Supporting Information for: Cabiddu, F., Bott, L., Jones, G., & Gambi, C. Classic-Utterance-Boundary: A chunking-based model of early naturalistic word segmentation. Article accepted in Language Learning on 23 November 2022.

Appendix S1: Computational models

Transitional probability models

The implementation of forward and backward transitional probability models (FTP and BTP, respectively) followed previous studies (e.g., Frank et al., 2010; Larsen et al., 2017; Saksida et al., 2016), in which the transitional probability of a phoneme/syllable pair is computed as:

$$Forward\ TP(U_{t-1}, U_t) = \frac{F(U_{t-1}, U_t)}{F(U_{t-1})} \quad ; \quad Backward\ TP(U_{t-1}, U_t) = \frac{F(U_{t-1}, U_t)}{F(U_t)}$$

Where $F(U_{t-1}, U_t)$ is the frequency of a pair of units (two phonemes or syllables), while $F(U_{t-1})$ and $F(U_t)$ are the frequencies of the first and second unit respectively. We used a strictly incremental version of these models, in which transitional probabilities are updated at every utterance. A word boundary is placed within a phoneme/syllable target pair if the transitional probabilities of the surrounding pairs are both greater than the target pair transitional probability (i.e., relative threshold). Utterance boundaries are used as additional units available to the models, therefore for the phoneme pair ↲*h* in [↲*hellobaby*↲], $F(U_{t-1})$ would correspond to the frequency of the utterance-initial marker ↲ and $F(U_t)$ to the frequency of the phoneme *h*.

Although using an absolute threshold has been shown to increase models' performance at Precision and Recall measures (Gambell & Yang, 2006; Saksida et al., 2016), we instead used a relative threshold where word boundaries are posited based on the transitional probabilities of the surrounding biphones or syllable pairs. The choice of a relative threshold is consistent with studies showing that infants segment at local minima of

transitional probability (e.g., Saffran et al., 1996; 1999), while we are not aware of any experimental findings that provide direct evidence for an absolute threshold mechanism. Further, we used a strictly incremental version of the transitional probability models (i.e., word boundaries are set based on *current* transitional probabilities of surrounding pairs), to match CLASSIC-UB and PUDDLE's incremental way of learning. Note that one could apply the same incremental principle to an absolute threshold, by updating a running average; indeed, absolute transitional probabilities can fall out of predictive incremental learning models (e.g., Baayen et al., 2013; Harmon & Kapatsinski, 2021).

PUDDLE

PUDDLE (Monaghan & Christiansen, 2010) parses utterances phoneme by phoneme, searching for a matched string in its lexicon (we also adapted the original model to process the input syllable by syllable). At the start of the segmentation process, whole sentences are stored in the lexicon as this is initially empty. Items in the lexicon are ranked by absolute frequency of occurrence (which guides further string matching). The frequency of an item is updated every time it is discovered in the input, making the model strictly incremental. The lexicon in PUDDLE stores chunks that can begin or end utterances and these can comprise phonemes, phoneme pairs, or longer sequences of phonemes up to whole utterances. When PUDDLE finds a match in the lexicon, it only recognizes the item if (1) there is an item on its left which ends with a previously encountered ending, and (2) there is an item on its right which begins with a previously encountered beginning.

Random baseline

We chose to implement a fully random baseline which relies on a random coin toss to place a boundary after each input unit (Lignos, 2012). This baseline represents a scenario in which a

child would segment the input by making random guesses on word boundary locations and tends to mostly segment short and frequent words as the input gives more opportunities to correctly segment them (Grimm et al., 2017) – i.e., these words are more likely to be discovered by chance. Comparing to chance is informative because, ideally, we would want a more complex model, which implements a specific segmentation mechanism, to at least perform better than chance. A fully random baseline is also more informative than baselines which consider each utterance or each unit as a word (e.g., Bernard et al., 2020). These baselines would only discover a very low proportion of word types from the phonemic input (an utterance baseline would only discover types that appear as one-word utterances, while a unit baseline would only discover mono-phonemic word types). Finally, pseudo-random baselines are problematic because of their prior knowledge assumptions: For example, it is unlikely that infants have knowledge of the true probability of a word boundary to occur in the language (oracle baseline; e.g., Bernard et al., 2020), or the true average word length in cross-linguistic terms (Loukatou et al., 2019).

Appendix S2: Input preprocessing

The 7 CHILDES corpora used were: Belfast (Henry, 1995), Manchester (Theakston et al., 2001), Thomas (Lieven et al., 2009), Tommerdahl (Tommerdahl & Kilpatrick, 2013), Wells (Wells, 1981), Forrester (Forrester, 2002), Lara (Rowland & Fletcher, 2006). The corpora were imported into the R environment (R Core Team, 2018) using the package childesr (Braginsky et al., 2019), which guarantees a standardized procedure for obtaining the utterance samples. The corpora were phonetically transcribed using the CMU dictionary (Lenzo, 2007). The transcription process was carried out without considering word stress markers in the dictionary. Utterances containing one or more words not appearing in the CMU were discarded.

The advantage of using a transcription dictionary is that it allows automatic transcription of large input corpora into phonetic form. However, it has the important limitation of assuming that words always consist of the same phonemes in running speech. This is not the case as words undergo significant phonetic reduction in conversational speech (e.g., *until* [ʌntɪl] may be also realized via phoneme deletion [_ntil] or substitution [ʌntəl]; see Johnson, 2004). Addressing this limitation would require access to either phonetically transcribed corpora which include different word realizations (e.g., Schuppler et al., 2011), or systems that directly operate on raw speech (e.g., Arnold et al., 2017; Ten Bosch et al., 2022).

The corpora differed by MLU (Mean Length of Utterance; see Table S2A). If utterances are not shuffled, the models' performance oscillates depending on the corpus MLU. This happens because long sentences are more difficult to segment for all segmentation models. Given the input to different children is likely to show variability in MLU across time, we controlled for this variation by randomly shuffling the utterances' order; given this variation influenced all models equally, this choice should not affect comparisons between models.

When required, the syllabification of the input was performed using the WordSeg package (Bernard et al., 2020), which applies the maximal onset principle (Phillips & Pearl, 2015). Note we are not claiming such procedure corresponds to how the infant would segment the input into syllables (for work focused on this problem, see Räsänen et al., 2018), as by definition the maximal onset principle requires prior knowledge of word onsets. Rather, it is a convenient deterministic strategy for pre-syllabifying the corpora, which can then be used as input for the models under the assumption that infants might be already organizing speech as strings of syllabic constituents before they have started representing word forms (e.g., Bertoncini & Mehler, 1981; Bertoncini et al., 1988; Bijeljac-Babic et al., 1993). Further, it is worth noting that the maximal onset principle is not the only strategy that could

be used, as other factors can influence English syllabification (e.g., word-edge frequency, stress, vowel quality, sonority, morphology; Derwing, 1992; Derwing & Eddington, 2014; Olejarczuk & Kapatsinski, 2018).

The input for the models were all utterances from the 7 corpora that were directed to children of age 2 (see Table S2A). We believe this choice to focus on age 2 is justified for two reasons. Firstly, at age 2 a larger amount of data on children's own productions is available in the corpora. Since we evaluate our models on measures that are based on child productions (i.e., age of first production and word-level measures), focusing on age 2 allows us to test the models on a much larger sample of word types. At ages earlier than 2 years child productions decrease significantly in type frequency (e.g., at year 1 child word types are about 1/4 of year 2 word types) which would significantly limit the sample of words used to compute our evaluation measures.

Table S2A.

Descriptive statistics of phonetically transcribed CHILDES English corpora filtering for utterances directed to children of age 2. For each corpus, the table indicates the number of input utterances, Mean Length of Utterance (MLU, i.e., average number of words in an utterance), number of words including repetitions (Word tokens), number of different words (Word types).

Corpus	Utterances	MLU	Word tokens	Word types
Forrester	3,183	4.89	15,567	1,576
Tommerdahl	5,700	4.84	27,610	1,646
Wells	16,292	3.62	59,042	3,053
Belfast	17,923	5.52	99,004	3,922
Lara	59,598	3.68	219,184	4,316

Thomas	194,695	5.23	1,018,726	8,160
Manchester	307,079	3.96	1,215,740	9,587

Secondly, the corpora also contain many more utterances directed to 2 year olds than to younger children. The input available in CHILDES at year 0 or 1 is significantly smaller in size compared to input directed to year 2. For example, our 2-year-olds' input comprises 604,000 utterances, while 1-year-olds' input only contains 54,274 utterances. This is problematic because a smaller sample of utterances is more likely to be biased and less likely to preserve the characteristics of naturalistic speech directed to young children. Thus, focusing on age 2 represents a compromise: This is the youngest age group for which a large enough (and thus representative) sample of child-directed speech is available.

To illustrate this point further, we have generated Figure S2A below. In this figure (panels "Raw" of each lexical measure), one can see that the characteristics of the input change significantly from age 0 to 2, with presence of more long, infrequent, low-neighborhood and high-phonotactic words as age increases. Crucially, we can show that these differences are mostly due to differences in sample size (see also Montag et al., 2018). This is because the likelihood of finding long/infrequent/low-neighborhood/high-phonotactic words increases as sample size increases. Indeed, when we match input at different age bins by sample size (i.e., sampling the same number of utterances as in the smallest age sample) we see that the differences between corpora decrease substantially (see "Matched" panels of figure S2A). Therefore, the loss from choosing input directed to age 2 (i.e., maximising sample size at the expense of input age) is lower compared to choosing input smaller in size at earlier ages, which would instead grossly misrepresent the characteristics of the naturalistic input.

In conclusion, although some differences between speech at different ages remain when we control for input sample size and it is thus possible that the age 2 input is not entirely representative of the input at younger ages, the input that we have available for younger children is almost certainly not representative of naturalistic input directed to children of those ages either, because so little of it is available in existing corpora. Future studies may try and replicate these analyses using speech directed at earlier ages, once large-scale corpora of language input directed at such earlier ages become available to researchers.

