The cross-linguistic performance of word segmentation models over time

Andrew CAINES[†], Emma ALTMANN-RICHER*, Paula BUTTERY[†]

[†]Department of Computer Science & Technology, University of Cambridge, Cambridge, U.K.

*Faculty of Modern & Medieval Languages, University of Cambridge, Cambridge, U.K.

andrew.caines@cl.cam.ac.uk, emma.altmann.richer@hotmail.co.uk, paula.buttery@cl.cam.ac.uk

Address for correspondence:

Andrew Caines, Department of Computer Science & Technology, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, U.K.

Acknowledgements:

We thank Alejandrina Cristia and Mathieu Bernard for developing the open-source phonemizer and wordseg software packages, for their assistance in troubleshooting, clarifying the workings of each toolkit, and providing very helpful commentary on an earlier draft of this article. This work was supported by Cambridge Assessment English, University of Cambridge. We also thank the editors and anonymous reviewers for their feedback which helped to greatly improve this article.

Keywords:

word segmentation, CHILDES, statistical learning

Abstract

We select three word segmentation models with psycholinguistic foundations – transitional probabilities, the diphone-based segmenter, and PUDDLE – which track phoneme co-occurrence and positional frequencies in input strings, and in the case of PUDDLE build lexical and diphone inventories. The models are evaluated on caregiver utterances in 132 CHILDES corpora representing 28 languages and 11.9m words. PUDDLE shows the best performance overall, albeit with wide cross-linguistic variation. We explore the reasons for this variation, fitting regression models to performance scores with linguistic properties which capture lexico-phonological characteristics of the input: word length, utterance length, diversity in the lexicon, the frequency of one-word utterances, the regularity of phoneme patterns at word boundaries, and the distribution of diphones in each language. These properties together explain four-tenths of the observed variation in segmentation performance, a strong outcome and a solid foundation for studying further variables which make the segmentation task difficult.

Introduction

The ability of the human infant to acquire language from the “great blooming, buzzing confusion” they are confronted with in their linguistic input (James, 1890) is a remarkable feat. At the very least there is the question of how they manage to identify words in a stream of speech in which words are not consistently, overtly delimited (the same question applies to the identification of signs in sign language, though we do not address it here). It has been shown that infants start to use familiar names for speech segmentation at the age of six months (Bortfield, Morgan, Golinkoff & Rathbun, 2005) and acquire at least a receptive lexicon of frequently heard words by the age of one, even if they may not yet know the meaning of all those words and are not able to produce them properly (Hallé & de Boysson-Bardies, 1994; Vihman, dePaolis, Nakai & Hallé, 2004). But how have they managed to pick these words out of the speech stream? This is the *word segmentation* problem, and the subject of our paper. Evidently, children are simultaneously acquiring knowledge of phonology, morphology, syntax, semantics, pragmatics, and more, but here we isolate the task of vocabulary identification and acquisition using text-based transcriptions of speech, as others have done (e.g. Larsen, Cristia & Dupoux, 2017).

Computational word segmentation involves an input transcription in which the speech stream is represented in phonemic form (‘phonemized’) in some way – for instance using the International Phonetic Alphabet (IPA). Basic units are delimited (with spaces in the example below) and presented to the model, which is required to insert hypothesised word boundaries between the basic units where appropriate. The success of the segmentation model is judged by evaluation of these proposed word boundaries against the true gold-standard boundaries in the original transcription.

INPUT: t w ɒ z b r ɪ l ɪ g a n d ð ə s l ɪ ð i t o v z
OUTPUT: twas brill lig and the slithytoves
GOLD: twas brillig and the slithy toves

We evaluate several computational models of word segmentation which have a cognitive basis, operate with phonemes as the basic unit, and are cross-linguistically implementable. We select three approaches – firstly a model which inserts word boundaries on the basis of transitional probabilities between sound sequences (Saksida, Langus & Nespor, 2017), secondly DiBS, the Diphone-Based Segmenter (Daland & Pierrehumbert, 2011), which tracks diphone (phoneme pair) co-occurrence probabilities – and thirdly a frequency-based approach which builds diphone and lexicon inventories, ‘Phonotactics from Utterances Determine Distributional Lexical Elements’, or PUDDLE (Monaghan & Christiansen, 2010).

These models essentially learn language-specific information without pre-specifying what that information should be, as some have done (Brent & Cartwright, 1996; Gambell & Yang, 2005; Mattys, White & Melhorn, 2005). Instead the bottom-up approach is grounded in work showing that infants eventually learn to use language-specific stress patterns alongside statistical cues, and develop a proto-lexicon in the first year (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003; Vihman *et al* 2004; Ngon, Martin, Dupoux, Cabrol, Dutat & Peperkamp, 2013). Our three chosen approaches (transitional probabilities, DiBS, PUDDLE) all make use of phoneme pair co-occurrence frequencies in some way and thereby extract language-specific phonotactic regularities. This choice is grounded in child psychology studies showing that infants are sensitive to bigram frequencies (Goodsitt, Morgan & Kuhl, 1993) and develop an early awareness of phonotactic cues (Mattys & Jusczyk, 2000).

In total we process 132 corpora representing 28 language varieties and 11.9 million word tokens. The data come from the CHILDES database (MacWhinney, 2000) and are freely available for research use. To the best of our knowledge this is the largest and most multilingual word segmentation experiment to date. Of our three target models, PUDDLE shows the best performance as measured by type (vocabulary), token (strict) and boundary (lax) metrics, albeit with considerable cross-linguistic variation. To account for this variation, we examine linguistic properties which could make a language more or less straightforward to learn to segment: mean word length, mean utterance length, type-token ratio, the proportion of one-word utterances, word boundary diphone entropy, and the 'Zipfian-ness' of diphone distributions. These *quantitative* measures are to some extent representative of the *quality* of the linguistic input children encounter: they represent the diversity of the lexicon, the regularity of the phonotactic system, and the difficulty of the task. They indicate how an input's quality may be scored for properties relating to learnability – our assumption that not all information sources and systems are equally easy to decode and to learn. They speak to the orderliness of the input and the ways in which starter hints are given to the child as stepping stones towards full decoding.

We use these six linguistic properties as fixed effects in a mixed-effects regression model with each individual corpus as a random effect and token segmentation as the dependent variable. The model explains 91% of the variation seen, albeit with the majority accounted for by the random language, speaker and topic idiosyncrasies of the corpora. This demonstrates why cross-linguistic work matters – namely that models proposed and evaluated with certain languages may not apply equally well to other languages. Basque, for instance, proves to be particularly hard for PUDDLE to segment, while showing by far the best performance on Mandarin. We attribute this difference to the orderliness of each language's diphone systems, or the strength of signal given by each diphone with regards its juxtaposition to a word boundary. In Mandarin, there is a smaller diphone system than in Basque, and the vast majority of diphones have a clear role as boundary or non-boundary signalling: 97% of diphones have a probability of at least 90% or 10% of occurring next to a boundary. For Basque, on the other hand, the equivalent statistic is 53%, and the number of diphones is almost 1.5 times as many as are found in Mandarin.

Meanwhile there appears to be a positive bias in previous work towards Germanic languages, albeit with a small sample of different language families generally. However, this provisional finding suggests that diphones matter more as segmentation cues in Germanic languages than others. Hence, the *quality* of the speech stream in terms of typological differences is shown to affect model performance and present challenges to theory-building going forward.

We also introduce a temporal dimension in our assessment of the segmentation models, exposing them to increasing amounts of data as a proxy for time. PUDDLE is again the outstanding model in this setting, showing improvement with increasing amounts of input data, whereas the other models deteriorate in performance. This aspect of the experiment demonstrates the benefit of PUDDLE's inventory-based approach, and suggests that purely statistical approaches are not sufficient for fully functioning word segmentation. We conclude that distributional cues combined with memory devices to accumulate knowledge are a successful method for the segmentation of phonemic transcriptions of speech.

Ideally of course we would have access to the original recordings so that we could make use of further prosodic features such as word stress and rhythm, as infants have been shown to also pay attention to these cues (Jusczyk, Cutler & Redanz, 1993; Curtin, Mintz & Christiansen, 2005; Curtin, 2009), or to accurate automatic syllabifiers so that the model can learn from syllables as well as phonemes. However, with only transcriptions to work with at this stage, we can still perform statistical sequence analyses, following the proposal that for infants the strategy of frequency tracking and probabilistic calculation can act as a successful initial approach to word segmentation before more accurate language-specific cues are learned (Thiessen & Saffran, 2003). It is likely that allowing the models to additionally learn from suprasegmental, semantic and multimodal information would further boost performance and be closer to full ecological validity.

Related Work

Many have attempted to explain word segmentation on the basis of various probabilistic cues available in the input speech stream, such as lexical stress and prosodic patterns (Cutler & Carter, 1987; Johnson & Jusczyk, 2001), transitional probabilities between syllables (Saffran, Aslin & Newport, 1996), and phonotactic constraints of the target language (Mattys, White & Melhorn, 2005). Computational models of word segmentation have so far tended to take an approach based on detection of frequent sequences (e.g. Goldwater, Griffiths & Johnson, 2009), transitional probabilities (e.g. Saffran, Aslin & Newport, 1996), or a combination of the two (e.g. Swingley, 2005). This is based on the repeated observation that infants are sensitive to distributional information in the input (Aslin, Saffran & Newport, 1998; Thiessen & Saffran, 2003; Pelucchi, Hay & Saffran, 2009a).

