LANGUAGE LEARNING AS LANGUAGE USE: STATISTICALLY-BASED CHUNKING IN DEVELOPMENT

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Stewart M. McCauley

January, 2017

LANGUAGE LEARNING AS LANGUAGE USE: STATISTICALLY-BASED CHUNKING IN DEVELOPMENT

Stewart M. McCauley, Ph. D.
Cornell University 2017

While usage-based approaches to language development enjoy considerable support from computational studies, there have been few attempts to answer a key computational challenge posed by usage-based theory: the successful modeling of language learning as language *use*. I present a usage-based computational model of language acquisition which learns in a purely incremental fashion, through on-line processing based on chunking, and which offers broad, cross-linguistic coverage while uniting comprehension and production processes within a single framework. The model's design reflects memory constraints imposed by the real-time nature of language processing, and is inspired by psycholinguistic evidence for children's sensitivity to the distributional properties of multi-word sequences and for shallow language comprehension based on local information. It learns from corpora of child-directed speech, chunking incoming words together to incrementally build an item-based "shallow parse." When the model encounters an utterance made by the target child, it attempts to generate an identical utterance using the same chunks and statistics involved during comprehension. In Chapter 2, I show that the model achieves high performance across over 200 single-child corpora representing 29 languages from the CHILDES database. It also succeeds in capturing findings from children's production of complex sentence types. In Chapter 3, I show that the model captures key developmental psycholinguistic findings on children’s language learning and use. Chapter 4 investigates the use of the model for understanding the different outcomes of child first-language learning versus second-language learning in adults, providing evidence that adult learners may rely on more fine-grained linguistic units. Together, the modeling results presented in this dissertation suggest that much of children's early linguistic behavior may be accounted for by item-based learning through on-line processing of simple distributional cues, consistent with the notion that acquisition can be understood as learning to process language.

BIOGRAPHICAL SKETCH

Stewart McCauley graduated from the University of Texas at Austin in 2007 with BAs in English and Archaeological Studies. In 2009, he received an MA in Linguistics and Cognitive Science from the University of Delaware. He completed his PhD in Psychology at Cornell University in 2016 and will join the Institute of Psychology, Health, and Society at the University of Liverpool as a post-doctoral researcher.

ACKNOWLEDGMENTS

Thanks first and foremost to my committee for being patient with me. Thanks to Morten Christiansen for guidance as my advisor, and for his incredible generosity as a collaborator. Most of the core ideas driving the studies in this dissertation came from Morten. Thanks to Shimon Edelman for sharing key insights, for introducing me to reinforcement learning, and for challenging me to explore new perspectives on language learning. His influence will surely be apparent in my future work. Thanks to Mike Goldstein for his enthusiasm and encouragement, his unique perspectives on development, and for opening my eyes to the importance of social cues in language learning. I hope to incorporate insights gained from Mike into my future research. Thanks to Barb Finlay for always bringing a fresh perspective, for asking challenging, thoughtful questions, and for encouraging me to explore fields outside my own. The range of her knowledge of psychology and related fields is truly something to aspire to.

Thanks also to collaborators past, present, and future: Erin Isbilen, Fabio Trecca, Melody Dye, Michael Ramscar, Inbal Arnon, Padraic Monaghan, and Nick Chater. Thanks to Barend Beekhuizen, Reza Shahbazi, Afra Alishahi, Marytn Lloyd-Kelly, Colin Bannard, Dan Freudenthal, Ben Ambridge, and Julian Pine for fruitful discussion on language development and computational modeling.

Thanks also to Holger Diessel, Pierre Perruchet, and Jenny Saffran for sharing additional details and data from their studies which were used in this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

TOWARDS A USAGE-BASED MODEL OF ON-LINE LANGUAGE LEARNING

By four years of age, most children have mastered the basic grammatical structures of their native language, an achievement marking the transition to a seemingly unbounded capacity for communicating novel information. How can children acquire such productivity—in the span of a few short years—given input that is both noisy and necessarily finite in nature? For over half a century, the dominant theoretical approach to this question has assumed that linguistic productivity can only be achieved through a system of abstract grammatical rules working over word classes, scaffolded by a considerable amount of innate knowledge (e.g., Chomsky, 1957). This has primarily taken the form of a *words-and-rules* approach, which posits a strict separation between grammar and the lexicon (e.g., Fodor, 1983; Pinker, 1999). Language is viewed as memory-based at the word level, but algorithmic at the multi-word level: syntactic rules define the skeletal structure of a sentence, which is fleshed out by lexical-semantic representations (e.g., Radford, 1988). At its most extreme, the words-and-rules approach takes the form of a model in which the lexicon contains only those units which cannot be computed from their component parts—morphemes, simple words, irregular words, and idioms—while grammar consists of operations for the hierarchical composition of lexical elements into complex words, phrases, and sentences (e.g., Pinker, 1991; Pinker, 1999).

In recent decades, linguists have proposed a number of theoretical alternatives to the words-and-rules approach which hold that grammar and the lexicon form a single system: grammar is taken to consist in stored units, similar to simpler symbols, such as words and morphemes (e.g., Croft, 2001; Fillmore, 1985; Goldberg, 1995, 2006; Langacker, 1987, 2008). What these approaches hold in common is the notion of the inseparability of the lexicon from grammar: the two can only be distinguished in so far as they constitute polar ends of a continuum of linguistic units ranging from simple to more complex. The mapping between form and meaning composing a more complex unit, such as an abstract grammatical construction (e.g., Goldberg, 2006), is not qualitatively different than the mapping between a word's meaning and its phonological form. Although this can be argued to involve a hierarchical system with rule-like properties, grammar itself under this approach is inherently meaningful and thus cannot be reduced to the type of operations over separately stored elements assumed by words-and-rules models. In other words, grammar cannot be dissociated from the lexicon.

It is with this latter set of theoretical approaches to language acquisition—which necessarily hold that language learning is *usage-based* and arises from actual usage events—that this dissertation is concerned. The studies described herein involve the development of a computational model of language acquisition which takes usage-based approaches to their natural conclusion, eliminating the sharp distinction between language learning and language processing. The model—dubbed the Chunk-Based Learner (CBL) because it operates primarily through a simple chunking mechanism—takes real corpora of child-directed speech as input and builds linguistic knowledge solely through attempts to perform concrete aspects of comprehension and production. The model provides strong performance across a typologically diverse array of 29 Old

World languages. Importantly, the model achieves this through incremental, on-line processing that adheres to psychological principles and constraints deriving from the real-time nature of language use.