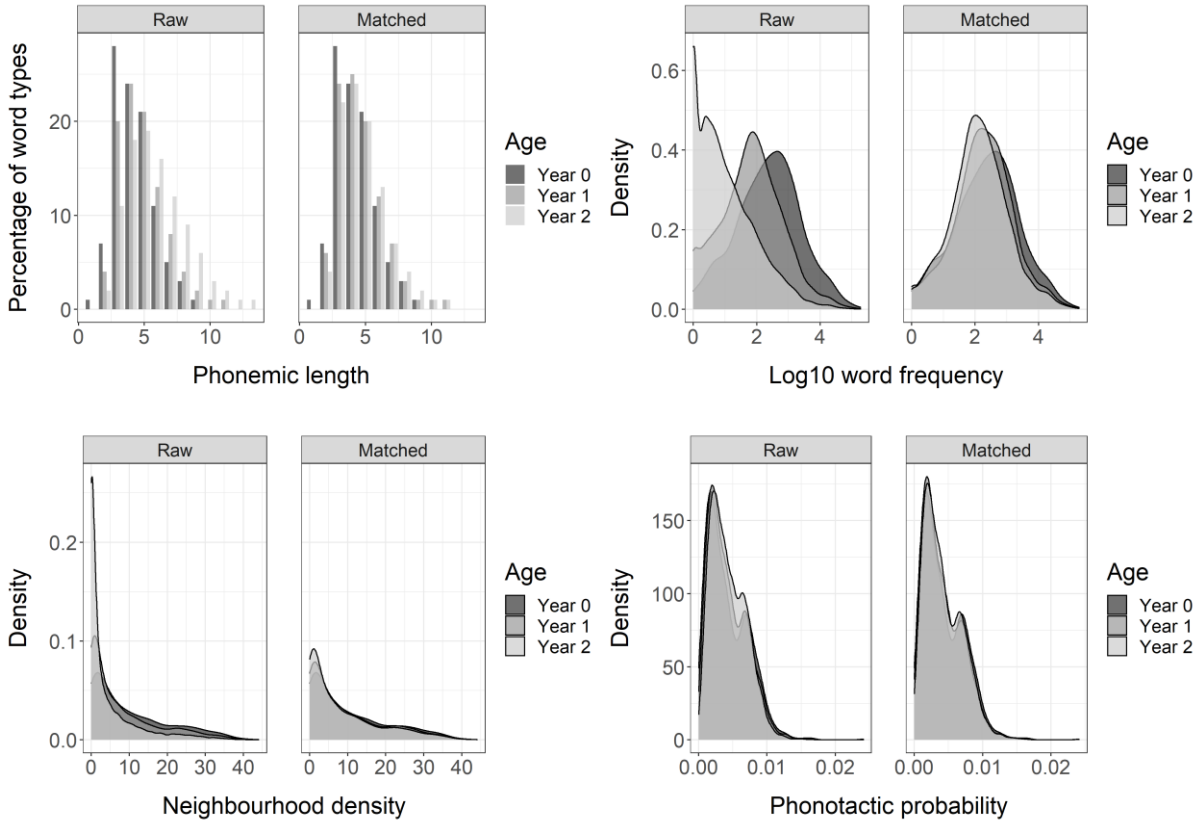

Figure S2A. Word characteristics of word type distributions for input directed at year 0, 1, and 2. "Raw" panels show word characteristics when considering all utterances available at each age (age 0 = 11,745; age 1 = 54,274; age 2 = 604,000). Age 1 and 2 utterances were taken from the same corpora used in the main manuscript, while age 0 utterances were taken from the Korman corpus (Korman, 1992), which contains maternal speech directed to infants

aged between 4 and 16 weeks. “Matched” panels refer to word type distributions when each input is matched by age 0 sample size, therefore randomly sampling 11,745 utterances from age 1 and age 2 corpora. Also note that results do not depend on the particular random samples computed, as repeating the sampling procedure produces identical distributions.

Appendix S3: Word age of first production estimation

Word age of first production has been used in Grimm et al. (2017; 2019) as an index of word learning. If a word is first produced early in development, it is assumed that this is in part because it is easy to learn. To compute word age of first production estimates, we used Grimm et al.’s (2017; 2019) procedure, as its validity was assessed in two ways: corpora age of first production estimates showed a fairly strong correlation with American English CDI parent-report measures of child expressive vocabulary (*Spearman’s rho* = .50, p <.001) and a stronger correlation with the only estimates for British English that are directly derived from children (i.e., from a picture-naming task; Morrison et al., 1997; *Spearman’s rho* = .65, p <.001).

To estimate word age of first production, MLU was used as a proxy of the developmental stage at which a word is acquired. For a given word, we first computed MLU for each transcript via bootstrapping (to compensate for differences in number of utterances). The lowest MLU across transcripts was then taken as age of first production value in order to correct for inflation (as it is likely that children knew a target word before they produced it in the recordings). This method also avoided having to find a set of common words across corpora to calculate a mean stage; the latter would mean discarding a high amount of low frequency words that do not appear in all corpora, resulting in a skewed set of high-frequency words.

Appendix S4: Comparison of Precision and Recall measures.

A narrative account of the findings in Table S4A and S4B is included in the paper at section *Results and Discussion / Precision and Recall*.

Table S4A

Comparison of Precision and Recall measures when phonemic input is used. Pairwise comparisons via Welch's t-test for unequal variances. *p* values and bootstrap 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). The table shows: type of comparison and accuracy measure, mean accuracy of first and second model considered (M1 and M2; e.g., in the first row, M1 = BTP and M2 = PUDDLE), difference between mean accuracy values (ΔM), t value, corrected *p* value, degrees of freedom (df), lower and upper cut-offs of corrected bootstrap 95% confidence intervals around the difference (Lower and Upper Bci). FTP = Forward Transitional Probability; BTP = Backward Transitional Probability

Comparison	Measure	M1	M2	ΔM	t	p	df	Lower Bci	Upper Bci
BTP vs. PUDDLE	Recall	.45	.79	-.33	-73.05	<.001	19,044.5	-.348	-.32
BTP vs. PUDDLE	Precision	.42	.73	-.32	-66.73	<.001	19,681.9	-.333	-.304
BTP vs. Baseline	Recall	.45	.17	.28	60.43	<.001	19,391.0	.266	.293
BTP vs. Baseline	Precision	.42	.14	.27	59.57	<.001	19,094.8	.258	.284
BTP vs. CLASSIC-UB initial/final	Precision	.42	.5	-.09	-16.96	<.001	19,989.6	-.101	-.071
BTP vs. CLASSIC-UB final	Precision	.42	.49	-.07	-13.26	<.001	19,936.1	-.084	-.055
BTP vs. FTP	Recall	.45	.51	-.05	-10.31	<.001	19,998.0	-.066	-.038
BTP vs. FTP	Precision	.42	.47	-.05	-10.26	<.001	19,997.8	-.067	-.037
BTP vs. CLASSIC-UB initial/final	Recall	.45	.5	-.04	-8.58	<.001	19,988.7	-.058	-.028
BTP vs. CLASSIC-UB final	Recall	.45	.45	.01	1.17	.243	19,907.3	-.004	.016
FTP vs. Baseline	Recall	.51	.17	.33	71.56	<.001	19,381.2	.318	.346
FTP vs. Baseline	Precision	.47	.14	.32	70.78	<.001	19,073.4	.31	.338
FTP vs. PUDDLE	Recall	.51	.79	-.28	-61.57	<.001	19,032.8	-.294	-.266
FTP vs. PUDDLE	Precision	.47	.73	-.26	-55.74	<.001	19,668.1	-.279	-.251
CLASSIC-UB final vs. Baseline	Precision	.49	.14	.34	72.09	<.001	18,639.9	.326	.355
CLASSIC-UB final vs. PUDDLE	Recall	.45	.79	-.34	-71.29	<.001	18,476.8	-.355	-.321

CLASSIC-UB final vs. Baseline	Recall	.45	.17	.27	56.75	<.001	18,895.5	.259	.289
CLASSIC-UB final vs. PUDDLE	Precision	.49	.73	-.25	-50.59	<.001	19,362.6	-.265	-.231
CLASSIC-UB final vs. FTP	Recall	.45	.51	-.06	-11.13	<.001	19,911.2	-.073	-.042
CLASSIC-UB final vs. CLASSIC-UB initial/final	Recall	.45	.5	-.05	-9.44	<.001	19,955.8	-.066	-.034
CLASSIC-UB final vs. CLASSIC-UB initial/final	Precision	.49	.5	-.02	-3.36	.003	19,973.2	-.032	-.004
CLASSIC-UB final vs. FTP	Precision	.49	.47	.02	3.28	.003	19,942.2	.004	.028
CLASSIC-UB initial/final vs. Baseline	Precision	.5	.14	.36	77.5	<.001	18,935.1	.345	.374
CLASSIC-UB initial/final vs. Baseline	Recall	.5	.17	.32	68.94	<.001	19,245.2	.311	.339
CLASSIC-UB initial/final vs. PUDDLE	Recall	.5	.79	-.29	-62.62	<.001	18,872.7	-.304	-.274
CLASSIC-UB initial/final vs. PUDDLE	Precision	.5	.73	-.23	-47.98	<.001	19,575.9	-.247	-.215
CLASSIC-UB initial/final vs. FTP	Precision	.5	.47	.03	6.79	<.001	19,991.8	.02	.047
CLASSIC-UB initial/final vs. FTP	Recall	.5	.51	-.01	-1.63	.206	19,989.9	-.02	.002
PUDDLE vs. Baseline	Recall	.79	.17	.61	149.07	<.001	19,950.6	.598	.624
PUDDLE vs. Baseline	Precision	.73	.14	.59	138.79	<.001	19,825.3	.574	.601

Table S4B

Comparison of Precision and Recall measures when syllabified input is used (columns refer to the same variables shown in Table S4A).

