Beyond frequency-based parameters in computational models of early word segmentation

Francesco Cabiddu (Cardiff University), Lewis Bott (Cardiff University), Gary Jones (Nottingham Trent University), Chiara Gambi (Cardiff University)

CabidduF@cardiff.ac.uk

How do infants form word representations despite the lack of reliable cues to word boundaries in the input? A statistical learning approach sees the infant as equipped with a dedicated learning mechanism that tracks statistical regularities in the input (Saffran et al., 1996). This statistical knowledge – represented by frequency-based parameters in word segmentation models - should allow infants to estimate the location of word boundaries in continuous speech. However, it is unclear whether infants constantly update frequency-based parameters or these simply represent a re-description of frequency effects that constrain language development (e.g., word frequency in caregiver's speech; Perruchet, 2019).

A chunk learning mechanism has been proposed as a more plausible explanation for infants' word segmentation (Christiansen, 2019). Under this hypothesis, infants represent co-occurring items as single entities via associative learning, allowing to represent speech using fewer units. Although chunking accounts have been able to segment corpora of child-directed speech and explain infants' segmentation of artificial languages (Monaghan & Christiansen, 2010; Perruchet & Vinter, 1998), they still rely on frequency-based parameters to decide what constitutes a chunk (e.g., chunk frequency is constantly updated and guides recognition, representing changes in chunk familiarity).

In this work, we present CLASSIC-Utterance-Boundary (CLASSIC-UB), a more parsimonious chunking model that learns to represent speech using larger (or fewer) chunks without relying on any frequency-based parameters. CLASSIC-UB - a modified version of the existing model CLASSIC (Jones & Macken, 2018) - tags chunks that appear in utterance-final position and these chunks are more likely to be segmented as words by the model. The developmental plausibility of this utterance-final cue is justified by the presence of acoustically salient utterance boundaries in child-directed speech that represent strong segmentation cues used by infants (Seidl & Johnson, 2006).

The model learned from a large corpus of child-directed speech (7 corpora, 604,000 utterances directed to 2-year-olds; CHILDES, MacWhinney, 2000). We evaluated the developmental plausibility of the model in two ways: (1) the number of times the model correctly segments a word in the input should predict that word's age of first production in children's speech (Larsen et al., 2017); (2) the model's vocabulary should approximate the distribution of words in children's speech in terms of word length, word frequency, neighbourhood density and phonotactic probability.

CLASSIC-UB explains the same amount of variability in age of first production as a model which discovers chunks as a function of their absolute frequency (PUDDLE; Monaghan & Christiansen, 2010; see Table 1), while it outperforms models based on forward and backward transitional probability (e.g., Hay et al., 2011). Further, CLASSIC-UB represents a closer fit to children's vocabularies than all other models in terms of all word-level characteristics (see Figure 1). In sum, we show that including an utterance-final cue within a simple associative learning model could be a significant factor in explaining infants' word segmentation during the second year of life.

To further validate this new chunking approach, CLASSIC-UB results will be discussed in light of previous work on statistically-based utterance-boundary cues (Diphone-Based Segmentation; Daland & Pierrehumbert, 2011) and the grammatical characteristics of the chunks discovered by CLASSIC-UB will be examined.

Table 1. Adjusted R squared for linear regression models predicting word age of first production (estimated from children's speech using the method in Grimm et al., 2017) as a function of Log10 of the number of times a word was correctly segmented by each model. We used heteroskedasticity-robust standard errors and report bootstrap confidence intervals around the estimated R squared (based on 1000 iterations).

Model	Adjusted R squared	Lower Bci	Upper Bci
CLASSIC-Utterance-Boundary	0.08	0.06	0.10
PUDDLE	0.07	0.06	0.09
Backward transitional probability	0.04	0.03	0.06
Forward transitional probability	0.04	0.03	0.05

