INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Relating unsupervised word segmentation to reported vocabulary acquisition

Elin Larsen[1]*, Alejandrina Cristia*[1]*, Emmanuel Dupoux*[1]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Études Cognitives, ENS, EHESS, CNRS, PSL Research University, Paris, France.

elin.larsen@ens.fr, alejandrina.cristia@ens.fr, emmanuel.dupoux@gmail.com

Abstract

A range of computational approaches have been used to model the discovery of word forms from continuous speech by infants. Typically, these algorithms are evaluated with respect to the ideal 'gold standard' word segmentation and lexicon. These metrics assess how well an algorithm matches the adult state, but may not reflect the intermediate states of the child's lexical development. We set up a new evaluation method based on the correlation between word frequency counts derived from the application of an algorithm onto a corpus of child-directed speech, and the proportion of infants knowing those words, according to parental reports. We evaluate a representative set of 4 algorithms, applied to transcriptions of the Brent corpus, which have been phonologized using either phonemes or syllables as basic units. Results show remarkable variation in the extent to which these 8 algorithm-unit combinations predicted infant vocabulary, with some of these predictions surpassing those derived from the adult gold standard segmentation. We argue that infant vocabulary prediction provides a useful complement to traditional evaluation; for example, the best predictor model was also one of the worst in terms of segmentation score, and there was no clear relationship between token or boundary F-score and vocabulary prediction.

Index Terms: language acquisition, word segmentation, infant vocabulary, speech units, computational modeling.

1. Introduction

Segmenting – identifying words in fluent speech – is a key step in language acquisition and especially in lexical development. However, the absence of systematic silences between words in continuous speech makes this task a particularly difficult one, especially early in development where other key components of language (phonology, morphology, syntax, etc.) are not yet fully known to infants. Thus, a key research question becomes: how do infants get word segmentation off the ground? Two strands of research have addressed this issue: one using laboratory experiments in infants, [1, 2, 3, 4, 5], and another one using computational modeling (a review in [6]). Here, we focus on the latter approach with a view to comparing the *adequacy* of computational systems as models of infant word segmentation.

Most of the computational work on word segmentation has taken the view that a good model, when fed with enough data, will reach the optimal performance of a human adult. This is usually quantified by defining as 'gold standard' the segmentation corresponding to the blank spaces between words in the orthographic transcription of the input data. Provided this ideal segmentation, objective metrics can be defined, typically, type, token and boundary precision, recall and f-scores (see below). While this gold standard can be argued to represent the adult state, it almost certainly does not capture intermediate states of knowledge in infants. Indeed, there is now clear evidence that infants make segmentation errors, in that they extract from continuous speech a "proto-lexicon" containing many items that adults would view as non-words [4, 7].

A number of alternatives to gold standard metrics have been discussed in previous work. One is that the proposed model should be *cognitively plausible* given what is known about infant cognition (e.g., [8]). Another is that models should reproduce documented patterns of successes and errors found in development (e.g., [9, 10]). One limitation of such proposals is that it is difficult to rank the relative merit of algorithms on these dimensions because there is no agreed upon list of criteria or patterns of results that could be turned into an objective metric. For instance, [8] argues that infants focus on segmenting rather than on storing potential words, whereas [11] argues that it is more cognitively plausible that infants try to learn word-like units rather than learning segmentation strategies.

We propose to use one source of evidence that has not yet been exploited for this purpose, and which may provide a way to determine the relative merit of algorithmic proposals in a quantifiable way: parental reports of word comprehension. These reports are typically collected using the MacArthur Communicative Development Inventory (CDI for short [12]), a standardized questionnaire containing more than 400 items. The CDI has been collected from a large quantity of families, and stored for re-use in the WordBank repository [13]. Recently, CDI comprehension data have been used to look at the earliest words that infants acquire and age of acquisition for each of these words has been estimated across several languages. Input-related factors – frequency, mean length utterance – and conceptual factors – concreteness, babiness : measure of association with infancy – have been found to predict vocabulary age of acquisition [14]. Of course, parental report may not reflect the true state of the infant's comprehension lexicon. The ideal source of data would use experiments that bypass the parent and measure lexical knowledge in the child, like word-to-meaning paradigms [15, 16] or segmentation experiments [1, 17]. Yet, these studies are limited to a small number of preselected test items and are difficult to deploy across a large number of infants. This is why, despite its many drawbacks, parental reports remain a good proxy of word knowledge in infants.