Comparison	Measure	M1	M2	ΔM	t	p	df	Lower Bci	Upper Bci
BTP vs. PUDDLE	Recall	.38	.89	-.51	-116.28	<.001	15,269.446	-.528	-.503
BTP vs. PUDDLE	Precision	.46	.85	-.4	-87.41	<.001	17,137.092	-.411	-.384
BTP vs. CLASSIC-UB initial/final	Precision	.46	.66	-.2	-39.55	<.001	19,734.761	-.217	-.185
BTP vs. CLASSIC-UB initial/final	Recall	.38	.58	-.2	-37.96	<.001	19,913.453	-.219	-.186
BTP vs. CLASSIC-UB final	Precision	.46	.57	-.11	-20.43	<.001	19,984.445	-.123	-.093
BTP vs. CLASSIC-UB final	Recall	.38	.48	-.11	-19.2	<.001	19,996.639	-.12	-.088
BTP vs. Baseline	Recall	.38	.46	-.08	-14.24	<.001	19,997.407	-.095	-.063
BTP vs. Baseline	Precision	.46	.51	-.06	-10.37	<.001	19,997.833	-.072	-.041
BTP vs. FTP	Precision	.46	.49	-.04	-6.79	<.001	19,987.974	-.052	-.022
BTP vs. FTP	Recall	.38	.41	-.03	-5.52	<.001	19,997.237	-.047	-.016
FTP vs. PUDDLE	Recall	.41	.89	-.48	-109.94	<.001	15,324.968	-.499	-.471
FTP vs. PUDDLE	Precision	.49	.85	-.36	-80.7	<.001	17,360.737	-.376	-.346
FTP vs. Baseline	Recall	.41	.46	-.05	-8.75	<.001	19,997.989	-.064	-.033
FTP vs. Baseline	Precision	.49	.51	-.02	-3.69	.001	19,990.389	-.034	-.006
CLASSIC-UB final vs. PUDDLE	Recall	.48	.89	-.41	-93.02	<.001	15,343.657	-.422	-.394
CLASSIC-UB final vs. PUDDLE	Precision	.57	.85	-.29	-64.64	<.001	17,397.080	-.302	-.272
CLASSIC-UB final vs. CLASSIC-UB initial/final	Precision	.57	.66	-.09	-18.51	<.001	19,838.308	-.107	-.079

CLASSIC-UB final vs. CLASSIC-UB initial/final	Recall	.48	.58	-.1	-18.31	<.001	19,933.371	-.112	-.081
CLASSIC-UB final vs. FTP	Precision	.57	.49	.07	13.78	<.001	19,997.734	.057	.088
CLASSIC-UB final vs. FTP	Recall	.48	.41	.08	13.71	<.001	19,997.914	.059	.091
CLASSIC-UB final vs. Baseline	Precision	.57	.51	.05	9.95	<.001	19,987.279	.037	.068
CLASSIC-UB final vs. Baseline	Recall	.48	.46	.03	4.95	<.001	19,997.843	.013	.042
CLASSIC-UB initial/final vs. PUDDLE	Recall	.58	.89	-.31	-74	<.001	15,872.605	-.325	-.297
CLASSIC-UB initial/final vs. PUDDLE	Precision	.66	.85	-.19	-46.49	<.001	18,269.313	-.208	-.179
CLASSIC-UB initial/final vs. FTP	Precision	.66	.49	.16	32.86	<.001	19,825.279	.151	.18
CLASSIC-UB initial/final vs. FTP	Recall	.58	.41	.17	32.39	<.001	19,928.617	.157	.189
CLASSIC-UB initial/final vs. Baseline	Precision	.66	.51	.15	28.66	<.001	19,747.424	.129	.161
CLASSIC-UB initial/final vs. Baseline	Recall	.58	.46	.12	23.37	<.001	19,926.899	.108	.14
PUDDLE vs. Baseline	Recall	.89	.46	.44	98.97	<.001	15,318.378	.422	.448
PUDDLE vs. Baseline	Precision	.85	.51	.34	75.27	<.001	17,165.941	.326	.351

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

A narrative account of the phonemic analysis is available in section *Results and Discussion / Word Age of First Production* of the main manuscript. Instead, here we focus on findings when models were run on syllabified input (see Table S5A). Models run on syllabified input did not perform better than the baseline. PUDDLE performed better than CLASSIC-UB initial-final ($\Delta Adj\ R^2$ =.023 [.008, .041]) and CLASSIC-UB final ($\Delta Adj\ R^2$ =.040 [.024, .062]) at predicting children's word age of first production. However, the difference between PUDDLE and the baseline was not significant ($\Delta Adj\ R^2$ = .020 [-.001, .039]) and neither was the difference between the baseline and CLASSIC-UB initial-final ($\Delta Adj\ R^2$ = .003 [-.012, .020]).

CLASSIC-UB initial-final performed better than CLASSIC-UB final ($\Delta Adj\ R^2$ = .017 [.010, .026]) at predicting children's word age of first production.

PUDDLE explained a significantly higher proportion of variance in word age of first production than forward ($\Delta Adj\ R^2$ = .048 [.023, .068]) and backward transitional probability models ($\Delta Adj\ R^2$ = .061 [043, .082]).

Table S5A

Frequency-weighted age of first production analyses: Pairwise differences between models' Adjusted R^2 when phonemic or syllabified input is used. The table shows model comparisons by type of input (phonemic vs. syllabified). For each pairwise comparison we report difference in Adjusted R^2 values (ΔR^2), and lower and upper limits of bootstrap confidence intervals (based on 1000 iterations and corrected using Holm's correction).

Comparison	Input type	ΔR^2	Lower Bci	Upper Bci
BTP vs. Baseline	Phoneme	.008	-.011	.026
FTP vs. Baseline	Phoneme	.010	-.011	.031
FTP vs. BTP	Phoneme	.002	-.013	.016
CLASSIC-UB final vs. Baseline	Phoneme	.043	.020	.069
CLASSIC-UB final vs. BTP	Phoneme	.035	.013	.059
CLASSIC-UB final vs. FTP	Phoneme	.033	.011	.057
CLASSIC-UB final vs. PUDDLE	Phoneme	.001	-.015	.018
CLASSIC-UB initial/final vs. Baseline	Phoneme	.048	.024	.073
CLASSIC-UB initial/final vs. BTP	Phoneme	.04	.010	.061
CLASSIC-UB initial/final vs. FTP	Phoneme	.038	.016	.059
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.006	-.016	.029
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.005	-.006	.016
PUDDLE vs. Baseline	Phoneme	.042	.021	.064
PUDDLE vs. BTP	Phoneme	.034	.011	.054
PUDDLE vs. FTP	Phoneme	.032	.014	.054
Baseline vs. BTP	Syllable	.041	.024	.058
Baseline vs. FTP	Syllable	.028	.012	.047
Baseline vs. CLASSIC-UB final	Syllable	.020	.005	.040
Baseline vs. CLASSIC-UB initial/final	Syllable	.003	-.012	.020
FTP vs. BTP	Syllable	.013	.005	.022
CLASSIC-UB final vs. BTP	Syllable	.021	.011	.032
CLASSIC-UB final vs. FTP	Syllable	.008	-.004	.020
CLASSIC-UB initial/final vs. BTP	Syllable	.038	.025	.054
CLASSIC-UB initial/final vs. FTP	Syllable	.025	.010	.044
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	.017	.010	.026
PUDDLE vs. BTP	Syllable	.061	.043	.082
PUDDLE vs. FTP	Syllable	.048	.023	.068
PUDDLE vs. CLASSIC-UB final	Syllable	.040	.024	.062

PUDDLE vs. CLASSIC-UB initial/final	Syllable	.023	.008	.041
PUDDLE vs. Baseline	Syllable	.020	-.001	.039

Appendix S6: Frequency-unweighted age of first production analyses

Interestingly, Larsen et al. (2017) found that a forward transitional probability model run on syllabified input showed the best performance, predicting 19% of variance in word age of acquisition. In contrast, we found that a forward transitional probability model run on syllabified input predicts a low proportion of variance (*Adj* R^2 = .013 [.007, .021]; see Table 1 in the main paper): We suggest this difference is related to differences in the predictor measure: we weighted the predictor measure by the frequency of a target word in the input, while Larsen used raw counts. Accordingly, when we used raw counts and syllabified input, we were able to replicate Larsen's finding (see Table S6A and S6B), with the forward transitional probability model showing the best performance (*Adj* R^2 = .311 [.284, .338]), followed by CLASSIC-UB final (*Adj* R^2 = .301 [.274, .327]). Importantly, however, even in this analysis we found that no model outperformed the baseline (*Adj* R^2 = .340 [.310, .364]), with the baseline performing significantly better than forward transitional probability (Δ*Adj* R^2 = .029 [.010, .047]). Note that Larsen did not include a comparison to a random baseline. We also obtained the same result when using raw counts and phonemic input, with CLASSIC-UB final showing the best performance (*Adj* R^2 = .227 [.205, .250]) but not being able to outperform the Baseline (*Adj* R^2 = .273 [.252, .295]; Δ*Adj* R^2 = .046 [.026, .069]). These results indicate that controlling for input word frequency and including a random baseline are both important to draw conclusions about the developmental plausibility of different segmentation models. A discussion on the role of the random baseline is included in Appendix S13.

Table S6A

Adjusted R^2 for linear regression models predicting word age of first production as a function of *unweighted* 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.

	Phonemic input			Syllabified input		
Model	Adjusted R^2	Lower Bci	Upper Bci	Adjusted R^2	Lower Bci	Upper Bci
Baseline	.273	.252	.295	.340	.310	.364
BTP	.153	.140	.167	.225	.202	.249
FTP	.168	.151	.185	.311	.284	.338
CLASSIC-UB final	.227	.205	.250	.301	.274	.327
CLASSIC-UB initial/final	.196	.176	.219	.278	.252	.302
PUDDLE	.195	.175	.214	.217	.194	.238

Table S6B

Pairwise differences between Adjusted R^2 of *unweighted* age of first production models. The table shows models comparison, input type considered, difference in Adjusted R^2 values, lower and upper limits of bootstrap confidence intervals (based on 1000 iterations and corrected using Holm's correction).

