci indicate lower and upper bounds of bootstrap confidence intervals around the estimate (based on 1000 iterations). Holm's correction was applied by expanding the confidence intervals.

	Phonemic input			Syllabified input		
Model	Adjusted R^2	Lower Bci	Upper Bci	Adjusted R^2	Lower Bci	Upper Bci

Baseline	.036	.023	.052	.041	.027	.057
BTP	.044	.030	.059	.000	.000	.002
FTP	.046	.030	.060	.013	.007	.021
CLASSIC-UB final	.079	.062	.100	.021	.012	.030
CLASSIC-UB initial/final	.084	.066	.103	.038	.025	.051
PUDDLE	.078	.060	.097	.061	.043	.078

Word-level Measures

In line with the previous analysis, the models approximated children's vocabularies better than the baseline only when run on phonemic input. Therefore, in the following sections we report results for the phonemic analysis. Results of the syllabic analysis are included in Appendix S7-S10, and we also discuss this finding in Appendix S13.

Phonemic Length. Qualitatively, all models learned more short than long words (see Figure 3) as children do (e.g., Storkel, 2009). However, CLASSIC-UB (both initial and initial-final) approximated the proportion of long words learned by children better than either PUDDLE or transitional probability models. The two CLASSIC-UB models were also the only ones to outperform the baseline (see Table S7B). We discuss this result in the General Discussion.

Finally, note that PUDDLE's performance at approximating children's vocabularies by phonemic length did not differ from forward and backward transitional probability models.

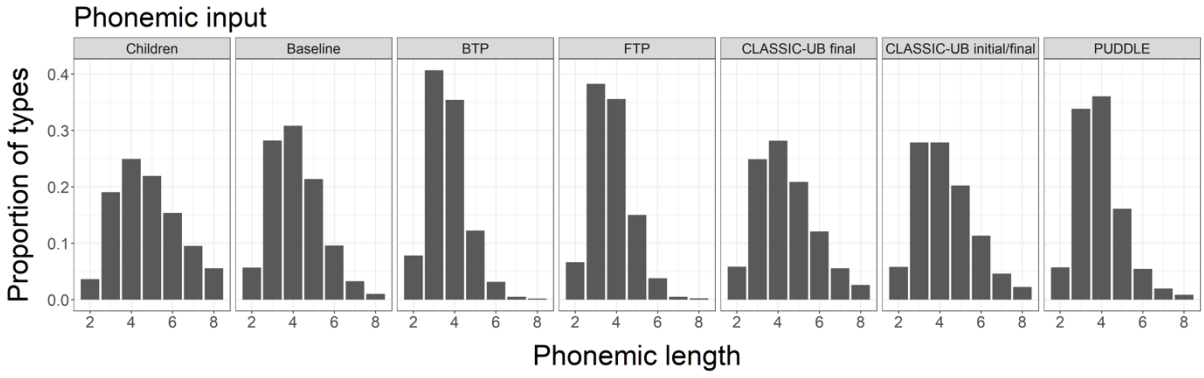


Figure 3. Proportion of word types produced by children and discovered by each model by phonemic length, when phonemic input is used.

Word Frequency. Children's vocabularies are Zipfian like the input they receive (e.g., Hendrickson & Perfors, 2019), and as such comprise more low frequency than high frequency words.

No significant difference was found between PUDDLE and CLASSIC-UB at approximating child vocabularies by word frequency (see Figure 4 and Table S8B), but chunking models outperformed transitional probability models. This result is in line with empirical evidence showing that chunking models are better than transitional probability models at capturing lexical effects (e.g., Frank et al., 2010).