In this paper, we introduce a new measure derived from the correlation between the frequency of occurrence of a word form in the output of an algorithm (applied onto a large corpus of transcribed child-directed speech), on the one hand and the proportion of infants reported to comprehend that word form on the other. We evaluate 8 algorithm-unit combinations, as follows. The corpus was represented using either the phoneme or the syllable as basic unit. The algorithms were drawn from two main classes of algorithms: one class tracks *local statistical cues* at a sub-lexical level in order to find where to segment speech, the other builds a *word form lexicon* to represent or capture the corpus. Both classes make the assumption that phrase bound-

http://dx.doi.org/10.21437/Interspeech.2017-937

aries are known, and contain distributional information useful for segmenting words [18]. We use two algorithms from each class, introduced in more detail in Section 2.

2. Methods

2.1. Data

We used as input to the word segmentation algorithms the Brent-Siskind corpus [19]. This corpus is the longest one in the CHILDES repository [20], containing orthographic transcriptions for more than 100 hours of recordings, gathered from 16 American English-speaking mother-infant pairs. Table 1 gives some details about the corpus. For evaluating the prediction from word segmentation algorithms of infants' reported comprehension scores, we used American English data available from the WordBank repository [13], corresponding to the "Words and Gestures" form of the CDI [21, 12, 22, 23]. Most of these items are nouns (e.g., *ball*) but there are also other classes, such as verbs ($watch$), function words (you), adjectives (big) and even onomatopeia (baa). There were different numbers of parental reports at different ages, ranging from 66 to 761. In order to maximize our chances of having sufficient sensitivity, we focus on the 761 parental reports for infants aged 13 months.

Table 1: *Descriptive statistics of the Brent-Siskind corpus — NU: number of utterances, AUL: average utterance length, TTR: token-type ratio, AWL-ph: average word length in number of phonemes, AWL-syl: average word length in number of syllables, infants' age range in months*

NU	AUL	TTR	AWL-ph	AWL-syl	age
113363	3.59	60,4	3.06	1.23	9-15

2.2. Word segmentation algorithms

The first model using local statistical cues is the **Diphone-Based Segmentation** (DiBS) algorithm [8], which keeps track of the frequency of two phones occurring together and decides to place a boundary between them using Bayes' Theorem. The model assumes that the learner knows the phonetic categories, is able to detect utterance boundaries, assumes phonological independence across word boundaries, and tracks context-free distribution of diphones. All these assumptions are discussed in [8] from a psycholinguistic point of view indicating why it is plausible that infants might act as a DiBS learner when starting to segment speech. The only free parameter of the model is the context-free probability of a word boundary and is determined by average word length and number of words per utterance. This setting could be viewed as controversial since the model is thus partly supervised – word boundaries are given for a subset of the corpus to fix the context-free probability of a word boundary (in our case the first 200 utterances); but in actual practice it has little effect on performance (see [8]).

Another local statistical model is based on tracking **transitional probabilities (TPs) over syllables**. In our implementation (drawn from [24]), TPs posits a boundary between two syllables if their forward transitional probability is locally lowest [1]. This relative threshold was preferred over an absolute threshold (e.g., the average over all attested syllable pairs) as it seemed more plausible; the forward preferred over the backward TP or the mutual information one based on the idea that forward TP may be more appropriate for head-initial languages [24]. A number of experimental studies using artificial languages have shown that infants segment at local minima of TP (and probably also that they treat sequences of syllables with high TP as units [1, 2].