A seminal, purely probabilistic approach was that of Saffran and colleagues, who demonstrated that 8-month-old infants could learn to segment an artificial bisyllabic language (Saffran, Aslin & Newport, 1996). They accounted for the infant's segmentation abilities on the basis of transitional probabilities (TPs) between syllables, grounded in experimental evidence of children's statistical learning abilities (Goodsitt, Morgan & Kuhl, 1993; Krogh, Vlach & Johnson, 2012). This model did not transfer successfully to an artificial language which varied in the number of syllables per word (Johnson & Tyler, 2010) but Saksida and colleagues (2017) then demonstrated good segmentation performance on natural language child corpora by using absolute thresholding to determine word boundaries per Gervain & Guevara Erra (2012) and Swingley (2005), rather than the relative thresholding approach used by Saffran and colleagues. With this modification, the mean bigram co-occurrence probability (where a gram may be either a syllable or a phoneme) in the corpus is used as a threshold below which word boundaries are posited, whereas with the relative TP method boundaries are proposed wherever a bigram's co-occurrence probability is less than both its neighbours. Further support for the TP approach comes from studies of Italian (Pelucchi, Hay & Saffran, 2009b; Hay *et al*, 2011).

Others have incorporated more than input frequencies in their model – for instance, prosodic or phonotactic information – with some success. A unique stress constraint was added to transitional probabilities in Gambell & Yang's model, stating that there should only be one primary stress per word in English (Gambell & Yang, 2005). As a result segmentation accuracy improved from a baseline of 41.6% to 73.5%, but the model has been criticised on the grounds that a pronouncing dictionary provides idealised representations of stress and moreover the constraint is not cross-linguistically generalisable (Phillips, 2015).

Previous work has focused mainly on artificial languages or on English (Saffran, Aslin & Newport, 1996; McCauley, Monaghan & Christiansen, 2015; Larsen, Cristia & Dupoux, 2017), understandably, since the former offers greater control, and the latter is by far the best-resourced in corpus terms. Exceptions include segmentation of Sesotho (Johnson, 2008), French (Boruta *et al*, 2011), Hungarian and Italian (Gervain & Erra, 2012), Japanese (Fourtassi *et al*, 2013), and eight languages besides English (Saksida *et al* 2017). We extend the cross-linguistic evaluation of word segmentation models, covering as many languages as our chosen instruments allow – the CHILDES database and the `phonemizer` toolkit to convert the corpora into phonemic form. The languages in our study include the nine from Saksida *et al*'s with the exception of Tamil and Polish.

First Cairns and colleagues (1997) then Hockema (2006) introduced DiBS, the Diphone-Based Segmentation model, which was initially considered unlearnable since the original DiBS algorithm relied on a *supervised* approach transcriptions in which boundaries were explicitly marked – an unrealistic scenario for infants who are in fact attempting to discover these boundaries in an *unsupervised* fashion. As a tractable solution, and in common with much work in the computational literature, Daland & Pierrehumbert (2011) incorporated Bayes' theorem to estimate parameters on the basis of distributional information. This model learns a combination of word-edge phone probabilities, overall boundary probability and diphone transitional probabilities in order to make segmentation decisions, which fits with evidence that infants are aware of phonotactic regularities by the age of one (Mattys & Jusczyk, 2000). Another model to make use of diphone sequences, though this time through frequency counts and inventory building, is PUDDLE (Monaghan & Christiansen, 2010). This approach tallies with experimental work showing that infants have a receptive lexicon by the age of one, listening longer to words frequently found in child-directed speech (e.g. *ballon* 'ball', *canard* 'duck') than rare words such as *busard* 'harrier' (Hallé & de Boysson-Bardies, 1994; Vihman, dePaolis, Nakai & Hallé, 2004). We made our selections as they offer alternative probabilistic and frequency-based approaches and are grounded in behavioural work. We discuss the three chosen models further below.

Word segmentation models

Here we describe and detail the word segmentation models employed in this study. The models are implemented with the `wordseg` library (Bernard *et al*, in press) and represent contrasting probabilistic and frequentist approaches.

Baselines

The baseline models in `wordseg` insert word boundaries with a given probability *P*(#), implementing the proposal by Lignos (2012). Variations include the *utterance baseline* in which whole utterances are treated as words, *i.e. P*(#)=0, and the *basic unit baseline* in which every phoneme (or syllable, if that is the basic unit) is treated as a word, *i.e. P*(#)=1. We also use a *random baseline*, in which *P*(#)=0.5, so that boundaries are inserted by chance, and an *oracle baseline* which is informed with the probability of a boundary in each corpus (*P*(#) = n.boundaries / n.phonemes). Our basic unit is the phoneme rather than syllable. We include the baselines as another way to evaluate the performance of our three chosen models by comparison with the random segmentation approach.

Transitional probabilities

The idea of the transitional probabilities approach (TPs) is to identify complex units of relatively high probability, given the distribution of unit sequences in the data. To calculate TPs, we firstly use the *forward TPs* model in `wordseg` according to which the transitional probability of a sequence *XY* is the count of *XY* divided by the count of *X* (Frank, Goldwater, Griffiths & Tenenbaum, 2010). We also use the *backward TP* and *mutual information* model variants: the former calculates transitional probabilities as the count of *XY* divided by the count of *Y* (Pelucchi *et al* 2009a); the latter is the binary logarithm of the count of *XY* over the product of *count(X)* and *count(Y)* (Gervain & Guevara Erra, 2012). For instance if the sequence /ba/ occurs 5 times, while /b/ occurs 100 times and /a/ occurs 500 times, the forward TP will be 5/100, the backward TP will be 5/500=1/100, and mutual information is log2(5/(100*500))=log2(5/50000)=-13.29.

In our implementation of TP models, word boundaries are inserted between basic units where TPs fall below an *absolute* threshold – where the transitional probability of *XY* is less than the mean TPs of all phoneme pairs in the corpus – an unsupervised method introduced by Swingley (2005). This is instead of a *relative* threshold used in classic TP work, where a boundary would be inserted wherever the TP of *XY* falls below that of both its neighbours *WX* and *YZ* (Saffran, Aslin & Newport, 1996), since it has been shown that absolute thresholding consistently outperforms relative thresholding (Saksida *et al* 2017). TPs were introduced with a focus on syllables, but here we use phonemes as our basic unit to allow for comparison across models. Note that the `wordseg` toolkit employs the TP code and refinements implemented by Saksida and colleagues (2017).

DiBS

DiBS models the word segmentation process as a matter of learning diphone collocations, leading to a bottom-up acquisition of phonotactic constraints. That is, the learner observes that, in English for example, [b] and [a] often co-occur within a word whereas [p] and [d] only co-occur *across* word boundaries. This proposition is founded on multiple reports in the literature (Friederici & Wessels, 1993; Jusczyk, Luce & Charles-Luce, 1994; Mattys & Jusczyk, 2001).

DiBS assigns a value between 0 and 1 to utterance-medial diphones indicating the probability that there is a word boundary between that pair of phonemes. The probabilistic information comes from a training corpus which provides frequencies of phonemes in word-initial and word-final positions, as well as their overall counts.

In our case the training data are the transcribed CHILDES corpora, filtered down to the utterances of non-child speakers only. We opt not to use the so-called DiBS-*gold* setting which gives access to the word-delimited corpora as training data, on the grounds that this is ecologically far from reality: it is precisely because children are not 'fed' one word at a time that they are faced with the word segmentation problem. Nor do we use DiBS-*lexical*, which involves a seed lexicon – even if children may have learned one by the age of 2 years – because it is not clear what that lexicon should be without individual testing of each of the children involved. Instead we use DiBS-*phrasal* which gives the learner access to utterance boundaries as training data, as this is ecologically a more valid scenario and does not give DiBS an unfounded advantage.

Phoneme frequencies are used to calculate the joint probability of a given phoneme *X* occurring in word-final position and another phoneme *Y* occurring in word-initial position; in other words occurring either side of a word boundary # in the sequence *X#Y*. This is expressed as *P(#|XY)* and defined by applying Bayes' rule as below:

EQUATION 1 HERE

Since *P(#)* over *P(XY)* is constant, the learner needs only calculate *P(XY|#)*, which approximates formally to the following:

EQUATION 2 HERE

Note the independence assumption taken here: that the probability of word-initial *Y* is not conditioned by a preceding word-final *X*. Testing this assumption through *n*-gram models where the grams are phonemes and where $n \geq 2$ is a matter for future work. In `wordseg`, as in Daland & Pierrehumbert (2011), word boundaries are placed where probabilities are greater than 0.5.

PUDDLE

In the PUDDLE model utterances are initially treated as whole lexical items but are broken down into smaller units if any already-stored lexical items are encountered, consistent with what adults are known to do when faced with a novel artificial language (Dahan & Brent, 1999). There is an implicit reliance on the occurrence of single-word utterances for this approach to be at all successful, but these are used as a way to bootstrap a lexicon rather than a sole learning strategy. This is based on the observations that (a) many utterances in child-directed speech contain one word only, this being a strategy caregivers use to introduce novel objects and concepts, and (b) single-word utterances alone are insufficient for fully-functional speech segmentation (Brent & Cartwright, 1996).

As an example, consider the following three utterances:

1. kitty
2. thatsrightkittyyes
3. lookkitty

We see that the first utterance is a single word, *kitty*, which is stored as a lexical entry along with its initial and final diphones: kɪ, ti (we phonemize words here according to a southern British English standard). An activation function keeps count of word occurrences since the lexicon is frequency ranked. The lexical entry *kitty* allows us to segment utterance 2, giving us a second instance of *kitty*, as well as new entries *thatsright* and *yes*. The same process segments *look* and *kitty* in utterance 3, and by now we have three instances of *kitty* a list of hypothesised word-initial diphones (kɪ, ða, jɛ, lʊ), and another list for word-final diphones (ti, ait, ɛs, ʊk).

