McCauley, S.M., Monaghan, P. & Christiansen, M.H. (2015). Language emergence in development: A computational perspective. In B. MacWhinney & W. O'Grady (Eds.), The handbook of language emergence (pp. 415-436). Hoboken, NJ: Wiley-Blackwell.

Language Emergence in Development:

A Computational Perspective

Stewart M. McCauley[1], *Padraic Monaghan*[2], *Morten H. Christiansen,*[1,3*]

1. Department of Psychology, Cornell University, USA
2. Department of Psychology, Lancaster University, UK
3. Haskins Laboratories, USA

September 2013

Running head: Language emergence in development

Word count (main text): 8,198

***Correspondence**:

Dr. Morten H. Christiansen
Department of Psychology
Cornell University
228 Uris Hall, Psychology
Cornell University
Ithaca, NY 14853-7601
Email: christiansen@cornell.edu
Tel: 607-255-3834
Fax: 607-255-8433

Language Emergence in Development: A Computational Perspective

Introduction

Emergentist accounts of language typically seek to explain acquisition as being rooted in simple learning mechanisms. Such an approach requires not only that the input to children be rich with cues to the structure of the language being learned, but also that children's learning mechanisms be well-attuned to these cues. Languages themselves must be, in turn, well-tailored to the learner through processes of cultural evolution (e.g., Christiansen & Chater, 2008).

While a number of developmental psycholinguistic findings have lent support to such an approach, computational work has allowed researchers to directly uncover a wealth of potentially useful statistical information in corpora of child-directed speech, as well as to test explicit models of language acquisition using representative input. Computational modeling provides a means to directly test the capacity of simple learning mechanisms to give rise to complex linguistic knowledge, and therefore represents a core methodology in emergentist work on language development.

Many of the early strides in emergentist modeling of language development were achieved by researchers working within the connectionist framework. In recent years, the computational limitations of connectionist techniques have lead to a shift away from artificial neural networks, towards higher-level models capable of scaling up to deal with input in the form of entire corpora of child-directed speech. A number of such models have made contact with emergentist principles, demonstrating that learning structure from input through general cognitive mechanisms can account for specific developmental patterns in syntactic production (Freudenthal, Pine, & Gobet, 2006, 2007; Gobet, Freudenthal, & Pine, 2004; Jones, Gobet, & Pine, 2000), the acquisition of constructions and construction-like units (Chang, 2008; Solan,

Horn, Ruppin, & Edelman, 2005), and semantic role learning (e.g., Alishahi & Stevenson, 2010), in addition to capturing the emerging complexity of children's grammatical knowledge more generally (e.g., Bannard, Lieven, & Tomasello, 2009; Borensztajn, Zuidema, & Bod, 2009).

Despite their considerable success in implementing the emergentist approach to language, contemporary models of acquisition do not typically seek to capture the on-line processes involved in learning incrementally (cf. Monaghan & Christiansen, 2010). Thus, a remaining challenge for computational accounts of language development is to take account of the constraints imposed by the fact language processing must take place in the here-and-now (Christiansen & Chater, in preparation). Modeling work filling this gap would serve to strengthen emergentist accounts, lending understanding to the ways in which language emergence takes place on multiple time scales: the time-scale of seconds, in which utterances are comprehended and produced; the time-scale of years, in which linguistic knowledge emerges; and the time-scale of centuries, in which languages themselves are transformed through processes of cultural evolution (cf. Christiansen & Chater, 2008). A comprehensive emergentist approach must ultimately take each of these time-scales into account within a unified framework, and it is to the connections between first two that we turn our attention in the present chapter.

Here, we discuss two recently developed computational models of language development which bring us closer to answering the computational challenge described above, each of which instantiates emergentist principles and demonstrates the capacity for simple on-line learning mechanisms to give rise to complex linguistic knowledge. The first model captures children's discovery of linguistic units in the speech stream, while the second simulates learning to use such units and their distributional properties to comprehend and produce speech. Both models go beyond previous computational approaches to acquisition, both in terms of their simplicity and

their ability to capture developmental data through on-line, incremental processing.

Importantly, each model is driven by item-based learning, consistent with emergentist and usage-based approaches to acquisition (e.g., MacWhinney, 1975; Tomasello, 2000). The first model, the Phonotactics from Utterances Determine Distributional Lexical Elements (PUDDLE; Monaghan & Christiansen, 2010) segments speech based on simple information derived from utterance boundaries. Phonotactic information learned from phoneme bigrams found at the boundaries of previously discovered units is used to constrain the model's segmentation of subsequent utterances. Despite its simplicity, PUDDLE performs well in comparison to more complex models of word segmentation, in addition to capturing developmental psycholinguistic effects. The second model, the Chunk-Based Learner (CBL; McCauley & Christiansen, 2011, submitted), begins with word-by-word exposure to the utterances in a corpus. Using simple frequency-based statistics, it learns to group words together on-line to create larger units, thereby segmenting utterances into phrase-like chunks, constructing an item-based "shallow parse." The very same chunks and statistics thus used for comprehension are also employed to produce new utterances that are compared to utterances produced by the target-child of the corpus. CBL achieves high performance in addition to successfully modeling psycholinguistic data from previous developmental studies.

Both models are inspired by recent psycholinguistic studies that have served to spark a renewed interest in "chunking" phenomena (cf. Miller, 1956). In both models, linguistic knowledge is represented as an inventory of "chunks" consisting of one or more words, which is gradually attuned through incremental learning while simultaneously playing an on-line role in the processing of incoming input. The two models demonstrate different ways of creating and exploiting chunks: PUDDLE "starts big," storing large chunks which are gradually broken down

into parts, while CBL learns to combine words into larger sequences. Both approaches make contact with usage-based approaches (e.g., Tomasello, 2003), in which multiword units play a significant role in the child's linguistic development, an idea that is bolstered by recent developmental psycholinguistic evidence (e.g., Arnon & Clark, 2011; Bannard & Matthews, 2008). While PUDDLE initially captures multiword units arrived at through under-segmentation of frequent phrases (cf. Arnon, 2009), CBL makes contact with a more widespread approach to chunking, following classic studies in psychology (e.g., Miller, 1956). We demonstrate that both approaches to chunking can be viewed as complementary, and, when instantiated in simple models such as PUDDLE and CBL, can give rise to a considerable amount of linguistic knowledge. We show that, despite contrasting computational approaches, the emergentist principles of both these models result in consistent and overlapping information about how and which linguistic structures are developed during language acquisition.

In what follows, we describe the computational architectures of PUDDLE and CBL, summarize the outcomes of simulations using corpora of child-directed speech as input to the models, and report results from the application of both models to a single dense corpus of child-directed speech (Maslen, Theakston, Lieven, & Tomasello, 2004) in order to better illustrate how the models relate to and complement one another, as well as to test the models' ability to scale up to a substantially larger corpus. We then discuss the implications of our findings for emergentist approaches to language more broadly.

The PUDDLE Model of Word Segmentation

A fundamental motivation for PUDDLE was to capture the incremental nature of learning and the online character of input processing in a simple, psychologically plausible computational model. The model began with a consideration about the input available to the language learner

and a reduction in the assumptions required of the learner in terms of internal processes brought to bear to the problem of speech segmentation. Other models of segmentation have tended to consider learning as idealized and optimal, with information from the whole data set potentially available to the learner with no memory limitations or other cognitive constraints on detecting ideal information present in these systems (Batchelder, 2002; Brent, 1999; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Venkataraman, 2001). In contrast, the PUDDLE model aimed to determine how much structure was available in the input itself and the extent to which a general purpose learning mechanism based on chunking could result in accurate segmentation of child-directed speech, as well as reflect psycholinguistic properties of the early stages of language acquisition.