The lexicon-based strategy needs learners to have larger memory and wider knowledge on its language than local statistical strategies. The **PUDDLE** (Phonotactics from Utterances Determine Distributional Lexical Elements) algorithm developed by [9] builds incrementally a lexicon as follows: For each utterance, phoneme subsequences are looked up in the long-term lexicon; if a match is found, then the two sounds preceding the match are searched for among possible word endings, and the two sounds following the match among possible word beginnings. If both are found, then the matched item is increased in frequency in the lexicon, and the two remainders are added to it. If not, then the whole utterance is stored. In both cases, the beginning and end diphones are added to the list of possible onsets and offsets, respectively. These "phonotactic" constraints avoid over-segmentation (to the level of phonemes). The model promotes frequent words by counting the occurrences of words added in the proto-lexicon and then by sorting the list of these words by frequency. Since time (number of mathematics operations) and space (memory needed) requirements were heavy in the original awk script kindly provided by Padraic Monaghan, we made a minimally different version in python using the collection modules, resulting in a 120-fold decrease in computation time and optimized space use.

The last algorithm used is the **unigram Adaptor Grammar (AGu)** [25], [26]. The framework consists of two modules: a lexicon generator and an adaptor. The first one generates a lexicon of items that are likely to be found in the corpus and the second assigns frequencies to the items. Importantly, the unigram AG assumes that lexicon items are generated independently from each other and that the stochastic process is chosen so that items' frequencies follow a power-law distribution, similar to natural language.

2.3. Data processing and algorithm evaluation

The scripts used for all processing steps are available on the second author's github: `https://github.com/alecristia/CDSwordSeg`. The corpus processing steps consisted of cleaning up annotations and converting orthography into surface phonological forms. Following [9], phonologization was achieved with the American English voice of the Festival Text-to-Speech system [27], which provides syllable boundaries.

A few alterations were necessary when changing the unit of input representation from the original ones used by each algorithm (phonemes for all except for TPs). For AGu-syllable, we created a unigram grammar whose terminals are all the syllables found in the corpus. For PUDDLE-syllable, we modified the boundary constraints, keeping a constraint spanning one syllable (rather than 2, as in the case of diphone constraints). Notice that applying a condition on bisyllables would effectively prevent segmentation when this results in monosyllabic chunks. Therefore, the boundary constraints pertained only to the previous and following syllable rather than the previous and following *pair* of syllables. No special modifications were necessary for the TPs and DiBS algorithms.

Evaluation of the final output of each algorithm for each input representation was assessed by using the traditional token F-score, i.e., the harmonic mean of precision (ratio of true positives to the number of segmented items by the algorithm) and recall (ratio of true positives to the number of segmented items in the input). Type and boundary F-scores were also used for comparison purposes. Since PUDDLE is incremental, we assessed its performance by a 5-fold cross-validation with the hope of measuring asymptotic performance (which is likely given the size of the Brent-Siskind corpus). Finally, for AGu, 8 parses of 2000 iterations (of which the first 100 were removed before parse reduction) were used. Because this algorithm is not deterministic (unlike all other algorithms), the process was repeated five times and averaged.

3. Relating segmentation output to infant vocabulary

For each algorithm-unit combination as well as the gold corpus, we used the scikit-learn library in python [28] to look at the potential relationship between (a) the number of times each word form was found in the segmented corpus, and (b) the proportion of infants reported to understand that word. In preliminary analyses, we used both a simple linear regression and a logistic regression. Since results were similar, we focus here on the linear regression; interested readers can find the results for the logistic models in our Open Science Framework site.

In our linear regression, the predictor was the logarithm of the number of times a word was correctly segmented; the outcome was the proportion of infants reported to understand that word. We checked that the residuals were normally distributed and had homogeneous variance.