PUDDLE builds phonotactic awareness in a bottom-up fashion, collecting word-initial and word-final diphones (two adjacent phonemes) for any item added to the lexical inventory. In this way the model learns lists of permitted word beginnings and endings from the input but does not need a fully specified phonotactic rule set. This knowledge is put to use to prevent over-segmentation: a matching lexical entry in the input is only accepted and processed if the diphone preceding the given segment also ends at least one word in the lexicon, and the diphone following the given segment also begins at least one word in the lexicon. As a consequence, whole utterances tend to be added to the lexicon holistically in the early stages of input processing. Subsequently, a single-word utterance may be encountered which enables the bootstrapping effect to kick in.

Obvious issues here include (a) rare words which the infant only hears once or twice in their early years, and (b) formulaic multi-word utterances (such as ‘thank you’, ‘bath time’, ‘eat up’, etc) whose unity is preserved by their frequency of occurrence. These outcomes are problematic only if perfect word segmentation is the target; however, in reality word segmentation is an imperfect skill, for adults as well as children, and indeed one might question whether it is important to segment words so rare in the input that they only occur a few times in a few years (of language samples), while also acknowledging that children do treat some frequent collocations as whole items, at least in the early stages of language acquisition (MacWhinney, 1982; Tomasello, 2000) – and perhaps long after (Siyanova-Chanturia *et al*, 2017).

The incremental design of PUDDLE, along with the activation function, means that over time only frequent collocations continue to play a role in segmentation. Indeed we can imagine that beyond these first three utterances the child might encounter examples which allow her to properly segment *thatsright*, so that eventually real words overtake its current activation of 1 in the lexicon. A decay parameter can be introduced to simulate learner forgetfulness over time. Pilot studies with PUDDLE did not clarify what the optimal value for this parameter should be. Therefore, Monaghan & Christiansen set the decay parameter to zero, as we do here.

Corpora

To test the efficacy of the four word segmentation models cross-linguistically, we obtained as many suitable corpora from the CHILDES database as were found to meet the following four criteria, which are justified in detail below:

a. The corpus is monolingual;
b. The corpus contains at least 10,000 speech utterances spoken by conversation participants other than the target child;
c. The target child is aged 2 years or younger at the start of the corpus;
d. The corpus language can be processed by the `eSpeak` Next Generation (NG) speech synthesizer (https://github.com/espeak-ng/espeak-ng) or `segments` grapheme-to-phoneme transformer (https://github.com/cldf/segments).

We do not attempt to process multilingual corpora – even though the multilingual environment is a highly frequent setting for children in reality – as it requires word-by-word language identification at scale; we see this as a challenge to be undertaken in future work. Previously it has been shown that word segmentation models, using PUDDLE with English data at least, stabilise after 10,000 input utterances (McCauley, Monaghan & Christiansen, 2015). We therefore set this as a minimum size for corpus selection. For reasons of practicality, we define our test and training data as all non-child utterances in CHILDES corpora, based on the assumption that the child could be paying attention, and therefore trying to segment, all of it. This may be an exaggeration, but there is a case for children learning from overheard speech (Mani & Patzold, 2016) and therefore, rather than manually identifying which of the transcribed utterances are indeed child-directed (which in itself will be error-prone without access to video recordings), we make this simplifying assumption. Note that we exclude diary corpora as they are usually focused on child utterances rather than child-directed speech.

The reason to control the starting age of the target child is that if children are much older than 2 years, they have normally begun using 2-word utterances (at least) and therefore learning to segment words, while still relevant, is confounded with the acquisition of multi-word constructional frames or schema (which may begin sooner than 2 years but is not apparent until then).

Finally, the current version of `eSpeak` NG (1.49.3) is capable of handling 100 languages (where 'language' in this context includes varieties such as different Englishes, Belgian-French, north and south Vietnamese, *etc*), while `segments` uses grapheme-to-phoneme rules to process languages such as Japanese which have been transcribed in the Roman script, as is the case with the CHILDES corpora. At least 37 languages are represented by the corpora contained in CHILDES at the time of writing. The true figure is likely to be greater than 37, as various languages may be concealed in the multilingual corpora contained in CHILDES. The effect of applying these criteria is that we are able to represent 28 language varieties, as listed in Table 1 along with the source corpora, the number of child participants represented, the range of their ages at the start of recording, and the number of word tokens in the corpora at their biggest (i.e. at 10,000 utterances). In total, we have data from 66 studies constituting 132 child corpora and 11.9 million word tokens.

Table 1: List of CHILDES corpora used in this study; ages are expressed as years;months.

Language	**N. studies**	**N. corpora**	**Child age at start**	**N. other participants**	**N. utterances**	**Word tokens**
Basque	1	1	2;0	60	10,000	117,255

Cantonese	1	6	1;10-1;11	49	60,000	420,836
Croatian	1	2	0;10-1;10	35	20,000	217.453
Danish	1	2	0;8-0;11	14	20,000	131,075
Dutch	3	10	0;11-1;9	52	100,000	841,863
English (N.Am.)	16	28	0;6-1;11	241	280,000	2,296,785
English (U.K.)	6	18	0;1-2;0	95	180,000	1,411,017
Estonian	2	2	1;3-1;7	8	20,000	285,143
Farsi	1	2	2;0	11	20,000	217,358
French	1	2	1;9-1;11	28	20,000	151,966
German	5	10	0;1-1;11	288	100,000	1,072,694
Greek	1	1	1;7	6	10,000	82,150
Hungarian	3	3	1;5-2;0	38	30,000	330,200
Icelandic	1	1	2;0	12	10,000	106,253
Indonesian	1	4	1;6-2;0	230	40,000	301,847
Irish	1	1	1;5	3	10,000	106,648
Italian	1	1	1;5	2	10,000	108,926
Japanese	3	7	0;6-1;5	65	70,000	590,789
Korean	2	3	1;3-2;0	11	30,000	501,680
Mandarin	2	2	0;1-1;7	14	20,000	364,552
Norwegian	1	1	2;0	5	10,000	94,410
Portuguese (Br.)	1	1	1;8	9	10,000	98,624
Portuguese (Pt.)	1	3	1;5-1;6	22	30,000	276,743
Romanian	1	1	1;5	7	10,000	76,147
Serbian	1	8	1;6	186	80,000	720,024
Spanish	6	7	0;10-1;10	61	70,000	628,713
Swedish	1	4	1;3-1;11	24	40,000	244,877
Turkish	1	1	2;0	11	10,000	140,462
Total	66	132	-	1587	1,320,000	11,936,490

As noted above, we opt to use the phoneme as our basic unit, rather than the syllable as some have done (e.g. Saksida *et al* 2017), either because the syllable is a requisite part of their segmentation model, or for theoretical reasons. There is a debate as to the validity of the syllable in perceptual and learning terms (Mehler, Dommergues, Frauenfelder & Segui, 1981; Ladefoged, 2003; Ziegler & Goswami, 2005; Räsänen, Doyle & Frank, 2018), which we do not address here but instead fall back on a structurally less complex unit – the phoneme, which we understand to be meaningful abstractions over clusters of similarly-realised phones – and leave comparison between phoneme and syllable inputs as a matter for future work. On a more practical note, *reliable* automated syllabifiers are not available for many of the languages in our sample, if any, and manual syllabification is infeasible for the size of data we work on without many-labs type collaboration: Saksida and colleagues (2017) did it for 3300 utterances in each of the nine languages they work with, and make their annotated data available, whereas we work with 10,000 utterances for each of 132 corpora.

Note also that the phonemic representations we work with are idealised speech productions based on dictionary pronunciation of words. We assume that the human transcribers have successfully recognised the words produced by participants in corpus recordings, though we acknowledge that there is likely to have been some degree of error in the recognition and transcription stages. In addition, human speakers of course produce phonemes with great variability from prototypical dictionary forms, whether for reasons of idiosyncrasy, dialect, reduction in connected speech, allophonic variation and errors. Listeners are adept at 'repairing' speech inputs according to their expectations, despite such variation and despite extraneous factors such as background noise (Dupoux *et al* 2011). Therefore it should be made clear that our analyses of word segmentation based on the speech transcriptions we extract from CHILDES are an idealisation of the natural scenario. However, in mitigation we emphasise that the transcriptions are a tiny sample of a child's language input, and that these utterances can reasonably be considered as a representation of 'good' exemplars they might encounter. In addition it has been shown that caregivers make an extra effort to produce speech in a clear fashion where possible (Hartman, Bernstein Ratner & Newman, 2017). Besides word segmentation, the child of course has to learn to group phone realisations as phonemes, though this is not a task we tackle here, as we do not have access to all the original recordings of the target corpora.

Language properties

We do not assume that all languages are equally easy to learn to segment. Therefore we hypothesise that there are properties intrinsic to each language which make word segmentation more or less straightforward. The properties we consider here – though we do not claim this to be an exhaustive list – are mean word length, mean utterance length, type-token ratio, the proportion of one-word utterances, word boundary diphone entropy, and the 'Zipfian-ness' of diphone distributions.

The first variable to some extent represents the difficulty of the word segmentation task in each language. Mean word length, expressed as the number of phonemes per word token in the corpus, indicates how much phoneme agglomeration occurs in the language: we hypothesise that the closer this value to one, the more straightforward segmentation will be as words will be phonologically simpler. On the other hand, the mean number of word tokens per utterance in each corpus indicates how much segmentation tends to be required for utterances in a given language.