In what follows, I discuss the theoretical motivation for the CBL model's emphasis on chunking: evidence that multiword patterns of varying sizes function as linguistic units in their own right. I then discuss some of the practical constraints on the psychological mechanisms for acquiring these useful chunks of linguistic information. Finally, I provide an overview of the dissertation and a breakdown of the following chapters.

A Unified Grammar and Lexicon

The studies described in this dissertation are motivated in part by a number of psycholinguistic findings from recent decades which have served to blur the lines between lexicon and grammar, consistent with the aforementioned usage-based approaches to language. These findings not only cast doubt on the existence of separate and distinct learning and processing mechanisms, but highlight a role for multiword linguistic units. A number of these findings center on the time course of lexical vs. grammatical development. Bates and Goodman (1997) provide a review of longitudinal studies demonstrating the connection between grammatical abilities and vocabulary size in both typically and atypically developing children. For instance, one early study identifies a strong correlation between grammatical proficiency and vocabulary size, with vocabulary at 20 months emerging as the best predictor of syntactic abilities at 28 months (Bates, Bretherton, & Snyder, 1988). This general pattern is replicated and extended in a number of further studies (e.g., Caselli, Casadio, & Bates, 1999; Fenson et al., 1994; Singer-Harris, Bellugi, Bates, Rossen, & Jones, 1998).

The mechanisms whereby lexical and grammatical abilities might emerge in unison are beginning to be explored in studies inspired by the usage-based approach to language development (Lieven, Pine, & Baldwin, 1997; Tomasello, 1992, 2000a, 2000b, 2003), which builds upon previous lexically-oriented theories of grammatical development (e.g., Bowerman, 1976; Braine, 1976) and is largely consistent with previous theoretical alternatives to dual-system models, proposed by linguists (as discussed above; e.g., Fillmore, 1985; Langacker, 1987). Such work has emphasized the item-based nature of children's early grammatical knowledge, yielding cross-linguistic findings of item-specific patterns in early verb usage (e.g., Gathercole, Sebastián, & Soto, 1999; Rubino & Pine, 1998; Tomasello, 1992; see also: Berman, 1982; MacWhinney, 1975; Pizutto & Caselli, 1992), as well as studies of children's production of novel verbs (e.g., Tomasello & Brooks, 1998; Akhtar, 1999), use of determiners (e.g., Mariscal, 2008; Pine & Lieven, 1997), case marking errors (e.g., Kirjavainen, Theakston, & Lieven, 2009), production of complex sentence types (e.g., Diessel & Tomasello, 2005), and question formation (e.g., Dabrowska, 2000). Such widespread developmental findings of item-specific usage patterns dovetail nicely with the intertwined developmental trajectories of lexical and grammatical abilities identified by Bates and colleagues.

A number of recent findings suggest that children's item-based linguistic behavior is rooted in the formation of representations of differing granularities, with stored linguistic units ranging from the fine-grained level of morphemes and words to the more coarse-grained level of word sequences comprising one or more phrases. This notion has received support from naturalistic observation (e.g., Lieven, Behrens, Speares, & Tomasello, 2003; Peters, 1983) and has recently been the supported by empirical work. The finding of Bannard and Matthews (2008) that young children repeat phrases faster and more accurately when they form a frequent chunk may have provided the first direct evidence not only that multiword chunk storage takes place, but that this storage can actively facilitate processing. Controlling for substring frequency, they contrasted repetition of four-word phrases in which the fourth word was of either high or low frequency, given the preceding trigram. Two and 3-year-olds were more likely to repeat a phrase correctly when its fourth word combined with the preceding trigram to form a frequent chunk, while 3-year-olds were significantly faster to repeat the first three words. Further evidence comes from children's production of irregular plurals: Arnon and Clark (2011) found that the overregularization errors are significantly reduced when irregular plurals are produced in the context of lexically-specific frames (e.g., "*brush your* teeth").

With respect to multiword linguistic units, usage-based approaches have focused primarily on the importance of stored sequences as exemplars in the abstraction of schemas and grammatical regularities; at the heart of usage-based theory lies the idea that productivity develops gradually through abstraction over multiword sequences of the sort identified in the above studies (e.g., Abbot-Smith & Tomasello, 2006). Nevertheless, children's apparent use of multiword units during on-line processing (as shown by both Arnon & Clark, 2011 and Bannard & Matthews, 2008) highlights an active role for such units in comprehension and production, suggesting the possibility that multiword sequences retain their significance throughout development. Such a finding would blur the lines between lexicon and grammar even further, presenting an even worse problem for words-and-rules approaches.

Indeed, this seems to be the case: a number of findings indicate that the storage and active use of multiword units persists beyond early acquisition and into adulthood. Bannard and Ramscar (2007) describe an effect of overall sequence frequency on reading times for units ranging from 4 to 7 words in length, while Reali and Christiansen (2007) showed chunk frequency effects in the processing of complex sentence types. Arnon and Snider (2010) found the same general pattern using a phrasal-decision task, whereby four-word expressions were classified as possible or impossible strings in English (in a vein similar to lexical-decision tasks). Importantly, Arnon and Snider's study explored multiple frequency bins; reaction times decreased as a function of phrase frequency. Caldwell-Harris, Berant, and Edelman (2012) extended this finding to a broader frequency spectrum, showing a continuous effect of frequency and providing evidence against a “threshold” beyond which frequency effects are exhibited. Additional evidence for adults' sensitivity to multiword sequence frequency has been gained from self-paced reading and sentence recall tasks (Tremblay, Derwing, Libben, & Westbury, 2011), eye-tracking data (Siyanova-Chanturia, Conklin, & van Hueven, 2011), and event-related brain potentials (Tremblay & Baayen, 2010). A similar pattern of results has been found in studies of adult production, demonstrating a decrease in naming latencies with increasing phrase frequency (Janssen & Barber, 2012) as well as reduced phonetic duration for frequent multiword sequences in elicited and spontaneous speech (Arnon & Cohen Priva, 2013).

While the above studies have focused primarily on distributional properties, there are likely to be additional semantic contributions to the ways in which language users represent and draw upon multiword units. For instance, Jolsvai, McCauley, and Christiansen (2013) investigated the processing of compositional multiword sequences alongside idioms which were matched for whole- and sub-string frequency. In a paradigm identical to that of Arnon and Snider (2010), subjects exhibited indistinguishable processing latencies for the idioms and compositional sequences. More importantly, norming scores collected from a separate set of subjects, who were asked to rate strings based on how meaningful they were "as a single unit," were a better predictor of reaction times than frequency or compositionality, for the entire set of test items. These results support the notion that multiword sequences are stored and processed in ways that relate as much to their semantics as to their distributional features, further undermining the exclusion of compositional forms from the lexicon.