4. Results

Figure 1 shows the proportion of infants reported to know a word in the CDI as a function of the number of times a word occurs in the Brent-Siskind corpus, for every word appearing both in the CDI and the Brent-Siskind corpus. This correlation gives us an insight into how well a perfect segmentation output could correlate with infants' reported comprehension scores. Thus, it serves as a baseline for comparison of the different word segmentation algorithm correlations in Table 2. This table indicates that only AGu-syllable, AGu-phoneme, and TPs-syllable have a correlation score above the baseline. In other words, frequencies estimated from these outputs predicts comprehension scores better than frequencies from the input corpus, where boundaries correspond to orthographic words.

Table 2 also gives the detail segmentation F- scores – lexicon, token and boundary F-scores – for each algorithm. At first glance, the best performances are: AGu-phoneme when evaluated with Type F-score, PUDDLE-syllable when evaluated on Token F-score and Boundary F-score, and TPs-syllable when evaluated on its correlation with the infant reported vocabulary. This is quite surprising, leading us to think that the traditional F-score metrics are not correlated with the coefficient of determination R^2 of the regression predicting reported comprehension scores. The values of the Pearson correlation coefficient between the different F-scores and R^2 support this assumption. Token and Boundary F-scores have no correlation with R^2: $\rho_{Token-R^2} = -0.083$ (see Figure 2), $\rho_{Boundary-R^2} = 0.04$. However, Type F-score and R^2 are slightly correlated: $\rho_{Type-R^2} = 0.35$.

With regard to the effect of the unit of representation, Figure 2 indicates that sub-lexical algorithms (DiBS and TPs) are more affected by the change of unit than lexical-based algorithms. Moreover, TPs, DiBS and PUDDLE show surprisingly better segmentation performance with their non-initial input: phoneme for TPs, syllable for DiBS and for PUDDLE. AGu, on the contrary, is clearly better with phonemes, but it might be due to the low complexity of the grammar itself: Taking into account one or two levels of collocation – groups of words that tend to appear together – might favor the syllabic unit over the phoneme. In fact, conclusions must be drawn carefully as the unit presented introduces some bias in the segmentation. Indeed, since phonemes are thrice as common as syllables in the corpus (see Table 1), the number of possibles boundaries is much larger with phonemes than with syllables, hence the boundary precision must be lower for algorithms using phonemes than syllables, which is actually found in Table 2.

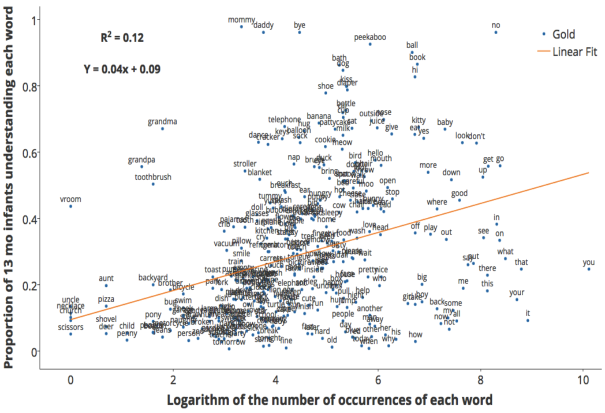

Figure 1: *Linear regression between the infant lexicon at 13 months old and the frequency of each word in the Brent corpus and in the CDI*

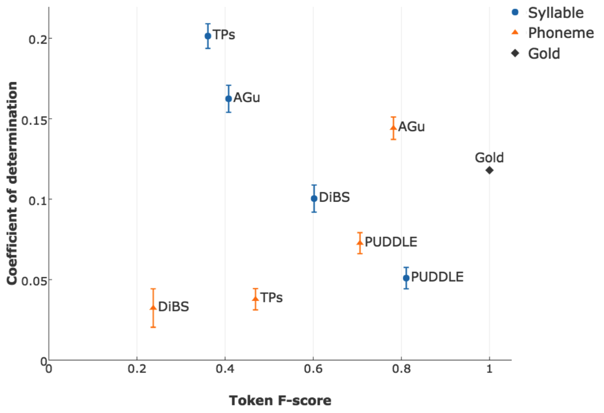

Figure 2: *Token F-score plotted against the coefficient of linear regression predicting infant lexicon from word segmentation algorithms' true positives. Error bars are twice the standard error, estimate of the standard deviation.*