Type-token ratio (TTR) is a size-normalised measure of lexical diversity in language samples. It is calculated by dividing the number of unique word types found in a corpus by the number of word tokens in that corpus. The outcome lies between 0 and 1, with a higher TTR indicating greater lexical diversity in the corpus relative to its size (i.e. many different words with low repetition). We hypothesise that TTR has a changing role over time, with high TTR in early phases of language acquisition serving to expose the child to more lexical items more quickly, followed by decreasing TTR in later stages moving the input closer to everyday discourse, with greater repetition of a few word types and a long tail of infrequent words. Therefore we model TTR in interaction with the size of the corpus, as a proxy for developmental stage.

The proportion of one-word utterances (OWU) – the count of one-word utterances over the count of all utterances in the corpus – is a measure of the salience given to words in speech directed to children. Word boundaries do not need to be identified in one-word utterances (though they still might be) and are potentially a rich source of lexicon entries for the learning child. We might expect that a high frequency of OWUs helps the child in the task of word segmentation, and of our models, aids PUDDLE in particular since it is specifically designed to absorb the information provided by utterance boundaries, along with DiBS since we have opted for its training setting based on utterance boundaries alone.

We can think of entropy (H) as the amount of uncertainty in a random variable, with higher values of H for a probability distribution D representing greater disorder (3).

EQUATION 3 HERE

In our case D is a vector of probabilities for each diphone in our corpus occurring at a word boundary (*i.e.* the count of *XY#* and *#XY*, over the count of *XY*). Thus we can assume that a higher entropy word boundary diphone system, H(D), is harder to learn because the diphones are more randomly distributed. A lower entropy diphone system, on the other hand, represents a more ordered phonotactic system in which certain diphones are highly indicative of word boundaries. That is, imagine a scenario where every diphone is equally likely to occur at a word boundary or not: the high amount of uncertainty as the learner encounters each diphone means that there are no straightforward cues to word segmentation. This scenario would have a high entropy value. Now imagine that there are certain diphones which are highly likely to occur at word boundaries, and there are some which hardly ever occur at word boundaries. This would be quite an orderly diphone system, in which there is less uncertainty, better cues, and lower entropy. We expect that our models will favour such systems, as they appeal directly to the phonotactics represented by diphones (Fourtassi *et al*, 2013).

Finally, we use the zipfR library for Large-Number-of-Rare-Events modelling (LNRE), designed for the analysis of power law distributions of the kind found in natural language corpora (Zipf, 1949; Baayen, 2001; Evert & Baroni, 2007). For each corpus we extracted ranked lists of diphone counts and constructed a Zipf-Mandelbrot LNRE model (Evert, 2004). We store the shape parameter (ZM.α) and goodness-of-fit values from a chi-squared test (ZM.X^2) as representations of the Zipfian-ness of each distribution. We presume that the better the fit of the LNRE model to a diphone distribution, the more learnable the phonotactic system will be (Kurumada, Meylan & Frank, 2013; Bentz et al, 2017; Hendrickson & Perfors, 2019). Higher values of α and lower values of X^2 are indicators of a better Zipf-Mandelbrot fit.

Evaluation

We report each segmentation model's performance using precision, recall and *F*-measure measured in three ways: *type*, *token* and *boundary*. The type metric assesses the lexicon accumulated by the model at the end of corpus processing, and compares it to the expected lexicon (the true set of word types in each corpus). The token metric indicates how many whitespace-delimited character strings (tokens) have been correctly segmented. This is the strictest measure, as it requires precise placement of boundaries both at the start and end of words.

The boundary metric, on the other hand, is more forgiving. It measures how many delimiters have been correctly placed, whether a word has been correctly segmented or not; thus it is a laxer performance measure than the token metric, as it rewards correct word boundaries even where the word token has not been correctly segmented. Furthermore every utterance gives two boundary matches for free, at the start and end, a matter addressed by an alternative 'no edges' metric which does not score utterance initial or final boundaries. We do not use this as it cannot be appropriately used for the utterance baseline model.

To calculate precision (*P*) and recall (*R*) we require three measures obtained by comparing segmentation hypotheses and *ground-truth* transcriptions. Firstly, true positives (*tp*) are hypothesised predictions in accord with the ground-truth. False positives (*fp*), on the other hand, are predictions in conflict with the ground-truth. False negatives (*fn*) are those boundaries in the ground-truth transcriptions which were not predicted by the model. Precision is then, tp / ($tp + fp$), a measure of positive predictive value – in our case how often the model correctly hypothesises a word boundary. Recall is, tp / ($tp + fn$), a measure of sensitivity, affected by the coverage of the model's predictions compared to the ground-truth. Finally, the *F*-measure represents both precision and recall by taking the harmonic mean of the two; that is, $F = 2 \times ((P \times R) / (P + R))$.

As an example of how each metric works, consider again the example utterance from the introduction, now extended along with a proposed segmentation, true segmentation, and various evaluation scores.

INPUT: t w ɒ z b r ɪ l ɪ g a n d ð ə s l ɪ ð i t o v z d ɪ d g aɪ ə a n d g ɪ m b l ɪ n ð ə w eɪ b
OUTPUT: twas brill lig and the slithytoves didgyre and gimblein the wabe
GOLD: twas brillig and the slithy toves did gyre and gimble in the wabe

Here, the model proposes 9 word types (*twas, brill, lig, and, the, slithytoves, didgyre, gimblein, wabe*) of which 4 are correct. Therefore type precision is 4/9=0.44. The model has returned 4 of the 11 true word types and thus type recall is 4/11=0.36. The *F*-measure is 0.4.

Only 6 tokens have been correctly segmented (*twas, and, the, and, the, wabe*). The model proposes 11 tokens in total, therefore token precision is 6/11=0.54, and there are 13 tokens in the gold utterance so token recall is 6/13=0.46. The *F*-measure is 0.5.

Finally, the laxer boundary evaluation metric counts 11 correct delimiters out of 12 proposed delimiters: precision is 11/12=0.92. There are 14 true delimiters, thus boundary recall is 11/14=0.79. Boundary *F* is 0.85. It should be clear from this example that these measures, applied at scale to whole corpora, give an indication of vocabulary learning (type scores), strict segmentation performance (token scores) and lax segmentation performance (boundary scores).

Method

For each language listed in Table 1 we processed the identified CHILDES corpora using the following steps:

i. Downloaded the corpora in XML format from http://childes.talkbank.org/data-xml and loaded the corpora using NLTK's CHILDES reader (Bird, Klein & Loper, 2009);
ii. Filtered any utterances spoken by the target child (marked CHI);
iii. Removed name anonymisers ('xxx, yyy, zzz'), filled pauses and paralinguistic speech tokens such as 'hm+, mm+, pft, uh+uh'; transformed all words to lower case;

iv. Passed the resultant text to the `phonemizer` with the appropriate language flag and `eSpeak` or `segments` backend speech synthesizer (Bernard, 2018);
v. Applied each word segmentation algorithm to the phonemized texts using `wordseg`, retrieving corpus statistics and evaluation metrics at 1000-utterance increments up to a limit of 10,000.

Our code is freely available for others to use at https://github.com/ANON/wordsegmentation

We note here that a child might begin to address the problem of word segmentation gradually, paying attention to the speech signal now-and-then rather than continuously. Researchers conventionally idealise the process by assuming that there is a moment from which the child begins to pay perfect attention to the available speech stream. Thus we can incrementally count the incoming utterances, and model word segmentation with the child as a perfect learner. This idealisation is no doubt far-fetched, but it is the best way to work with the data at present.

Results & Discussion

In Table 2 we report mean evaluation metrics for our nine segmentation models averaged across the 132 child corpora after 10,000 utterances. We show the four baseline models – utterance, random, unit and oracle – three implementations of TPs, along with DiBS and PUDDLE. The metrics are type, token and boundary with the highest-scoring model for each measure being highlighted in bold type.

Table 2: Performance of nine word segmentation models after 10,000 utterances: mean type, token and boundary precision (*P*), recall (*R*) and *F*-measure (*F*), where the first four models are the baselines, TP_FTP uses forward transitional probabilities, TP_BTP uses backward transitional probabilities, and TP_MI uses mutual information.

Model	**Type**			**Token**			**Boundary**		
	P	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Utterance	.230	.066	.102	.068	.178	.095	**1.00**	.423	.592
Random	.070	.122	.088	.075	.275	.115	.439	.712	.541
Unit	.018	.056	.027	.232	.005	.010	.356	**1.00**	.523
Oracle	.099	.118	.108	.064	.329	.106	.522	.599	.557
TP_FTP	.252	.297	.271	.102	.160	.119	.664	.756	.704
TP_BTP	.220	.241	.229	.070	.144	.092	.631	.692	.657
TP_MI	.294	.172	.216	.094	.308	.141	.848	.564	.676
DiBS	.445	.285	.346	.148	.421	.215	.914	.649	.758
PUDDLE	**.579**	**.616**	**.592**	**.360**	**.431**	**.378**	.825	.902	**.854**

It is apparent from Table 2 that two of the naive baseline strategies by definition show perfect boundary precision (the utterance baseline) and boundary recall (the unit baseline) since they were set to only place boundaries at utterance boundaries and after every unit, respectively. Conversely, they suffer for lack of recall and precision, as reflected in their boundary *F* scores. Otherwise, PUDDLE shows the best performance, out-scoring the other models on all remaining measures. The fact that it performs so much better on the type scores – a comparison of proposed and observed vocabularies – is a reflection of its specific lexical storage device. Its token scores are relatively high though still indicate that this is the strictest measure, since these are the metrics on which it performs worst. Finally, it has the best boundary scores overall (*F*), indicating that 8 times out of 10 it correctly places a boundary (*P*), and that it recovers 9 in 10 of all true boundaries (*R*).