There are also direct parallels between the learning and processing of multiword units and individual words with respect to age-of-acquisition (AoA) effects. In a variety of tasks, adults exhibit processing advantages for words that are acquired earlier in childhood. Arnon, McCauley, and Christiansen (2017) show that multiword sequences, like individual words, display AoA effects when AoA is determined using either corpus-based metrics or subjective AoA ratings. The authors also show that the effect cannot be reduced to frequency, semantic plausibility, or lexical AoA. By underscoring this further parallel between words and multiword patterns, this study builds strong support the notion of coarse linguistic units as key building blocks for language learning and use.

Thus, the importance of multiword linguistic units extends beyond merely serving as exemplars for the formation of item-based schemas or the abstraction of grammatical regularities; multiword sequences play an active role in on-line processing, one that persists into adulthood. These findings clearly require the expansion of the mental lexicon in dual-system models to contain compositional strings. Indeed, a small number of approaches in generative linguistics have incorporated storage for larger compositional sequences (e.g., Culicover & Jackendoff, 2005). Nevertheless, the psycholinguistic findings described above indicate that compositional forms are not stored merely as exceptions (i.e., when falling above a certain frequency threshold), but by default, which casts doubt on a sharp boundary between lexicon and grammar.

Shallow Processing Based on Local Information

The item-based nature of children's and adults' processing of multiword strings nicely complements a body of work suggesting that sentence processing is often shallow with respect to syntactic structure, relying heavily on semantic knowledge and global discourse structure. Such findings provide another key motivation for the studies presented in this dissertation. Evidence for shallow processing of linguistic input has lead some researchers to question the centrality of rule-based structures of the sort required by dual-system approaches, as well as the standard assumption that syntactic processes are carried out completely and automatically. Shallow processing has been shown to be a widespread phenomenon through psycholinguistic research (Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002), and while the vast majority of this work has dealt with adult subjects, the theoretical implications extend from adult processing to the study of language acquisition.

What is perhaps the most well-known thread of evidence for shallow processing comes from work on text-change blindness (following work on change blindness in visual processing; e.g., Simons & Levin, 1998) to demonstrate the extent to which several factors modulate depth of processing, including focus (Sanford, 2002; Sturt, Sanford, Stewart, & Dawydiak, 2004) and computational load (Sanford, Sanford, Filik, and Molle, 2005). Perhaps more relevant is work demonstrating subjects' interpretation of nonsensical sentences as coherent (Fillenbaum, 1974; Wason & Reich, 1979) as well as the processing of semantically anomalous sentences in ways that directly contradict the interpretations that would be made according to a full syntactic parse (Ferreira, 2003), demonstrating the on-line use of background world knowledge and pragmatic expectations.

Evidence for shallow processing makes good contact with Ferreira and Patson's (2007) Good Enough approach to sentence processing, in which it is argued that the goal of language comprehension is to establish representations which are merely “good enough” to suit the needs of a listener or reader in a given situation, as opposed to representing communicated information in full detail. In the above-mentioned cases, readers seem to rely on local linguistic information and global background knowledge rather than compositional meanings derived from fully articulated syntactic representations. Thus, evidence for shallow processing based on local information also makes close contact with Sanford and Garrod's (1981; 1998) Scenario Mapping theory of comprehension, in which background knowledge of situations and scenarios is used on-line to interpret linguistic input as it is encountered. During on-line interpretation, incoming linguistic input is mapped onto schemas of events, situations, or scenarios which have been established based on previous contexts or input; interpretation of the overall message is therefore heavily influenced by the background information which linguistic input is mapped onto. While the computational modeling work described in this dissertation does not yet incorporate semantic representations, it is compatible with the aforementioned perspectives on shallow processing.

The Real-time Constraint on Language Learning and Processing

Evidence for multiword linguistic units and the ubiquity of shallow processing over local information converge on a recent theoretical proposal by Christiansen and Chater (2016; see also Chater, McCauley, & Christiansen, 2016), earlier forms of which provided much of the inspiration for the studies presented in this dissertation. The proposal rests on the uncontroversial observation that language takes place in the "here and now." The consequences of this real-time constraint—which Christiansen and Chater refer to as the "Now-or-Never bottleneck"—are rarely considered, however.

The fleeting nature of signal and memory presents something of an enigma: At a normal rate of speech, humans produce between 10 and 15 phonemes per second (Studdert-Kennedy, 1986). Nevertheless, the ability to process discrete sounds seems limited to about 10 items per second (Miller and Taylor, 1948). To make matters worse, the auditory trace is limited to about 100 ms (Remez, Ferro, Dubowski, Meer, Broder, & Davids, 2010). Moreover, memory for arbitrary sequences seems to be limited to about four items (Cowan, 2001; Warren, Obusek, Farmer, & Warren, 1969).

Thus, the signal—and human memory for it—are incredibly short-lived. If all of these findings are taken at face value, language learning should be impossible. A key strategy for overcoming these sensory and memory limitations lies in *chunking*: incoming items can be rapidly grouped and passed to successively higher levels of representation, with higher-level representations allowing input to be dealt with before it is overwritten by the onslaught of incoming information at a lower level. It is fairly intuitive and uncontroversial, for instance, that the raw acoustic signal is rapidly packaged into phoneme- or syllable-like representations, which can in turn be chunked into word- and phrase-like representations, and so on. The consequences of applying this general approach to sentence-level processing and grammatical development are, however, less obvious, as discussed by Christiansen and Chater (2016).

While receiving new emphasis under this perspective, chunking has been regarded as a key learning and memory mechanism in human cognition for over half a century (e.g., Feigenbaum & Simon, 1962; Miller, 1956; Simon, 1974). While verbal theories have been more common, computational models of chunking have been present in the literature for over four decades (e.g., Ellis, 1973; French, Addyman, & Mareschal, 2011; Perruchet & Vintner, 1998; Servan-Schreiber & Anderson, 1990; Simon & Gilmartin, 1973). Previous computational approaches to chunking have had a significant impact on approaches to language development, particularly with respect to the area of speech segmentation (e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010). The CBL model extends the use of chunking beyond word segmentation to aspects of sentence comprehension and production, while stopping short of capturing grammatical abstraction.