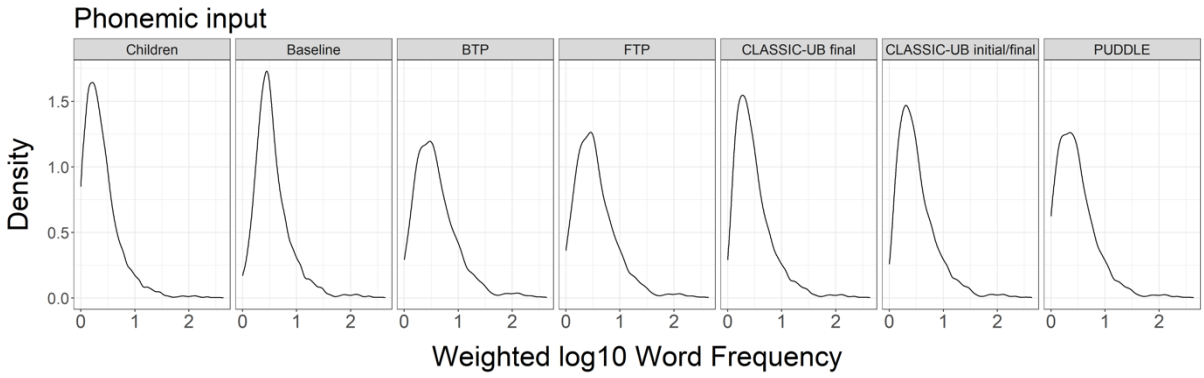


Figure 4. Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by Log10 word frequency (weighted by dividing a word frequency value by its phonemic length). Phonemic input is used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

Neighborhood Density. In line with the fact that the majority of words in the language have zero or few lexical neighbors (e.g., Vitevitch, 2008), child vocabularies are populated by a high number of low-neighborhood words. In this measure, only CLASSIC-UB final outperformed the baseline at approximating child vocabularies by neighborhood density, and this model performed significantly better than all other models (see Figure 5 and Table S9B). We return to this result in the General Discussion.

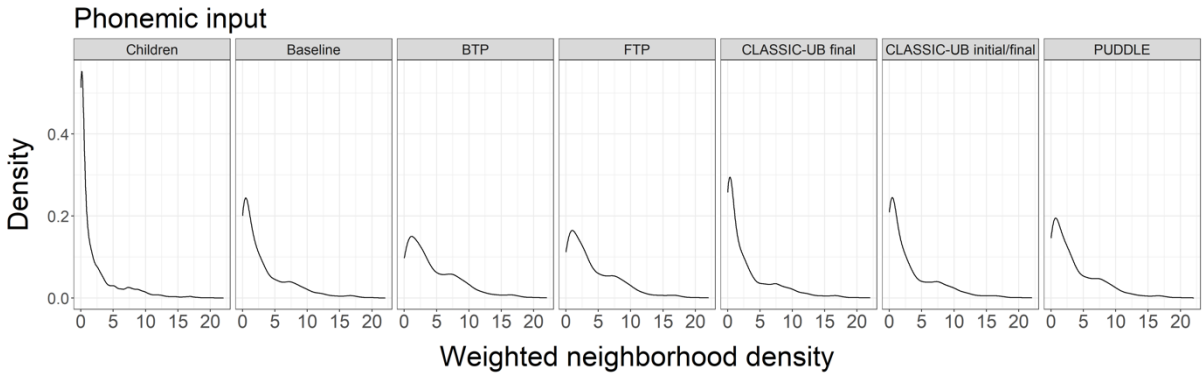


Figure 5. Distribution of unique words in child speech (Children) and discovered by each model, by neighborhood density (weighted by dividing a word neighborhood density value by its phonemic length). Phonemic input is used.

Phonotactic Probability. As shown in Figure 6, child vocabularies are populated by words with low internal predictability (e.g., Storkel, 2009). All models were equally good at approximating child vocabularies, in line with evidence showing that both chunking and transitional probability models are sensitive to sublexical regularities in the speech input. However, the models' performance did not differ statistically from the baseline model (see Table S10B), suggesting that this measure might not provide sufficient sensitivity for evaluating segmentation models.