Aside from the anomalous boundary scores for the utterance and unit approaches, the baselines are on the whole the lowest scoring models, as expected given their naive and random methods. The TP models do not perform much better than the baselines on token scores, but are much better for type and boundary scores. Of the three, the forward transitional probability and mutual information variants outperform the backward transitional probability variant, with FTP out-scoring MI on type and boundary *F* but the reverse being true for token *F*. This indicates firstly that the frequency information from the first phoneme in a diphone gives a slightly better cue to segmentation than the second, at least in the languages examined here; and secondly that the information from both phonemes helps more with the higher standards of exact token segmentation. DiBS is intermediate to TPs and PUDDLE in performance terms, though out-scoring PUDDLE on boundary precision, performing almost as well on this score as the utterance baseline. Recall that DiBS learns diphone boundary probabilities from utterance-delimited corpora and the high score on this measure, but low boundary recall, suggests it is relatively conservative in attempting to segment beyond utterance boundaries.

Pairwise Welch's *t*-tests (with Bonferonni adjustment for multiple comparisons) between the set of scores for each model and each of the three evaluation *F*-measures return a very low probability that the scores come from the same distribution ($p < 0.001$). We infer from the means in Table 2 that the overall performance ranking of our models is therefore [PUDDLE > DiBS > TPs > baselines].

Note that the mean values reported in Table 2 disguise a lot of inter-corpus variation, with standard deviations of .194, .111 and .081 for the type, token and boundary *F*-measures respectively. Moreover, seventy percentage points separate the maximum and minimum type *F* scores, 52% separate maximum and minimum token *F*, and there is 32% between top and bottom boundary *F*. Figure 1 illustrates this variation with boxplots representing *F*-measures for each of the nine models.

FIGURE 1 APPROX HERE

Figure 1: *F*-measure boxplots for each of nine word segmentation models showing performance on 132 CHILDES corpora, where the thick horizontal bar is the median, the upper and lower limits of the box represent the 3rd and 1st quartiles respectively, the whiskers extend as far as 1.5 times the inter-quartile range (3rd minus 1st quartile) from the box limits, and any other points are outliers beyond this range (plotted with an alpha of 0.25, meaning that it requires 4 points to be overlaid before they appear as solid black).

It is clear from Figure 1 that the model with the most performance variation is PUDDLE, which may be a symptom of its multi-part architecture – with accumulating inventories dependent on lexical and diphone frequencies, order of presentation, and corpus or speaker idiosyncrasies. In other words, the knowledge it accumulates is more sensitive to language-specific corpus distributions than the other models. Otherwise the points are relatively tightly packed, with TP backward transitional probabilities and DiBS showing the next most variance after PUDDLE.

Meanwhile performance variation by corpus language is another focus of these experiments, and thus in Table 3 we report mean *F*-measures from PUDDLE averaged for the corpora in each language, with language family classifications from Glottolog (Hammarström, Forkel & Haspelmath, 2018), using the first or second classification level, whichever is the more specific[1]. The table is presented in descending order by token *F*-measure since it is the strictest evaluation metric.

Table 3: Mean PUDDLE type, token and boundary *F*-measures per language after 10,000 utterances, with the number of corpora per language and standard deviations in parentheses for languages with multiple corpora.

Language	Family	N.Corpora	Type F	Token F	Boundary F
Mandarin	Sinitic	2	.901 (.030)	.697 (.051)	.970 (.009)
German	Germanic	10	.787 (.026)	.497 (.016)	.927 (.012)
Danish	Germanic	2	.788 (.014)	.494 (.024)	.936 (.002)
Swedish	Germanic	4	.625 (.019)	.477 (.032)	.876 (.006)
Icelandic	Germanic	1	.652 (-)	.477 (-)	.879 (-)
Greek	Graeco-Phrygian	1	.581 (-)	.426 (-)	.861 (-)
Norwegian	Germanic	1	.603 (-)	.420 (-)	.854 (-)
Dutch	Germanic	10	.659 (.028)	.418 (.034)	.882 (.011)
English (N.Am.)	Germanic	28	.716 (.026)	.414 (.034)	.902 (.011)
English (U.K.)	Germanic	18	.702 (.098)	.411 (.054)	.897 (.043)
Farsi	Indo-Iranian	2	.539 (.048)	.394 (.049)	.848 (.027)
Estonian	Finnic	2	.534 (.006)	.392 (.015)	.821 (.001)
Serbian	Balto-Slavic	8	.412 (.025)	.362 (.017)	.797 (.008)
Romanian	Italic	1	.476 (-)	.360 (-)	.829 (-)
Cantonese	Sinitic	6	.848 (.018)	.351 (.024)	.950 (.006)
Turkish	Common Turkic	1	.377 (-)	.350 (-)	.780 (-)

[1] For instance, we opt to label German as Germanic (its 2nd level classification) rather than Indo-European (its 1st level classification); Hungarian on the other hand has Uralic as its 1st level classification but no further sub-classifications. Basque meanwhile is of no known family grouping: https://glottolog.org/resource/languoid/id/basq1248

Irish	Celtic	1	.544 (-)	.337 (-)	.831 (-)
Croatian	Balto-Slavic	2	.380 (.008)	.324 (.004)	.761 (.002)
Indonesian	Malayo-Polynesian	4	.275 (.033)	.303 (.027)	.763 (.023)
Italian	Italic	1	.362 (-)	.302 (-)	.767 (-)
French	Italic	2	.556 (.016)	.294 (.010)	.840 (.004)
Portuguese (Br.)	Italic	1	.497 (-)	.265 (-)	.826 (-)
Portuguese (Pt.)	Italic	3	.478 (.021)	.261 (.023)	.819 (.010)
Spanish	Italic	7	.359 (.024)	.239 (.018)	.770 (.016)
Hungarian	Uralic	3	.295 (.031)	.236 (.030)	.720 (.020)
Korean	Koreanic	3	.212 (.030)	.233 (.047)	.675 (.042)
Japanese	Japanesic	7	.272 (.058)	.210 (.044)	.711 (.040)
Basque	Unknown	1	.173 (-)	.174 (-)	.650 (-)

It is apparent from Table 3, for PUDDLE at least, that Mandarin Chinese stands out as the most successfully segmented after 10,000 utterances, by all three measures, whereas Basque is segmented least successfully. By examining the diphone systems of these two languages, we find that Mandarin has a smaller (n=351) set of diphones which are almost completely organised into those which do and do not occur next to a word boundary: 255 of those diphones have a boundary probability greater than 0.9; 87 have a probability less than 0.1, leaving just 9 in the intermediate zone. Basque, on the other hand, has a larger (n=503) set which is organised in a more variable way as concerns boundary juxtaposition: 88 have a boundary probability greater than 0.9, 179 have a boundary probability less than 0.1, meaning 236 have probabilities intermediate to those two thresholds. We propose that much of the performance difference between these two language may be ascribed to these two very different diphone systems, as PUDDLE accumulates a frequency list of boundary-occurring diphones as it encounters them. To further verify this proposal, we would need to undertake psycholinguistic work with infants to see if they can learn novel words positioned between diphones varying in their strength of boundary cue, and in languages varying in orderliness of diphone systems.

Another notable outcome is how Danish can be segmented relatively well, given the literature indicating that it is hard to segment because of its high ‘vocoid’ content (Basbøll, 2005). Vocoids are segments produced without vocal tract constriction: *i.e.* vowels, semi-vowels and non-lateral approximants. They are contrasted with ‘contoids’ which do involve vocal tract constriction: obstruents, nasals and lateral approximants. Vocoid-only utterances are not uncommon in Danish (*e.g. jeg er ude,* [jɑ ɑ u:ðə] ‘I am out’) and it has been suggested that contoids are stronger cues for segmentation than vocoids, thereby making Danish hard to segment (Nespor, Peña & Mehler, 2003; Bleses *et al,* 2008; Basbøll, 2012). Trecca and colleagues (in press) simulate a high vocoid artificial language only to find that human subjects do not show the expected difficulty with segmentation of such a speech stream. We remain agnostic but intrigued by this matter, and anticipate that vowel-consonant and vocoid-contoid annotation of the corpora will offer new insight in future work. We do note, however, that phonemic transcriptions and artificial languages offer an idealised version of the stimulus in which all units are equally and independently recognisable to the learner. It may be that the renowned difficulty of human speech recognition and learning in, for instance, Danish (Bleses, Basbøll & Vach*,* 2011; Schüppert, Hilton & Gooskens, 2016; Trecca, Bleses, Madsen & Christiansen, 2018) is a result of phonetic or suprasegmental factors, or the accumulatve effect of vocoid sequences. According to one observer, Danish is characterised by "an abundance of vowels, weak syllable codas, unstressed syllables without any vowel sound, and fairly inexpressive prosody"; perceptually, then, Danish is a "hard nut to crack" (Grønnum, 2003). In these experiments, phonemes and diphones are treated independently: a sequence of six contoids or vocoids is dealt with as any other sequence of six units, even if in reality they may be difficult to perceive, or unlicensed in the language.

The Germanic and Italic languages cluster around separately-similar levels of performance. This could simply be symptomatic of the high number of languages from these families in our set of corpora, but it does seem to indicate that the segmentation algorithms particularly favour the Germanic family – whether as a result of their design or the properties of this language family is open to debate. The Italic languages cluster to a fairly similar level of performance as well, though less successfully than the Germanic group. As explained above, our sample of corpora from CHILDES was opportunistic, though maximal, and therefore we did not control for language family, nor do we have sufficient representatives of families other than Germanic or Italic to come to strong conclusions. However, it does appear that language families group together, even if the performance of PUDDLE on Mandarin and Cantonese offers an obvious counter-example.

