Words in Puddles of Sound:

Modelling Psycholinguistic Effects in Speech Segmentation

Padraic Monaghan[1] *& Morten H. Christiansen*[2]

[1]Department of Psychology and Centre for Research in Human Development and Learning, Lancaster University, Lancaster UK

[2]Cornell University, Ithaca NY, USA

Accepted for publication in Journal of Child Language

Corresponding author:

Padraic Monaghan

Department of Psychology

Lancaster University

Lancaster

LA1 4YF

UK

E-mail: p.monaghan@lancaster.ac.uk

Tel: +44 1524 593813

Fax: +44 1524 593744

Abstract

There are numerous models of how speech segmentation may proceed in infants acquiring their first language. We present a framework for considering the relative merits and limitations of these various approaches. We then present a model of speech segmentation that aims to reveal important sources of information for speech segmentation, and capture psycholinguistic constraints on children's language perception. The model constructs a lexicon based on information about utterance boundaries and deduces phonotactic constraints from the discovered lexicon. Compared to other models of speech segmentation, our model performs well in terms of accuracy, computational tractability, and the number of components of the model. Finally, our model also reflects psycholinguistic effects of language learning, in terms of the early advantage for segmentation provided by the child's name, and by revealing the overlap in usefulness of information for segmentation and for grammatical categorisation of the language.

Introduction

The speech that infants hear is generally produced in a continuous stream, without pauses that reliably indicate where words begin and end. Indeed, if pauses do occur, then this can be at misleading points in speech, occurring within words before consonants with long voice onsets (Slis, 1970), though pauses are also frequent between phrases in speech (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). The problem of speech segmentation has therefore been characterised as words occurring in a 'sea of sound' (Saffran, 2001) from which lexical items have to be identified and extracted. Consequently, an array of subtle, interacting, probabilistic indicators to word boundaries have been proposed as cues that assist in solving the segmentation problem, including cues such as lexical stress and prosodic patterns across utterances (Curtin, Mintz, & Christiansen, 2005; Cutler & Carter, 1987; Johnson & Juszcyk, 2001), transitional probabilities between syllables (Saffran, Aslin, & Newport, 1996), and phonotactic constraints between phonemes (Hockema, 2006; Mattys, White, & Melhorn, 2005).

Several computational models have been proposed to account for the developmental processes involved in early speech segmentation. Some of these models take as input raw speech, and such approaches have produced up to 54% accuracy on very small corpora (e.g., Roy & Pentland, 2002). An alternative approach is to take as input unsegmented phonological transcriptions of speech (e.g., Batchelder, 2002; Brent, 1999; Brent & Cartwright, 1996; Frank, Goldwater, Mansinghka, Griffiths, & Tenenbaum, 2007). These latter models considerably simplify the complexities of the raw speech input in identifying phonemes or phoneme features, but they do highlight the potential statistical sources of information useful for reflecting word boundaries in child-directed speech (CDS), and have been successfully related to psycholinguistic studies of children's language acquisition.

Previous developmental models of speech segmentation differ substantially across a umber of parameters, including whether the model builds a lexicon, segments words by

clustering smaller units or breaking down larger units, or incorporates external constraints on performance (see Batchelder, 2002, for a review). From a developmental psycholinguistics perspective, it is not clear which model(s) should be preferred. In this paper, we therefore first propose a set of psychologically-motivated criteria for assessing developmental models of speech segmentation before presenting our own computational model.

Criteria for Assessing Developmental Models of Speech Segmentation

Precision and recall

Previous work on speech segmentation has quite rightly focused on assessing computational models in terms of their ability to correctly segment a corpus into words, as determined by an objective parse of the speech. The best performance of developmental models of speech segmentation appear to be converging to approximately three quarters of words in CDS corpora. However, it is unclear what level of segmentation performance best reflects the child's ability. Nonetheless, all else being equal, a model that shows it can exploit information in a way that maximises the correct segmentation of a CDS corpus is to be preferred. When all else is not equal, then roughly similar performance to comparable models provides a useful benchmark level.

Computational tractability

The second criterion concerns the plausibility of the model as a reflection of the cognitive processing of the infant learning the language. The model should be computationally tractable – memory limitations should be observed, and optimal learning should not be assumed. Critical for computational tractability is whether the model is incremental or whether the whole corpus must be considered in segmenting a particular utterance. Thus, an incremental model – in which the segmentation of a target utterance depends only on what has preceded the utterance in the child's exposure – is to be preferred. However, there may be incremental

approximations of models that process the whole corpus, and thus preferring an incremental model as a decision criterion requires proof that a 'batch' model would not operate effectively in an incremental mode. Moreover, everything else being equal, a model that requires small memory capacity, and limited search and computational resources, is preferable. Models that require close approximation to optimal learning conditions – where all the input can be stored and accessed simultaneously – should be rejected as models of the infant's cognitive process, though they may have substantial value in reflecting the potential information present in the child's language input.

External components

Some models may include external components that do not emerge from the basic processing principles of that model. As an example, Frank et al. (2007) and Brent and Cartwright (1996) use a vowel constraint, whereby a candidate lexical item must contain a vowel to be considered. For these specific models, this qualifies as an external constraint, as it is a constraint applied to the model, and which cannot be inferred from the language exposure alone. We suggest that, all else being equal, a model with few external components is to be preferred for reasons of parsimony.