Modeling Language Learning as Language Use

The modeling approach described in this dissertation attempts to satisfy the constraints imposed by the real-time nature of language: all processing (and therefore, all learning) takes place in a fully incremental and on-line fashion. This is made possible by the reliance on chunking and the use of shallow linguistic processing of local information tied to multiword sequences as opposed to fully hierarchical representations of structure. In **Chapter 2**, the CBL model is presented in full detail. In this chapter, I explore the capabilities and limitations CBL's fully item-based approach of language development. Importantly, there is no distinction between learning and processing in the model: language learning *is* learning to process language. The model's approach to chunking is explored in the context of ongoing debates in statistical learning over recognition-based versus probabilistic processing of sequences. The CBL model is shown to compare favorably to purely recognition-based models of chunking as well as purely statistically-based approaches, through the use of statistical cues in a recognition-based framework. Despite its extreme simplicity, the model is shown to be capable of strong performance for specific aspects of comprehension and production across a typologically-diverse array of 29 Old World languages. By taking usage-based approaches to the extreme in this way, I explore the use of the model as a foundation for a future approach also capable of learning abstract linguistic units, in the vein of computationally more complex models of grammatical development which acquire partially abstract, productive constructions (e.g., Solan, Horn, Ruppin, & Edelman, 2005).

While Chapter 2 shows that CBL can discover and use building blocks for language learning, in **Chapter 3** the psychological validity of these building blocks is explored. I explore simulations of empirical data covering several key developmental psycholinguistic findings regarding children's distributional and item-based learning. It is shown that CBL can capture developmental data from a range of findings spanning from child artificial language learning (Saffran, 2002) to child sensitivity to multiword sequence frequency (Bannard & Matthews, 2008) to morphological development (Arnon & Clark, 2011). These results are discussed in the context of their significance for understanding formulaic language use over the course of development, spanning into adulthood.

In **Chapter 4**, I explore the notion that children rely more heavily on multiword units in language learning than do adults learning a second language. To this end, I take an initial step towards using large-scale, corpus-based computational modeling as a tool for exploring differences between the linguistic units of different learner types. The CBL model is used to compare the usefulness of chunk-based knowledge in accounting for the speech of second-language learners vs. children and adults speaking their first language. In the same vein as the TraceBack method (Lieven et al., 2003), we compare the CBL model's ability to use chunks discovered in the speech of single learners to generalize to the on-line production of unseen utterances from the same learners. This modeling effort thus aims to provide the kind of "rigorous computational evaluation" of the Traceback Method called for by Kol, Nir, & Wintner (2014). Moreover, I explore the nature of the chunk inventories acquired by the model for each learner type, using a network-theoretic approach. Together, these findings suggest that while multiword units are likely to play an important role in second-language learning, adults may learn less useful chunks, rely on them to a lesser extent, and arrive at them through different means than children learning a first language.

In **Chapter 5**, I discuss the limitations and future directions for the modeling approach outlined in this dissertation. In particular, the need for incorporating extra-linguistic input and semantic representations is explored. I also discuss successful attempts to test the predictions of the modeling work through behavioral studies conducted with human adults.

REFERENCES

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language, 26,* 261–278.

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. & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56, 349-371.*

Arnon, I., McCauley, S.M. & Christiansen, M.H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. Journal of Memory and Language.

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

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

Bannard, C., & Ramscar, M. (2007). Reading time evidence for storage of frequent multiword sequences. Abstract in *Proceedings of the Architectures and Mechanism of Language Processing Conference (AMLAP-2007), Turku, Finland*.

Bates, E., Bretherton, I., & Snyder, L. S. (1991). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge, MA: Cambridge University Press.

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia, and real-time processing. *Language and Cognitive Processes*, *12*, 507-584.

Berman, R. (1982). Verb-pattern alternation: The interface of morphology, syntax, and semantics in Hebrew child language. *Journal of Child Language, 9*, 169–191.

Bowerman, M. (1976). Semantic factors in the acquisition of rules for word use and sentence construction. In *Directions in normal and deficient language development* (pp. 99-179). University Park Press.

Braine, M. D. (1976). Children’s first word combinations. Monographs of the Society for Research in Child Development, 41, 104.

Caldwell-Harris, C.L., Berant, J.B., & Edelman, S. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In S. T. Gries & D. Divjak (Eds.), *Frequency effects in cognitive linguistics (Vol. 1): Statistical effects in learnability, processing and change.* The Hague, The Netherlands: De Gruyter Mouton.

Caselli, M.C., Casadio, P. & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language, 26*, 69–111.

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244-254.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. New York: Oxford University

Press.

Dabrowska, E. (2000). From formula to schema: The acquisition of English questions. *Cognitive Linguistics, 11*, 83-102.

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

Ellis S. H. (1973). Structure and Experience in the Matching and Reproduction of Chess Patterns. Doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Feigenbaum, E. A., & Simon, H. A. (1962). A theory of the serial position effect. *British Journal of Psychology*, *53*, 307-320.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... & Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, 1-185.

Ferreira, F. (2003). The misinterpretation of non-canonical sentences. *Cognitive Psychology*, *47*, 164–203.

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

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*, 574-578.

Fillmore, C. (1985). Syntactic intrusions and the notion of grammatical construction. In Mary Niepokuj et al. (Eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society* (pp. 73-86). Berkeley: Berkeley Linguistics Society.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: The MIT Press.

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

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614.

Gathercole, V., Sebastian, E., & Soto, P. (1999). The early acquisition of Spanish verbal morphology: Across-the-board or piecemeal knowledge? *International Journal of Bilingualism, 3*, 138–182.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* New York: Oxford University Press.

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

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kirjavainen, M., Theakston, A., & Lieven, E. (2009). Can input explain children's me-for-I errors? *Journal of Child Language, 36*, 1091-1114.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language*, *41*, 176-199.

Langacker, R. (1987). *The foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Palo Alto: Stanford University Press.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333–370.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187-219.

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

Mariscal, S. (2008). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *Journal of Child Language, 35,* 1-29.

Miller, G. A., & Taylor, W. G. (1948). The perception of repeated bursts of noise. *The Journal of the Acoustical Society of America*, *20*, 171-182.

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

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

Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Pine, J. M., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123-138.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530-535.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Harper Collins.

Pizutto, E. and Caselli, C. (1992). The acquisition of Italian morphology. *Journal of Child Language, 19*, 491–557.

Radford, A. (1988). *Transformational grammar: A first course* (Vol. 1). Cambridge: Cambridge University Press.