The PUDDLE model starts with a single utterance heard by the child. This utterance is stored as a chunk in the model's chunk inventory. Then, the next utterance the child hears is interpreted. If the next utterance contains the first utterance as a part, then the model considers this as indicating the potential chunk boundaries in the speech. Thus, utterances are stored as potential words in the language, and the model searches against its memory for ways in which following utterances can be segmented. The model has some similarities to the PARSER model of speech segmentation (Perruchet & Vinter, 1998). In PARSER, frequently co-occurring syllables become chunked together, and these chunks, if frequently occurring in the speech input, become candidates for lexical items. PARSER has simulated artificial language learning experiments that have previously been described in terms of transitional probability learning (e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997) except that PARSER is instead sensitive to frequency of co-occurrences rather than transitional probabilities. PUDDLE also utilizes frequent sequences to

segment utterances, but goes beyond models like PARSER in two important ways. Firstly, PUDDLE operates at the phoneme level, rather than the syllable level. The phonemes belonging to each syllable may not be a given in the child's language system, so from the speech stream "*thedog*," the child must learn that "*d*" attaches to the second syllable and not the first. As such, processing at the phoneme level adds considerable difficulty to the task of speech segmentation. Secondly, the model is successfully applied to natural language corpora, to which PARSER has not yet been effectively extended. PUDDLE is also an advance on other models of speech segmentation (e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Norris & McQueen, 2008) in that it embodies natural constraints on language processing, and does not presume that the child is an ideal learner. As such its principal advantage is to highlight the developmental process of language acquisition as learning progresses, making contact with the emergentist idea that simple learning mechanisms can give rise to complex linguistic knowledge.

Architecture of the Model

The model is illustrated in Figure 1. It has two components: 1) an input buffer containing the utterance, and 2) a chunk inventory. At the first stage, the utterance that the child hears is entered into the input buffer. Then, from the beginning of the utterance, the chunk inventory is searched for matches to sequences of phonemes contained in the utterance. If there is no match, the entire utterance is entered into the chunk inventory. If there is a match, then the matching lexical item is extracted from the utterance and the phoneme sequence preceding the extracted lexical item is entered into the chunk inventory. The phoneme sequence succeeding the extracted lexical item is then searched for matching lexical items. Each item in the chunk inventory has an associated activation level, which is boosted each time the lexical item is matched in the input buffer. Lexical items are searched in order of their activation level.

[Insert Figure 1 about here]

As an example, consider a situation in which the chunk inventory contains the sequences "the" and "peter" and the current input buffer contains the sequence "thedogbelongstopeter," as in Figure 1. At stage 2, the model begins the search at the first phoneme "th" (search point indicated by the dashed-line box). There is a match straight away with "the," and the activity of the lexical item "the" increases from 3 to 4. Then, at stage 3, "dogbelongstopeter" is searched for a match, beginning at the phoneme "d," and so on until a match is found at stage 14 with "peter." The activity of the lexical entry "peter" is increased, and finally the end of the utterance is reached and so the remaining sequence "dogbelongsto" is added to the chunk inventory.

To avoid potential oversegmentation, an additional constraint was introduced: words could only be extracted from the speech stream if the word was preceded by a legal word ending and succeeded by a legal word beginning, where a sequence is legal if it belongs to at least one item in the chunk inventory. Beginnings and endings are defined as phoneme bigrams. Consequently, in the above example, "the" will only be extracted from the sequence "thedogbelongstopeter" if "do-" begins one of the words already in the model's chunk inventory. This word boundary constraint is cognitively plausible due to the sensitivity of language learners to bigram sequences that occur at word boundaries (e.g., Mattys, White, & Melhorn, 2005).

Learning to Segment English

Monaghan and Christiansen (2010) report simulations in which PUDDLE is exposed to six child-specific corpora of English child-directed speech from the CHILDES database (MacWhinney 2000). Baseline performance was determined by randomly assigning the correct number of segmentations within each utterance to positions within the utterance, as used by Brent (1999). Therefore, the baseline had information not available to the PUDDLE model about

how many segmentations to make, but information about where the segmentations should be placed was not available to the baseline. The model performed at levels similar to those of more complex models of segmentation (e.g., Batchelder, 2002; Brent, 1999; Venkataraman, 2001), with 69.0% accuracy (10.9% random baseline), and 73.3% completeness (9.4% random baseline), both significantly higher than baseline across the six children's corpora, p's < .001. This was the case even though the comparison models were not constrained by cognitive limitations (for a more detailed description of the input corpora and simulations, see Monaghan & Christiansen, 2010).

We next describe a further simulation of PUDDLE on a dense corpus of English child-directed speech (Maslen et al., 2004). We begin by describing the input corpus and its preparation before reporting and discussing PUDDLE's segmentation performance in the context of the emergentist approach to language development.

Corpus and corpus preparation. We began by taking the most recent version of the dense corpus described by Maslen et al. (2004) from the CHILDES database (MacWhinney, 2000), and extracted all speech spoken in the child's presence from the age 2;0 to 4;11. Following the removal of tags, markers, and speaker identifiers, the corpus contained just over 650,000 utterances with a total of 2.6 million words. We converted the orthographic transcription of the corpus to phoneme sequences by running the corpus through the FESTIVAL speech synthesizer (Black, Clark, Richmond, King, & Zen, 2004) and encoded the resulting phoneme sequences for each utterance. This method allowed us to capture phonemic variation stemming from part-of-speech context, rendering the input more representative of the child's actual experience.

Scoring. The model was run once through the corpus. Testing was incremental and on-line, in that the model was tested as it progressed through the corpus on sequences it had not seen before.

The model's performance was analyzed for the number of correctly identified words (true positives), number of wrongly identified words (false positives), and number of words not segmented (false negatives). Precision of segmentation was determined by true positives divided by true plus false positives. Recall was determined by true positives divided by true positives plus false negatives. We used the same baseline as for previous PUDDLE simulations by assigning word boundaries randomly in each utterance, but with the number of boundaries being equal to the actual number of word boundaries in the correctly segmented corpus. Mean baseline values were determined from 10 random assignments of word boundaries.

Segmentation results. The model's performance is illustrated in Figure 2, showing levels of precision and recall at different points throughout processing the 658,000 utterances of the corpus. The baseline level is for baseline precision. Baseline recall was at a very similar level, a consequence of providing the number of words in an utterance to the baseline and averaging over a large number of utterances for each data point. As with the PUDDLE model's previous simulations of small corpora, the model is able to segment the dense corpus at high levels of precision and recall. After 10,000 utterances, a similar size to previous simulations reported in Monaghan and Christiansen (2010), precision was .739 and recall was .746, both significantly higher than the random baseline, $Z = 182.8$ and $Z = 242.9$, respectively, both $p < .001$, and at levels consistent with previous simulations on distinct child-directed speech corpora.

[Insert Figure 2 about here]

Furthermore, the model was robust to extended training. Over long exposures, the model maintained consistently high levels of precision and recall even though training on a large corpus increased the possibility of the model learning to over-segment speech due to many more exposures to possible words. At the end of exposure to the corpus, the model has precision and

recall at .683 and .779, respectively. The slight dip in precision at the expense of recall is because the model is making more segmentations at later points in exposure to the corpus which is a consequence of the denser population of possible word beginnings and endings in the model – the model has more opportunity to segment the speech, but does not catastrophically over-segment. The tendency to over-segment could likely be countered if the model was augmented to also include associations between words and meanings (e.g., objects or events in the environment).