Psycholinguistic features

Perhaps the most important criterion of all for the assessment of the models is the extent to which they can reflect psycholinguistic observations of the infant learning to segment speech. For example, Brent (1999) demonstrated that certain predictions of a computational model of segmentation can be tested in experimental studies of language learning (e.g., Dahan & Brent, 1999), and Perruchet and Vinter (1998) explicitly tested the artificial languages of Saffran et al. (1996) to determine whether a chunking strategy, elicited by transitional probabilities, could account for participants' segmentation performance based on these materials. The particular psycholinguistic effects we feature for our modelling are reported in the next section, where we outline the basic principles of our model's functioning.

Sources of Information in Child-Directed Speech

Our model aims to advance on previous models with respect to these criteria for assessing developmental models of speech segmentation, though there is a large degree of overlap between our approach and previous models of speech processing. One advantage is that we provide a model that is computationally tractable, in that it does not assume a large lexicon, nor does it require multiple, competing decisions about the match between the lexicon and the utterance string. Furthermore, the model is incremental in its processing of utterances. Along with the Perruchet and Vinter (1998) PARSER model, the memory resources and computational requirements are minimal. However, unlike PARSER, our model can process at the phoneme level, and does not require the syllable structure to be provided to the model. The second advantage we claim for our approach is that it does not require additional constraints that lie outwith the model's discovery of the lexicon itself. The third advantage of our modelling approach is an attempt to draw together the modelling approach with features of infant speech processing that highlights what may be the important aspects of CDS that are formative for language learning (though see also Batchelder 2002). In particular, we focus on two features of CDS that we believe are critical for language learning: utterance boundaries and the interspersal of high frequency words in speech.

Utterance boundaries provide a rich source of information about word boundaries, represented either by physical pauses in speech, or indicated by alternations between conversational partners. Though MacWhinney and Snow (1985) estimated that only about one in seven words were spoken in isolation in CDS, this still presents a potentially large number of words that can then be bootstrapped into segmenting multi-word utterances. From the English CDS section of the CHILDES corpus, of 1,369,574 utterances, 358,397 (26.2%) are single word utterances; Table 1 shows proportions of utterances of various lengths in words. Relying solely on utterance boundaries to indicate word boundaries, however, is likely

to be insufficient for infant speech segmentation (Brent & Cartwright, 1996). First, though a large proportion of utterances consist of a single word, the majority of utterances are multi-word sequences and there are no proposed methods for distinguishing between single- and multi-word utterances (Christophe, Dupoux, Bertoncini, & Mehler, 1994). Second, many words very rarely occur as single-word utterances, such as determiners (e.g., 'the' only occurs 129 times as a single word utterance in the combined CHILDES corpus of English CDS).

< Insert Table 1 around here >

Although highly frequent function words seldom occur as single word utterances, other high frequency words may occur in isolation a substantial number of times. Proper names, for instance, can occur frequently as single-word utterances in CDS, and have been proposed to be important for assisting the learning of other words from the child's speech input. In the set of corpora we use for the analyses in this paper, the child's own name occurred as a single-word utterance a total of 1.3% of all utterances in the combined corpora. Importantly, though, as much as 23.7% of the occurrences of the proper name were in a single-word utterance.

But what contribution do utterance boundaries make alongside the wealth of other cues to indicate word boundaries available in speech? Though accurate speech segmentation clearly does not involve processing each utterance as a separate lexical item, this does not preclude the possibility that learning to segment speech may at least be facilitated by such information. Several models of speech segmentation have included utterance boundary information as input to the model (Aslin, Woodward, LaMendola & Bever, 1996; Batchelder, 2002; Brent, 1999; Brent & Cartwright, 1996; Christiansen, Allen, & Seidenberg, 1998), whereas other models incorporate it as an upper bound on the possible length of a candidate word (Perruchet & Vinter, 1998).

Our model utilises utterance boundaries to determine, in an incremental fashion, word boundaries in continuous speech, we term this the 'Phonotactics from Utterances Determine Distributional Lexical Elements' (or PUDDLE) model of speech segmentation. The PUDDLE

model initially treats each utterance as a lexical item, but breaks up longer utterances into shorter lexical items if another stored lexical item is a part of the longer utterance. Indeed, Dahan and Brent (1999) showed that, for adults listening to an artificial language, a novel utterance will be processed as a lexical item providing it contains no known words. The segmented sections of the longer utterance are then each entered as separate lexical items.

However, matching utterances within other utterances is not sufficient for a model of segmentation, as short, frequently occurring utterances are likely to be segmented within larger word-level chunks resulting in an over-segmentation of words into their segmental phonology. As an example, given the utterances: ‘oh’, and ‘no’, the unconstrained model will store ‘oh’ as a candidate lexical item, and then divide up ‘no’ into ‘n’ and ‘o’, as, in terms of their phonological transcription, the ‘o’ matches the stored utterance ‘oh’. Then, all future occurrences of utterances containing ‘n’ will be divided, resulting eventually in a set of lexical candidates that are the individual phonemes of English. To overcome such over-segemtation, our model incorporates a boundary constraint derived from its lexicon (as described below).

There were several, related aims to our computational model of segmentation in terms of connecting with the developmental literature on language learning. First, we wanted to indicate that single-word utterances are identifiable in speech, and can be extracted as lexical items from CDS corpora. Second, we wanted to explore which words emerge as those earliest identified, and which are consequently the most useful indicators of word boundaries. If a small set of frequent words can be accurately identified by the model, then these may be useful for carving up the rest of the speech stream into its constituent words, just as frequent words are useful for determining grammatical categories of the content words that surround them (Monaghan, Christiansen, & Chater, 2007). In this respect, too, we wanted to determine whether the child’s name is one of these early-identified words. Third, we wanted to plot the model’s discovery of words over time. Children learn language in an item-based manner

where frequently co-occurring words are initially processed as single words (MacWhinney, 1982; Tomasello, 2002), and only later are they distinguished into their constituents (see also Bannard & Matthews, 2008, for an empirical demonstration of this phenomenon).