5. Discussion and conclusion

In this paper, we identified a new way to evaluate computational models of word segmentation on the basis of standardized word comprehension reports. Three main conclusions can be drawn from the results obtained. First, some word segmentation algorithms predict reported word comprehension much better than

Table 2: *F-score, Precision and Recall for Type, Token and Boundary statistics of four word segmentation algorithms on the Brent-Siskind corpus as a function of input units, R^2 prediction score of 13 month-old-infants' vocabulary and standard error SE . Boldface indicates the best statistics on each column.*

Algorithm	Unit Input	Type			Token			Boundary			Regression	
		F-score (rank)	Prec.	Rec.	F-score (rank)	Prec.	Rec.	F-score (rank)	Prec.	Rec.	R^2 (rank)	SE
TPs	Syllable	0.188 (6)	0.111	0.607	0.361 (7)	0.476	0.291	0.602 (7)	0.947	0.441	**0.186** (1)	0.0076
	Phoneme	0.166 (7)	0.124	0.253	0.468 (5)	0.432	0.512	0.657 (5)	0.590	0.742	0.034 (7)	0.0066
DiBS	Syllable	0.417 (2)	0.523	0.347	0.602 (4)	0.570	0.638	0.801 (4)	0.745	0.866	0.093 (4)	0.0084
	Phoneme	0.057 (8)	0.035	0.159	0.236 (8)	0.234	0.240	0.467 (8)	0.459	0.475	0.029 (8)	0.0119
PUDDLE	Syllable	0.315 (5)	0.234	0.479	**0.811** (1)	**0.821**	**0.802**	**0.903** (1)	0.918	**0.889**	0.046 (6)	0.0066
	Phoneme	0.380 (3)	0.306	0.501	0.706 (3)	0.682	0.733	0.820 (3)	0.782	0.862	0.067 (5)	0.0065
AGu	Syllable	0.340 (4)	0.232	**0.634**	0.408 (6)	0.532	0.331	0.637 (6)	**0.985**	0.471	0.148 (2)	0.0084
	Phoneme	**0.517** (1)	**0.585**	0.464	0.782 (2)	0.787	0.777	0.889 (2)	0.897	0.881	0.135 (3)	0.0070
Adult gold standard		–	–	–	–	–	–	–	–	–	0.118	0.0065

others, and some do so better than the gold lexicon. In other words, there is variation in performance that appears useful for evaluating the relative merit of algorithms. Second, it is not the word segmentation algorithms that perform best when evaluated on adult segmentation that best predict reported word comprehension. In fact, the algorithm that predicts word comprehension best has very poor segmentation performances: TPs-syllables is ranked 6 or 7 over the 8 algorithms when inspecting F-scores, and yet TPs-syllables predicts infants' reported comprehension better than AGu, which is sometimes considered as the benchmark for adult word segmentation. Finally, a comparison of the unit of input representation has been done systematically over several algorithms (cf. [29]). We found that our two algorithms based on local regularities (DiBS and TPs) were more sensitive to the change of unit than others. Nonetheless, there was no clear advantage for one unit over another on all F-scores, nor on the correlation with infants' reported comprehension.