We also investigated how model performance varies over time – where the number of input utterances acts as a proxy for time (since we maintain the temporal order in the transcriptions) – by carrying out segmentation experiments for each of our nine models at every 1000 utterance increase in corpus size. We opt to use the token *F*-measure to illustrate change over time, since this is the strictest comparison with the gold standard. Figure 2 illustrates how token *F* changes for each segmentation model as they are exposed to the corpora in one thousand token increments, with boxplots summarising all scores. What becomes apparent from this visualisation is that the models on the whole *deteriorate* in token segmentation with increasing amounts of input, with the exception of PUDDLE which takes advantage of its knowledge accumulation device – its diphone and lexical inventories – and improves with increasing amounts of input. Similar improvements over time may be found in type and boundary *F*-measures.

FIGURE 2 APPROX HERE

Figure 2: Token *F*-measures for nine word segmentation models on 1000-utterance CHILDES corpus increments, with boxplots summarising scores as explained in Figure 1.

For the baseline and TP models, their purely probabilistic approaches do not benefit from exposure to increasing amounts of data. Instead their deteriorating performance may be accounted for by the increasing number of word types presented to the learner as corpus size increases, as shown in Figure 3, which prompts a higher number of errors in statistics-based approaches. DiBS, on the other hand, does have the opportunity to learn over time thanks to its training device which encounters utterance-delimited data in order to learn its set of probabilities. However, the serendipity involved in the diphones which will occur in utterance initial and final positions in any given language sample in fact means that improvements are empirically hard to come by even if in principle they are feasible with the DiBS architecture.

FIGURE 3 APPROX HERE

Figure 3: Type (vocabulary) size, type-token ratios, the proportion of one-word utterances, word boundary diphone entropy (H), the shape (α) of a Zipf-Mandelbrot model of diphone frequencies, and the mean (μ) number of phonemes per word in the 132 CHILDES corpora in 1000-utterance increments, with boxplots summarising values as explained in Figure 1.

In Figure 3 we show how the corpora change as they grow, in terms of our six chosen language properties – mean number of phonemes per word, mean tokens per utterance, type-token ratio, the proportion of one-word utterances, word boundary diphone entropy, and the Zipfian-ness of diphone distributions. Firstly we see that the average length of words in phonemes remains fairly stable as the corpora grow, with the high word length outlier points being the Korean and Turkish datasets. Utterance length tends to increase slowly with most corpora having utterance lengths lower than 10.

TTR rapidly falls away after the first 1000 utterances – that is, relative lexical diversity reduces over time. The proportion of one-word utterances is fairly stable throughout, but note that the outliers with high proportions are consistently the Japanese and Swedish corpora. This could be an idiosyncrasy of the speakers involved but given the consistency of the effect it could be culture or language specific; a matter for further investigation.

The entropy of the word boundary diphone system – a measure of uncertainty in the distribution of diphones at word boundaries – increases as the corpora grows, indicating that there is more certainty about the diphones occurring at word edges in the early stages, which gradually becomes more disordered as the vocabulary size grows. Meanwhile the shape of the Zipf-Mandelbrot model fit to diphone frequencies drops sharply, showing that the fit deteriorates as the corpora grow: that is, the frequency of diphones becomes decreasingly distributed in a Zipfian way.

Error analysis

As referred to above, it may be that segmentation of the speech stream does not have to be *word-perfect*, but could involve some under- or over-segmentation which is not necessarily harmful – for example, undersegmenting multi-word expressions or oversegmenting compound nouns. Here we undertake some small post-hoc analysis of the different models' segmentation of the opening ten lines of the English (U.K.) corpus, Lara (Rowland & Fletcher, 2006). In Table 4 we show the gold-standard transcription in orthographic and phonemic form, alongside outputs from our three target models. We select TP_MI to represent the TPs approach since it was the most successful of the three TP models in terms of token *F*.

Table 4: Segmentation of the first ten utterances from the Lara corpus by TP_MI, DiBS and PUDDLE; under-segmentation is marked by a tilde (~) and over-segmentation is marked by a reference mark (※).

Lara Corpus	Gold	TP_MI	DiBS	PUDDLE
that's the machine	ðats ðə məʃiːn	ðats~ðə~m ※əʃiːn	ðats ðə~mə ※ʃiːn	ðats ðə~məʃiːn
you can listen if you're good	juː kan lɪsən ɪf jɔː gʊd	juː~kan lɪs ※ ən~ɪf~jɔː~gʊd	juː~kan lɪsən~ɪf~jɔː~gʊd	juː kan lɪsən ɪf~jɔː gʊd
that	ðat	ðat	ðat	ðat

oh dear	əʊ diə	əʊ~diə	əʊ diə	əʊ diə
never mind	nɛvə maɪnd	n ※ɛvə~maɪnd	nɛvə~maɪnd	nɛvə maɪnd
grandma will wipe that off in a minute	gɹandmɑː wɪl waɪp ðat ɒf ɪn ɐ mɪnɪt	gɹandmɑː wɪl~waɪp ðat ɒf~ɪn~ɐ~mɪnɪt	gɹandmɑː wɪl waɪp ðat~ɒf~ɪn~ɐ~mɪnɪt	gɹandmɑː wɪl waɪp ðat ɒf ɪn ɐ~mɪnɪt
you do your jigsaw	juː duː jɔː dʒɪgsɔː	juː~duː~jɔː~dʒɪg & sɔː	juː~duː~jɔː~dʒɪgsɔː	juː duː jɔː dʒɪgsɔː
grandma will wipe it in a minute	gɹandmɑː wɪl waɪp ɪt ɪn ɐ mɪnɪt	gɹandmɑː wɪl~waɪp~ɪt~ɪn~ɐ~mɪnɪt	gɹandmɑː wɪl waɪp~ɪt~ɪn~ɐ~mɪnɪt	gɹandmɑː wɪl waɪ & p~ɪt ɪn ɐ~mɪnɪt
shall i go and get a cloth and wipe it	ʃal aɪ gəʊ and gɛt ɐ klɒθ and waɪp ɪt	ʃal~aɪ~gəʊ and~gɛt~ɐ~klɒθ~and~waɪp~ɪt	ʃal~aɪ~gəʊ and~gɛt~ɐ~klɒθ~and waɪp~ɪt	ʃal~aɪ gəʊ and gɛt ɐ~klɒθ and waɪ ※ p~ɪt
alright then	ɔːlɹaɪt ðɛn	ɔːlɹaɪt~ðɛn	ɔːlɹaɪt ðɛn	ɔːl ※ɹaɪt ðɛn

The differing levels of success of the three approaches may be seen in this small sample of their segmentation outputs. TP_MI's errors are mainly ones of undersegmentation, though there are some damaging oversegmentations of the words *machine, listen, never* and *jigsaw.* Some of the undersegmentations are more problematic than others: for instance, *you can* being run together is a feasible collocation, but *and get a cloth and wipe it* is not.

For DiBS there is again an oversegmentation of *machine,* which combines with undersegmentation of *the* to produce an utterance like, 'that's thema sheen'. Otherwise there are again undersegmentation sequences which produce extraordinary conglomerated strings: *that off in a minute, you do your jigsaw,* for example. PUDDLE is also guilty of undersegmentation, but omitting only one boundary at a time, in these utterances at least. For example there is *the_machine, if_you're, a_minute,* and *a_cloth.* These are reasonable collocations of determiners and nouns, or two grammatical items ('if you're'). Meanwhile the effect of the accumulating lexicon is seen in real word oversegmentations such as 'all right' for *alright* and 'why pit' for *wipe it.*

It seems then that our type, token and boundary evaluation measures do fairly represent the performance of each model. Examination of the severity of errors, however, does sometimes offer mitigation and indicate how the models fail or can be improved. Phillips and Pearl (2015) have proposed 'utility-based' evaluation metrics which take into account how useful the segmentations are in facilitating further language acquisition, or which tally with known prosodic patterns. With sufficient annotation of these corpora and the output of the segmentation models, utility-based evaluation metrics may reveal further performance differences between TPs, DiBS, PUDDLE – and any other approach for that matter – in further work.

The effect of language properties on word segmentation models

To account for the observed cross-linguistic variance in word segmentation performance, we turn to regression modelling of token *F*-measures for PUDDLE using our defined linguistic properties (mean word length, mean utterance length, type-token ratio, the proportion of one-word utterances, word boundary diphone entropy, and the Zipfian-ness of diphone distributions) as independent variables. We opt to use performance measures for PUDDLE as it is our best model, and we choose token rather than type or boundary *F* as the dependent variable because it is the strictest of the three evaluation methods.

At first we fit a linear regression model in R (R Core Team, 2018) with fixed effects: the six language properties listed above, instantiated as six variables (representing Zipfian fit as the alpha and chi-squared values from Zipf-Mandelbrot models), and scaled to values between 0 and 1. This model has an *R*-squared value of .252, meaning that it accounts for 25.2% of the variance in the performance measures. For this first model, the Akaike Information Criterion (AIC), a measure of goodness-of-fit for which lower values are better (Akaike, 1974), is -3055.

On the basis that we expect TTR to have a changing role over time, given that it decreases with increasing corpus size (Figure 3), we next fit a model with TTR as an interaction term with corpus size. This new model has an R^2 of .310 and an AIC of -3160, meaning it has more explanatory power than our initial model.

Having observed that there appears to be some kind of association between segmentation performance and language family (Table 3), we now introduce a typological factor. However, there are too few examples of most families in our sample, therefore to avoid overfitting we only add logical true/false variables for the Germanic and Italic families, ignoring the others. The addition of these two variables ('is Germanic' and 'is Italic') sees R^2 increase to .377 and AIC decrease to -3290.