Psycholinguistic effects. Most instructive, however, is the model's ability to reflect the development of the lexicon while learning to segment speech. The following results resonate with the previous results reported for the 6 individual child corpora reported in Monaghan and Christiansen (2010). The 20 items with highest activity in the model's chunk inventory for the dense corpus are shown in Table 1.

[Insert Table 1 about here]

Note that after 1000 utterances, several closed class words have already been identified, while two lexical items are multiword chunks consisting of two words: "ahat" and "*whatsthis*." Such multiword chunks that co-occur frequently together have been taken to be hallmarks of item-based learning (Bannard & Matthews, 2008; Clark, 1977; Tomasello, 2003). However, the multiword chunks disappear from the top twenty activated chunks in the PUDDLE model by 10,000 utterances. This mimics later development by children as multiword chunks are deconstructed into their constituent words with extensive exposure (cf. Arnon, 2009). The model shows a similar developmental trajectory, but illustrates this in an accelerated development. A version of the model that requires multiple exposures to a chunk before it is entered as a candidate word in the model's chunk inventory would simulate this slower development of

multiword dechunking.

The high activation of the child's own name (Thomas)—both at 1000 utterances and again at 10,000 utterances—makes close contact with psycholinguistic results demonstrating benefits for words occurring immediately after a child's name, as opposed to words occurring other contexts (Bortfeld, Morgan, Golinkoff, & Rathburn, 2005). This is because highly activated words in the chunk inventory can be used by the model to successfully highlight boundaries of other words preceding and succeeding them in utterances (with approximately 70% accuracy and completeness after training even on very short corpora of speech).

In addition, these highly activated words overlap with words previously proposed to be useful as distributional markers of grammatical category. Monaghan, Chater, and Christiansen (2005) examined the 20 most frequent words in child-directed speech corpora for their ability to distinguish open from closed class words and nouns from verbs. For instance, nearly all of the words shown in Table 1 were found to be useful for learning both the closed/open class and noun/verb category distinctions. The proper noun *Thomas* was not included in Monaghan et al.'s (2005) study, as it combined speech from all the sub-corpora of CHILDES and so this name occurred only a small fraction of the time in the larger data set incorporating speech to hundreds of different children. However, proper nouns may also be useful as preceding markers of verbs, or, with possessives, other nouns. Such usefulness would be revealed in distributional analyses for categorization over single-child corpora.

A further developmental psycholinguistic feature of the model is its learning of phonotactic constraints on word boundaries. Behavioral work with infants indicates a sensitivity to phonotactic information during word segmentation (e.g., Mattys & Jusczyk, 2000) and word learning (e.g., Graf Estes, Edwards, & Saffran, 2011) alike. As noted previously, pilot studies of

the model without the word boundary constraint heavily over-segmented the speech input. The addition of this simple constraint allowed phonotactic information to emerge naturally from the discovered chunk inventory, rather than being externally imposed. Imposed constraints, such as the constraint that legal segmentations must contain a vowel (featured in previous models, such as that of Brent & Cartwright, 1996) were found to be less effective than the simple boundary constraint in further pilot runs.

In summary, the model instantiates an item-based perspective on learning, through its chunk-based knowledge representation and under-segmentation errors: such "errors" result in multiword units, which may be useful in grammatical development, as suggested by usage-based approaches (e.g., Bannard & Matthews, 2008; Tomasello, 2003) and demonstrated empirically by the CBL model (as described below).

Discussion

PUDDLE successfully demonstrates the ability of simple learning mechanisms to make considerable headway in solving the non-trivial problem of identifying linguistic units in a continuous speech stream. The model shows that segmenting speech can proceed with great accuracy by simply isolating highly frequent words in speech. The words do not need to be given to a language learner in advance, but rather can be constructed and identified in principled ways, based only on the child's speech input. Moreover, a number of interesting psycholinguistic features emerge over the course of the model's incremental, on-line processing of the input.

Despite its simplicity, the model was able to attain accuracy and completeness scores comparable to more sophisticated models that involve greater memory demands and computational complexity. This serves to highlight the richness of the stimulus; the child's input contains a wealth of potentially useful distributional, acoustic, contextual, and socio-pragmatic

information, and PUDDLE demonstrates what can be done on the basis of just two types of distributional cue (boundary information and phonotactics of word boundaries). The success of the model suggests that future modeling work exploiting simple mechanisms for learning that in addition computes information from multiple cues will be met with similarly encouraging results. We return to this point in the general discussion.

The CBL Model of Language Acquisition

An important feature of the PUDDLE model is that it actively uses the units it has discovered to find other units in an online, incremental fashion. Similarly, the CBL model uses the word chunks that it discovers to comprehend and produce subsequent utterances that it comes across. Unlike the PUDDLE model, however, which breaks large chunks down into smaller units, the CBL model combines smaller units (words) to create larger ones (multiword chunks). The CBL model is inspired by recent psycholinguistic findings which suggest that the task facing the language learner is better characterized as one of "learning by doing" than as one of grammar induction. It draws upon evidence for children's and adults' use of multiword linguistic units during comprehension (e.g., Arnon & Snider, 2010) and production (e.g., Bannard & Matthews, 2008; Janssen & Barber, 2012) alike, as well as evidence that language processing involves a reliance on local information rather than the construction of a full syntactic parse (e.g., Frank & Bod, 2011; Sanford & Sturt, 2002). Thus, the model can be viewed as providing a test of the usage-based hypothesis that children's language use may be explained in terms of a reliance on multiword chunks (e.g., Tomasello, 2003), while also testing the idea that sequential learning can give rise to complex linguistic knowledge in the absence of hierarchical representations (cf. Frank, Bod, & Christiansen, 2012).

CBL is further designed with several key psychological and computational properties in

mind. As with PUDDLE, the model learns from naturalistic input, in a purely incremental rather than batch fashion, and uses simple, frequency-based statistics. In the case of CBL, learning is based on backward transitional probabilities (BTPs), which 8-month-olds can track (Pelucchi, Hay, & Saffran, 2009). All learning is based on on-line processing of the input as it is encountered on a word-by-word basis. More generally, language acquisition in CBL involves learning to perform two tasks: comprehension of child-directed speech through the statistical discovery and use of chunks as building blocks, and sentence production utilizing the same chunks and statistics involved in comprehension. Comprehension is approximated in terms of the model's ability to segment utterances into phrasal units as they unfold (thereby arriving incrementally at a shallow parse; cf. Sanford & Sturt, 2002), and production is approximated in terms of the model's ability to construct utterances which are identical to those produced by the target child. While comprehension and production processes in the model are two sides of the same coin, for the sake of simplicity we describe them separately in what follows.

Architecture of the Model

Comprehension. As the model processes utterances word-by-word, it tracks frequency information for words and word-pairs, which is used on-line to track the BTP between words and maintain a running average BTP for previously encountered pairs. When the model calculates a BTP that is greater than expected, based on the running average, it groups the word-pair together such that it forms part (or all) of a chunk. When the calculated BTP falls below the running average, a boundary is placed and the chunk thereby created (consisting of one or more words to the left of the inserted boundary) is added to the chunk inventory. The model uses the chunk inventory to make on-line predictions for which words will form a chunk, based on previously learned chunks. When a word-pair is encountered, it is checked against the chunk inventory; if it

has occurred before as a complete chunk or as part of a larger chunk, the words are grouped together and the model moves on to the next word. If the word-pair is not found in the chunk inventory, the BTP is compared to the running average, with the same consequences as before.