Comparison	Input type	ΔR^2	Lower Bci	Upper Bci
Baseline vs. BTP	Phoneme	.12	.101	.14
Baseline vs. FTP	Phoneme	.105	.088	.126
Baseline vs. PUDDLE	Phoneme	.078	.06	.098
Baseline vs. CLASSIC-UB initial/final	Phoneme	.077	.056	.097
Baseline vs. CLASSIC-UB final	Phoneme	.046	.026	.069
FTP vs. BTP	Phoneme	.015	.00	.029
CLASSIC-UB final vs. BTP	Phoneme	.074	.051	.098
CLASSIC-UB final vs. FTP	Phoneme	.059	.031	.083
CLASSIC-UB final vs. PUDDLE	Phoneme	.032	.009	.056
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	.031	.02	.042
CLASSIC-UB initial/final vs. BTP	Phoneme	.043	.022	.065
CLASSIC-UB initial/final vs. FTP	Phoneme	.028	.008	.048
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.001	-.014	.017
PUDDLE vs. BTP	Phoneme	.042	.022	.06
PUDDLE vs. FTP	Phoneme	.027	.009	.047
Baseline vs. PUDDLE	Syllable	.123	.101	.145

Baseline vs. BTP	Syllable	.115	.095	.137
Baseline vs. CLASSIC-UB initial/final	Syllable	.062	.042	.08
Baseline vs. CLASSIC-UB final	Syllable	.039	.02	.058
Baseline vs. FTP	Syllable	.029	.01	.047
BTP vs. PUDDLE	Syllable	.008	-.015	.03
FTP vs. PUDDLE	Syllable	.094	.067	.122
FTP vs. BTP	Syllable	.086	.064	.109
FTP vs. CLASSIC-UB initial/final	Syllable	.033	.012	.056
FTP vs. CLASSIC-UB final	Syllable	.01	-.007	.026
CLASSIC-UB final vs. PUDDLE	Syllable	.084	.061	.109
CLASSIC-UB final vs. BTP	Syllable	.076	.053	.1
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.023	.013	.033
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.061	.041	.08
CLASSIC-UB initial/final vs. BTP	Syllable	.053	.023	.079

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

Analyses were run on both phonemic and syllabified input. A narrative account of the phonemic-input analysis is available in section *Results and Discussion / Word-level Measures / Phonemic length* of the main manuscript; below we focus on syllabified input.

When syllabified input was used (see Fig S7, Table S7A and S7B), the model with the best performance was CLASSIC-UB final (X^2 = 14.62 [7.11, 56.59]), but even this model did not outperform the baseline (X^2 = 16.62 [6.1, 64.2]; ΔX^2 = 2.00 [-25.63, 31.97]) at approximating children's vocabularies by phonemic length. A discussion on the role of the random baseline is included in Appendix S13.

CLASSIC-UB initial-final (ΔX^2 = 420.11 [274.95, 581.06]) and CLASSIC-UB final (ΔX^2 = 424.83 [299.34, 576.60]) showed a better performance than PUDDLE at approximating children's vocabularies by phonemic length.

No significant difference was found when comparing CLASSIC-UB final and CLASSIC-UB initial-final (ΔX^2 = 4.72 [-33.69, 43.77]).

Finally, PUDDLE performance did not differ statistically from backward (ΔX^2 = 38.38 [-211.69, 295.25]) and was significantly worse than forward transitional probability models (ΔX^2 = 309.26 [78.34, 508.42]).

Table S7A

Child-model comparison by phonemic length. We compared the probability of observing words of different phonemic lengths in the models' vocabularies against the expected probability of words being of a given phonemic length in children's vocabularies. Comparisons were tested via a Chi-Square Goodness of Fit Test. The X^2 statistic always compares the distance of a model's distribution from children's. The table shows the type of comparison, the input type used, the Chi-squared statistic (X^2), degrees of freedom (df), *p* value and cut-offs of 95% bootstrap confidence interval of the statistic. Holm's correction was applied to *p* values and confidence intervals.

Comparison	Input type	X^2	df	*p* value	Lower Bci	Upper Bci
Children vs. Baseline	Phoneme	528.99	6	<.001	421.46	691.44
Children vs. BTP	Phoneme	1314.99	6	<.001	1112.25	1552.42
Children vs. FTP	Phoneme	1274.04	6	<.001	1107.59	1486.48
Children vs. CLASSIC-UB final	Phoneme	244.9	6	<.001	167.47	357.26
Children vs. CLASSIC-UB initial/final	Phoneme	311.02	6	<.001	223.76	440.03
Children vs. PUDDLE	Phoneme	1178.97	6	<.001	969.29	1406.66
Children vs. Baseline	Syllable	16.62	6	.022	6.1	64.2
Children vs. BTP	Syllable	401.07	6	<.001	268.08	598.9
Children vs. FTP	Syllable	130.19	6	<.001	67.29	244.6
Children vs. CLASSIC-UB final	Syllable	14.62	6	.023	7.11	56.59
Children vs. CLASSIC-UB initial/final	Syllable	19.34	6	.011	6.62	74.01
Children vs. PUDDLE	Syllable	439.45	6	<.001	317.47	604.34

Table S7B

Pairwise differences between the Chi-squared statistics reported in Table S7A, comparing how well two models' observed probabilities of phonemic lengths fit children's expected probabilities, when phonemic or syllabified input is used. Therefore, the ΔX^2 measure examines whether two models' distributions are at the same distance from children's. The

order of each pairwise difference was set as in the column Comparison (e.g., in *Baseline vs. CLASSIC-UB final*, the CLASSIC-UB final X^2 estimate is subtracted from the Baseline X^2 estimate). The table shows models comparison, difference in Chi-squared statistics (ΔX^2), input type considered, lower and upper limits of bootstrap confidence intervals (based on 1000 iterations and corrected using Holm's correction).

Comparison	Inout type	ΔX^2	Lower Bci	Upper Bci
Baseline vs. CLASSIC-UB final	Phoneme	284.09	146.62	416.98
Baseline vs. CLASSIC-UB initial/final	Phoneme	217.97	69.46	393.14
BTP vs. CLASSIC-UB final	Phoneme	1070.09	874.65	1291.03
BTP vs. CLASSIC-UB initial/final	Phoneme	1003.97	813.07	1255.92
BTP vs. Baseline	Phoneme	785.99	564.44	983.90
BTP vs. PUDDLE	Phoneme	136.01	-75.75	368.09
BTP vs. FTP	Phoneme	40.94	-133.07	239.68
FTP vs. CLASSIC-UB final	Phoneme	1029.14	856.78	1260.98
FTP vs. CLASSIC-UB initial/final	Phoneme	963.02	736.17	1207.71
FTP vs. Baseline	Phoneme	745.05	551.63	946.67
FTP vs. PUDDLE	Phoneme	95.07	-111.00	287.36
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	66.12	-41.65	172.99
PUDDLE vs. CLASSIC-UB final	Phoneme	934.07	717.61	1153.22
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	867.95	679.12	1084.11
PUDDLE vs. Baseline	Phoneme	649.98	423.18	858.81
Baseline vs. CLASSIC-UB final	Syllable	2.00	-25.63	31.97
BTP vs. CLASSIC-UB final	Syllable	386.45	203.26	578.85
BTP vs. Baseline	Syllable	384.45	223.81	556.98
BTP vs. CLASSIC-UB initial/final	Syllable	381.73	200.61	570.78
BTP vs. FTP	Syllable	270.88	122.27	419.30
FTP vs. CLASSIC-UB final	Syllable	115.57	31.55	219.54
FTP vs. Baseline	Syllable	113.57	31.45	212.32
FTP vs. CLASSIC-UB initial/final	Syllable	110.85	13.61	226.86
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	4.72	-33.69	43.77
CLASSIC-UB initial/final vs. Baseline	Syllable	2.72	-36.82	44.92
PUDDLE vs. CLASSIC-UB final	Syllable	424.83	299.34	576.60
PUDDLE vs. Baseline	Syllable	422.83	253.88	609.58
PUDDLE vs. CLASSIC-UB initial/final	Syllable	420.11	274.95	581.06
PUDDLE vs. FTP	Syllable	309.26	78.34	508.42
PUDDLE vs. BTP	Syllable	38.38	-211.69	295.25

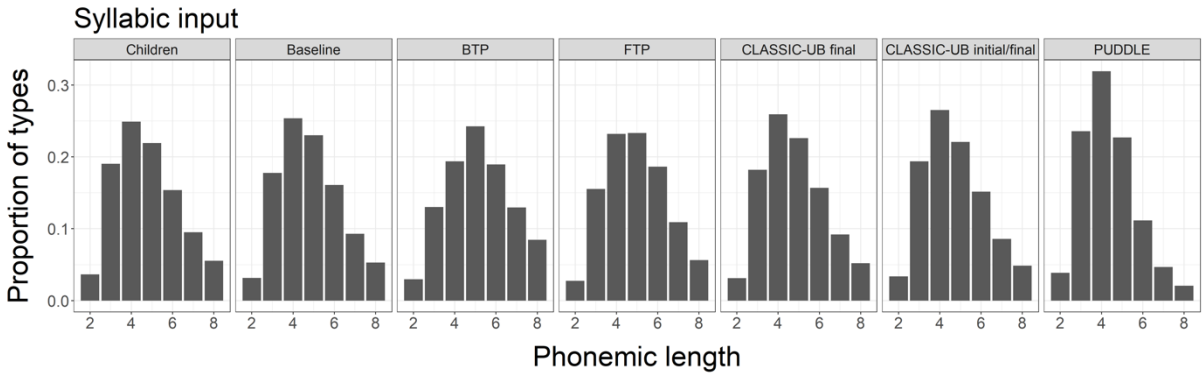

Fig. S7A. Proportion of unique words (types) produced by children and discovered by each model by phonemic length, when syllabified input is used.

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

Analyses were run on both phonemic and syllabified input. A narrative account of the phonemic-input analysis is available in section *Results and Discussion / Word-level Measures / Word Frequency* of the main manuscript; below we focus on syllabified input.

When syllabified input was used (see Fig S8A, Table S8A and S8B), PUDDLE outperformed CLASSIC-UB initial-final (ΔD = .031 [.006, .058]) and CLASSIC-UB final (ΔD = .053 [.028, .079]) at approximating children's vocabularies by weighted Log10 word frequency. However, neither PUDDLE nor CLASSIC-UB initial-final were able to outperform the baseline (D = .05 [.03, .07], p = <.001) (baseline *vs*. PUDDLE, ΔD = .003 [-.017, .028]; CLASSIC-UB initial-final vs. baseline, ΔD = .028 [.000, .052]). A discussion on the role of the random baseline is included in Appendix S13.

CLASSIC-UB final did not differ statistically from CLASSIC-UB initial-final (ΔD = .022 [-.002, .045]).

Finally, PUDDLE outperformed forward (ΔD = .052 [.028, .081]) and backward transitional probability (ΔD = .066 [.039, .091]) at approximating children's vocabularies by weighted Log10 word frequency.

Table S8A.

Child-model comparison by weighted Log10 word frequency. We compared models' distributions of unique words by weighted Log10 word frequency to child distribution. Comparisons were tested via Kolmogorov–Smirnov test statistic. The table shows the type of comparison, the input type used, the Kolmogorov–Smirnov test statistic (*D*), p value and cut-offs of 95% bootstrap confidence interval of the statistic. Holm's correction was applied to p values and confidence intervals.