Reali, F. & Christiansen, M.H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60,* 161-170.

Remez, R. E., Ferro, D. F., Dubowski, K. R., Meer, J., Broder, R. S., & Davids, M. L. (2010). Is desynchrony tolerance adaptable in the perceptual organization of speech?. *Attention, Perception, & Psychophysics*, *72*, 2054-2058.

Rubino, R. and Pine, J. (1998) Subject–verb agreement in Brazilian Portuguese: What low error rates hide. *Journal of Child Language, 25*, 35–60.

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

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations of comprehension beyond the sentence*. New York: Wiley.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, *26*, 159–190.

Sanford, A. J. S., Sanford, A. J., Filik, R., & Molle, J. (2005). Depth of lexical-semantic processing and sentential load. *Journal of Memory and Language*, *53*, 378–396.

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

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592.

Simon, H. A. (1974). How big is a chunk?. *Science*, *183*, 482-488.

Simon H. A. & Gilmartin K. J. (1973). A simulation of memory for chess positions. Cognitive Psychology, 5, 29-46.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*, 644–649.

Singer-Harris, N., Bellugi, U., Bates, E., Jones, W., & Rossen, M. (1997). Contrasting profiles of language development in children with Williams and Down Syndromes. In D. Thal & J. Reilly, (Eds.), Special issue on origins of communication disorders: *Developmental Neuropsychology, 13*, 345-370.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*, 251-272.

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

Studdert-Kennedy, M. (1986). Some developments in research on language behavior. *Behavioral and Social Science: 50 Years of Discovery*, 208.

Sturt, P., Sanford, A. J., Stewart, A., & Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, *11*, 882–888.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge, MA: Cambridge University Press.

Tomasello, M. (2000a). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*, 61-82.

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

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

Tomasello, M. and Brooks, P. (1998) Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics, 9,* 379–395.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.) *Perspectives on formulaic language: Acquisition and communication* (pp. 151-173). London: Continuum International Publishing Group.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing Advantages of Lexical Bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*, 569-613.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science*, *164*, 586-587.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*, 591–597.

CHAPTER 2

A CROSS-LINGUISTIC MODEL OF CHILD LANGUAGE DEVELOPMENT[1]

The ability to comprehend and produce an unbounded number of novel utterances has long been regarded as a hallmark of human language. How does a child acquire such productivity, given input that is both noisy and finite? For over half a century, generative linguists have argued that such open-endedness can only be explained by a system of abstract grammatical rules operating over word classes, scaffolded by innate, language-specific knowledge (e.g., Chomsky, 1957; Pinker, 1999). In recent years, however, an alternative theoretical perspective has emerged in the form of usage-based approaches (e.g., Croft, 2001; Goldberg, 2006; Tomasello, 2003), which hold that children's language development is initially item-based. Rather than being guided by system-wide abstract principles, productivity is taken to emerge gradually, beginning with concrete items in the child's input. This perspective is motivated in part by analyses of child-directed speech, showing that there is considerably more information available in the input than previously assumed (e.g., Redington, Chater, & Finch, 1998; Monaghan & Christiansen, 2008), as well as a wide range of observational and empirical work showing that children can use such information in an item-based manner. Such evidence includes cross-linguistic findings of item-specific patterns in early verb usage (e.g., Berman, 1982; MacWhinney, 1975; Gathercole, Sebastián, & Soto, 1999; Pizutto & Caselli, 1992; Rubino & Pine, 1998),

[1] This chapter is under review as: McCauley & Christiansen (submitted), Language learning as language use: A cross-linguistic model of child language development.

as well as studies of children's production of novel verbs (e.g., Tomasello & Brooks, 1998; Akhtar, 1999), use of determiners (e.g., Mariscal, 2008; Pine & Lieven, 1997), case marking errors (e.g., Kirjavainen, Theakston, & Lieven, 2009), production of complex sentence types (e.g., Diessel & Tomasello, 2005), and question formation (e.g., Dabrowska, 2000).

In addition to this wealth of observational and empirical evidence, a number of computational modeling studies have provided a source of complementary support for usage-based approaches, using item-based learning to successfully capture specific developmental patterns (Freudenthal, Pine, & Gobet, 2006, 2007; Gobet, Freudenthal, & Pine, 2004; Jones, Gobet, & Pine, 2000), the acquisition of item-based constructions and schemas (Chang, 2008; Solan, Horn, Ruppin, & Edelman, 2005), and semantic role learning (e.g., Alishahi & Stevenson, 2010), as well as tracing the emerging complexity of children's grammatical knowledge more generally (e.g., Bannard, Lieven, & Tomasello, 2009; Borensztajn, Zuidema, & Bod, 2009).

Despite the considerable success of item-based computational approaches to acquisition, there have been few computational accounts of the on-line processes driving children's attempts to comprehend and produce speech, or the ways in which these specific usage events incrementally contribute to the child's emerging linguistic abilities. This lack seems to stem, in part, from traditional ways of idealizing the task facing the language learner: from a computational standpoint, the issue of linguistic productivity tends to be approached primarily as a problem of grammar induction; to attain open-ended productivity, the learner must first identify a target grammar on the basis of exposure to a sample of sentences generated by that grammar (Gold, 1967).

While computational approaches to acquisition have largely moved beyond Gold's formal learnability approach, incorporating a variety of different sources of linguistic information, the idealization of the task facing the learner as one of grammar induction has remained largely intact. As a consequence, computational work within the usage-based tradition has continued to focus on grammar induction (e.g., Borensztajn et al., 2009).

Usage-based theory suggests the possibility of sidestepping the grammar induction approach altogether, focusing instead on the ways in which linguistic knowledge is built up and reinforced through specific usage events (the child's attempts to comprehend and produce speech). This perspective has recently been bolstered by a number of complementary experimental results which suggest that the task facing learners is better characterized as one of "learning by doing" than as one of grammar induction (see also Chater & Christiansen, 2010; Christiansen & Chater, 2016a). Evidence for the psychological reality of multiword linguistic units has served to blur the lines between grammar and lexicon, demonstrating the storage of "compositional" phrases as well as their use in comprehension and production (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008). Moreover, work on associative learning (e.g., Perruchet, Vinter, Pacteau, & Gallego, 2002) and statistical learning (e.g., Thompson & Newport, 2007) suggests that computationally simple mechanisms may be sufficient to identify the boundaries of such units in the speech stream.