As a simple example, consider the model's processing of the utterance *"the dog chased the cat."* At Time 1, the model encounters *the* in the input. At Time 2, the model calculates the BTP between *the* and *dog*, which exceeds the average BTP threshold, resulting in the two words being grouped together. Since the next word has not yet been encountered, the two words are not yet stored in the chunk inventory as a chunk. At Time 3, the BTP between *dog* and *chased* falls below the running average, so *chased* is not grouped together with the preceding material and *the dog* is then stored in the chunk inventory. At Time 4, the BTP between *chased* and *the* falls below the running average, so the two words are not grouped together and *chased* is added to the chunk inventory as a single-word chunk. At Time 5, the BTP between *the* and *cat* rises above the average threshold and because a pause follows the sequence, *the cat* is joined together and stored in the chunk inventory.

Because there are no *a priori* limits on the number or size of the multiword building blocks that can be learned, the resulting chunk inventory will contain a mix of words and multiword chunks. While some of the multiword chunks stored by the model may contain overlapping words or sequences, each chunk is treated as a wholly distinct unit (the model possesses no information regarding such overlap). Importantly, decay does not feature in the current version of the model; chunks do not weaken or cease to exist because of disuse.

We evaluate the model's comprehension performance against shallow parsers (tools widely used in the field of natural language processing), which identify and segment out non-embedded phrases in a text. The shallow parsing method was chosen because it is consistent with

evidence for the relatively underspecified nature of human sentence comprehension (e.g., Ferreira & Patson, 2002; Frank & Bod, 2011; Sanford & Sturt, 2002) and provides a mechanistic approximation of the item-based way in which children are hypothesized to process sentences in usage-based theories (e.g., Tomasello, 2003).

Production. Production relies on the very same chunks and statistics involved in comprehension. Each time the model encounters a multiword child utterance, it is required to recreate the utterance using only building blocks discovered in the previously encountered input. Following Chang, Lieven, and Tomasello (2008), we assume that the overall message, which the child wants to convey, can be roughly approximated by treating the utterance as a randomly-ordered set of words: a "bag-of-words." The task for the model, then, is to output these words in the correct order (as originally produced by the child). Following usage-based approaches, the model utilizes building blocks from its chunk inventory to reconstruct the child's utterances. In order to model retrieval of stored chunks during production, word combinations from the utterance that are represented as multiword chunks in the chunk inventory are placed in the bag-of-words. The model then has to reproduce the child's utterance using the unordered chunks in the bag. We model this as an incremental, chunk-to-chunk process rather than one of whole-sentence optimization. Thus, the model begins by removing from the bag the chunk with the highest BTP given the start-of-utterance tag (which marks the beginning of each utterance in the corpus), and outputs it as the start of its new utterance. The chunk is removed from the bag before the model selects and produces its next chunk, the one with the highest BTP given the most recently produced chunk. In this manner, the model uses chunk-to-chunk BTPs to incrementally produce the utterance, outputting chunks one-by-one until the bag is empty. The overall percentage of correctly produced utterances (those identical to the original target

utterance) is then used as a measure of sentence production performance for a given corpus. We adopt this rather conservative performance measure, by which a single ordering error renders the whole utterance incorrect, to avoid inflating the success rate of the model.

Importantly, comprehension and production processes in the model are further intertwined through the assumption that a child's linguistic behaviors may be reinforced by its own utterances: following each production attempt, the comprehension side of the model is exposed to an identical utterance, thereby reinforcing existing frequency information in the chunk inventory.

Learning to Comprehend and Produce English

In what follows, we describe the results of a CBL simulation in which the model is exposed to the same dense corpus of child-directed speech (Maslen et al., 2004) as was PUDDLE. The results are described in the context of previous CBL simulations using additional English corpora, which are briefly summarized alongside the dense corpus findings. Before reporting our findings, we describe the input corpora, the models used as baselines against which to evaluate specific aspects of CBL, and the system for automatically scoring comprehension and production performance.

Corpora. While in the current chapter we focus primarily on the dense corpus (described in the PUDDLE section above), we initially used CBL to simulate child comprehension and production using corpora of child and child-directed speech taken from the English language section of the CHILDES database (MacWhinney, 2000). We selected all of the English CHILDES corpora that contained at least 50,000 words and spanned at least 6 months in terms of the target-child's age across the entire corpus. These criteria were met by 42 corpora (US: 25, UK: 17).

Baseline models. Prior to evaluating the model's performance, we created three baseline models, in order to explore a 2 x 2 design contrasting the direction of TP calculation (forward vs. backward) and the presence or absence of stored chunks (chunk-based models vs. trigram models). For the forward chunk-based model, we simply constructed a model that was identical to CBL aside from all TP calculations being in the forward direction (referred to hereafter as the FTP-Chunk model). The non-chunk-based models were standard trigram models (cf. Manning & Schütze, 1999), which were adapted to work in the same incremental, on-line fashion as CBL. Trigram models were chosen as a baseline in light of findings that they are quite robust, comparing favorably to more complex models such as probabilistic context-free grammars (PCFGs).

Gold standards for scoring comprehension and production. To evaluate comprehension performance, the model and its baselines were evaluated against a state-of-the-art shallow parser (Punyakanok & Roth, 2010). After shallow parsing the corpora, phrase labels (VP, NP, etc.) were removed and replaced with boundary markers of the sort produced by CBL and its baselines. Each boundary marker placed by the model was scored as a *hit* if it corresponded to a boundary marker created by the shallow parser, and as a *false alarm* otherwise. Each boundary placed by the shallow parser but not by the model was scored as a *miss*. This allowed us to calculate both accuracy (*hits/false alarms*) and completeness (*hits/hits+misses*) for each simulation. To quantify overall comprehension performance, we relied on the *F*-score, which combines accuracy and completeness. We used the general F_β formula (van Rijsbergen, 1979), which weights the completeness score according to β. In our case, β was the ratio of gold standard phrase boundaries (boundaries placed by the shallow parser) to the total number of word pairs (possible slots for phrase boundaries) in a given corpus. Thus, β is always less than 1.

This measure protects against score inflation due to trivial factors, such as over-segmentation.

To assess production performance, each production attempt made by the model and its baselines was scored against the target child's original utterance in the corpus. If the model's utterance was identical to the child utterance, it received a score of 1. Otherwise, it received a score of 0. This allowed us to calculate production performance as the overall percentage of correctly produced multiword utterances. Note that this is a conservative measure of production performance, as even grammatical utterances (e.g., *the cat chased the dog*, when the target utterance was *the dog chased the cat*) will receive a score of 0 if they do not match the target child's utterance.

Comprehension Results. Over the course of the dense corpus simulation, CBL attained an overall F-score of 70.3, with a shallow parsing accuracy of 77.7%. CBL outperformed its baselines, with the FTP-Chunk, BTP3G, and FTP3G models attaining F-scores of 60.2, 63.0, and 66.9, respectively, and accuracy rates of 61%, 66.2%, and 64.8%. These results are consistent with the results of the previous simulations: as can be seen in Figure 3 (left-hand side of each column), throughout the additional simulations, CBL not only outperformed its baselines, but achieved a tighter, more uniform distribution of scores.