Comparison	Input type	*D*	p value	Lower Bci	Upper Bci
Children vs. Baseline	Phoneme	.29	<.001	.27	.32
Children vs. BTP	Phoneme	.26	<.001	.23	.30
Children vs. FTP	Phoneme	.23	<.001	.20	.27
Children vs. CLASSIC-UB final	Phoneme	.13	<.001	.11	.15
Children vs. CLASSIC-UB initial/final	Phoneme	.16	<.001	.14	.19
Children vs. PUDDLE	Phoneme	.13	<.001	.11	.16
Children vs. Baseline	Syllable	.05	<.001	.03	.07
Children vs. BTP	Syllable	.11	<.001	.09	.14
Children vs. FTP	Syllable	.10	<.001	.08	.12
Children vs. CLASSIC-UB final	Syllable	.10	<.001	.08	.12
Children vs. CLASSIC-UB initial/final	Syllable	.07	<.001	.06	.10
Children vs. PUDDLE	Syllable	.04	<.001	.03	.06

Table S8B

Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S8A, comparing how closely two models' distributions of unique words are to children's productions by weighted Log10 word frequency, when phonemic or syllabified input is used. The table shows models comparison, input type, difference in Kolmogorov–Smirnov test statistics (ΔD), lower and upper limits of bootstrap confidence intervals (based on 1000 iterations and corrected using Holm's correction).

Comparison	Input type	Δ*D*	Lower Bci	Upper Bci
Baseline vs. CLASSIC-UB final	Phoneme	.169	.136	.198
Baseline vs. PUDDLE	Phoneme	.163	.127	.192
Baseline vs. CLASSIC-UB initial/final	Phoneme	.132	.087	.159
Baseline vs. FTP	Phoneme	.064	.028	.101
Baseline vs. BTP	Phoneme	.03	-.006	.069
BTP vs. CLASSIC-UB final	Phoneme	.138	.098	.175
BTP vs. PUDDLE	Phoneme	.133	.096	.175
BTP vs. CLASSIC-UB initial/final	Phoneme	.101	.063	.139
BTP vs. FTP	Phoneme	.034	-.001	.072
FTP vs. CLASSIC-UB final	Phoneme	.105	.067	.145
FTP vs. PUDDLE	Phoneme	.099	.063	.132
FTP vs. CLASSIC-UB initial/final	Phoneme	.068	.028	.106
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.037	.006	.072
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.032	-.006	.069
PUDDLE vs. CLASSIC-UB final	Phoneme	.005	-.020	.036
Baseline vs. PUDDLE	Syllable	.003	-.017	.028
BTP vs. PUDDLE	Syllable	.066	.039	.091
BTP vs. Baseline	Syllable	.063	.028	.089
BTP vs. CLASSIC-UB initial/final	Syllable	.035	.009	.060
BTP vs. FTP	Syllable	.015	-.008	.034
BTP vs. CLASSIC-UB final	Syllable	.014	-.010	.038
FTP vs. PUDDLE	Syllable	.052	.028	.081
FTP vs. Baseline	Syllable	.048	.023	.073
FTP vs. CLASSIC-UB initial/final	Syllable	.021	-.002	.042
CLASSIC-UB final vs. PUDDLE	Syllable	.053	.028	.079
CLASSIC-UB final vs. Baseline	Syllable	.049	.019	.074
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.022	-.002	.045
CLASSIC-UB final vs. FTP	Syllable	.001	-.017	.016
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.031	.006	.058
CLASSIC-UB initial/final vs. Baseline	Syllable	.028	.000	.052

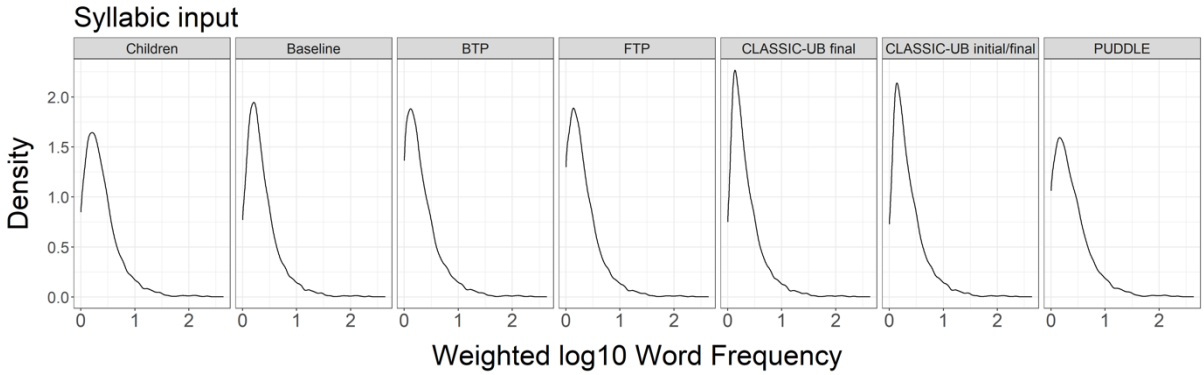

Fig. S8A. Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted Log10 word frequency.

Syllabified input is used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

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

Analyses were run on both phonemic and syllabified input. A narrative account of the phonemic-input analysis is available in section *Results and Discussion / Word-level Measures / Neighborhood Density* of the main manuscript; below we focus on syllabified input.

When syllabified input was used (see Fig. S9A, Table S9A and S9B), CLASSIC-UB final showed the best performance (D = .03 [.02, .05], p = .005) at approximating children's vocabularies by weighted neighborhood density, but it was not able to outperform the baseline (D = .03 [.02, .04], p = .029; ΔD = .006 [-.007, .02]. A discussion on the role of the random baseline is included in Appendix S13.

CLASSIC-UB final did not differ statistically from CLASSIC-UB initial-final (ΔD = .019 [-.002, .038]).

Finally, PUDDLE did not differ statistically from backward transitional probability (ΔD = .049 [-.002, .090]) and performed significantly worse than forward transitional probability (ΔD = .122 [.079, .155]) at approximating children's vocabularies by weighted neighborhood density.

Table S9A

Child-model comparison by weighted neighborhood density. We compared models' distributions of unique words by weighted neighborhood density to child distribution. Comparisons were tested via Kolmogorov–Smirnov test statistic. The table shows the type of

comparison, the input unit used, the Kolmogorov–Smirnov test statistic (*D*), p value and cut-offs of 95% bootstrap confidence interval of the statistic, adjusted using Holm's correction.

Comparison	Input type	*D*	p value	Lower Bci	Upper Bci
Children vs. Baseline	Phoneme	.2	<.001	.18	.23
Children vs. BTP	Phoneme	.37	<.001	.34	.4
Children vs. FTP	Phoneme	.34	<.001	.32	.37
Children vs. CLASSIC-UB final	Phoneme	.14	<.001	.12	.17
Children vs. CLASSIC-UB initial/final	Phoneme	.18	<.001	.16	.21
Children vs. PUDDLE	Phoneme	.29	<.001	.26	.32
Children vs. Baseline	Syllable	.03	.029	.02	.04
Children vs. BTP	Syllable	.12	<.001	.10	.14
Children vs. FTP	Syllable	.05	<.001	.03	.07
Children vs. CLASSIC-UB final	Syllable	.03	.005	.02	.05
Children vs. CLASSIC-UB initial/final	Syllable	.05	<.001	.03	.07
Children vs. PUDDLE	Syllable	.17	<.001	.14	.19

Table S9B

Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S9A, comparing how closely two models' distributions of unique words are to children's productions by weighted neighborhood density, when phonemic or syllabified input is used. The table shows models comparison, input type, difference in Kolmogorov–Smirnov test statistics (Δ*D*), lower and upper limits of bootstrap confidence intervals (based on 1000 iterations and corrected using Holm's correction).

Comparison	Input type	Δ*D*	Lower Bci	Upper Bci
Baseline vs. CLASSIC-UB final	Phoneme	.063	.034	.092
Baseline vs. CLASSIC-UB initial/final	Phoneme	.021	-.007	.050
BTP vs. CLASSIC-UB final	Phoneme	.228	.191	.261
BTP vs. CLASSIC-UB initial/final	Phoneme	.185	.147	.222
BTP vs. Baseline	Phoneme	.164	.127	.197
BTP vs. PUDDLE	Phoneme	.081	.049	.110
BTP vs. FTP	Phoneme	.028	-.002	.058
FTP vs. CLASSIC-UB final	Phoneme	.199	.168	.229
FTP vs. CLASSIC-UB initial/final	Phoneme	.157	.122	.191
FTP vs. Baseline	Phoneme	.136	.101	.171
FTP vs. PUDDLE	Phoneme	.053	.023	.083
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.042	.007	.074
PUDDLE vs. CLASSIC-UB final	Phoneme	.146	.115	.178
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	.104	.065	.135

PUDDLE vs. Baseline	Phoneme	.083	.049	.118
BTP vs. Baseline	Syllable	.092	.055	.123
BTP vs. CLASSIC-UB final	Syllable	.086	.044	.119
BTP vs. FTP	Syllable	.073	.045	.100
BTP vs. CLASSIC-UB initial/final	Syllable	.066	.028	.103
FTP vs. Baseline	Syllable	.019	-.01	.044
FTP vs. CLASSIC-UB final	Syllable	.013	-.019	.039
CLASSIC-UB final vs. Baseline	Syllable	.006	-.007	.02
CLASSIC-UB initial/final vs. Baseline	Syllable	.026	.001	.046
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	.019	-.002	.038
CLASSIC-UB initial/final vs. FTP	Syllable	.007	-.023	.036
PUDDLE vs. Baseline	Syllable	.141	.113	.164
PUDDLE vs. CLASSIC-UB final	Syllable	.135	.110	.155
PUDDLE vs. FTP	Syllable	.122	.079	.155
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.115	.087	.142
PUDDLE vs. BTP	Syllable	.049	-.002	.090

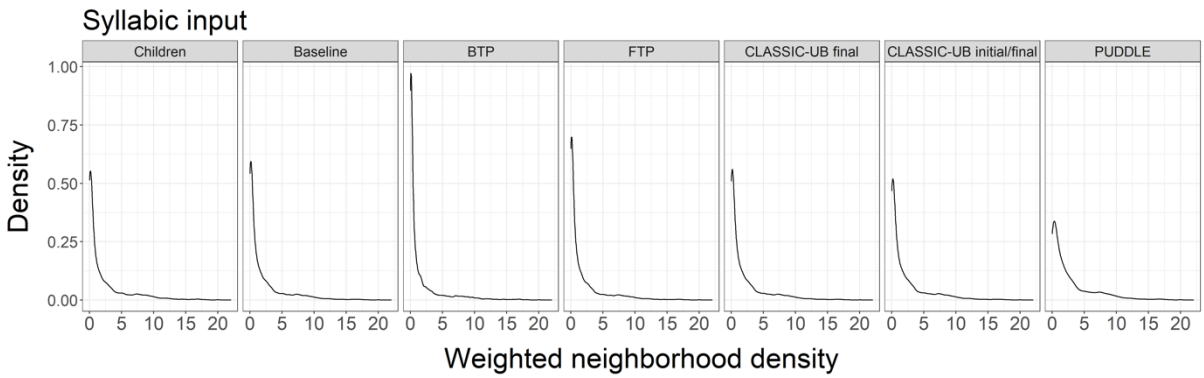

Fig. S9A. Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted neighborhood density. Syllabified input is used.