A highly relevant—though previously unconnected—line of research has focused on the issue of syntactic processing depth, providing evidence that comprehension processes are often shallow and underspecified (e.g., Sanford & Sturt, 2002). Taken together with evidence for the primacy of local information during processing (e.g., Tabor, Galantucci, & Richardson, 2004), this suggests that children and adults form representations which are merely “good enough” for the communication task at hand (e.g., Ferreira & Patson, 2007). Evidence for multiword linguistic units, shallow processing, and the use of local information makes contact with other work emphasizing the importance of sequential as opposed to hierarchical linguistic structure (e.g., Frank & Bod, 2011; see Frank, Bod, & Christiansen, 2012, for a review).

Despite the importance of these complementary areas of research for strengthening item-based approaches, as well as their implications for re-characterizing the task facing language learners, they have remained largely unconnected. A recent theoretical proposal by Christiansen and Chater (2016b) unites these seemingly disparate strands of evidence. The proposal rests on the uncontroversial acknowledgement that language takes place in the “here and now.” The consequences of this real-time constraint—which Christiansen and Chater refer to as the “Now-or-Never bottleneck”—are rarely considered, however. The fleeting nature of signal and memory have implications for how we approach human language: At a normal rate of speech, humans produce between 10 and 15 phonemes per second (Studdert-Kennedy, 1986). Nevertheless, the ability to process discrete sounds appears to be limited to about 10 items per second (Miller and Taylor, 1948), beyond which they are perceived to fuse into a single buzzing sound. To make matters worse, the auditory trace is limited to about 100 ms (Remez, Ferro, Dubowski, Meer, Broder, &

Davids, 2010). Moreover, memory for arbitrary sequences seems to be limited to about four items (Cowan, 2001; Warren, Obusek, Farmer, & Warren, 1969).

Thus, the signal—and human memory for it—are incredibly short-lived. On the surface, the Now-or-Never bottleneck would seem to render language learning and use impossible. A key strategy for overcoming these sensory and memory limitations lies in *chunking*: incoming items can be rapidly grouped and passed to successively higher levels of representation, with higher-level representations allowing input to be dealt with before it is overwritten by the onslaught of incoming information at a lower level. It is fairly intuitive and uncontroversial, for instance, that the raw acoustic signal is rapidly packaged into some sort of sound-based unit (e.g., phoneme- or syllable-like representations), which can in turn be chunked into word-like representations, and so on. The consequences of applying this general approach to sentence-level processing and grammatical development are, however, less obvious, as discussed by Christiansen and Chater (2016a, b).

Though gaining renewed emphasis under this perspective, chunking has been regarded as a key learning and memory mechanism in human cognition for over half a century (e.g., Feigenbaum& Simon, 1962; Miller, 1956; Simon, 1974). While verbal theories have been more common, computational models of chunking have been present in the literature for over four decades (e.g., Ellis, 1973; French, Addyman, & Mareschal, 2011; Jones, 2012; Perrchet &Vinter, 1998; Servan-Schreiber & Anderson, 1990; Simon & Gilmartin, 1973). Previous computational accounts of chunking have had a significant impact on approaches to language development, particularly with respect to the area of speech segmentation (cf. Frank, Goldwater, Griffiths, & Tenenbaum, 2010).

In what follows, we present a computational framework which extends the real-time use of chunking beyond word segmentation to aspects of sentence comprehension and production, uniting evidence for multiword linguistic units and shallow processing within a simple, developmentally motivated model of acquisition that learns through on-line processing. We begin by discussing these lines of research as they pertain to our computational approach, before introducing the model and its inner workings. We then report results on the acquisition of English as well as the simulation of a key psycholinguistic experiment on children's sentence processing. Finally, we demonstrate that our approach extends beyond English to cover the acquisition of a broad array of typologically diverse languages.

The Psychological Reality of Multiword Linguistic Units

Our computational approach to acquisition begins with the idea that language learners form representations of differing granularities, with linguistic units ranging from the fine-grained level of morphemes and words to the more coarse-grained level of word sequences comprising one or more phrases. This perspective emerges straightforwardly from item-based approaches to acquisition; at the heart of usage-based theory lies the idea that linguistic productivity develops gradually through abstraction over multiword sequences (e.g., Abbot-Smith and Tomasello, 2006; Tomasello, 2003), requiring that storage of multiword units (chunks) occurs. In contrast, generative approaches have traditionally remained faithful to a words-and-rules perspective, in which learning and processing are supported by separate systems for lexicon and grammar (e.g., Pinker, 1999)[2].

While the assumption that children in some sense store multiword sequences has received support from naturalistic observation (e.g., Lieven, Behrens, Speares, & Tomasello, 2003; Peters, 1983), it is only recently that its validation has been made the target of empirical work. The finding of Bannard and Matthews (2008) that young children repeat phrases faster and more accurately when they form a frequent chunk may have provided the first direct evidence not only that multiword chunk storage takes place, but that this storage can actively facilitate processing. Controlling for substring frequency, they contrasted repetition of four-word phrases in which the fourth word was of either high or low frequency, given the preceding trigram. Two and 3-year-olds were more likely to repeat a phrase correctly when its fourth word combined with the preceding trigram to form a frequent chunk, while 3-year-olds were significantly faster to repeat the first three words. Further evidence comes from children's production of irregular plurals: Arnon and Clark (2011) found that the overregularization errors are significantly reduced when irregular plurals are produced in the context of lexically-specific frames (e.g., "*brush your*teeth").

The importance of such findings to usage-based approaches is underscored by previous computational modeling work demonstrating that the alignment and comparison (cf. Edelman, 2008) of multiword sequences can give rise to a considerable amount of linguistic productivity, through the abstraction of partially

[2] More recent accounts, however, have allowed for storage of multiword sequences within a generative framework (Culicover & Jackendoff, 2005; Jackendoff, 2002).

item-based grammatical constructions (Kolodny, Lotem, & Edelman, 2015; Solan et al., 2005).