To further assess the usefulness of simple item-based information—as opposed to abstract information tied to grammatical categories—we also ran a version of each simulation in which words were replaced by their grammatical categories (using a simple word tagger; Schmid, 1995). As an example, the sentence *"the dog chased the cat"* became *"DET N V DET N,"* allowing the models to learn statistics tied entirely to form classes. This set of simulations was motivated both by the importance placed on form class information in many computational approaches to language development, as well as claims made in the statistical learning literature

about the usefulness of transition probabilities when calculated over classes as opposed to concrete items. For the dense corpus, the model's performance when exposed to form classes fell to 9.7 (previously 70.3), with the baseline models following the same pattern. Results for the dense corpus simulation were again highly consistent with those of the additional set of simulations, the results of which are shown in Figure 3 (right-hand side of each column). As can be seen, CBL and its chunk-based baseline achieved far better performance when exposed to items as opposed to form classes, while the item-based 3G baselines outperformed their class-based counterparts to a significant though less dramatic extent. The implications of these results are discussed below.

[Insert Figure 3 about here]

Production Results. Throughout the course of the dense corpus simulation, the CBL model successfully produced the majority of the multiword child utterances encountered, attaining an overall sentence production performance of 63%, with scores for the FTP-Chunk, BTP3G, and FTP3G baselines reaching just 57%, 47.3% and 47.1%, respectively. This was yet again consistent with the additional simulations involving smaller corpora, as can be seen in Figure 4.

[Insert Figure 4 about here]

Discussion. In our simulations of child comprehension and production processes, CBL's performance exceeded that of its baselines. CBL was able to identify phrasal groupings, approximating the performance of a shallow parser with high accuracy and completeness, by learning in an on-line, incremental fashion. Moreover, learning was tied to a single distributional

cue. This result is striking when one considers that the challenges posed by shallow parsing are regarded as nontrivial in the field of natural language processing (e.g., Hammerton, Osborne, Armstrong, & Daelemans, 2002), as well as the complexity of shallow parsing algorithms (which involve separate training phases, tagging, and so forth). The model's performance highlights the usefulness of the distributional information available in the input to children, while also demonstrating how well-tailored such information is to simple learning mechanisms: the model achieves high performance on the basis of just one out of many potentially useful sources of information.

The fully item-based nature of the model further underscores its simplicity. When compared to simulations involving learning based on grammatical categories, the item-based model still achieved the highest performance. This result is significant, as a great deal of computational work on language development has focused on word classes as opposed to concrete items. It also runs counter to predictions made in the statistical learning literature, involving the greater usefulness of transition probabilities when calculated over form classes as opposed to specific items (e.g., Thompson & Newport, 2007).

The model's ability to produce the majority of the sentences produced by the target children in our simulations is also striking. In addition to underscoring the wealth of information provided by simple distributional cues, this finding also serves to demonstrate that the same sources of information may be useful for learning about language at multiple levels: a single distributional statistic can be used to find word boundaries when calculated over syllables (such as BTP; Pelucchi et al., 2009), to discover phrasal-like units when calculated over individual words, and to help produce utterances when calculated over phrase-like units themselves (as demonstrated by our production results).

Developmental Psycholinguistic and Cross-linguistic Coverage

CBL successfully demonstrates the potential usefulness of incremental, on-line learning from simple distributional cues in language acquisition. But does the type of learning performed by the model hold any psychological validity? This does indeed appear to be the case, as suggested by the model's ability to fit data from a number of developmental psycholinguistic studies spanning a range of topics (detailed in McCauley & Christiansen, submitted), including: child artificial language learning (Saffran, 2002), phrase frequency effects in children's repetition (Bannard & Matthews, 2008), the facilitation of children's irregular plural production by lexically specific frames (Arnon & Clark, 2011), and children's production of complex sentence types (Diessel & Tomasello, 2005).

Having successfully applied CBL to the simulation of language development in children learning English, as well as developmental psycholinguistic studies involving English-speaking children, we also sought to determine whether the model could offer cross-linguistic coverage. McCauley and Christiansen (submitted) performed 174 additional CBL simulations involving single-child corpora (found in CHILDES; MacWhinney, 2000) from 28 additional languages, representing 15 genera from 9 different language families. Because we lacked shallow parsers for all of the languages represented in our sample, we highlight the production results here. Consistent with the English results, CBL once more correctly produced the majority of the child utterances encountered and outperformed its baselines, achieving the highest mean sentence production performance score (55.2%), followed by the FTP-Chunk model (52.3%). The 3G baselines followed the same general pattern of greater accuracy for BTPs (48.4% and 45.5% accuracy for the BTP3G and FTP3G baselines, respectively).

The superior performance of CBL relative to the 3G baselines for even the

morphologically complex languages suggests that multiword units of the type learned by CBL are likely useful across a wide array of typologically diverse languages. Taken together with previous findings of item-based patterns in children's acquisition of morphologically rich languages (e.g., MacWhinney, 1978), this result is promising in the context of future cross-linguistic item-based modeling work. In addition to offering cross-linguistic support for CBL, this finding lends support to the view that simple distributional learning may underlie a large part of children's early linguistic behavior. This result also serve to bolster previous psycholinguistic evidence for chunk-based learning, which has primarily involved English-speaking children (e.g., Arnon & Clark, 2011; Bannard & Matthews, 2008), suggesting that multiword units may play a role cross-linguistically, in analytic and synthetic languages alike.

Complementary ways of building a chunk inventory: PUDDLE and CBL

We have shown that through incremental, online learning of simple distributional information, the PUDDLE and CBL models are able to acquire a considerable amount of linguistic knowledge. PUDDLE succeeds in learning to segment continuous streams of phonemes into words, while CBL learns to combine words to comprehend and produce sentences. Rather than viewing the segmentation and combination of words as a two-step process, we hold that the models instantiate two types of chunk processing which likely take place side-by-side in early acquisition: the discovery of progressively smaller units through the breakdown of larger, unanalyzed chunks (as in PUDDLE), and the discovery of larger units through the combination of smaller units into larger chunks (as in CBL). As the child’s knowledge of words increases, the breakdown of chunks is likely to play a smaller role, whereas the creation of larger chunks from smaller units is likely to continue throughout development and into adulthood, as suggested by psycholinguistic findings of sensitivity to the properties of multiword sequences in adult

comprehension (e.g., Arnon & Snider, 2010) and production (e.g., Janssen & Barber, 2012). In the present section, we seek to support such a view by showing that despite their contrasting computational approaches to building chunk inventories, the models are able to acquire complementary and overlapping linguistic information.

Table 2 provides a snapshot of CBL's early chunk inventory during the dense corpus simulation, listing the model's top 20 most highly activated chunks after exposure to just 1000 utterances, and again at 10,000 utterances. An examination of these top chunks in the context of PUDDLE's entire discovered chunk inventory (at identical time points during its simulation involving the same corpus) reveals a considerable amount of overlap. Across the span of just the first 10,000 utterances, the models appear to converge on complementary information: at 1000 utterances, PUDDLE has already discovered 15 of the unique words (covering 47% of the word tokens) appearing in CBL's top 20 multiword chunks from the same timepoint; at 10,000 utterances, PUDDLE has already discovered 88% of the word tokens (all but 5) appearing in CBL's corresponding top 20 chunks. Thus, many of the words CBL learns to combine with other words early on are among PUDDLE's earliest correctly identified words. Moreover, the models demonstrate that the same multiword sequences learned through chunking (as in CBL) can be arrived at through undersegmentation of the speech stream (as in PUDDLE): five of CBL's earliest multiword sequences listed in Table 2 also appear in PUDDLE's chunk inventory at the same time point. Thus, the models suggest that both types of chunk (those arrived at through undersegmentation and those arrived at through combination of previously segmented units) may play an active role in children's early comprehension and production, which may help to highlight the robustness of previous developmental psycholinguistic findings of multiword utterances and their role in language acquisition (e.g., Arnon & Clark, 2011; Bannard &

Matthews, 2008).