It is noteworthy that the algorithm with the highest predictive value for infants' comprehension is also the algorithm which received the most attention in experiments using artificial languages made up of strings of syllables [1], and more recently in a study re-analyzing post-hoc the potential sources of some false positives contained in French-learning infants' proto-lexicon [7]. It is also noteworthy that TPs fails to predict infant comprehension when its input is specified in terms of phonemes, rather than syllables. However, before jumping to conclusions and claiming that our analysis provides *proof* that TPs-syllables is *the* algorithm-representation used by infants, two caveats are in order. First, it would be important to replicate this study with other corpora and other languages in order to assess the robustness and generality of our finding. Second and more importantly, our analysis only provides evidence for half of the predictions of the algorithm. Indeed, with the CDI, we can only test the predicted correct segmentation, not the errors. All of the 8 algorithms make systematic segmentation errors which can be considered as predictions regarding the content of the infant lexicon. In order to prove that an algorithm is truly used by infants, one would therefore ideally also check that infants' proto-lexicon contains same false positives that are predicted by the algorithm, using, for instance, the paradigm developed by [7].

Conducting large-scale experimental work in infants is costly. Fortunately, our approach could still be used to as a first pass to quickly test a large set of algorithms on their correct predictions before it becomes worth setting up the experiments on their incorrect ones. Here, we only scratched the surface in that many other segmentation algorithms have been proposed, some sublexical [30], some lexical [29, 31] and others using a combination of both strategies [32, 33]. It would be interesting to add to the mix algorithms that work from raw speech instead of phonetic transcriptions [34, 35].

Finally, our findings provide a new angle to understand the strategy that infants may be using to kick-start their lexical segmentation: The fact that a very rudimentary algorithm like TP outperforms the gold segmentation in predicting infants' reported comprehension scores indicates that infants may not use, at least initially, an optimal segmentation strategy, but rather a simple heuristics that gives them a first proto-lexicon, to be cleaned up at a later stage. One could therefore get a hint at such a heuristics by studying the way in which the initial vocabulary, as assessed by parental reports, systematically deviates from what could be expected based on the gold segmentation.

To conclude, we presented evidence that word segmentation algorithms can be distinguished through their correlation with reported infant word comprehension, providing a novel way of integrating cognitive considerations in modeling approaches to early language acquisition.

6. Acknowledgements

This project is supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*, ANR-14-CE30-0003 MechELex), Ecole de Neurosciences de Paris, Region Ile de France (DIM cerveau et pense) and AWS in Education Research Grant award.

7. References

[1] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *Science*, vol. 274, pp. 1926–1928, 1996a.

[2] J. Saffran, E. Johnson, Aslin, R.N., and E. Newport, "Statistical learning of tone sequences by human infants and adults." *Cognition*, vol. 70, pp. 27–52, 1998.

[3] P. W. Jusczyk, E. A. Hohne, and A. Bauman, "Infants' sensitivity to allophonic cues for word segmentation." *Perception & Psychophysics*, vol. 68, pp. 1465–76, 1999.

[4] P. W. Jusczyk, D. M. Houston, and M. Newsome, "The beginnings of word segmentation in English-learning infants." *Cognitive Psychology*, vol. 39, pp. 159–207, 1999.

[5] S. L. Mattys and P. W. Jusczyk, "Do infants segment words or recurring contiguous patterns?" *Journal of Experimental Psychology. Human Perception and Performance.*, vol. 27, pp. 644–55, 2001.

[6] O. Rasanen, "Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions," *Speech Communication*, vol. 54, no. 9, pp. 975–997, 2012.

[7] C. Ngon, A. Martin, E. Dupoux, D. Cabrol, M. Dutat, and S. Peperkamp, "(Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life," *Developmental Science*, vol. 16, no. 1, pp. 24–34, 2013.

[8] R. Daland and J. Pierrehumbert, "Learning Diphone-Based Segmentation," *Cognitive Science*, vol. 35, pp. 119–155, 2011.

[9] P. Monaghan and M. H. Christiansen, "Words in puddles of sound: modelling psycholinguistic effects in speech segmentation," *Journal of child language*, vol. 37, pp. 545–564, 2010.

[10] L. Phillips and L. Pearl, "The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation," *Cognitive science*, vol. 38, pp. 1824–1854, 2015.