While usage-based theory has focused primarily on the importance of stored sequences as exemplars in the abstraction of grammatical regularities, children's apparent use of multiword units during on-line processing (Arnon & Clark, 2011; Bannard & Matthews, 2008) highlights an active role for such units in comprehension and production, suggesting the possibility that multiword sequences retain their significance throughout development. Indeed, a number of findings indicate that the storage and active use of multiword units persists beyond early acquisition and into adulthood. Bannard and Ramscar (2007) found an effect of overall sequence frequency on reading times for units ranging from 4 to 7 words in length, while Reali and Christiansen (2007) showed chunk frequency effects in the processing of complex sentence types. Arnon and Snider (2010) found the same general pattern using a phrasal-decision task, whereby four-word expressions were classified as possible or impossible strings in English (in a vein similar to lexical-decision tasks). Importantly, Arnon and Snider's study explored multiple frequency bins; reaction times decreased as a function of phrase frequency. Caldwell-Harris, Berant, and Edelman (2012) extended this finding to a broader frequency spectrum, showing a continuous effect of frequency and providing evidence against a frequency “threshold” beyond which a sequence is unitized. Additional evidence for adults' sensitivity to multiword sequence frequency has been gained from self-paced reading and sentence recall tasks (Tremblay, Derwing, Libben, & Westbury, 2011), eye-tracking data (Siyanova-Chanturia, Conklin, & van Hueven, 2011), and event-related brain potentials

(Tremblay & Baayen, 2010). A similar pattern of results has been found in studies of adult production, demonstrating a decrease in naming latencies with increasing phrase frequency (Janssen & Barber, 2012) as well as reduced phonetic duration for frequent multiword sequences in elicited and spontaneous speech (Arnon & Cohen Priva, 2013)[3].

There are also direct parallels between the learning and processing of multiword units and individual words with respect to age-of-acquisition (AoA) effects. In a variety of tasks, adults exhibit processing advantages for words that are acquired earlier in childhood (for reviews, see Ghyselinck, Lewis, & Brysbaert, 2004; Johnston & Barry, 2006; Juhasz, 2005). Arnon, McCauley, and Christiansen (2017) show that multiword sequences, like individual words, display AoA effects when AoA is determined using either corpus-based metrics or subjective AoA ratings. They also show that the effect cannot be reduced to frequency, semantic plausibility, or lexical AoA. By underscoring a further parallel between words and multiword patterns, this study builds strong support the notion of stored multiword sequences as key building blocks for language learning and use.

Thus, the importance of multiword linguistic units extends beyond merely serving as exemplars for the formation of item-based schemas or the abstraction of grammatical regularities; multiword sequences play an active role in on-line processing, and this persists into adulthood. Accordingly, the on-line discovery and

[3] While the above studies have focused primarily on distributional properties, there are additional semantic and prosodic contributions to the ways in which language users represent and draw upon multiword units (e.g., Jolsvai, McCauley, & Christiansen, 2013)

use of multiword sequences during comprehension and production forms one of the key features of the present computational approach.

The use of multiword linguistic units also leads us to explore the possibility that children's language development does not inevitably arrive at the use of fully articulated, hierarchical phrase structure, as assumed in many previous computational studies. In blurring the lines between lexicon and grammar, the active use of multiword sequences points to a potential role for relatively “flat” syntactic structures, suggesting that a more shallow form of processing may persist throughout development and into adulthood. This radically changes the problem facing the learner; instead of being forced to learn global hierarchical structures tied to a target grammar, local sequential structure moves to the fore. In what follows, we explore this idea more closely, reviewing evidence that shallow processing based on local information represents the norm rather than the exception in language use.

The Ubiquity of Shallow Processing in Language Use

Evidence for shallow processing of linguistic input has led some researchers to question the centrality of hierarchical phrase structure as well as the standard generativist assumption that syntactic and semantic processes are carried out completely and automatically. Yet for over half a century, hierarchical phrase structure has been viewed as fundamental to most accounts of language acquisition and processing (e.g., Chomsky, 1957). Consequently, the idea that the meaning of a sentence need not stem from a fully articulated syntactic structure remains controversial. Nevertheless, shallow processing has been shown to be a widespread phenomenon through psycholinguistic research (for reviews, see Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002), and while the vast majority of this work has dealt with adult subjects, the theoretical implications extend from adult processing to the study of language acquisition. Here, we briefly discuss the evidence for shallow processing in adult language users before turning our attention to similar (though much more limited) evidence from developmental studies, and finally outlining an account of shallow sentence processing which forms part of the motivation for the computational approach to acquisition put forth in this paper.

What is perhaps the most well-known thread of evidence for shallow processing comes from the failure of readers to notice semantically anomalous words and phrases in texts, indicating that processes of semantic integration have not been fully completed by readers who nevertheless form coherent semantic representations based on the sentences in question (e.g., Barton & Sanford, 1993; Erickson & Mattson, 1981). Other work has focused on text-change blindness (following work on change blindness in visual processing; e.g., Simons & Levin, 1998) to demonstrate the extent to which several factors modulate depth of processing, including focus (Sanford, 2002; Sturt, Sanford, Stewart, &Dawydiak, 2004) and computational load (Sanford, Sanford, Filik, and Molle, 2005). Perhaps more relevant is work demonstrating subjects' interpretation of nonsensical sentences as coherent (Fillenbaum, 1974; Wason& Reich, 1979) as well as the processing of semantically anomalous sentences in ways that directly contradict the interpretations that would be made according to a full syntactic parse (Ferreira, 2003), demonstrating the on-line use of background world knowledge and pragmatic expectations.

The above-mentioned evidence for shallow processing meshes naturally with work highlighting readers' tendencies to form “underspecified” representations of sentences, in which no commitment is made to any one of a number of possible analyses, clearly indicating that fully articulated syntactic processing has not taken place. Evidence for underspecification comes from work involving ambiguous relative clause attachment (Swets, Desmet, Clifton, & Ferreira, 2008), quantifier scope (Tunstall, 1998), metonymy (Frisson & Pickering, 1999), ambiguous nouns (Frazier & Rayner, 1990), and anaphoric reference (Koh, Sanford, Clifton, & Dawydiak, 2008). Like shallow processing more generally, underspecified representations are at odds with theories of processing that assume full completion of syntactic and semantic analyses. Much of the evidence for underspecification makes good contact with Ferreira and Patson's (2007) Good Enough approach to sentence processing, in which it is argued that the goal of language comprehension is to establish representations which are merely “good enough” to suit the needs of a listener or reader in a given situation, as opposed to representing communicated information in full detail[4].