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

Analyses were run on both phonemic and syllabified input. A narrative account of the phonemic-input analysis is available in section *Results and Discussion / Word-level Measures / Phonotactic Probability* of the main manuscript.

When syllabified input is used, no significant differences were found between models' performance at approximating children's vocabulary by weighted phonotactic probability (see Fig. S10A, and Table S10A and S10B).

Table S10A

Child-model comparison by weighted phonotactic probability. We compared models' distributions of unique words by weighted phonotactic probability to child distribution. Comparisons were tested via Kolmogorov–Smirnov test statistic. The table shows the type of comparison, the input type used, the Kolmogorov–Smirnov test statistic (*D*), p value and cut-offs of 95% bootstrap confidence interval of the statistic, adjusted using Holm's correction.

Comparison	Input type	*D*	p value	Lower Bci	Upper Bci
Children vs. Baseline	Phoneme	.05	.002	.03	.08
Children vs. BTP	Phoneme	.08	<.001	.05	.12
Children vs. FTP	Phoneme	.08	<.001	.05	.11
Children vs. CLASSIC-UB final	Phoneme	.07	<.001	.05	.10
Children vs. CLASSIC-UB initial/final	Phoneme	.09	<.001	.06	.12
Children vs. PUDDLE	Phoneme	.05	<.001	.03	.08
Children vs. Baseline	Syllable	.02	.77	.01	.04
Children vs. BTP	Syllable	.01	.77	.01	.03
Children vs. FTP	Syllable	.02	.368	.01	.05
Children vs. CLASSIC-UB final	Syllable	.02	.368	.01	.04
Children vs. CLASSIC-UB initial/final	Syllable	.03	.044	.02	.06
Children vs. PUDDLE	Syllable	.02	.77	.01	.04

Table S10B

Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S10A, comparing how closely two models' distributions of unique words are to children's productions by weighted phonotactic probability, when phonemic or syllabified input is used. The table shows models comparison, input type, difference in Kolmogorov–Smirnov test statistics (ΔD), lower and upper limits of bootstrap confidence intervals (based on 1000 iterations and corrected using Holm's correction).

Comparison	Input type	Δ*D*	Lower Bci	Upper Bci
BTP vs. Baseline	Phoneme	.031	-.001	.073
BTP vs. PUDDLE	Phoneme	.028	-.021	.072
BTP vs. CLASSIC-UB final	Phoneme	.007	-.029	.045
BTP vs. FTP	Phoneme	.002	-.027	.033
FTP vs. Baseline	Phoneme	.029	-.007	.064
FTP vs. PUDDLE	Phoneme	.026	-.018	.078
FTP vs. CLASSIC-UB final	Phoneme	.005	-.032	.034
CLASSIC-UB final vs. Baseline	Phoneme	.024	-.014	.059
CLASSIC-UB final vs. PUDDLE	Phoneme	.021	-.028	.075
CLASSIC-UB initial/final vs. Baseline	Phoneme	.042	.008	.081
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.038	-.011	.098
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.017	-.014	.05
CLASSIC-UB initial/final vs. FTP	Phoneme	.012	-.022	.045
CLASSIC-UB initial/final vs. BTP	Phoneme	.01	-.032	.05
PUDDLE vs. Baseline	Phoneme	.003	-.038	.042
Baseline vs. BTP	Syllable	.004	-.016	.024
FTP vs. BTP	Syllable	.009	-.014	.03
FTP vs. Baseline	Syllable	.005	-.017	.023
FTP vs. PUDDLE	Syllable	.003	-.018	.019
CLASSIC-UB final vs. BTP	Syllable	.01	-.014	.028
CLASSIC-UB final vs. Baseline	Syllable	.006	-.017	.022
CLASSIC-UB final vs. PUDDLE	Syllable	.003	-.017	.02
CLASSIC-UB final vs. FTP	Syllable	.001	-.015	.015
CLASSIC-UB initial/final vs. BTP	Syllable	.017	-.017	.037
CLASSIC-UB initial/final vs. Baseline	Syllable	.013	-.008	.034
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.011	-.012	.032
CLASSIC-UB initial/final vs. FTP	Syllable	.008	-.015	.029
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	.008	-.014	.027
PUDDLE vs. BTP	Syllable	.006	-.021	.027
PUDDLE vs. Baseline	Syllable	.002	-.017	.022

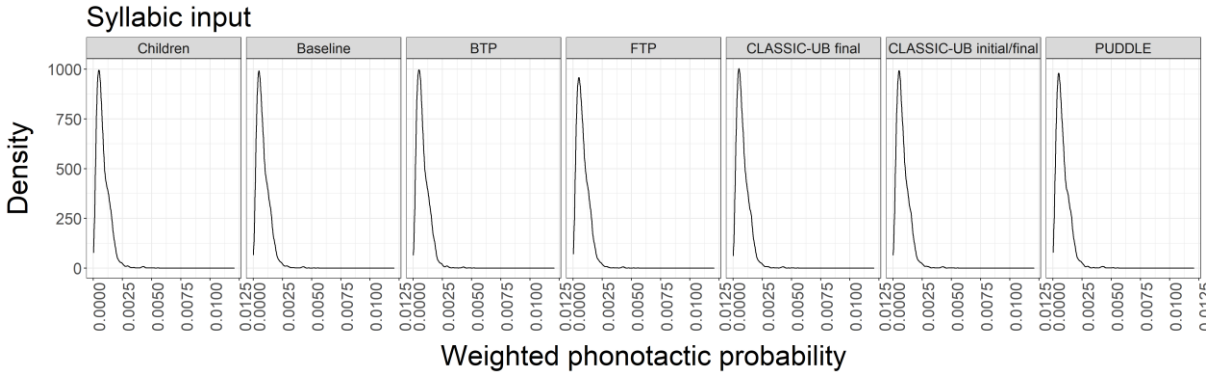

Fig. S10A. Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted phonotactic probability. Syllabified input is used.

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

In this section, we briefly discuss whether the comparison between CLASSIC-UB initial-final and CLASSIC-UB final (on all the measures considered in our study) suggests that the addition of utterance-initial markers improves model performance.

As can be seen in Fig. 2 and Appendix S4, CLASSIC-UB initial-final shows a better performance than CLASSIC-UB final in the traditional measures, reaching .50 Precision and .50 Recall with phonemic input (vs. .49 Precision and .45 Recall), and .66 Precision and .58 Recall with syllabified input (vs. .57 Precision and .48 Recall). This suggests that the inclusion of initial (in addition to final) utterance-boundary markers is useful in segmenting the speech input, as other studies have shown (Seidl & Johnson, 2006; 2008).

However, results for the developmental measures suggest that utterance-initial markers do not significantly improve model performance. CLASSIC-UB initial-final does not explain more variability in child age of first production compared to CLASSIC-UB final, suggesting that an initial utterance-boundary marker might not be necessary to predict word age of first production (see Table S6A). Similarly, adding utterance-initial markers does not significantly improve the model's ability to capture any of the word-level characteristics of children's vocabularies (see Table S7B-S10B).

When considering measures that are not weighted by input frequency (i.e., traditional measures, unweighted age of first production, word-level measures) (see Appendix S4 and S6-S10), this result is likely due to the ratio of type to token frequency of the words present

in the input at utterance-initial and final position. Namely, token frequency (i.e., frequency of a word including repetitions) is lower for words appearing at the end of utterances (M = 305.35, SE = 28.81) than words appearing at the start of utterances (M = 652.05, SE = 62.81). At the same time, the input contains higher type frequency (i.e., more different words) at the end of utterances (N = 5,485) than at the beginning (N = 786). This suggests that CLASSIC-UB's segmentation accuracy increases when provided with utterance-initial markers because there are more repeated words that the model will be able to segment correctly at the start of utterances, but their role becomes marginal for building a lexicon as the majority of novel words appear at utterance ends (e.g., Fernald & Mazzie, 1991).

This result provides evidence in support of previous work (e.g., Pearl et al., 2010) suggesting that utterance-initial words might be segmented with higher accuracy because they have a higher token frequency (e.g., pronouns, determiners) than more variable utterance-final ones (e.g., nouns, verbs). Additionally, using measures based on child data we showed that the high type frequency of utterance-final words might be important in the process of building a lexicon from the segmented words. In other words, even if the perceptual salience of word boundaries at utterance-initial and final edges equally facilitates word extraction in the lab (Seidl & Johnson, 2006; 2008), their role in the naturalistic environment might be moderated by frequency information. The repeated presentation of few different words in utterance-initial position might increase the likelihood of segmenting those words correctly. Conversely, encountering a large number of different words at utterance ends might increase the chance of building a more diverse (i.e., larger) vocabulary. Finally, this also means that facilitatory effects of utterance boundaries in naturalistic settings might be different for languages where, for example, new words do not tend to be placed at utterance ends as in English child-directed speech (e.g., Dutch, Japanese; Han et al., 2021).