General Discussion

The successes of PUDDLE and CBL are rooted in emergentist accounts of language development, demonstrating empirically that computationally simple learning mechanisms can exploit distributional cues to successfully segment words, as well as learn how to combine words to comprehend and produce sentences. We have shown that PUDDLE, which is based on a simple mechanism for learning boundary information, not only compares well to more sophisticated models of segmentation, but also captures a number of important developmental psycholinguistic effects throughout the course of incremental learning, including segmentation effects from proper nouns, the emergence of phonotactic information, the importance of high-frequency, closed class words for segmentation, and multiword units. We have also shown that CBL's incremental, on-line processing of a single frequency-based statistic is able to not only identify components of an utterance useful for arriving at that utterance's meaning (phrase-like units), but is also able to account for important aspects of children's early linguistic behavior through its simulation of sentence production. Furthermore, it is able to directly model a number of developmental psycholinguistic results ranging from studies of morphological development to studies of complex sentence production.

These findings ultimately serve to highlight the richness of the input. The emergentist view that language development begins with simple learning mechanisms assumes that the stimulus, far from being impoverished (e.g., Chomsky, 1965, p. 58), is replete with potentially useful information that is well-tailored to the learner. This perspective serves to remove much of the explanatory burden from the learner, placing it instead on the cultural evolution of languages

themselves (e.g., Christiansen & Chater, 2008). If very simple learning mechanisms support language development, as suggested by the results presented in this chapter, they are able to do so by virtue of rich input, which has been, shaped to fit the learner (Christiansen, 2013; Chater & Christiansen, 2010).

The work described in this chapter also lays a foundation for future modeling work focused on the role of incremental, on-line processing in development. A comprehensive emergentist approach must ultimately describe language on multiple timescales while making clear the connections between them; the goal is not only to understand how language is acquired and used, or how languages change over time, but also to understand why languages possess the particular properties they do. We posit that to reach such an understanding, researchers must not view acquisition, processing, and cultural evolution as easily separable and distinct phenomena. Future modeling work may serve as the foundations of such a unified approach, but only by capturing constraints imposed by the fact that language takes place in the here-and-now (Christiansen & Chater, in preparation). In this vein, PUDDLE and CBL demonstrate the viability of a perspective in which acquisition and processing are tightly interconnected.

Limitations and Directions for Future Work

Despite the successes enjoyed by our simple frameworks for modeling language development, PUDDLE and CBL are not without limitations. Perhaps most importantly, the models deal solely with distributional information. While both models are designed to demonstrate that much can be done with very simple cues and mechanisms, the environment of the language-learning child is replete with potentially useful contextual, socio-pragmatic, and acoustic information. These sources of information not only provide additional support for learning but also are likely to interact in unexpected ways. For instance, previous behavioral

work involving simultaneous word-referent mapping and segmentation tasks suggests that visual input may often serve to reduce rather than compound the difficulty of word segmentation (Cunillera, Laine, Càmara, & Rodríguez-Fornells, 2010). Thus, future modeling efforts should aim to incorporate learning from multiple cues within simple frameworks like PUDDLE and CBL.

Along these lines, one of CBL's greatest limitations lies in the model's lack of learning from semantic information – the model never learns "meanings" tied to the chunks it discovers. This means that while the model simulates important aspects of language comprehension, it is never called to interpret the meanings of utterances during processing; and while the model is able to correctly sequence the words in a sentence during the production task, the random bag-of-words approach represents a very poor approximation of an overall meaning the model is attempting to convey.

Future work will focus on moving from chunks to constructions, allowing models based on simple mechanisms like those of PUDDLE and CBL to learn from idealized semantic information that is paired with utterances in the input. While the models are currently capable of discovering something along the lines of "candidate constructions" (words and multiword units) in the input, they are devoid of meaning. An important challenge for simple developmental models such as those presented here is to move beyond distributional information, towards an approach that involves simulating the learning not only of word-referent mappings, but the meanings of sentences.

Conclusion

We have described two simple, developmentally-motivated computational models of language acquisition which instantiate emergentist principles. The first, PUDDLE, successfully

discovers words in a stream of phonemes by learning simple distributional information tied to utterance boundaries. Phonotactic knowledge emerges naturally from the discovered chunk inventory, based on learning of phoneme bigrams at the boundaries of segmented units. The second model, CBL, learns to chunk words together to form multiword units that are actively used to simulate aspects of comprehension and production across a set of typologically diverse languages. Both models go beyond previous computational approaches to acquisition by engaging in incremental, on-line processing using simple learning mechanisms and psycholinguistically-motivated knowledge representation. The models compare well to more computationally complex models, while also capturing developmental psycholinguistic effects and successfully modeling results from developmental experiments.

Taken together, PUDDLE and CBL demonstrate that simple learning mechanisms are able to account for a surprising amount of children's early linguistic knowledge and behavior. This, in turn, serves to bolster the emergentist approach to language by highlighting the wealth of information available in learners' input and drawing closer connections between the multiple timescales on which language emergence takes place.

References

Alishahi, A., & Stevenson, S. (2010). Learning general properties of semantic roles from usage data: A computational model. *Language and Cognitive Processes, 25*, 50-93.

Arnon, I. (2009). Starting Big: The role of multiword phrases in language learning and use. Unpublished doctoral dissertation. Stanford University, Palo Alto.

Arnon, I., & Clark, E. (2011). Why brush your teeth is better than teeth: Children's word production is facilitated by familiar frames. *Language Learning and Development*, *7*, 107-129.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language*, *62*, 67–82.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*, 17284–17289.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, *19*, 241.

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

Black, A. W., Clark, R., Richmond, K., King, S., & Zen, H. (2004). Festival speech synthesizer (Version 1.95). *Edinburgh: University of Edinburgh.*

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age: Evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, *1*, 175–188.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-Stream segmentation. *Psychological Science*, *16*,

298-304.

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.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, *9*, 198–213.

Chang, N. C. L. (2008). *Constructing grammar: A computational model of the emergence of early constructions.* Unpublished doctoral dissertation. University of California, Berkeley.

Chater, N. & Christiansen, M.H. (2010). Language acquisition meets language evolution. *Cognitive Science, 34,* 1131-1157.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, USA: The MIT Press.

Christiansen, M.H. (2013). Language has evolved to depend on multiple-cue integration. In R. Botha & M. Everaert (Eds.), *The evolutionary emergence of language: Evidence and Inference* (pp. 42-61). Oxford: Oxford University Press.

Christiansen, M.H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral & Brain Sciences, 31,* 489-558.

Christiansen, M. H. & Chater, N. (in preparation). Now or never: A fundamental constraint on language.

Clark, R. (1977). What's the use of imitation? *Journal of Child Language*, *4*, 341-358.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology*, *63*, 260-

274.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*, 882-906.

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, *1*, 71–83.

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

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829.

Frank, S.L., Bod, R. & Christiansen, M.H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences, 297*, 4522-4531.

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in Dutch and English using MOSAIC. *Cognitive Science, 30*, 277-310.

Freudenthal, D., Pine, J. M., & Gobet, F. (2007). Understanding the developmental dynamics of subject omission: The role of processing limitations in learning. *Journal of Child Language*, *34*, 83.

Gobet, F., Freudenthal, D., & Pine, J. M. (2004). Modelling syntactic development in a cross-linguistic context. *Proceedings of the First Workshop on Psycho-computational Models of Human Language Acquisition* (pp. 53-60).