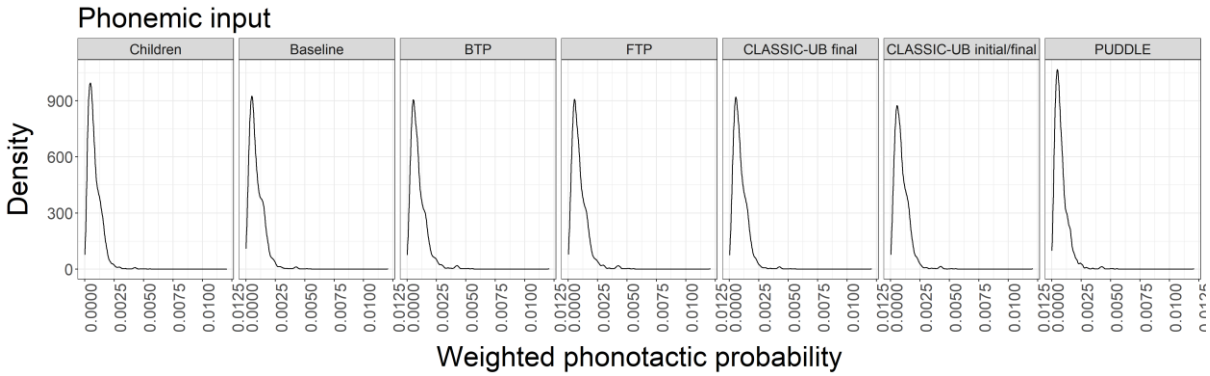


Figure 6. Distribution of unique words in child speech (Children) and discovered by each model, by phonotactic probability (weighted by dividing a word phonotactic probability value by its phonemic length). Phonemic input is used.

General Discussion

We compared CLASSIC-UB, a word segmentation model that uses naturalistic input, to another chunking model, PUDDLE, as well as to non-chunking accounts of word segmentation. We broadened the assessment of model developmental plausibility by introducing new measures that relate model performance to child corpus data. We found that CLASSIC-UB acquired a vocabulary which more closely captures child vocabularies than all other models – for example, both children and CLASSIC-UB learn a higher proportion of long and low-neighborhood words compared to other models. We discuss each of these findings in turn below.

Measures of Developmental Plausibility

In line with Larsen et al. (2017), we found that traditional evaluation measures can be inconsistent with measures based on child speech. In fact, overall, CLASSIC-UB performed better than PUDDLE at predicting measures based on child speech, despite segmenting approximately 30% fewer word tokens. One reason for this finding might be that traditional

measures represent an adult benchmark. Infants might not segment speech into the same units as adults, but might – at least initially - segment and store a protolexicon made of both word and frequent non-word units (Ngon et al., 2013). This is also consistent with different accounts (e.g., Cutler et al., 2012; Pinker, 1994) that predict that learners should commit segmentation errors based on the same cues that allow them to segment speech (e.g., rhythmic structure of the language, possible-word constraint, phonotactic constraints). Although we still do not know which specific errors - and more importantly in which proportion - infants make when segmenting naturalistic speech over the course of development, the present findings nevertheless suggest that carrying out an in-depth examination of the kind of vocabulary built by models might be a first step toward assessing models' developmental plausibility.

In Larsen et al. (2017), transitional probability models explained a higher proportion of variance in age of acquisition than chunking models. Using our adapted production measure, we showed that this result might depend on controlling for the role of word frequency. Namely, if one controls for frequency, transitional probability models do not actually perform above chance (see TP models *vs.* Baseline in Table S5A). This means that the higher performance of transitional probability models might be largely driven by input frequency. Note that this finding is not dependent on using a production measure; in a supplementary analysis (see *CDI addendum* in OSF), we examined the models' ability to predict age of acquisition based on the UK CDI (a comprehension-based measure). When the comprehension measure is not frequency-weighted, we replicate Larsen et al. (2017). But importantly, when the measure is frequency-weighted, CLASSIC-UB again performs better than other models (consistent with the production-based analyses reported here).

We suggest that our proposed set of word-level measures might provide a richer and more nuanced method to evaluate the developmental plausibility of segmentation models.

First, findings from word-level measures were in line with the age of first production results, with chunking models outperforming transitional probability, and models run on syllabified input performing at chance (see Table S7B-S10B and Figures S7A-S10A). In line with previous findings capturing in-lab data (e.g., French et al., 2011; Kurumada et al., 2013), word-level measures also showed that while both transitional probability and chunking models closely approximated child vocabularies at the sublexical level (phonotactic probability), chunking models performed better when lexical measures were considered (word length, word frequency, neighborhood density).