We now present the PUDDLE model of speech segmentation, and report its performance on 6 corpora of English CDS. Testing the model on several CDS corpora presents an advance on previous models of speech segmentation that have typically focused on a single corpus (e.g., models reviewed in Brent, 1999), and provides insight into the generalisability of the model's performance across corpora, as well as highlighting distinctive properties of CDS in terms of their influence on speech segmentation performance, such as the use of proper nouns.

The PUDDLE Model of Speech Segmentation

Method

Algorithm. The model has two components: a lexicon, and a list of beginning and ending phoneme pairs, generated from the lexicon. The model begins by inputting the first utterance into the lexicon. The model searches through the current utterance starting at the first phoneme, and testing whether there is a match with any of the stored lexical items. If there is a match then the word is extracted, the phonemes occurring before the matched word are taken to constitute a new lexical item, and search for the next lexical item in the utterance recommences at the first phoneme in the utterance following the matched word. If there is no match at a particular phoneme position, then the model proceeds to the next phoneme in the utterance string, until the end of the utterance is reached. If the end of the utterance is reached without a match, then the phonemes following the last match of a word in the utterance are taken to be a new lexical item. The next utterance is then presented to the model.

< Insert Figure 1 around here >

As an example, consider the set of utterances ‘kitty’, ‘that’s right kitty yes’, ‘look kitty’, illustrated in Figure 1. The model will begin with the /k/ in the first utterance ‘kitty’. The lexicon is empty, so there are no matches, and the model will move on to consider /I/ from the first utterance. There is again no match, and so the model will proceed through to the end of the utterance with no matches and will code the entire utterance – in this case the string ‘kitty’ – as a lexical item. At the end of processing the first utterance, then, there is one item in the lexicon. Then the model proceeds to the second utterance, and attempts to match any of the lexical items with each phoneme in turn. There is just one match: ‘kitty’ matches at the /k/, and the string preceding the match – ‘that’s right’ – will be entered into the lexicon. Then, for the remaining phonemes in the second utterance, comprising the word ‘yes’, the model will attempt to match with the set of lexical items starting at each phoneme in turn. There are no matches, and so ‘yes’ will then be entered as a new lexical item. So, at the end of the second utterance there are three candidate lexical items. For the third utterance, the model will attempt to match all the lexical items ‘kitty’, ‘that’s right’, and ‘yes’ at each phoneme position. Once again, there is only one match at the second /k/, and so ‘look’ will also be entered as a new lexical item. (Note that utterances and lexical items are encoded as phoneme sequences, the terms in speech-marks and the transcriptions in Figure 1, indicate a short-hand version of these phoneme sequences for ease of interpretation.)

Each item in the model’s lexicon has associated with it an activity level, as in the PARSER model (Perruchet & Vinter, 1998). Each time a word is matched in an utterance its activity increases by 1, as shown in Figure 1 for the word ‘kitty’ when matched in the second utterance. For new lexical items, activity is initially set at 1. To simulate forgetting of the lexical items, a decay parameter can be used such that the activity of every lexical item reduced by a set amount each time a new utterance was presented. This has the effect of long utterances that are rarely repeated dropping out of the lexicon, but words that occur frequently maintaining a high activity level. Pilot studies indicated that setting the decay rate too high

resulted in a very small lexicon, and consequently under-segmentation of the corpus, hence precision was high but recall was low. In the following simulations, we report the results when decay is 0, indicating the model's performance when the learning capacity of word items was high.

A further parameter that influences the model's performance is the order in which the lexical items are searched for matches. We assume that the lexical items most available to be matched to input speech are those that occur with the highest frequency of identification in the child's previous exposure, and so we sorted the candidate lexicon according to the activity of each lexical item.

To reduce over-segmentation, phonotactic information about legal word boundaries was derived from the model's lexicon and used as a boundary constraint. Once a word produced a match in the utterance, the match was processed only if the phonemes around the matched segment were represented already within the lexicon as possible word endings or word beginnings. We implemented this by requiring that the two phonemes preceding the matched segment ended one of the candidate words in the lexicon and the two phonemes succeeding the matched segment began one of the candidate words. If the lexical item was shorter than two phonemes in length, then it did not contribute to the beginnings and endings list. Figure 1 illustrates that a list of all the beginning and ending phoneme pairs is constructed from the lexicon. Listeners are sensitive to whether pairs of phonemes are likely to occur within or across word boundaries (Mattys et al., 2005) and the distributions of within- and between-word phoneme bigrams is potentially valuable information for speech segmentation (Hockema, 2006). This constraint was important in order to prevent individual phonemes becoming candidate lexical items. In the example above, 'kitty' would only be matched in 'that's right kitty yes' if the last two phonemes of 'right' and the first two phonemes of 'yes' ended and began words in the lexicon, respectively. If there had been no input prior to the first utterance in this example then all three utterances would have been entered as lexical

items, and the beginnings and endings of these utterances only would be listed as potential word boundaries.