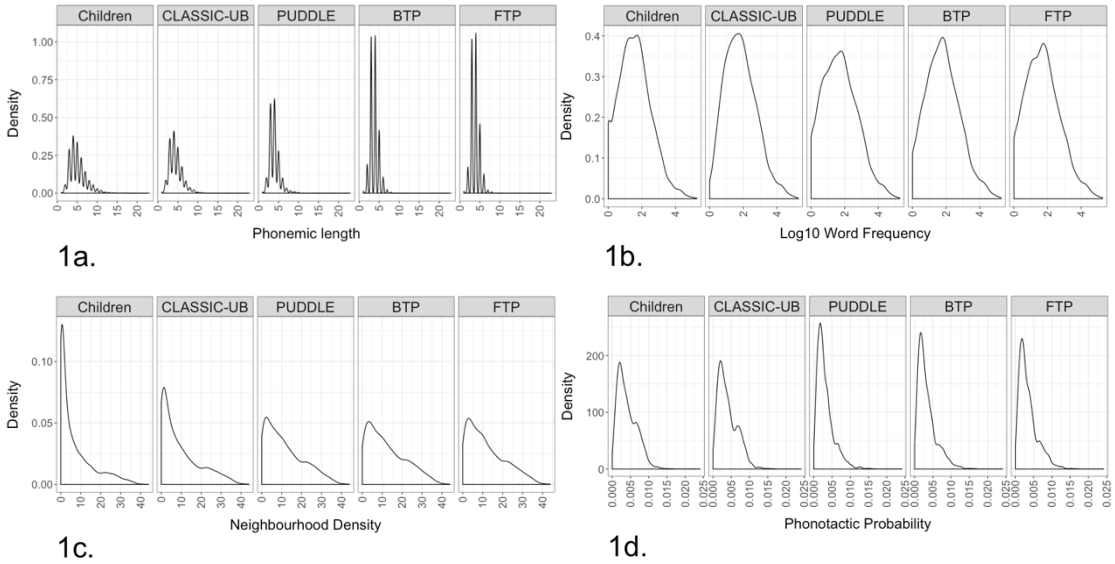

Figure 1. Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and of the words discovered by CLASSIC-Utterance-Boundary (CLASSIC-UB), PUDDLE, backward (BTP) and forward (FTP) transitional probability models, by phonemic length (1a), Log10 word frequency (1b), neighbourhood density (1c) and phonotactic probability (1d). The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

References

Christiansen, M. H. (2019). *Top Cogn Sci*, 11, 468–481.

Daland, R., & Pierrehumbert, J. B. (2011). *Cogn Sci*, 35, 119–155.

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). *Front Psychol*, 8, 555.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). *Cogn Psychol*, 63, 93–106.

Jones, G., & Macken, B. (2018). *Mem Cognit*, 46, 216–229.

Larsen, E., Cristia, A., & Dupoux, E. (2017). *Interspeech 2017*, 2198–2202.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Monaghan, P., & Christiansen, M. H. (2010). *J Child Lang*, 37, 545–564.

Perruchet, P. (2019). *Top Cogn Sci*, 11, 520–535.

Perruchet, P., & Vinter, A. (1998). *J Mem Lang*, 39, 246–263.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). *Science*, *274*, 1926–1928.

Seidl, A., & Johnson, E. K. (2006). *Dev Sci*, 9, 565–573.

Computational model: CLASSIC-Utterance-Boundary

The proposed chunking model is based on CLASSIC (Jones & Macken, 2018), which processes phonetically transcribed utterances and initially represents the standard British English phonemes. In this model, co-occurring items encountered in the input are gradually chunked together into increasingly larger sequences. For example, when presented with the phrase *W, EH, R, Z, D, AE, D* (i.e., *where's dad*), CLASSIC will first chunk adjacent phonemes into phoneme pairs (*WEH*, *EHR*, *RZ*, *ZD*, *DAE*, *AED*), being able to represent the utterance as *WEH, RZ, DAE, D* at the second presentation. The model will then progressively create larger and larger chunks at new utterance presentations (e.g., at the third presentation, the utterance would be represented by the following chunks: *WEHRZ*, *DAED*).

In the original version of CLASSIC, the input is pre-segmented into word units. We therefore modified CLASSIC to be able to find word boundaries in an unsupervised way. An utterance-final marker was added to chunks discovered at the end of the utterance, based on evidence that infants learn items presented at the utterance edges more easily (Seidl & Johnson, 2006). Only utterance-final markers were implemented, as adding an utterance-initial marker did not significantly contribute to model performance (*utterance-initial-final model* Adjusted R squared = .08 [.07, .10]; absolute difference in Adjusted R squared between *utterance-initial-final* and *utterance-final* models = .00 [.00, .01]).

In CLASSIC-Utterance-Boundary, when encoding a phoneme pair such as *AED* (in *D, AE, D* at the end of *Utterance 1*) the model will learn the chunk with an associated utterance-final marker (i.e., *AED*↲). If the chunk *AED* appears in later utterances, the model will then recognise the chunk with its final marker even if this is *not* at the end of the utterance (e.g., *dad is coming*; see *Utterance 3*):

Utterance 1 = *W, EH, R, Z, D, AE, D;* chunks discovered = *WEH*, *EHR*, *RZ*, *ZD*, *DAE*, *AED*↲

Utterance 2 = *WEH, RZ, DAE, D;* chunks discovered = *WEHRZ*, *DAED*↲

Utterance 3 = *DAED, IH, Z, K, AH, M, IH, NG;* chunks discovered = *DAED*↲*IH, IHZ, ZK, KAH, AHM, MIH, IHNG*

At later stages, the model will represent utterances using single chunks that include multiple speech boundaries (e.g., *I'll do it later* represented with 2 chunks: *I'll*↲*do*↲*it*↲, *later*↲). The model gives priority to chunks with associated final markers when parsing an utterance, as these represent strong cues to word boundaries. Equal chunks (i.e., sharing the same phonemic information) with and without associated final markers coexist in the model lexicon (e.g., *AED*, *AED*↲), as a chunk could also be initially found in the middle of an utterance and later as utterance final (e.g., if *Utterance 3* were encountered before *Utterance 1*, the sequence *AED* would have been discovered for the first time in the middle of the utterance). If a chunk with a final marker is not found, the model will then search in the lexicon for the same chunk with no associated marker.