Secondly, word-level measures provided a more detailed test of the models' lexical characteristics, highlighting performance differences that might be attributed to architectural differences across models. Indeed, CLASSIC-UB's learning mechanism facilitates the discovery of words that overlap phonologically with previously discovered words. This allowed the model to approximate a greater proportion of children's long/low-neighborhood words than competing models (see Figure 3 and 5). Therefore, uniquely relying on mechanisms that privilege highly probable sequences (e.g., PUDDLE, transitional probability models) makes it difficult to capture a portion of long/low-neighborhood words that are generally more difficult to learn, but that children nevertheless learn and that CLASSIC-UB can learn by exploiting phonological overlap. Interestingly, this feature of CLASSIC's learning mechanism also means the model can account for nonword repetition effects (Jones, 2016) that are due to phonological overlap across word and nonword sequences. Similarly, it is possible that CLASSIC-UB captures additional processes of storage and recall involved in word production (i.e., going beyond aspects of segmentation), and that this sensitivity explains its superior performance in approximating the characteristics of children's productions.

Although CLASSIC-UB more accurately represents the make-up of children's early lexicons, its accuracy in segmenting words is not quite as good as that of PUDDLE (i.e., PUDDLE has a larger vocabulary). One could therefore argue that at earlier stages in PUDDLE's learning, word-level characteristics may match those of CLASSIC-UB and that it is only the subsequent increase in PUDDLE's vocabulary that skews their distribution. We conducted additional analyses (see Appendix S12) to evaluate this possibility. These analyses show that differences in vocabulary size do not explain the differences in word-level measures (see Figure S12A).

Finally, to support our claim regarding the role of overlapping phonological sequences in CLASSIC-UB, we conducted an additional exploratory analysis showing that CLASSIC-UB's ability to better approximate children's vocabulary in word length and neighborhood density increases as word frequency increases (see Figure S12B). This is in line with recent work showing that frequent words are more likely to share phonological material with previously learned words, therefore boosting child learning compared to less frequent words (Jones et al., 2022). Our result is also in line with evidence showing an effect of overlapping phonological sequences on vocabulary learning at around 2 years of age (e.g., Jones et al., 2022; Stokes, 2010; Storkel, 2009), but no effect at 12–15 months (Swingley & Humphrey, 2018), suggesting that children first build a diverse repertoire of phonological chunks which later boost word learning (for a computational test of this idea using CLASSIC, see Jones & Rowland, 2017).

Overall, our results speak in favor of models that exploit phonological overlap between sequences in word segmentation (e.g., French et al., 2011; Perruchet et al., 1998) and add to previous work which highlights the significant role of this in word processing and acquisition (Gathercole, 1995; Jones et al., 2021).

Limitations and Future Directions

We have shown that chunking might play a significant role in early word segmentation by comparing our new chunking-based segmentation model CLASSIC-UB to two other influential models (transitional probability, PUDDLE). However, there are additional models we have not considered. One important class of Bayesian models assumes that infants formulate hypotheses on the possible segmentations of utterances, ultimately preferring those segmentations that contain few frequent and short chunks (e.g., Goldwater et al., 2006; 2009). Another account is that infants form chunks based on both frequency and transitional probabilities (forward and backward) of syllable sequences (i.e., mutual information-based clustering; Swingley, 2005). Given that these accounts are primarily driven by frequency information, future comparisons to CLASSIC-UB are important to support our conclusion that phonological overlap between sequences plays a role in the segmentation process in addition to frequency. Such comparisons would also be important because one influence does not exclude the other. As argued above, CLASSIC-UB's encoding efficiency uniquely increases when items become connected to others: i.e., the more opportunities to chunk sublexical items the faster lexical representations are formed. However, once CLASSIC-UB has extracted a word representation from the input, it could further benefit from tracking its frequency in the input (e.g., see Jones, Justice et al., 2020 on how a frequency-tracking mechanism might improve CLASSIC's performance).