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

The difference between CLASSIC-UB and PUDDLE in the word-level measures (i.e., with CLASSIC-UB better approximating children's vocabularies by phonemic length and neighborhood density) might be explained by differences in vocabulary size: at the end of learning, PUDDLE has a larger vocabulary than CLASSIC-UB, and might be taken to represent a child with more advanced vocabulary knowledge. Conversely, it is possible that an earlier stage of PUDDLE with smaller vocabulary may show similar performance to CLASSIC-UB on our developmental measures. To assess this possibility, we can look at models' developmental cascades, to see whether models' differences still hold when we consider the stage at which PUDDLE has reached a vocabulary equal in size to CLASSIC-UB's. We carry out this analysis only for phonemic input, because CLASSIC-UB develops a smaller vocabulary than PUDDLE only when using phonemic input, but not when using syllabified input (see Table S12A).

Table S12A

Raw number of word types learned by CLASSIC-UB models and PUDDLE when run on phonemic or syllabic input, ranked from largest to smallest.

Model	**Input type**	**Word types learned**
CLASSIC-UB final	Syllables	8,047
CLASSIC-UB initial/final	Syllables	7,451
PUDDLE	Syllables	5,903
PUDDLE	Phonemes	3,967
CLASSIC-UB final	Phonemes	3,611
CLASSIC-UB initial/final	Phonemes	3,049

In Fig. S12A, black vertical lines indicate the stage at which PUDDLE has reached a vocabulary size equal to CLASSIC-UB final's or CLASSIC-UB initial/final's (as indicated by the text labels). If differences between models are explained by vocabulary size, we should find that PUDDLE word-level distributions at the vertical lines become similar to CLASSIC-UB's distributions at stage 20 (i.e., at the end of its learning). Instead, for those measures that were found to show significant differences - i.e., phonemic length and neighborhood density - we can see that differences between PUDDLE and CLASSIC-UB models hold across stages, with PUDDLE's learning being always biased toward short (3-phoneme and 4-phoneme) words and high-neighborhood words compared to CLASSIC-UB models.

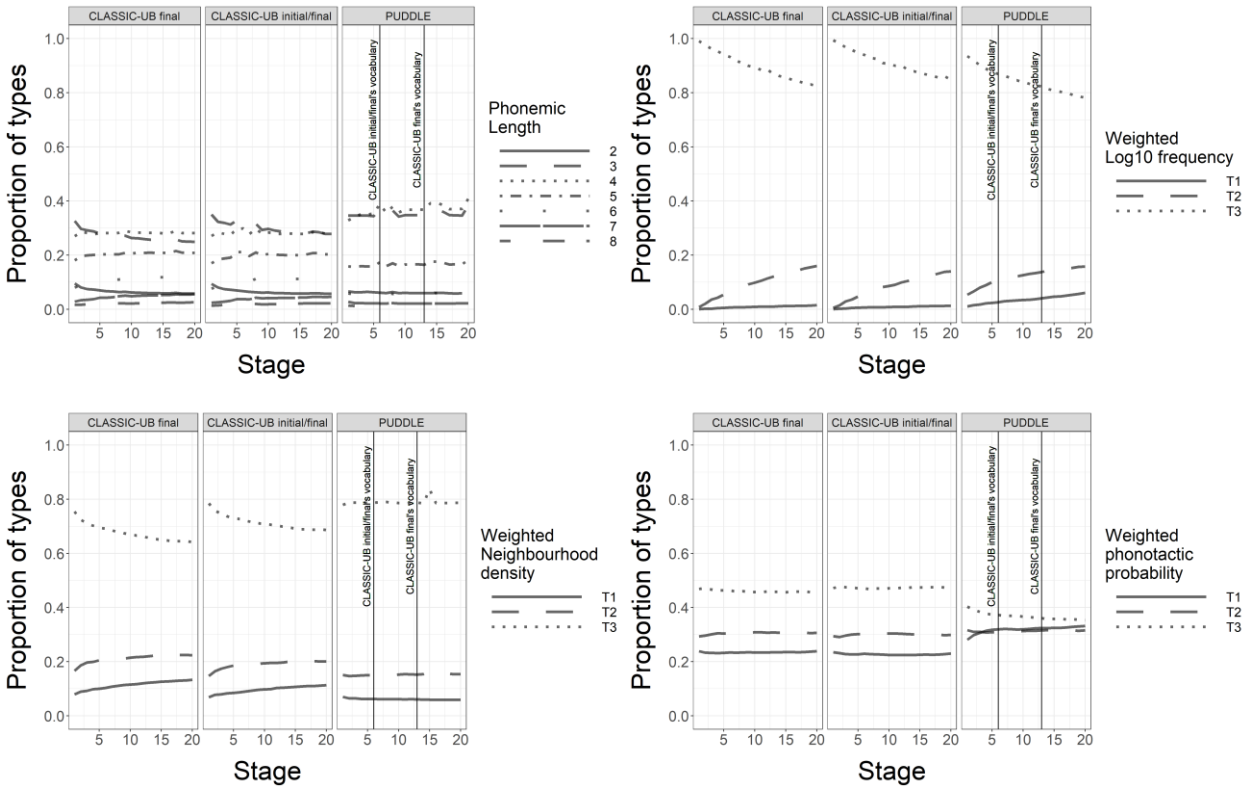

Fig. S12A. Proportion of types discovered at each input stage for each word-level measure. Proportion of types is computed by dividing the cumulative number of word types by the total number of types at a specific stage. Stage is computed by dividing the segmented

utterances into 20 equal stages (note that the 604 stages used for Precision and Recall were divided into wider stages, 20, because the probability of discovering new word types decreases substantially at later stages). For continuous word-level measures (i.e., word frequency, neighborhood density, and phonotactic probability), word types were divided into groups based on child-directed speech tertiles. For example, T1 in the word frequency measure identifies words that have a low frequency in child-directed speech (<= 33rd percentile), while T3 refers to high-frequency words in child-directed speech (> 66th percentile). Black vertical lines indicate the stages at which PUDDLE has reached a vocabulary size equal to CLASSIC-UB final or CLASSIC-UB initial/final.

As discussed in the main paper (see *Measures of Developmental Plausibility* section of the General Discussion), differences in performance can be explained by CLASSIC-UB's ability to learn words with overlapping phonological sequences (see Jones, 2016). Indication of this can be seen when we look at the length and neighborhood findings separately by word frequency. In Fig. S12B below, we can see that CLASSIC-UB becomes more accurate at capturing child vocabularies as frequency increases. This happens because frequent words are more likely to share phonological sequences with previously learned words, consequently boosting CLASSIC-UB's learning compared to other models which do not show such facilitation (as their learning mechanism is uniquely based on tracking target sequences' frequency). Namely, other models' performance at capturing child phonemic length and neighborhood density does not improve as word frequency increases.

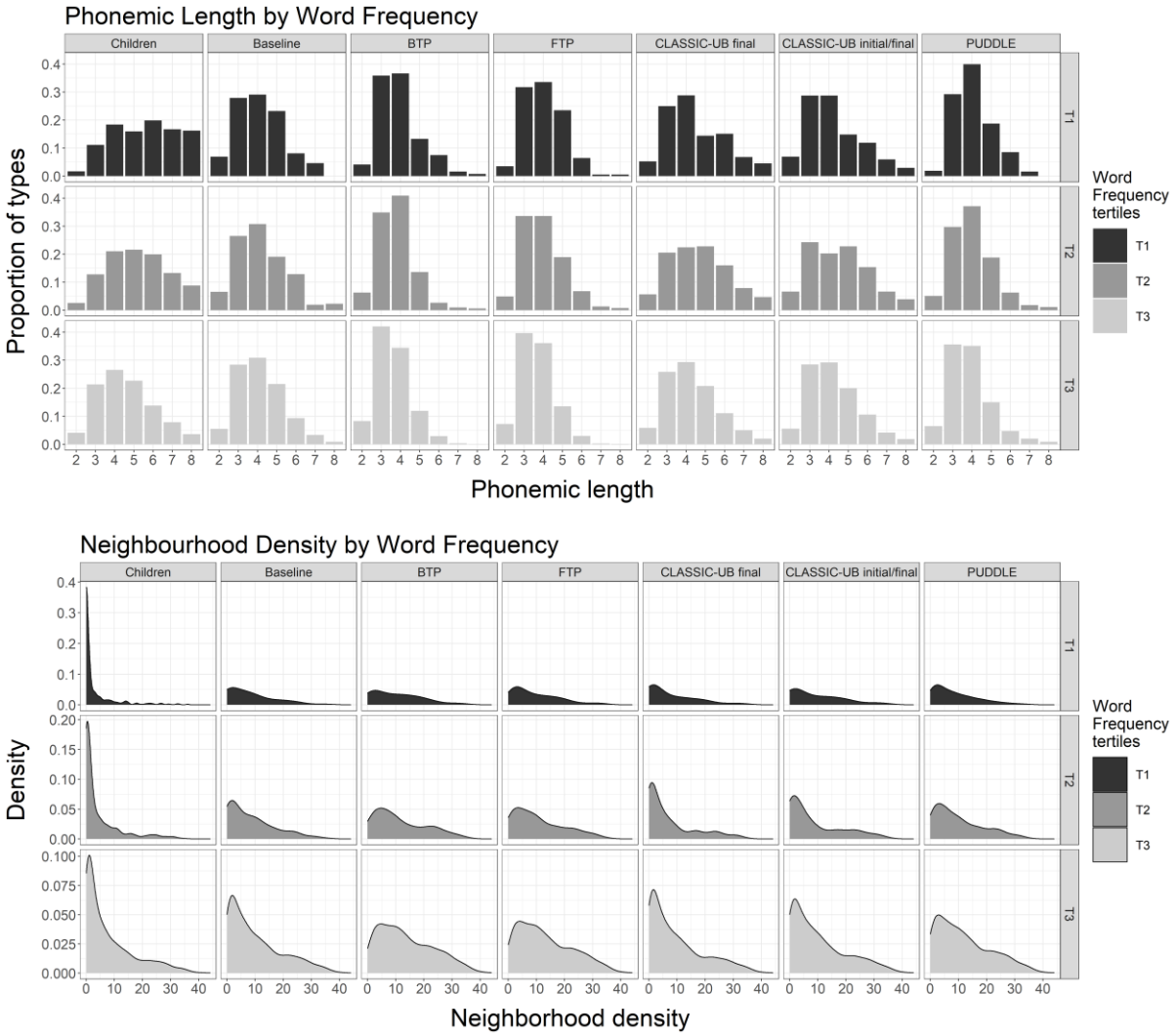

Fig. S12B. Child and models' phonemic length and neighborhood density distributions at different child-directed word frequency tertiles.