Corpus preparation. We selected six English CDS corpora from the CHILDES database (MacWhinney, 2000): Eve (Brown, 1973), Peter (Bloom, Hood, & Lightbown, 1974), Naomi (Sachs, 1983), Nina (Suppes, 1974), Anne and Aran (Theakston, Lieven, Pine, & Rowland, 2001). We only included speech spoken in the presence of children aged 2;6 or younger, and only adult speech was included. The corpora are orthographically transcribed in the CHILDES database, including indicators of speech pauses in the transcription. Pauses and changes in speaker were encoded as utterance boundaries. The numbers of utterances, words, and phonemes in each corpus are shown in Table 2.

< Insert Table 2 around here >

To generate the spoken form of the speech, we streamed the orthographic transcription through the Festival speech synthesiser (Black, Clark, Richmond, King, & Zen, 1990) which produced a sequence of phonemes for each utterance, together with a separate transcription that also included objective marking of which phonemes were generated for each word. This method of phonological transcription has the advantage that some phoneme variation according to part-of-speech context was encoded within the corpus, for instance, ‘a’ was pronounced either as /eI/ as a noun and /@/ when used as a determiner, similarly, ‘uses’ was pronounced with a /z/ as a verb and /s/ as a noun. The resulting input is therefore closer to the actual speech that children hear than what was used in most previous simulations of speech segmentation (e.g., Batchhelder, 2002; Brent, 1999; Brent & Cartwright, 1996; Christiansen et al., 1998; Hockema, 2006; Venkataraman, 2001), in which the same citation form (taken from a pronunciation dictionary) is used every time a word occurs independent of its context (though see Aslin et al., 1996, for a similar approach). Also, influences of lexical stress on vowel pronunciation were also encoded by Festival in the speech, so that when unstressed, vowels were often realised as schwa (see, e.g., Gerken, 1996).

Scoring. The model's performance was measured on blocks of 1000 utterances. The model's performance was scored on-line as the model proceeded through the corpus, so the model's performance was determined on portions of the corpus that it had not yet been exposed to. The model's segmentation was compared to the segmentation that reflected the orthographic transcription into words from the original corpus. We computed true positives, false positives, and false negatives in the model's segmentation. True positives were words that were correctly segmented by the model – a word boundary occurred immediately before and after the word but with no incorrect boundaries in between. False positives were sequences segmented by the model that did not match to individual words in the festival segmentation. False negatives were words in the Festival segmentation that were not correctly segmented by the model. To quantify the performance of the model we used the complementary measures of *precision* and *recall*, which have been used as conservative measures of model performance in previous research (e.g., Batchelder, 2002; Brent & Cartwright, 1996; Christiansen et al., 1998; Hockema, 2006; Venkataraman, 2001). Precision was computed as true positives divided by the sum of true positives and false positives. Recall was computed as true positives divided by the sum of true positives and false negatives. Thus, precision provides a measure of how many of the words that the model found are actual words, whereas recall indicates how many of the words in the corpus the model was able to find.

As a baseline, we created a 'word-length model' (e.g., Brent & Cartwright, 1996; Christiansen et al., 1998) that randomly inserted word boundaries into the speech stream given the correct number of word boundaries found across the whole corpus. Note that this baseline provides information about how many words there are in the corpus but not where the boundaries occur, so it is likely to perform better than a truly random baseline that lacks this information.

Results and Discussion

The model's performance was assessed for each 1000 utterance block in each corpus, up to the first 10,000 utterances had been processed. For corpora smaller than 10,000 utterances, performance for the final block of 1,000 utterances was repeatedly reported. Figure 2 reports the model's segmentation performance on each corpus with zero decay, compared to the word length segmentation baseline. At the 10,000 utterance block, the improvement over baseline performance was significant for both precision, $t(5) = 71.25$, $p < .0001$, and recall, $t(5) = 61.98$, $p < .0001$. The model was also highly significantly different from chance for precision and recall at all points in training, from 1000 to 10,000 utterance exposures, all $t >= 10$, all $p < .0005$.

< Insert Figure 2 about here>

The model's performance was consistent in its precision and recall across the different CDS corpora. The worst performance of the model was for the Naomi corpus, which was the smallest, so the model's training had completed by approximately 8,000 utterance exposures. Yet, the model's performance even on this corpus was .70 precision and .70 recall, compared to .11 precision and .09 recall baseline. The best performance of the model was on the Aran corpus, with .76 precision and .79 recall, compared to .11 and .10 baselines for precision and recall, respectively.

Though the model represented an extremely simple algorithm for discovering words, it compared well to more complex incremental models of speech segmentation (Batchelder, 2002; Brent, 1999; Venkataraman, 2001). After 10,000 words had been processed for a single CDS corpus, precision and recall was approximately .75-.80 for the BMDP1 model (see Figures 3 and 4 in Brent, 1999). In the same paper, other algorithms were also compared, and all performed substantially worse than the PUDDLE model: the SRN model of Christiansen et al (1998): precision .40-.45, recall .40-.45; Olivier (1968): precision .50-.55, recall .35-.40. Venkataraman (2001) reports slightly reduced precision and recall for the BMDP1 (Brent, 1999) model for a similarly derived CDS corpus of .65-.70 and .70-.75, respectively, and the

best performance of his model with a trigram-based algorithm performed with precision .70-.75 and recall .70-.75. Batchelder (2002) reported precision and recall in a similar range for her model with the word length constraint set to its optimal level on one corpora, but other corpora tested yielded slightly reduced precision and recall.