Moreover, it is highly likely that early naturalistic segmentation involves the use of a combination of cues. Indeed, the results of this study indicate that chunking alone might not be enough to discover items that are very long (Figure 3), occur very infrequently (Figure 4), receive no facilitation from word neighbors (Figure 5) and are made up of improbable sequences of sounds (Figure 6). This suggests that CLASSIC-UB might need to have access to additional cues to word boundary to be able to account for children's ability to learn these

words. We know that infants use a wide range of cues when segmenting speech, such as prosodic salience of phrase edges (Gout at al., 2004), alternative ways to pronounce specific phonemes (i.e., allophonic variation; Hohne & Jusczyk, 1994), stress patterns (Jusczyk et al., 1999), degree of coarticulation of speech sounds (Johnson & Jusczyk, 2001) and others. Such cues could be considered in future work.

An alternative (and non-mutually exclusive) possibility is that long, infrequent items with few neighbors might be learned via generalization of linguistic structures at different levels (e.g., syntactic, Lippeveld & Oshima-Takane, 2020). For example, in Abend et al. (2017) an ideal Bayesian learner performed one-shot learning (i.e., formation of new word representations from a single exposure) by leveraging the mapping of words to their syntactic categories. Examining the role of syntactic categories would be important in future work as evidence shows that infants' development of grammatical knowledge starts in parallel with the acquisition of phonology and the lexicon (e.g., Marino et al., 2020).

Aside from our focus on a single word segmentation cue, another limitation is that we have not considered the models' ability to capture the role of additional factors in word segmentation and learning. For example, Swingley and Humphrey (2018) showed that word concreteness, word frequency in isolation (i.e., frequency with which a word occurs in a single-word utterance), and syntactic category predict word learning at 12 and 15 months. These predictors could be included in the statistical models of age of acquisition/production – alongside our word-level predictors - to see how they moderate models' accuracy (i.e., number of correct word segmentations). Alternatively, our word-level evaluation measure could be extended to examine whether segmentation models can capture the distributions of these additional word-level features in children's vocabularies. We would expect models to better capture characteristics they are sensitive to (e.g., it is likely that chunking models would show sensitivity to word frequency in isolation, Kurumada et al., 2013).

Moreover, including these additional factors would be important because they differently impacted word comprehension and production in Swingley and Humphrey (2018): word concreteness only predicted word comprehension, and the effect of word frequency in isolation was moderated by syntactic category type only in word comprehension. Although we have highlighted limitations in using comprehension measures to investigate how well segmentation models perform, methods that look at comprehension and production should be considered complementary. Comparing comprehension and production would also allow us to test the extent to which CLASSIC-UB captures processes that are uniquely involved in production (such as recall and articulation).

We would also like to highlight limitations deriving from the use of phoneme-based input adopted in the current study. The models did not have to deal with the complex problem of gradually abstracting phonological categories. Under an early phonetic learning approach (e.g., Werker, 2018), infants have to learn the relations between different realizations of phonemes based on contextual variation or lexical contrast (e.g., aspirated stops and unreleased stops are allophones of the phoneme /t/). Addressing this limitation in future work is important to increase the developmental plausibility of the investigations. Alternatively, under more recent approaches, the goal of infant speech perception may not be learning discrete phonetic categories, but instead representing continuous dimensions of raw speech (e.g., spectral energy) that are relevant to the native language (i.e., perceptual space learning; Feldman et al., 2021; McMurray, 2022). This implies that future work would need to consider more gradient units of speech perception. For example, recent work by Schatz et al. (2021) showed that a distributional learner can learn to discriminate phonetic contrasts by clustering auditory features into categories which are significantly smaller and more variable than traditional phonetic categories. Finally, we acknowledge that an early phonetic learning approach used in the present work is also in contrast with other accounts that do not assume

phonemes as basic units of perception, for example work that argues for gradient units dependent on the temporal unfolding of speech (e.g., Browman & Goldstein, 1992; Bybee, 2001; Mowrey & Pagliuca, 1995; Port & Leary, 2005) or others that argue for features or morphophonemic forms (e.g., Chomsky & Halle, 1965; Postal, 1968).