[11] E. O. Batchelder, "Can a computer really model cognition? A case study of six computational models of infant word discovery," in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, 1998, pp. 120–125.

[12] L. Fenson, V. A. Marchman, D. J. Thal, P. S. Dale, J. S. Reznick, and E. Bates, *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.).* Baltimore, MD: Brookes., 2007.

[13] M. C. Frank, M. Braginsky, D. Yurovsky, and V. A. Marchman, "Wordbank: An open repository for developmental vocabulary data," *Journal of Child Language*, pp. 1–18, 2016.

[14] M. Braginsky, D. Yurovsky, V. A. Marchman, and M. C. Frank, "From uh-oh to tomorrow: Predicting age of acquisition for early words across languages." *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2016.

[15] E. Bergelson and D. Swingley, "The acquisition of abstract words by young infants," *Cognition*, vol. 127, no. 3, pp. 391–397, 2013.

[16] R. Tincoff and P. W. Jusczyk, "Some beginnings of word comprehension in 6-month-olds," *Psychological Science*, vol. 10, no. 2, pp. 172–175, 1999.

[17] P. Jusczyk and R. Aslin, "Infants' detection of the sound patterns of words in fluent speech," *Cognitive Psychology*, vol. 29, no. 1, pp. 1 – 23, 1995.

[18] R. N. Aslin, J. Woodward, N. LaMendola, and T. G. Bever, *Models of word segmentation in fluent maternal speech to infants.* L. Erlbaum Associates, 1996, pp. 117–134.

[19] M. R. Brent and J. R. Siskind, "The role of exposure to isolated words in early vocabulary development." *Cognition*, vol. 81, pp. 31–44, 2001.

[20] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk.*, L. E. Associates, Ed., Mahwah, NJ, 2000.

[21] K. Byers-Heinlein, "Words and Gestures CDI data contributed to WordBank," unpublished.

[22] M. C. Frank, "Words and Gestures CDI data contributed to WordBank," unpublished.

[23] D. Thal, V. Marchman, J. Tomblin, L. Rescorla, and P. Dale, "Late-talking toddlers: Characterization and prediction of continuing delay," in *Late talkers: Language development, interventions and outcomes*, L. Rescorla and P. Dale, Eds. Brookes Publishing, 2013, pp. 169–201.

[24] A. Saksida, A. Langus, and M. Nespor, "Co-occurrence statistics as a language-dependent cue for speech segmentation," *Developmental Science*, pp. 1–11, 2016.

[25] M. Johnson, T. L. Griffiths, and S. Goldwater, "Adaptor grammars: a framework for specifying compositional nonparametric Bayesian models," in *Advances in Neural Information Processing Systems 19*, B. Schlkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, p. 641648.

[26] S. Goldwater, T. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context." *Cognition*, vol. 112, pp. 21–54, 2009.

[27] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis system," 1998.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] L. Phillips and L. Pearl, "Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things," in *Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)@ EACL*, 2014, pp. 9–13.

[30] M. C. Frank, S. Goldwater, T. L. Griffiths, and J. B. Tenenbaum, "Modeling human performance in statistical word segmentation," *Cognition*, vol. 117, no. 2, pp. 107–125, Nov. 2010.

[31] A. Venkataraman, "A statistical model for word discovery in transcribed speech," *Computational Linguistics*, vol. 27, pp. 351–372, 2001.

[32] M. Johnson, "Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure." in *ACL*, 2008, pp. 398–406.

[33] D. Swingley, "Statistical clustering and the contents of the infant vocabulary," *Cognitive psychology*, vol. 50, no. 1, pp. 86–132, 2005.

[34] C. Bergmann, L. Boves, and L. Ten Bosch, "Measuring word learning performance in computational models and infants," in *IEEE International Conference on Development and Learning (ICDL)*, vol. 2. IEEE, 2011, pp. 1–6.

[35] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," *Procedia Computer Science*, vol. 81, pp. 67–72, 2016.