However, despite our model being quantitatively comparable to other methods in terms of precision and recall of segmentation performance, it is the qualitative behaviour of the model that we wish to focus on. Tables 3 and 4 show the model's performance in terms of the 20 most highly activated words in the lexicon for the six corpora after 1000 and 10000 utterance exposures. Evident from the Tables are that the model very early in training identifies words with a high degree of precision. After 1000 utterances, the model identifies no false positives in the top-20 words in each corpus, but includes several frequently occurring word sequences as potential word candidates (so counting as misses in the analyses), such as 'isn't it', 'you tell me', and 'that's right'. By 10000 word utterances nearly all of these have been correctly broken down into their constituent words, so that across the top-20 words in all 6 corpora, only three remain (2 for Aran and 1 for Naomi).

< Insert Tables 3 and 4 around here >

Also of note in the most highly activated words in each corpus is the presence of the child's name. Even after a small amount of exposure – to 1000 utterances – the child's name had been identified and occurred in the top-20 most highly activated words for 4 of the 6 corpora. For the other corpora, 'Aran' was identified in the lexicon but occurred as the 36th word, and 'Nina' was 44th. By 10000 utterances, the child's name maintains its prominence, and in 5 of the 6 corpora it occurs in the top-20 most highly activated words, and for the other corpus, 'Aran' occurs 33rd in the lexicon.

The identity of the other words that were most highly activated, and formed the basis of the model's segmentation performance, also provide qualitative data about the model's performance. They were generally the words that occurred with the highest frequency in the

corpus and were principally constituted of pronouns and determiners, but also included some prepositions, conjunctions, interjections, and high frequency verbs. Whereas some of these words can occur as single word utterances (such as proper names and interjections), other words, as noted earlier, such as determiners seldom occur in isolation. 'The', for instance, was reliably identified as a word in the model and occurred in the top-20 for all 6 corpora.

Though there is a correspondence with the frequency of the word's occurrence, this was not the only factor influencing it becoming highly activated, and a consequent basis for segmenting words that occur around it. The distributional pattern of the word is also important for determining whether it becomes highly activated in the lexicon. In previous work, we have found that certain words are more useful than others for indicating the grammatical category of words with which they co-occur. In order to determine whether the same words that are useful for indicating grammatical categories in language learning are the same as those identified early, and useful for, speech segmentation, we examined the top-20 words for each corpus in terms of whether they were words that significantly distinguished nouns and verbs (data from Monaghan et al., 2007). The words were: *he, we, are, no, your, that's, in, do, is, to, a, the*, and *you*. The analyses indicated that many of these words were highly activated in the PUDDLE model's lexicon. The final column of Tables 3 and 4 indicates how many of these thirteen word cues were in the top-20 highly activated word in each corpus. The words useful for segmentation substantially overlap with those distributional word cues useful for grammatical categorisation.

General Discussion

The PUDDLE model of speech segmentation was designed to provide an explicit test of how far utterance boundaries alone can provide a bootstrap into identifying word boundaries from continuous CDS. Considering each utterance as a potential word candidate, the model discovered which utterances were single-word utterances accurately, and the rarer multiple-

word occurrences were less highly activated as candidates for the lexicon. Using these identified words, the model was successful in segmenting the speech corpora to a level of precision and recall similar to other, more sophisticated models that assume substantially more computational complexity and memory load for the child. Another important component of the model was the boundary constraint, which required that, before a lexical candidate could be entered into the lexicon, the boundaries around the candidate in the speech must form a legal phonotactic context. In pilot modelling, we found that this constraint was necessary in order for the language to be segmented effectively. However, the boundaries were discovered by the model as a consequence of establishing a lexicon, and so the constraint was not external to the model's functioning, but emerged as a consequence of its functioning. No other constraints were found to be required in order for the model to learn to an effective level.

In terms of the five criteria for assessing developmental models of speech segmentation, our model fulfils many of them more effectively than other models of segmentation (see Table 5). This is not a simple consequence of the decisions we made about the assessment criteria, which we see as more generic principles that ought to apply to models of other domains of language acquisition (see, e.g., Brent, 1996; Christiansen & Chater, 2001). We intend these criteria as a snapshot of the current state of the field, rather than a criticism of previous models of segmentation. The extent to which each model meets the criteria in Table 5 was not always possible to deduce from the relevant papers, and thus we have sought to err on the side of caution in our ratings.

< Insert Table 5 around here >

For the reasonable precision and recall, the connectionist models (Christiansen et al., 1998) and Olivier's (1968) model fall short of the levels of performance of the other approaches. We imagine that the Perruchet and Vinter (1998) model will also perform poorly on this criterion if the model is given free reign on phoneme transcriptions rather than

syllables, for similar reasons to those we outlined for the PUDDLE model's failure without the boundary constraint. In terms of computational tractability, we contend the Batchelder (2002) and Venkatamaran (2001) models will require highly computationally intensive processing particularly after extensive training. This is because all possible individual segments and their combinations are stored in the lexicon (though in the case of Batchelder's model, a decay parameter can remove them from the lexical store). Such a store can become extremely large very quickly, and a model that does not consider all possible combinations of sequences of segments should be preferred. Brent and Cartwright's (1996) model and the Frank et al. (2007) models are both idealised learners, and so require simultaneous processing of the entire corpus, rather than taking an incremental approach, to achieve accurate speech segmentation performance. It is not, however, established that incremental versions of these models could not effectively learn to segment speech, and so the 'N' in Table 5 against computational tractability indicates that, in their current form, these models do not yet meet this criterion. In terms of external constraints, all the models except the connectionist models, Brent's (1999) model, Venkataraman's (2001) model, and the PUDDLE model have additional components that are not discovered by the model. For the final criterion, other models may be effective in simulating particular aspects of children's performance in speech segmentation, but such explicit tests have not always been reported, so it is as yet unclear whether these other models would simulate the psycholinguistic effects on which we have focused.