Graf Estes, K., Edwards, J., & Saffran, J. R. (2011). Phonotactic constraints on infant word learning. *Infancy*, *16*, 180-197.

Hammerton, J., Osborne, M., Armstrong, S., & Daelemans, W. (2002). Introduction to special

issue on machine learning approaches to shallow parsing. *The Journal of Machine Learning Research*, *2*, 551-558.

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PloS one, 7, e33202*.

Jones, G., Gobet, F., & Pine, J. M. (2000). A process model of children's early verb use. In L. R. Gleitman & A. K. Joshi (Eds.) *Proceedings of the 22nd Meeting of the Cognitive Science Society* (pp. 723–728). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1975). Pragmatic patterns in child syntax. *Stanford Papers and Reports on Child Language Development*, *10*, 153-165.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development, 43*, 1-123.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Maslen, R. J., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research*, *47*, 1319.

Mattys, S. L., & Jusczyk, P. W. (2000). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91-121.

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

McCauley, S.M. & Christiansen, M.H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1619-1624). Austin, TX: Cognitive Science Society.

McCauley, S. M. & Christiansen, M. H. (submitted). Language learning as language use: A computational model of children's language comprehension and production.

Monaghan, P., Chater, N. & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96,* 143-182.

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

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81-97.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological review*, *115*, 357-395.

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

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

Punyakanok, V., & Roth, D. (2001). The use of classifiers in sequential inference. In *Proceedings of NIPS 2001* (pp. 995-1001).

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*, 172–196.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606-621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101-105.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*, 382–386.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. Proceedings of the ACL SIGDAT-Workshop, March 1995.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, *102*, 11629-11634.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1-42.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, US: Harvard University Press.

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworth.

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

Table 1

The twenty most highly activated items in the PUDDLE chunk inventory after 1000 and 10,000 utterances during the course of the dense corpus simulation

1000 Utterances	**10000 Utterances**
no	*the*
oh	*a*
this	*no*
dear	*oh*
and	*and*
hat	*you*
thomas	*this*
it	*that*
what	*ee*
grapes	*it*
ahh	*to*
there	*are*
oops	*on*
where	*dear*
two	*what*
ooh	*thomas*
look	*two*
a_hat	*is*
blue	*what's*
what's_this	*there*

Table 2

The twenty most highly activated multiword items in the CBL chunk inventory after 1000 and 10,000 utterances during the course of the dense corpus simulation. Bold text denotes a word already discovered by the PUDDLE model at the corresponding time point, while asterisks denote PUDDLE's storage of an identical multiword chunk at the corresponding timepoint.

1000 Utterances	**10000 Utterances**
oh dear	***oh dear***
*is it**	***what's this***
the *giraffe**	***i*** *think*
*a hat**	***you like***
i ***think***	***look at***
the *basket*	***that's*** *right*
the *steps*	***is it***
the lion	***all*** *done*
what does	***and there's***
what do	***oh dear dear***
are they*	***this*** *morning*
its ***gone***	***and then***
you like	***what does***
and ***there's***	***going to***
the wind	***the bus***
what ***about under***	***a ride***
the boat	***what do***
on ***mummy's head***	***the door***
*that's thomas**	***a hat***
your ***birthday cards***	*as* ***well***

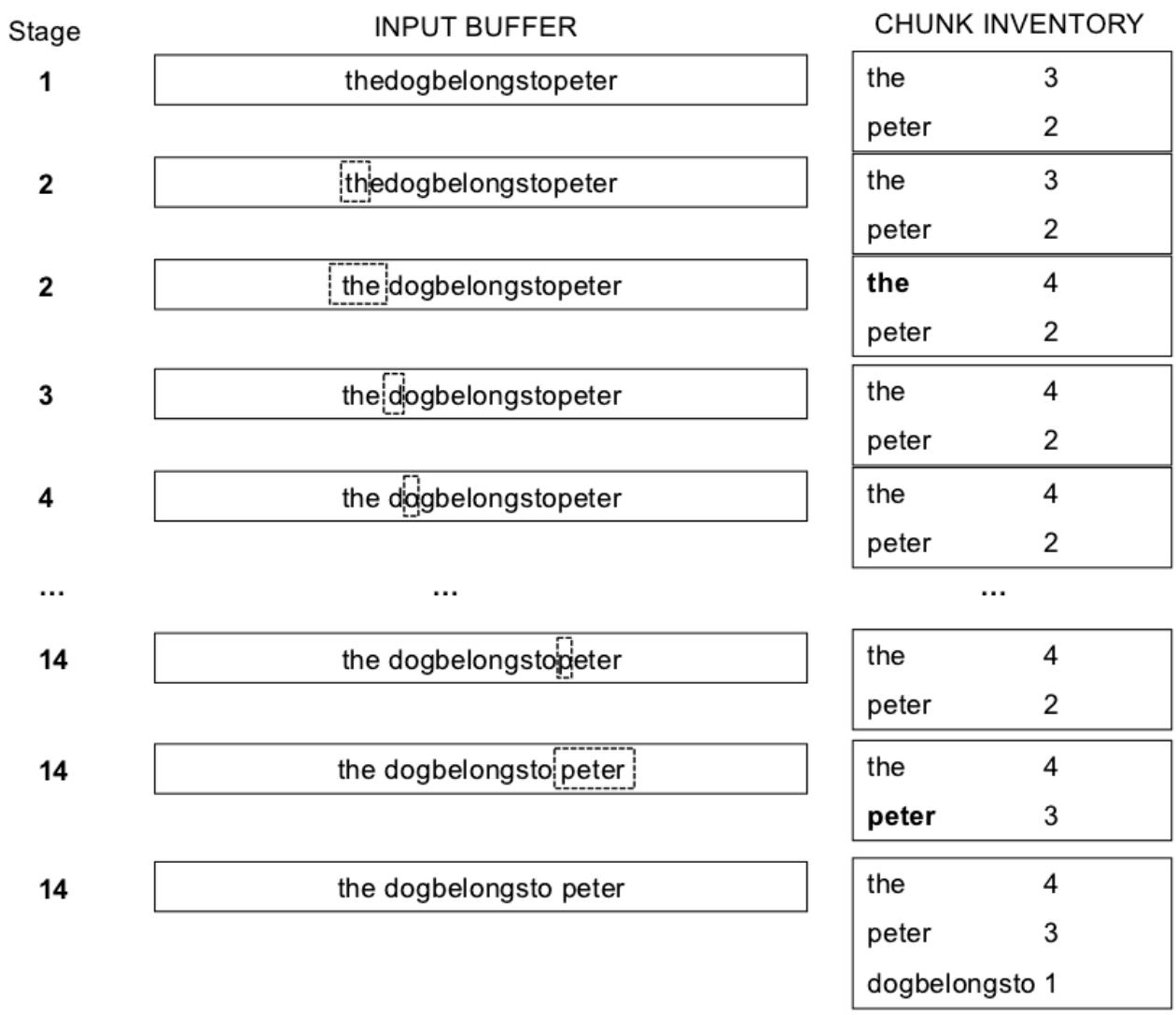

Fig. 1: The stages of segmentation for one utterance in the PUDDLE model of speech segmentation