Finally, recognising that there are likely to be idiosyncrasies in each corpus, whether through the selection of topics and vocabulary, or through speaker variation, we add an identifier for each of our 132 corpora as a random effect, based on literature showing such *mixed-effects* models to be preferable to ones in which speaker and corpus variation is averaged over (Baayen et al, 2008; Winter & Wieling, 2016). We used lme4 for R to fit this fourth model (Bates, Maechler, Bolker & Walker, 2015), which has a marginal pseudo-R^2 of .423 (taking only the fixed effects into account) and a conditional pseudo-R^2 of .905 (taking both fixed and random effects into account), calculated with the R library MuMIn (Bartoń, 2018). The mixed-effects model has an AIC of -4714 and is found to be significantly better than the other three models according to analyses of variance. A summary of the model is given in Table 5, showing coefficients, standard error and *t*-tests for each effect, mean-centred with jtools in R (Long, 2018).

Table 5: Mixed-effects model of PUDDLE token *F*, showing coefficients and standard errors for the fixed effects: mean word length (phonemes/token), mean utterance length (tokens/utterance), type-token ratio (TTR), number of utterances, proportion of one-word utterances (OWU proportion), word boundary diphone entropy (H(boundary diphones)), and the Zipfian-ness of diphone distributions (shape parameter α, goodness-of-fit χ^2), member of Germanic and Italic families.

Effect	Coefficient	Standard error
(Intercept)	0.47	0.02
phonemes/token	-0.02	0.01
tokens/utterance	0.03	0.01

TTR	0.37	0.03
n.utterances	0.25	0.02
TTR*n.utterances	0.20	0.01
OWU proportion	0.01	0.01
H(boundary diphones)	-0.10	0.01
Zipf-Mandelbrot α	-0.02	0.00
Zipf-Mandelbrot χ^2	0.00	0.00
is Germanic	0.04	0.02
is Italic	-0.07	0.02

As Table 5 shows, the strongest positive effects on segmentation performance are type-token ratio and the size of the corpus, an interaction between these two effects, and whether or not the language is of the Germanic family. Utterance length and the proportion of one-word utterances show small but positive effects on token *F.* The strongest negative effects are word boundary diphone entropy and whether the language is in the Italic family. The negative effect of increasing H(boundary diphones) confirms our hypothesis that higher entropy diphone systems are harder to learn as there is more uncertainty about the 'outcome', or the word boundary status, of each diphone. Note that these factors are analysed in order to account for inter-corpus differences in PUDDLE word segmentation performance rather than to explain why PUDDLE outperforms the other models. For instance, differences in OWU proportions do not correlate with per-corpus token *F* variation, but the facility to learn word boundaries from one-word utterances remains a crucial part of the PUDDLE algorithm. Or in other words, PUDDLE learns from the OWUs present in each corpus, rather than being affected by *how many* there are.

Conclusions & Future Work

We have evaluated three word segmentation models cross-linguistically and over time and found that (a) PUDDLE performs best overall; (b) there is wide linguistic variation with differences of 70% between minimum and maximum token *F* scores; (c) PUDDLE is the only model to show improvement over time. We went some way to accounting for cross-linguistic variation using six linguistic properties: word length, utterance length, type-token ratio, the proportion of one-word utterances, word boundary diphone entropy, and parameters from Zipf-Mandelbrot models of diphone distributions. These properties encode qualitative features of the child's linguistic input. How often words are presented 'as is' in a single utterance (OWU proportion), how many boundaries need to be identified per phoneme and per utterance (word and utterance lengths), and how much lexical repetition there is in the input (type-token ratio). The orderliness of the diphone system at word edges is represented by its entropy, and Zipfian parameters indicate how skewed the distribution of diphones is, our hypothesis being that the more Zipf-like the better for learnability though this was not upheld by the study.

This work has focused on the computational models of word segmentation as described in the literature and implemented in the `wordseg` toolkit. We have taken the perspective that the interaction between each model and properties of the input languages underpins the performance measures reported above. However, it is not necessarily the case that all models hold equal cognitive validity. Firstly, even though phonemes were the basic unit used here, TPs in its original formulation actually opts for syllables as the basic unit. The correct choice of unit is a debate we do not directly address here, primarily because syllabification resources are not available for all the languages in our sample. Therefore the development of such resources and the comparison of phonemes and syllables as the basic unit remains a matter for future work.

Moreover, as has been discussed, DiBS either infers diphone probabilities from segmented training data or bootstraps them from a supplied lexicon. We did not experiment with the latter setting, but instead tested the utterance-delimited training mode only, on the grounds that it is ecologically preferable to a fully word-delimited training setting. Meanwhile PUDDLE recognises the importance of one-word utterances as a means to build and confirm a lexicon which in turn aids subsequent speech segmentation. This, combined with its inventory of word-edge diphones, allows PUDDLE to out-perform the other models and furthermore show improvement over time. However, this does not mean that PUDDLE is a total representation of how infants learn to segment words. Rather, it is the best performing of the statistically-based models we test here.

Using mixed-effects regression modelling with our six language properties plus corpus size and language families as fixed effects, and the corpora as random effects, we find that we can account for 91% of the observed variation in PUDDLE's performance across corpora. But since only 42% of the explanation is assigned to the fixed effects it is clear that the majority of variance remains unexplained other than that it is on a per-corpus basis. We surmise that further explanation for variation in word segmentation success must come from speaker idiosyncrasies, topic selection and language factors other than the ones we have examined in this work. These are likely to be linguistic properties beyond the immediate scope of orthographic transcriptions – for instance, prosodic, multimodal and semantic cues which might combine with the lexico-phonological features we already consider.

It is likely that distributional information is one of many cues infants attend to in early language input, and what we learn from these experiments is that this strategy can at least offer a foothold in breaking down the speech stream. As the performance of this approach asymptotes, the learner needs to take account of other information sources in order to continue learning to segment words – for instance speech rhythm, as has been shown to be useful in the segmentation task (Saksida *et al* 2017). Such a model might then mimic the switch from statistical to stress based cues which has been observed (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). We do not propose that the switch is a matter of one or the other, with prosody ignored until it is needed – we know for instance that neonatals are alert to suprasegmental information (Moon, Panneton Cooper & Fifer, 1993; Mampe, Friederici, Christophe & Wermke, 2009; May, Byers-Heinlein, Gervain & Werker, 2011) – more that all available cues are attended to and processed in different ways at different developmental stages. A 'lexical shift' in processing skills towards the end of the first year has been observed in neurological work, for example, though its onset varies by individual (Kidd, Junge, Spokes, Morrison & Cutler, 2018). Word segmentation models accounting for individual differences is another area for exploration.

We already highlighted several potential areas for future work. These include but are not limited to: testing the models on additional CHILDES corpora and languages as they become available, and accessing multimodal files in order to incorporate suprasegmental and paralinguistic cues which might aid in word segmentation – at least prosody, gesture and semantics would be psycholinguistically grounded information sources to make available to the learner. Furthermore we would like to test whether the predicted order of acquisition of words from a given corpus correlates with observed orders in child language corpora (Braginsky *et al*, 2018), polysemous semantic networks learned from that input (Amatuni & Bergelson, 2017; Casas *et al*, 2018), how it interacts with the acquisition of morphology and syntax (Frank, Keller & Goldwater, 2013), and phonotactic knowledge (Linzen & Gallagher, 2017). In addition, how does the acquisition of a lexicon through learning to segment words facilitate further learning in turn? Since the precision and recall metrics are designed for single-point rather than time series reporting, we will need adapted or alternative metrics to address this pertinent question (Phillips & Pearl, 2015). We would also like to return to the matter of the decay parameter in PUDDLE, as humans are not perfect memorisers, and the decay parameter in these experiments was set to zero; instead we should find a set of values for the parameter which best reflects observed rates of forgetting.

Overall it remains a challenge to architects of word segmentation models to incorporate observations from human performance, especially relating to memory and resource limitations (Frank *et al*, 2010), prediction (Çöltekin, 2017), the tension between learnability and confusability (Dautriche *et al*, 2017), sensitivity to prosody and word location within prosodic structures (Graf Estes & Hurley, 2013; Butler & Frota, 2018), and a special role for input strategies such as reduplication (Ota & Skarabela, 2018), vowel harmony (Mintz *et al*, 2018) and sonority sequencing (Ettlinger, Finn & Kam, 2012). In addition we see promise in looking at high-frequency sound sequences which are in fact non-words, as these have been shown to be stored by infants in a 'protolexicon' along with real words (Ngon *et al*, 2013). If we can collaborate with native speakers of the many languages in our dataset, we can in future explore the role of non-words in the input and in segmentation. There are also extensions of this work in the area of second language acquisition, where it has been shown that learners may acquire a second phonological system in different ways depending on exposure to formal instruction or not (Shoemaker & Wauquier, 2019).

Finally, we note that the data are at the same time impoverished and idealised – impoverished because we have access to minimal samples of the linguistic input infants are exposed to, and idealised because we assume that the speech is perfectly produced, perceived and analysed. Ideally, we would have access to more data, and be able to test model predictions with research on vocabulary acquisition over time. The availability for research of more naturalistic and comprehensive child language corpora can only improve and accelerate our understanding of acquisition in general and word segmentation specifically (Tamis-LeMonda *et al*, 2017; Chin *et al*, 2018; Bergelson *et al*, 2019).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19, 716-723.

Amatuni, A., & Bergelson, E. (2017). Semantic networks generated from early linguistic input. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H., Davidson, D., & Bates, D. (2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language,* 59, 390-412.