The future benchmark that merits most development, we believe, is the extent to which models can reflect what is known about the psycholinguistic properties of segmentation in infants and the hypotheses they raise for future studies in this regard. The PUDDLE model has indicated how proper nouns, in particular the child's name, can emerge early in language processing as a word candidate, and can form a critical basis for speech segmentation of the words that occur around it. Bortfeld et al. (2005) showed that the words occurring

immediately after the child's name were attended to more than words that occurred in other contexts. Our model demonstrates that this benefit of the child's name can be discovered by a model of speech segmentation based on the distributional properties of the name. In the PUDDLE model, the child's own name can be discovered early and can then act as a basis for the segmentation of words around it. As a caveat, however, the Bortfeld et al. (2005) study also showed that co-occurrence of a word with 'mommy' also demonstrated an advantage in terms of the infant's learning of the target word. 'Mommy' occurs rarely in several of the CDS corpora in our study, and so was unlikely to emerge as useful cue for segmentation. In the Aran and Anne corpora it occurred zero times. It occurred most frequently in the Peter corpus, with a frequency of 1.3 per thousand words, and was identified as a word in the lexicon for the model but with low activation. However, the model is also sensitive to local aberrations in the occurrence of words, so if a lexical item occurs frequently in a portion of the corpus then it will increase in activation and consequently increase in its usefulness for segmenting other words with which it co-occurs.

An additional developmental psycholinguistic property of the model is its use of phonotactic information about legal word boundaries. Pilot studies of the model without this word boundary constraint resulted in oversegmentation of the speech into individual phonemes. An alternative to the boundary constraint would be to impose other constraints on legal segmentations of the speech, such as the vowel constraint, utilised by models such as Brent and Cartwright (1996). We prefer to use the boundary constraint, however, as this emerges from the discovered lexicon in the model itself rather than being an external property imposed on the model. Additionally, when we implemented the vowel constraint as a requirement that segmented words must contain at least one vowel, we found this constraint to be less effective as a supplement to the PUDDLE model than the boundary constraint. Introducing the vowel constraint, but omitting the boundary constraint, resulted in mean precision of .43 and recall of .52 after 10000 utterance exposures, so the boundary constraint

resulted in 20-30% better precision and recall in speech segmentation across the six corpora. Including both the vowel constraint and the boundary constraint resulted in performance similar to the boundary constraint alone: mean precision was .73, mean recall was .75. It may be that sensitivity to phonotactic constraints in language learners, in terms of word-internal phoneme pairs, may be a consequence of their importance as a constraint for constructing hypotheses about which phoneme sequences may constitute a word.

The errors that the model makes in its speech segmentation are also instructive in terms of the information potentially available to the child from their speech environment, and the range of computational processes that may react to this information. In particular, the undersegmentation of certain frequent phrases bears a resemblance to the item-based model of language learning proposed by Tomasello (2000). In the model's performance at early stages of learning, several candidate words were multiple-word utterances that occurred frequently in the speech. Whereas the model eventually learned to accurately decompose these frequent co-occurrences into their constituent words, this indicates that the PUDDLE model's utterance-based approach to segmentation is consistent with such developmental trends, which now have behavioural support (e.g., Bannard & Matthews, 2008).

One of the insights generated from this model was that the words that emerged as useful for segmenting speech are precisely those that are also useful for indicating grammatical category. Peña, Bonatti, Nespor, and Mehler (2002) claimed that segmentation and learning of grammatical structure are separable and sequential processes in language learning. They claimed, on the basis of results from artificial language learning tasks, that generalisation of the grammar cannot occur before the problem of speech segmentation has been accomplished. However, their artificial language did not contain the distributional properties of natural language that the PUDDLE model indicates are potentially extremely useful for both speech segmentation and grammar learning. Instead, their model only contained transitional probability information, in non-adjacent syllables, to indicate word and

language structure. The PUDDLE model cannot reveal whether speech segmentation precedes grammatical category learning, but it does indicate that the same high frequency words within a natural language corpus can serve both these tasks and are discoverable extremely early in language acquisition.

As may by now be apparent, the PUDDLE model we have presented is not inconsistent with other approaches to speech segmentation. Our aim was to highlight how models of segmentation can be informed, and in turn can inform, developmental studies of the cues that are useful and used by children in segmenting speech. In summary, this paper presents a novel framework for comparing developmental models of speech segmentation in qualitative as well as quantitative terms. Based on these criteria, the PUDDLE model performs comparably to other models in terms of precision and recall of segmentation and presents an advance on other models in its ability to reflect qualitative aspects of children's early speech segmentation performance. The PUDDLE model suggests that when a child embarks on language acquisition she does not have to swim through a vast sea of sound to discover the words of her native language but instead is faced with the relatively easier (though nontrivial) task of looking for words in small puddles surrounded by helpful boundary information to facilitate segmentation.

References

Aslin, R., Woodward, J., LaMendola, N, & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan and K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Bannard, C. & Matthews, D.E. (2008). Stored word sequences in language learning: The effect of familarity on children's repetition of four-word combinations. *Psychological Science, 19*, 241-248.

Batchelder, E.O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*, 167-206.

Black, A. W., Taylor, P., & Caley, R. (1990). *The Festival speech synthesis system*, available from http://www.cstr.ed.ac.uk/projects/festival.html, Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology, 6*, 380-420.

Bortfeld, 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.

Brent, M.R. (1996). Advances in the computational study of language acquisition. *Cognition, 61*, 1-38.