We also conducted a final exploratory analysis to support our claim that CLASSIC-UB captures long and low-neighborhood words from the child vocabularies better than PUDDLE. Specifically, we wanted to check whether CLASSIC-UB actually learns more long and low-neighborhood words than PUDDLE or rather it simply misses a portion of children's short, high-neighborhood words (that PUDDLE instead captures), producing in turn an increase in the relative proportion of long, low-neighborhood words in its vocabulary.

Thus, we looked at the absolute number of children's word types captured by each model, as shown in Figure S12C. In this figure, we plot the raw number of types produced by children, alongside the number of children's words that CLASSIC-UB final or PUDDLE have captured or missed (by phonemic length and neighborhood density). Note that this analysis excludes a portion of words that the models learned from the input but that were not produced by children; when including this set of words, the results we obtain are consistent with the analysis reported below. As can be seen in Figure S12C, differences in phonemic length are not only due to the fact that PUDDLE captures more 3- and 4-phoneme children's words than CLASSIC-UB, but also to the fact that CLASSIC-UB captures a higher absolute number of 5- to 8-phoneme words than PUDDLE. Similarly, although PUDDLE captures a higher number of high-neighborhood words (T3), it also captures a lower absolute number of words in the low and middle neighborhood range (T1 and T2) than CLASSIC-UB. In sum, this analysis supports our claim that CLASSIC-UB's learning mechanism facilitates the learning of words that are generally more difficult to learn (i.e., long and with a low number of similar words in the input) but that children nevertheless acquire.

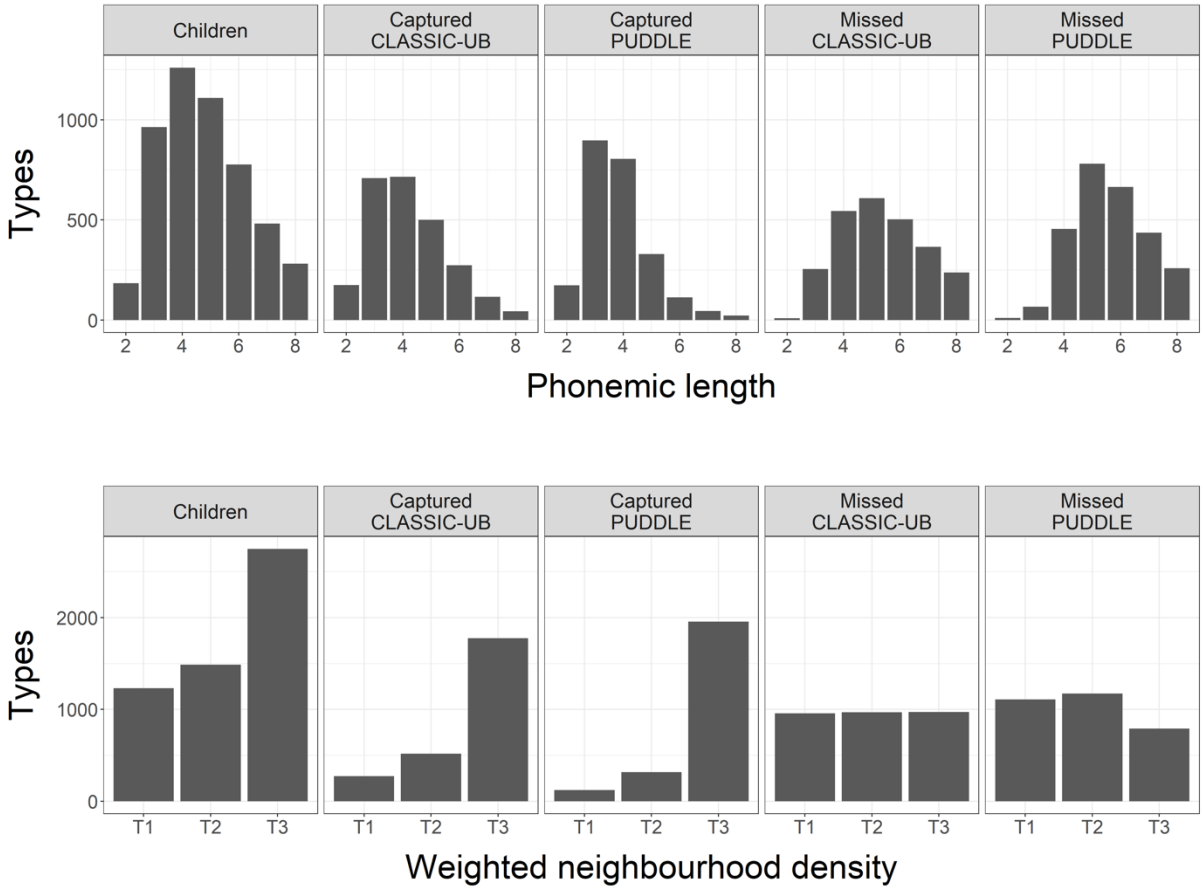

Figure S12C. The plot shows the raw number of word types produced by children, alongside the raw number of word types produced by children that CLASSIC-UB final and PUDDLE learned (captured) or not learned (missed). Phonemic length considers children's words from 2 to 8 phonemes, while weighted neighborhood density considers children's words in low (T1), middle (T2) and high (T3) neighborhood child-directed speech tertiles.

Appendix S13: Controlling for baseline segmentation performance

An unexpected finding of the present study is that, when we used syllabified input, no model was able to outperform the baseline in developmental measures. Providing a model with the input syllabic structure likely represents a strong facilitation which makes it difficult to compare competing models. First, given that models' input contains 81% of monosyllabic

tokens, syllabifying the input (i.e., avoiding oversegmentation of syllables) allows a model to discover – by chance - a large proportion of word types. For example, although models were exposed to limited input compared to what children receive, when processing syllabified input they discovered more word types (*M* = 7223, *min* = 5903, *max* = 8047) than Thomas (the child with the largest production vocabulary; *N* = 5899). The models also learned more low-frequency words than children when run on a syllabified input (see Fig. S8A in the Appendix), and this may be for the same reason.

Furthermore, previous computational work has shown that providing chunking models with the input syllabic structure might not be necessary, as models run on phonemic input only commit a small proportion of intra-syllabic segmentation error (Goldwater et al., 2009). To confirm this, more work that compares models and infants' actual segmentation performance is needed. For example, future work could investigate whether the issue with the syllabic baseline applies cross-linguistically or is only present in languages such as English that have a large number of mono-syllabic words.

References

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, *12*(4), e0174623. https://doi.org/10.1371/journal.pone.0174623

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the Combinatorial Explosion: An Explanation of n-gram Frequency Effects Based on Naive Discriminative Learning. *Language and Speech*, *56*(3), 329–347. https://doi.org/10.1177/0023830913484896

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

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology. General, 117*(1), 21–33. https://doi.org/10.1037//0096-3445.117.1.21

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology, 29*(4), 711–721. https://doi.org/10.1037/0012-1649.29.4.711

Braginsky, M., Sanchez, A., & Yurovsky, D. (2019). *childesr: Accessing the 'CHILDES' Database*. R package version 0.1.2. https://github.com/langcog/childesr

Derwing, B. L. (1992). A 'pause-break' task for eliciting syllable boundary judgments from literate and illiterate speakers: Preliminary results for five diverse languages. *Language and Speech, 35*(1–2), 219–235. https://doi.org/10.1177/002383099203500217

Derwing, B. L., & Eddington, D. (2014). The experimental investigation of syllable structure. *The Mental Lexicon, 9*(2), 170–195. https://doi.org/10.1075/ml.9.2.02der

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

Forrester, M. (2002). Appropriating cultural conceptions of childhood: Participation in conversation. *Childhood*, *9*, 255-278. https://doi.org/10.1177/0907568202009003043

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

Gambell, T. & Yang, C. (2006). *Word segmentation: Quick but not dirty* [Unpublished manuscript] http://www.ling.upenn.edu/ycharles/papers/quick.pdf

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

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

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2019). Children probably store short rather than frequent or predictable chunks: Quantitative evidence from a corpus study. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00080

Han, M., de Jong, N. H., & Kager, R. (2021). Language Specificity of Infant-directed Speech: Speaking Rate and Word Position in Word-learning Contexts. *Language Learning and Development*, *17*(3), 221–240. https://doi.org/10.1080/15475441.2020.1855182

Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*, *128*(6), 1112–1144. https://doi.org/10.1037/rev0000305

Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting.* New York: Oxford University Press.

Johnson, K. (2004). Massive reduction in conversational American English. In Spontaneous speech: Data and analysis. *Proceedings of the 1st session of the 10th international symposium* (pp. 29-54).

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

Korman, M. (1992). *CHILDES English Korman Corpus*. https://doi.org/10.21415/T59G7B

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.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*(3), 481-507. https://doi.org/10.1515/COGL.2009.022

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).

Loukatou, G. R., Moran, S., Blasi, D. E., Stoll, S., & Cristia, A. (2019). Is word segmentation child's play in all languages?. In *ACL (1)* (pp. 3931-3937).

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

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive science, 42*, 375-412. https://doi.org/10.1111/cogs.12592

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

Olejarczuk, P., & Kapatsinski, V. (2018). The metrical parse is guided by gradient phonotactics. *Phonology*, *35*(3), 367–405. https://doi.org/10.1017/S0952675718000106

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854. https://doi.org/10.1111/cogs.12217

R Core Team (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Rowland, C. F. & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language*, *33*, 859-877. https://doi.org/10.1017/S0305000906007537

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., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

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

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96–109. https://doi.org/10.1016/j.wocn.2010.11.006

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

Ten Bosch, L., Boves, L., & Ernestus, M. (2022). DIANA, a process-oriented model of human auditory word recognition. *Brain Sciences*, *12*(5), 681. https://doi.org/10.3390/brainsci12050681

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28,* 127-152. https://doi.org/10.1017/S0305000900004608

Tommerdahl, J., & Kilpatrick, C. (2013). Analyzing reliability of grammatical production in spontaneous samples of varying length. *Journal of Child Language Teaching and Therapy*, *29,2*, 171-183. https://doi.org/10.1177/0265659012459528

Wells, C. G. (1981). *Learning through interaction: The study of language development.* Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511620737