Conclusion

Our goal in this paper was to test whether a chunking-based mechanism that has previously been successful in capturing early vocabulary learning might play a significant role in infant word segmentation. We constructed CLASSIC-UB, which forms chunks of phonological and utterance-boundary material. Our simulations make three important contributions: (1) They offer an existence proof that utterance boundaries carry useful information for word segmentation (2) Age of production and word level measures can sensibly be used to evaluate model performance (3) CLASSIC can be augmented to form the segmentation model CLASSIC-UB, consistent with the hypothesis that chunking might be an important mechanism in early naturalistic word segmentation.

Notes

1. For ease of exposition, the example uses IPA phonetic transcription. However, note that in our simulations we used a transcription based on the CMU pronouncing dictionary (Lenzo, 2007; see an example in Figure 1).
2. However, CLASSIC's encoding does not allow partial activation of chunks (unlike in Baayen et al., 2011).
3. The CDI words and gestures includes 373 phonological words (not considering homophone duplicates) typically acquired by infants between 8 and 18 months of age. Our final sample comprised 330 words after filtering for those CDI words present in the

child-directed input the segmentation models received (i.e., CDI words that the models had the opportunity to learn).

4. A discussion about the effect of sample size reduction when using the age of acquisition measure from the CDI can be found in the file *CDI_addendum*, at https://doi.org/10.17605/OSF.IO/KBNEP
5. Adjusted R^2 estimates cannot typically be directly compared to R^2 estimates. However, because of our large sample size, adjusted R^2 and R^2 estimates and confidence intervals are actually identical, allowing us to compare our estimates to Larsen et al.'s (2017) R^2ones. In fact, as sample size increases expected R^2 estimates become less biased approaching adjusted R^2 unbiased estimates of the population explained variance (Karch, 2020).

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition, 164*, 116–143. https://doi.org/10.1016/j.cognition.2017.02.009

Alcock, K. (2020). *The UK communicative development inventory database: words and gestures ages 8-18 months 2012-2016*. [data collection]. UK Data Service. SN: 853687, http://doi.org/10.5255/UKDA-SN-853687

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9781315806822

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical

processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity, 2019,* e4895891. https://doi.org/10.1155/2019/4895891

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*(3), 438–481. https://doi.org/10.1037/a0023851

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*(2), 167–206. https://doi.org/10.1016/S0010-0277(02)00002-1

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods, 52*(1), 264–278. https://doi.org/10.3758/s13428-019-01223-3

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4,* 247–260. https://doi.org/10.1016/S0163-6383(81)80027-6

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. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*(3–4), 155–180. https://doi.org/10.1159/000261913

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics, 1*(2), 97–138. https://doi.org/10.1017/S0022226700001134

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. https://doi.org/10.1080/016909698386528

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2021). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods, 53*(3), 945–976. https://doi.org/10.3758/s13428-020-01356-w

Cinque, G. (1993). A null theory of phrase and compound stress. *Linguistic Inquiry*, *24*(2), 239–297.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120*(3), 235. https://doi.org/10.1037//0096-3445.120.3.235

Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics, 39*(3), 388–402. https://doi.org/10.1016/j.wocn.2011.02.004

Cristià, A., Seidl, A., & Gerken, L. (2011). Learning classes of sounds in infancy. *University of Pennsylvania Working Papers in Linguistics, 17*(1), 9.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.

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

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, *92*, 37–64. https://doi.org/10.1016/j.cogpsych.2016.11.004

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3). https://doi.org/10.1111/j.1467-9280.2007.01885.x

Fais, L., Werker, J., F., Cass, B., Leibowich, J., Barbosa, A., V., & Vatikiotis-Bateson, E. (2012). Here's looking at you, baby: What gaze and movement reveal about minimal pair word-object association at 14 months. *Laboratory Phonology*, *3*(1), 91–124. https://doi.org/10.1515/lp-2012-0007

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

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222. https://doi.org/10.1111/j.1467-8624.2011.01692.x

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. https://doi.org/10.1016/j.cognition.2010.07.005

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614. https://doi.org/10.1037/a0025255

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition, 23*, 83–94. https://doi.org/10.3758/BF03210559

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*, 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4.