Bartoń, K. (2018). MuMIn: Multi-Model Inference. R package version 1.42.1. https://cran.r-project.org/package=MuMIn

Basbøll, H. (2005). *The phonology of Danish.* Oxford: Oxford University Press.

Basbøll, H. (2012). Monosyllables and prosody: The sonority syllable model meets the word. In: T. Stolz, N. Nau, & C. Stroh (Eds.), *Studia typologica: Monosyllables: From phonology to typology.* Berlin: De Gruyter.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1-48.

Bentz, C., Alikaniotis, D., Cysouw, M., & i Cancho, R. F. (2017). The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19, 275.

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science,* 22, e12715.

Bernard, M. (2018). phonemizer-1.0. (http://doi.org/10.5281/zenodo.1045826)

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

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media.

Bleses, D. & Basbøll, H. (2004). The Danish sound structure: Implication for language acquisition in normal and hearing impaired populations. In: E. Schmidt, U. Mikkelsen, I. Post, J. B. Simonsen, & K. Frensgaard (Eds.), *Brain, hearing and learning. 20th Danavox Symposium.* Copenhagen: Holmen Center Tryk.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T., & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language,* 35, 619-650.

Bleses, D., Basbøll, H. & Vach, W. (2011). Is Danish difficult to acquire? Evidence from Nordic past-tense studies. *Language and Cognitive Processes,* 26, 1193-1231.

Bortfield, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science,* 16, 298-304.

Boruta, L., Peperkamp, S., Crabbé, B., & Dupoux, E. (2011). Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*.

Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. (2018). Consistency and variability in word learning across languages. PsyArXiv. doi: 10.31234/osf.io/cg6ah

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93-125.

Butler, J. & Frota, S. (2018). Emerging word segmentation abilities in European Portuguese-learning infants: New evidence for the rhythmic unit and the edge factor. *Journal of Child Language,* 45, 1294-1308.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33, 111-153.

Casas, B., Català, N., Ferrer-i-Cancho, R., Hernández-Fernández, A., & Baixeries, J. (2018). The polysemy of the words that children learn over time. *Interaction Studies,* 19, 389-426.

Çöltekin, Ç. (2017). Using predictability for lexical segmentation. *Cognitive Science*, 41, 1988-2021.

Chin, I., Goodwin, M., Vosoughi, S., Roy, D., & Naigles, L. (2018). Dense home-based recordings reveal typical and atypical development of tense/aspect in a child with delayed language development. *Journal of Child Language*, 45, 1-34.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: evidence from word segmentation. *Cognition,* 96, 233-262.

Curtin, S. (2009). Twelve-month-olds learn novel word-object pairs differing only in stress pattern. *Journal of Child Language,* 36, 1157-1165.

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.

Dahan, D., & Brent, M. (1999). An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, 128, 165-185.

Daland, R., & Pierrehumbert, J. (2011). Learning diphone-based segmentation. *Cognitive Science*
35, 119-155.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition,* 163, 128-145.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language*, 64, 199-210.

Ettlinger, M., Finn, A., & Kam, C. H. (2012). The effect of sonority on word segmentation: evidence for the use of a phonological universal. *Cognitive Science*, 36, 655-673.

Evert, S. (2004). A simple LNRE model for random character sequences. In *Proceedings of JADT*.

Evert, S., & Baroni, M. (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions.*

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment? In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*.

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117, 107-125.

Frank, S., Keller, F., & Goldwater, S. (2013). Exploring the utility of joint morphological and syntactic learning from child-directed speech. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Friederici, A., & Wessels, J. (1993). Phonotactic knowledge of word boundaries and its use in infant speech-perception. *Perception & Psychophysics*, 54, 287-295.

Gambell, T., & Yang, C. (2005). Word segmentation: Quick but not dirty. (Manuscript, Yale University. http://www.ling.upenn.edu/~ycharles/papers.html)

Gervain, J., & Guevara Erra, R. G. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, 125, 263-287.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, 112, 21–54.

Goodsitt, J. V., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language,* 20, 229-252.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy,* 18, 797-824.

Grønnum, N. (2003). Why are the Danes so hard to understand? In: H. Galberg Jacobsen, D. Bleses, T. O. Madsen & P. Thomsen (Eds.), *Take Danish – for instance: Linguistic studies in honour of Hans Basbøll presented on the occasion of his 60th birthday 12 July 2003.* Odense: University Press of Southern Denmark.

Hallé, P. A., & de Boysson-Bardies, B. (1994). Emergence of an early receptive lexicon: infants' recognition of words. *Infant Behavior and Development,* 17, 119-129.

Hammarström, H., Forkel, R., & Haspelmath, M. (2018). Glottolog 3.3, https://glottolog.org (accessed 2019-01-09).

Hartman, K., Bernstein Ratner, N., & Newman, R. (2017). Infant-direct speech (IDS) vowel clarity and child language outcomes. *Journal of Child Language,* 44, 1140-1162.

Hay, J., Pelucchi, B., Estes, K., & Saffran, J. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology,* 63, 93-106.

Hendrickson, A., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition,* 189, 11-22.

Hockema, S. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development*, 2, 119-146.

James, W. (1890). *The Principles of Psychology, Volume 1*. New York: Henry Holt and Company.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44, 548-567.

Johnson, E., & Tyler, M. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, 13, 339-345.

Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress patterns of English words. *Child Development,* 64, 675-687.

Jusczyk, P. W., Luce, P., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630-645.

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in infant speech segmentation: Achieving the lexical shift. *Infancy,* 23, 770-794.

Krogh, L., Vlach, H. A., & Johnson, S. P. (2012). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology,* 3, 598.

Kurumada, C., Meylan, S., & Frank, M. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127, 439-453.

Ladefoged, P. (2003). Commentary: some thoughts on syllables–an old-fashioned interlude. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI.* Cambridge: Cambridge University Press.

Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. In *Proceedings of INTERSPEECH.*

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

Linzen, T., & Gallagher, G. (2017). Rapid generalization in phonotactic learning. *Laboratory Phonology*, 8, 1-32.

Long, J. (2018). jtools: Analysis and presentation of social scientific data. R package version 1.1.1 https://cran.r-project.org/package=jtools

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language acquisition. volume 1: Syntax and semantics*. Hillsdale, NJ: Lawrence Erlbaum.

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

Mampe, B., Friederici, A. D., Christophe, A., & Wermke K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology,* 15, 1-4.

May, L., Byers-Heinlein, K., Gervain, J., & Werker J. F. (2011). Language and the newborn brain: Does prenatal language experience shape the neonate neural response to speech? *Frontiers in Psychology,* 2, 222.

Mani, N., & Pätzold, W. (2016). Sixteen-month-old infants' segment words from infant- and adult-directed speech. *Language Learning and Development*, 12, 499-508.

Mattys, S., & Jusczyk, P. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91-121.

Mattys, S., White, L., & Melhorn, J. (2005). Integration of multiple segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 134, 477-500.

McCauley, S., Monaghan, P., & Christiansen, M. (2015). Language emergence in development. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence*. Oxford: Blackwell.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20, 298-305.

Mintz, T., Walker, R., Welday, A., & Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition*, 171, 95-107.

Monaghan, P., & Christiansen, M. (2010). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37, 545-564.

Moon, C., Panneton Cooper, R., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavioral Development,* 16, 495-500.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e Linguaggio,* 2, 221-247.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science,* 16, 24-34.

Ota, M., & Skarabela, B. (2018). Reduplication facilitates early word segmentation. *Journal of Child Language*, 45, 204-218.

Pelucchi, B., Hay, J., & Saffran, J. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition,* 113, 244-247.

Pelucchi, B., Hay, J., & Saffran, J. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development,* 80, 674-685.

Phillips, L. (2015). The role of empirical evidence in modeling speech segmentation. (Dissertation, University of California, Irvine. http://eric.ed.gov/?id=ED568017)

Phillips, L., & Pearl, L. (2015). Utility-based evaluation metrics for models of language acquisition: a look at speech segmentation. In *Proceedings of the Sixth Workshop on Cognitive Modeling and Computational Linguistics*.

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. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171, 130-150.

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.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, 20, e12390.

Schüppert, A., Hilton, N. H., & Gooskens, C. (2016). Why is Danish so difficult to understand for fellow Scandinavians? *Speech Communication,* 79, 47-60.

Shoemaker, E., & Wauquier, S. (2019). The emergence of speech segmentation in adult L2 learners of French. *Language, Interaction and Acquisition*, 10, 22-44.

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & Van Heuven, W. (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language,* 175, 111-122.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology, 50, 86-132.

Tamis-LeMonda, C., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. (2017). Power in methods: language to infants in structured and naturalistic contexts. *Developmental Science,* 20, e12456.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology,* 39, 706-716.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4, 156-163.

Trecca, F., Bleses, D., Madsen, T. O., & Christiansen, M. H. (2018). Does sound structure affect word learning? An eye-tracking study of Danish learning toddlers. *Journal of Experimental Child Psychology,* 167, 180-203.

Trecca, F., McCauley, S. M., Andersen, S. R., Bleses, D., Basbøll, H., Højen, A., Madsen, T. O., Ribu, I. S. B., & Christiansen, M. H. (in press). Segmentation of highly vocalic speech via statistical learning: Initial results from Danish, Norwegian, and English. *Language Learning.*

Vihman, M., dePaolis, R., Nakai, S., & Hallé P. A. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language,* 50, 336-353.

Winter, B., & Wieling, M. (2016). How to analyze language change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution,* 1, 7-18.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley

Ziegler, J. C. and U. Goswami (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin* 131, 3-29.