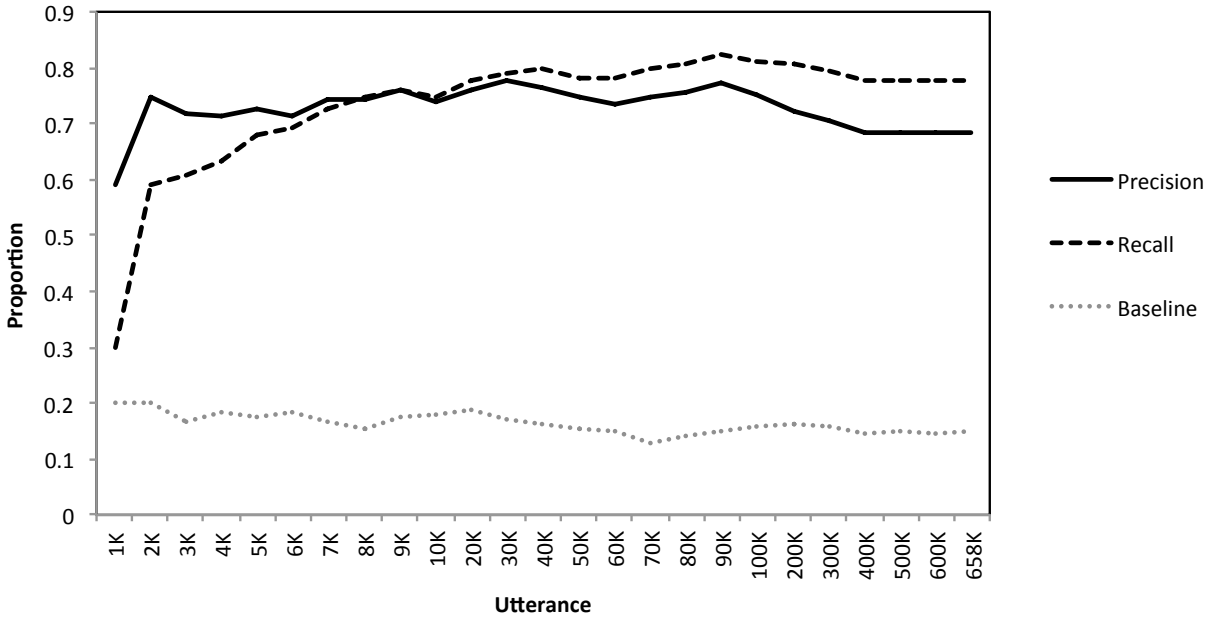

Figure 2. Performance of the PUDDLE model throughout the dense corpus simulation

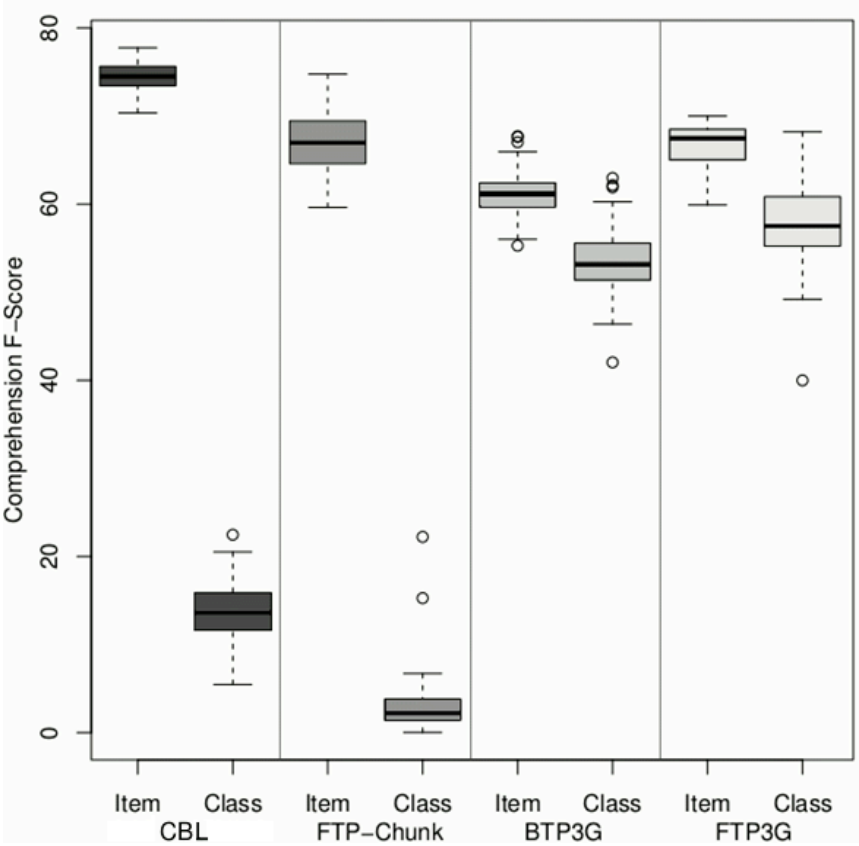

Fig. 3: Boxplots depicting comprehension performance (F-scores) for the CBL model and its baselines, comparing item- vs. class-based simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within the 1.5 IQR of the quartiles, while dots depict outliers.

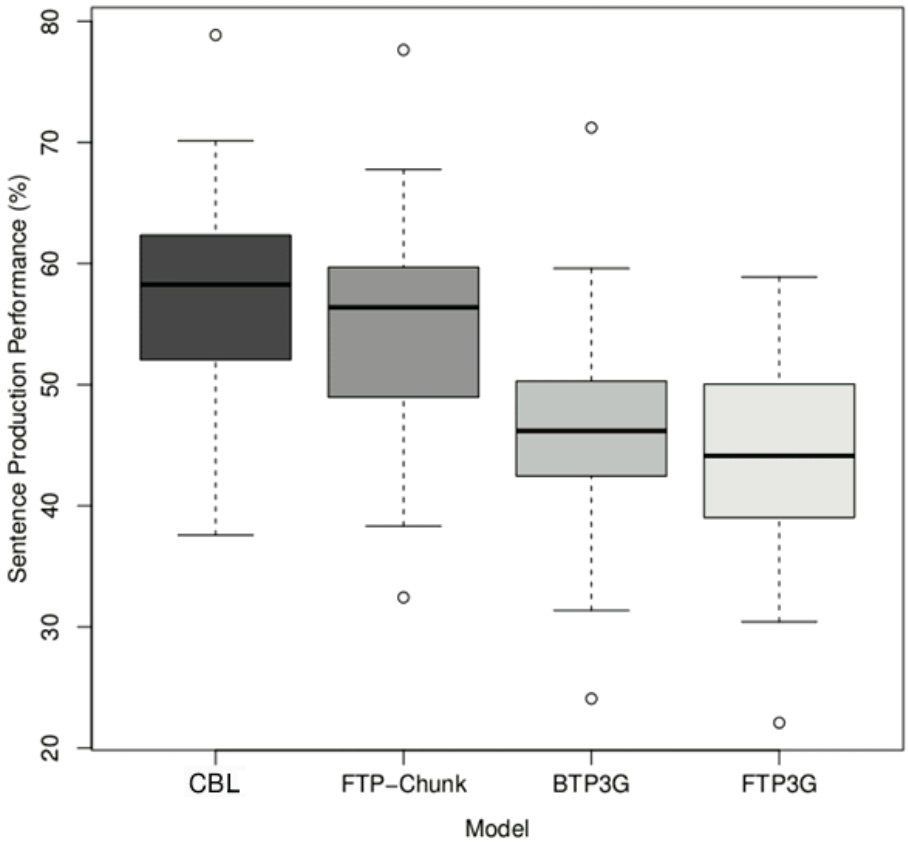

Fig. 4: Boxplots depicting Sentence Production Performance scores for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within the 1.5 IQR of the quartiles, while dots depict outliers.