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

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the annual meeting of the association for computational linguistics* (Vol. 44, p. 673).

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, *51*(4), 548–567. https://doi.org/10.1016/j.jml.2004.07.002

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00555

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. https://doi.org/10.1016/j.cogpsych.2011.06.002

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11-22. https://doi.org/10.1016/j.cognition.2019.03.005

Hoff, E., Core, C., & Bridges, K. (2008). Non-word repetition assesses phonological memory and is related to vocabulary development in 20- to 24-month-olds. *Journal of Child Language*, *35*(4), 903–916. https://doi.org/10.1017/S0305000908008751

Hohne, E. A., & Jusczyk, P. W. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, *56*(6), 613-623. https://doi.org/10.3758/BF03208355

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

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition*, *153*, 79–88. https://doi.org/10.1016/j.cognition.2016.04.017

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Jones, G., Cabiddu, F., & Avila-Varela, D. S. (2020). Two-year-old children's processing of two-word sequences occurring 19 or more times per million and their influence on subsequent word learning. *Journal of Experimental Child Psychology*, *199*, 104922. https://doi.org/10.1016/j.jecp.2020.104922

Jones, G., Cabiddu, F., Andrews, M., Rowland, C. (2021). Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *Journal of Memory and Language*, *119*, 104232. https://doi.org/10.1016/j.jml.2021.104232

Jones, G., Cabiddu, F., Barrett, D., Castro, A., & Lee, B. (2022). How the characteristics of words in child-directed speech differ from adult-directed speech to influence children's productive vocabularies [Manuscript submitted for publication]. *First Language*.

Jones, G., Justice, L. V., Cabiddu, F., Lee, B. J., Iao, L.-S., Harrison, N., & Macken, B. (2020). Does short-term memory develop? *Cognition*, *198*, 104200. https://doi.org/10.1016/j.cognition.2020.104200

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

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

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants′ Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language*, *33*(5), 630–645. https://doi.org/10.1006/jmla.1994.1030

Karch, J. (2020). Improving on Adjusted R-Squared. *Collabra: Psychology*, *6*(1), 45. https://doi.org/10.1525/collabra.343

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453. https://doi.org/10.1016/j.cognition.2013.02.002

Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. *Interspeech 2017*, 2198–2202. https://doi.org/10.31219/osf.io/86tu3

Lenzo, K. (2007). *The CMU pronouncing dictionary*. Carnegie Melon University.

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics* (Vol. 30, pp. 13–15).

Lippeveld, M., & Oshima-Takane, Y. (2020). Children's initial understanding of the related meanings of polysemous noun-verb pairs. *Language Learning and Development*, *16*(3), 244–269. https://doi.org/10.1080/15475441.2020.1737073

Mani, N., & Plunkett, K. (2010). Twelve-month-olds know their cups from their keps and tups. *Infancy*, *15*(5), 445–470. https://doi.org/10.1111/j.1532-7078.2009.00027.x

Marino, C., Bernard, C., & Gervain, J. (2020). Word frequency is a cue to lexical category for 8-month-old infants. *Current Biology*, *30*(8), 1380-1386.e3. https://doi.org/10.1016/j.cub.2020.01.070

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

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1–51. https://doi.org/10.1037/rev0000126

McGillion, M., Herbert, J. S., Pine, J., Vihman, M., dePaolis, R., Keren-Portnoy, T., & Matthews, D. (2017). What Paves the Way to Conventional Language? The Predictive Value of Babble, Pointing, and Socioeconomic Status. *Child Development, 88*(1), 156–166. https://doi.org/10.1111/cdev.12671

McMurray, B. (2022). The acquisition of speech categories: Beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. *Language, Cognition and Neuroscience*, *0*(0), 1–27. https://doi.org/10.1080/23273798.2022.2105367

Mersad, K., Kabdebon, C., & Dehaene-Lambertz, G. (2021). Explicit access to phonetic representations in 3-month-old infants. *Cognition, 213*, 104613. https://doi.org/10.1016/j.cognition.2021.104613

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*(3), 545–564. https://doi.org/10.1017/S0305000909990511

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A, 50*(3), 528-559. https://doi.org/10.1080/027249897392017

Mowrey, R., & Pagliuca, W. (1995). The reductive character of articulatory evolution. *Rivista di linguistica, 7*, 37-124.