Brent, M.R. (1999). An efficient probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*, 71-105.

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

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221-268.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America, 95*, 1570-1580.

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

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

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

Frank, M.C., Goldwater, S., Mansinghka, V., Griffiths, T., & Tenenbaum, J. (2007). Modeling human performance on statistical word segmentation tasks. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Gerken, L. A. (1996). Prosodic structure in young children's language production. *Language, 72*, 683-712.

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

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

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language acquisition: Vol. 1. Syntax and semantics* (pp. 73-136). Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12*, 271-296.

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

Monaghan, P., Christiansen, M.H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259-305.

Olivier, D.C. (1968). *Stochastic grammars and language acquisition mechanisms*. Unpublished PhD dissertation, Harvard University, Cambridge, MA.

Peña, M., Bonatti, L., Nespor, M., Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604-607.

Roy, D.K. & Pentland, A.P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*, 113-146.

Perruchet, P. & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246-263.

Sachs, J., (1983). Talking about the there and then: the emergence of displaced reference in parent-child discourse. In K.E. Nelson (Ed.), *Children's language*. Hillsdale, NJ: Lawrence Erlbaum

Saffran, J. R. (2001). Words in a sea of sound: The output of statistical learning. *Cognition, 81*, 149–169

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

Slis, I.H. (1970). Articulatory measurements on voiced, voiceless and nasal consonants. *Phonetica, 21*, 193-210.

Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29*, 103-114.

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

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics, 27*, 351-372.

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

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

Acknowledgements

Work with the Festival speech synthesiser was greatly assisted by Korin Richmond. We are grateful to Ronald Peereman for the suggestion of inputting text corpora through the speech synthesiser to generate a phonological transcription.

Table 1. Proportion of utterances from child-directed speech of different lengths of words.

Utterance Length (in Words)	Proportion of Corpus
1	.26
2	.14
3	.13
4	.12
5	.10
6	.08
7	.06
8	.04
>8	.09

Table 2. Size and characteristics of each child-directed speech corpus.

Corpus	Number of utterances	Mean words per utterance	Mean phonemes per word
Anne	27,474	3.37	3.07
Aran	27,794	3.81	3.07
Eve	17,327	3.55	3.05
Naomi	8.318	3.56	3.12
Nina	17,865	4.01	3.03
Peter	20,091	3.61	3.01

Table 3. The most highly activated words in each corpus after 1000 utterance exposures.

Corpus	Top 20 'Words'	Word overlap with grammatical cues
Anne	a, there, you, in, it, what, are, is, no, that, anne, on, and, shall, sit, look, whoops, then, pardon, thank	7
Aran	what, isn't_it, yes, hm, that, what's, come_on, a, are_you, and, it, is_it, one, there's, erm, didn't_we, that's, do_you, is, bang	5
Eve	what, you, yes, no, do, a, want, there, is_that, that, well, and, you_tell_me, where, would, eve	5
Naomi	naomi, say, achey, yes, blanket, what's_this, is, that's, what, the, no, what's, brush, birdie, that, broken, good, honey, goldie, yah	4
Nina	what, is, a, where, that's_right, yah, you, on, doing, are, do, shall, okay, that's, these, let's, happened, darlie*?, here, no	7
Peter	see, you, a, what, there, is, it, peter, right, say, that, that's, lets, are_you, the, yah, here, and, sit, look	6

Table 4. The most highly activated words in each corpus after 10000 utterance exposures.

Corpus	Top 20 'Words'	Word overlap with grammatical cues
Anne	you, a, it, the, are, there, what, that, to, we, on, in, do, is, yah, right, one, no, going, anne	10
Aran	you, it, are, is, what, to, that, the, I, we, a, going, there, and, on, hm, no, isn't_it, in, come_on	9
Eve	you, what, it, that, is, no, the, yes, I, a, on, eve, in, do, well, and, are, there, your, that's	9
Naomi	naomi, you, it, what, the, to, that, I, honey, on, no, ee*, yes, is, want, are, okay, do_you, that's, right	7
Nina	you, what, is, to, it, the, are, do, I, w*, on, did, where, in, that, a, put, no, this, nina	9
Peter	you, it, the, that, there, a, what, is, in, see, to, put, I, on, mhm, no, right, can, an, peter	7

Table 5. Criteria for assessing developmental models of speech segmentation.

Model	Reasonable precision and recall	Computationally tractable	No external components	Psycholinguistic effects
Batchelder (2002)	Y	N	N	Y
Brent (1999)	Y	Y	Y	Y
Brent & Cartwright (1996)	Y	N	N	Y
Christiansen et al. (1998)	N	Y	Y	Y
Olivier (1968)	N	N	?	?
Perruchet & Vinter (1998)	N	Y	N	Y
Frank et al. (2007)	Y	N	?	?
Venkataraman (2001)	Y	N	Y	?
PUDDLE model	Y	Y	Y	Y

Figure 1. The PUDDLE model operating on the first few utterances of a corpus.

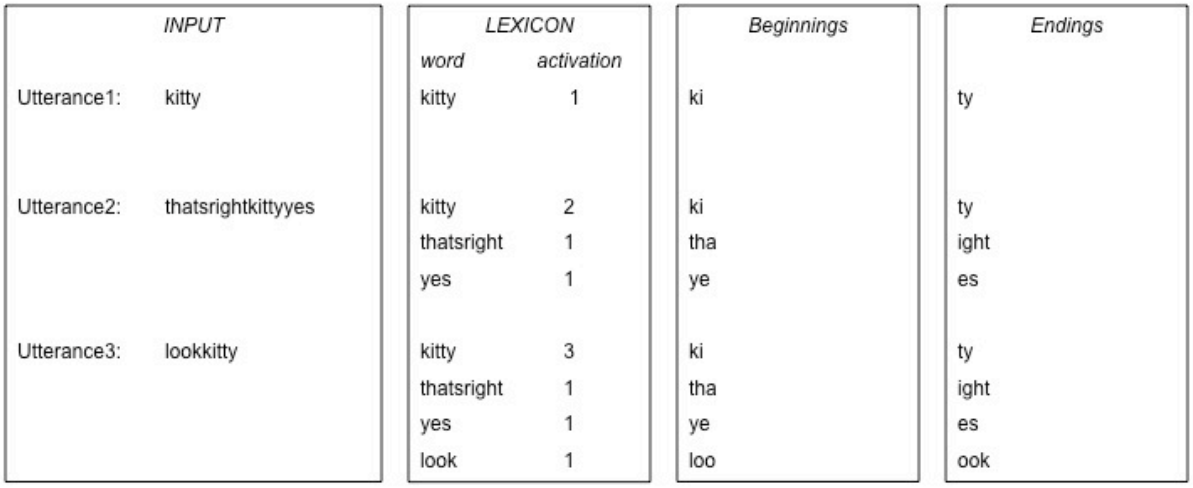
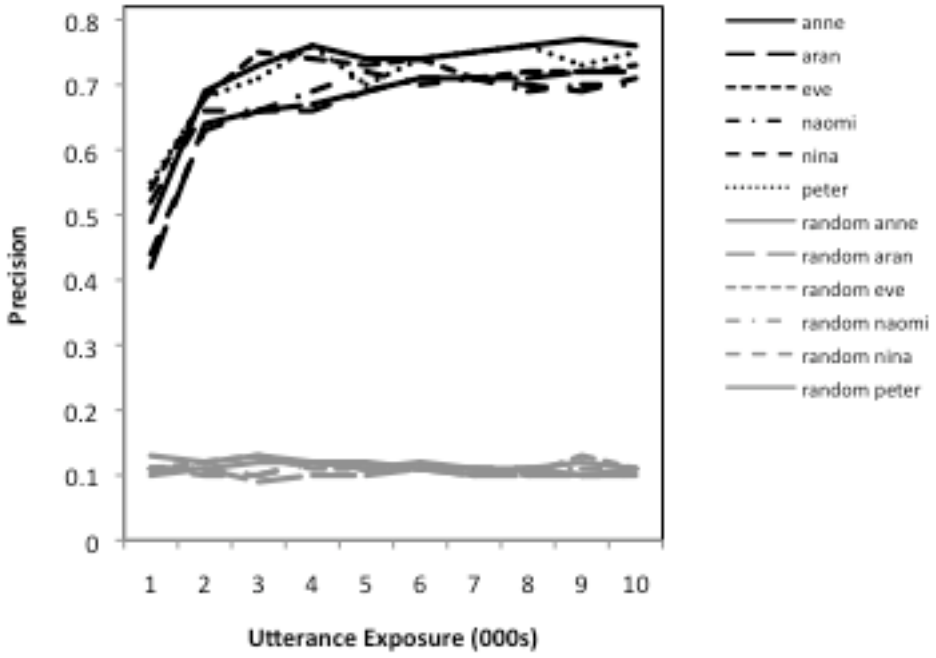
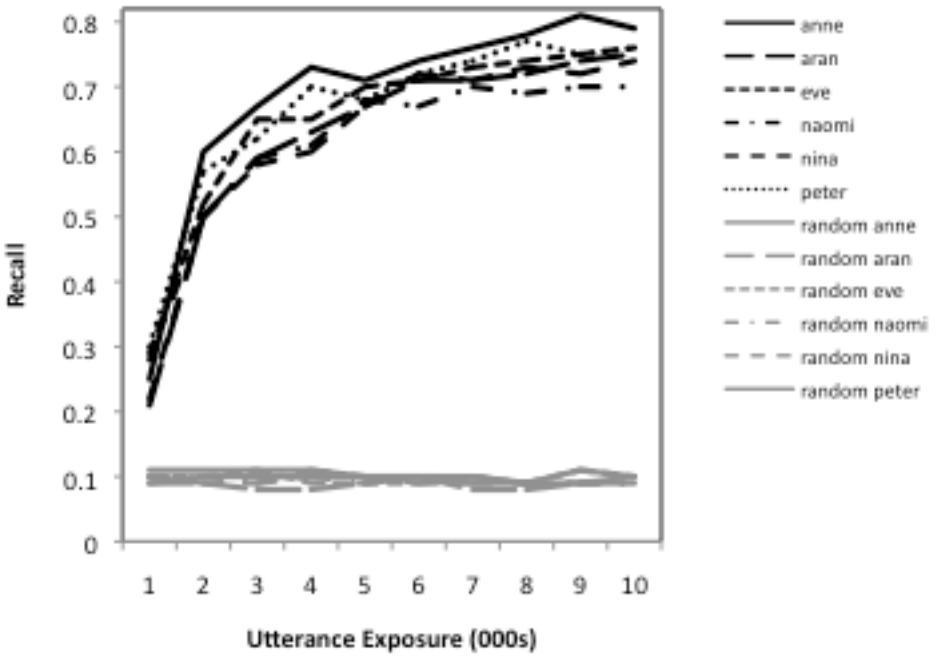


Figure 2. (A) Precision and (B) recall for the PUDDLE model of speech segmentation.

(A)



(B)