Newman, R. S., Rowe, M. L., & Bernstein Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language, 43*(5), 1158–1173. https://doi.org/10.1017/S0305000915000446

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology, 42*(4), 643–655. https://doi.org/10.1037/0012-1649.42.4.643

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science, 16*(1), 24–34. https://doi.org/10.1111/j.1467-7687.2012.01189.x

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: 8-month-old infants track backward transitional probabilities. *Cognition, 113*(2), 244–247. https://doi.org/10.1016/j.cognition.2009.07.011

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

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263. https://doi.org/10.1006/jmla.1998.2576

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

Pinker, S. (1994). *The language instinct*. Penguin Books.

Port, R. F., & Leary, A. P. (2005). Against Formal Phonology. *Language*, *81*(4), 927–964.

Postal, P. (1968) *Aspects of phonological theory*. Harper & Row.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). 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

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological science*, *8*(2), 101-105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Saksida, A., Langus, A., & Nespor, M. (2016). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science, 20*(3), e12390. https://doi.org/10.1111/desc.12390

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

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science, 9*(6), 565–573. https://doi.org/10.1111/j.1467-7687.2006.00534.x

Seidl, A., & Johnson, E. K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of child language, 35*(1), 1-24. https://doi.org/10.1017/S0305000907008215

Skarabela, B., Ota, M., O'Connor, R., & Arnon, I. (2021). 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition, 211*, 104612. https://doi.org/10.1016/j.cognition.2021.104612

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of speech, language, and hearing research, 53*(3), 670–683. https://doi.org/10.1044/1092-4388(2009/08-0254)

Stokes, S. F. (2014). The impact of phonological neighborhood density on typical and atypical emerging lexicons. *Journal of Child Language*, 41(3), 634-657. https://doi.org/10.1017/S030500091300010X

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech,*

Language, and Hearing Research, *47*(6), 1454–1468. https://doi.org/10.1044/1092-4388(2004/108)

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of child language*, *36*(2), 291-321. https://doi.org/10.1017/S030500090800891X

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

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific English words. *Child Development*, *89*(4), 1247–1267. https://doi.org/10.1111/cdev.12731

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*, 408–422. https://doi.org/10.1044/1092-4388(2008/030)

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325-329. https://doi.org/10.1111/1467-9280.00064

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728. https://doi.org/10.1017/S0142716418000152

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717. https://doi.org/10.1121/1.402450

Yeung, H. H., Chen, L. M., & Werker, J. F. (2014). Referential Labeling Can Facilitate Phonetic Learning in Infancy. *Child Development*, *85*(3), 1036–1049. https://doi.org/10.1111/cdev.12185

Zipf, G. (1936). *The psychobiology of language*. London: Routledge.

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Computational models.

Appendix S2. Input preprocessing.

Appendix S3. Word age of first production estimation.

Appendix S4. Comparison of Precision and Recall measures.

Appendix S5. Frequency-weighted age of first production analyses: Pairwise differences between models' adjusted R^2.

Appendix S6. Frequency-unweighted age of first production analyses.

Appendix S7. Approximation of child production vocabulary by phonemic length.

Appendix S8. Approximation of child production vocabulary by weighted log10 word frequency.

Appendix S9. Approximation of child production vocabulary by weighted neighborhood density.

Appendix S10. Approximation of child production vocabulary by weighted phonotactic probability.

Appendix S11. CLASSIC-UB initial-final vs. CLASSIC-UB final.

Appendix S12. Does PUDDLE represent a child with more advanced vocabulary knowledge?.

Appendix S13. Controlling for baseline segmentation performance.