Taken together, the evidence suggests that shallow, underspecified processing, far from representing a degenerate case or mere exception to normal full syntactic and semantic processing, is ubiquitous. It is worthy of note that current evidence for shallow processing comes from work with written texts, a medium which allows

[4] As pointed out by Sanford and Sturt (2002), the contrast between traditional notions of full syntactic processing and shallow, underspecified processing is mirrored in the fields of computational linguistics and natural language processing (NLP) by differences between the output of shallow parsers, which identify a subset of interpretations for a sentence, and full syntactic parsers, which build a fully articulated syntactic analysis. Even in the context of NLP, shallow parsing sometimes offers computational advantages over full parsing (e.g., Li & Roth, 2001). Recently, it has also been shown that shallow parsing is sufficient for semantic role labeling in a morphologically rich language (Goluchowski & Przepiorkowski, 2012).

subjects to process language without facing considerable challenges from 1) the highly noisy, variable nature of the speech signal and 2) the time constraints that come with not being able to control the speed at which input is encountered[5]. Thus, it is likely that much stronger evidence for shallow processing can be gained using speech stimuli (cf. Christiansen & Chater, 2016).

In the above-mentioned cases, readers seem to rely on local linguistic information and global background knowledge rather than compositional meanings derived from fully articulated syntactic representations. Thus, support for shallow processing makes close contact with the claim that adults process sentences by using small chunks of local information to arrive at a semantic representation (e.g., Ferreira &Patson, 2007), which is reflected by local coherence effects (e.g., Tabor et al., 2004).

Evidence that adults process sentences in this manner makes it more than plausible that children may rely on similarly shallow, underspecified processing in which local information is key. While the issue of syntactic processing depth remains largely unexplored in children, initial evidence suggests that young learners rely upon shallow, underspecified processing to an even greater extent than adults (e.g., Gertner& Fisher, 2012). Corpus analyses of child speech similarly suggest that children's earliest complex sentences featuring sentential complements (e.g., *I think I saw one*) represent the simple concatenation of a formulaic expression (*I think*) with a sentence (*I saw one*) in a shallow rather than hierarchical fashion (Diessel & Tomasello, 2000).

Evidence for shallow processing based on local information makes close

[5] Note that there may also be strict time pressures during normal fluent reading, when readers take in about 200 words per minute (see Chater & Christiansen, 2016, for discussion)..

contact with Sanford and Garrod's (1981; 1998) Scenario Mapping and Focus theory of comprehension, in which background knowledge of situations and scenarios is used on-line to interpret linguistic input as it is encountered. During on-line interpretation, incoming linguistic input is mapped onto schemas of events, situations, or scenarios which have been established based on previous contexts or input – interpretation of the overall message is therefore heavily influenced by the background information which linguistic input is mapped onto. It may therefore be fruitful to test the view of language comprehension as the attempt to map chunks of language input onto specific parts of a scenario or event schema (which can, of course, be quite abstract and need not correspond to concrete objects and events in the real world); shallow processing may be sufficient for accomplishing this task. This, in turn, helps us reframe the problem facing the language learner: multiword unit learning (which allows rapid and efficient retrieval of chunks of local information during comprehension and production) naturally dovetails with a shallow processing approach, allowing language learners to comprehend much of the input without the need for full global syntactic parsing of the sort assumed in the vast majority of approaches to language learning.

This perspective fits nicely with several threads of psycholinguistic and computational work which are beginning to converge on the view that language users rely on sequential rather than hierarchical structures (for a review, see Frank et al., 2012). For instance, Ferreira and Patson (2007) found that interpretations can be constructed on the basis of small numbers of adjacent items, at the expense of more global syntactic structures and meanings, suggesting that global hierarchical structures were either impeded by local information or were altogether less important. Along the same lines, Christiansen and MacDonald (2009) found that simple recurrent networks (Elman, 1990), which simply learn to predict upcoming items in sentences in a linear rather than hierarchical fashion, provide a close fit to the abilities of human subjects to process recursive constructions involving center-embedding and cross-dependency. Consistent with this finding, Frank and Bod (2011) demonstrated that models which learn linear, non-hierarchical sequential information about word classes provide a stronger fit to actual human eye movement data during reading than models which learn hierarchical phrase structures.

In line with the view that sentence processing relies heavily on sequential structures computed over chunks of local information, our computational approach is centered on simple mechanisms for the on-line discovery, storage, and sequencing of words and chunks through sensitivity to the local rather than global information contained in utterances. Before detailing our computational approach in greater depth, we briefly discuss the potential sources of information children might use to discover useful chunks of local information, and the relationships between them, during their attempts to comprehend and produce utterances.

Discovering Useful Multiword Sequences

Our computational account of children's on-line processing seeks to capture some of the mechanisms by which multiword units are learned and employed in language comprehension and production. For the sake of simplicity, we distinguish between two types of multiword units: 1) *unanalyzed chunks*, and 2) *phrasal chunks* (see also Arnon & Christiansen, in press; McCauley, Monaghan & Christiansen,2015).

Unanalyzed chunks are those that are stored and accessed holistically before segmentation of their parts has taken place, whereas phrasal chunks can be (and sometimes are) broken down into their component words. Chunks falling into the first category are most relevant to the study of very early language development (for a recent incremental, on-line computational model of word segmentation which captures processes whereby unanalyzed chunks can be discovered and gradually broken down into smaller units, see Monaghan & Christiansen, 2010). As an example, the chunk *look at this* may be treated as a holistic, unanalyzed unit by very young children (for a review of the literature on children's use of such "frozen" sequences, see Wray, 2005), while the same chunk would fall into the second category (as a *phrasal chunk*) for older children who are capable of breaking the chunk down into its component parts.

Beyond a certain point, chunks will rarely be treated as holistic units: Consider evidence that idioms, which even generative grammar-oriented approaches recognize as stored (Jackendoff, 1995; Pinker, 1999), prime, and are primed by, their component words (Sprenger, Levelt, &Kempen, 2006) as well as lead to syntactic priming (Konopka & Bock, 2009). Given that idioms would appear to form stored multiword units (their meanings are idiosyncratic and cannot be determined on the basis of component parts), we must allow, then, that a multiword unit can be accessed and used as a meaningful entity in its own right, even when activation of its individual parts occurs.

One well-studied source of information that infants might use to arrive at some of their earliest, unanalyzed multiword chunks lies in the acoustic correlates of clause and phrase boundaries. Pre-linguistic infants can use prosodic information to segment the speech stream into multiword units, and this ability has been shown to facilitate certain types of processing. Early work in this vein established that infants are sensitive to the prosodic correlates of clause boundaries (Hirsh-Pasek, Kemler Nelson, Jusczyk, Cassidy, Druss, & Kennedy, 1987). Further work demonstrated that infants are better able to recall phonetic information when it is packaged in a prosodically well-formed unit, and that infants can use the acoustic correlates of clause boundaries to form representations which are available for use in later segmentation of continuous speech (Mandel, Jusczyk, & Kemler-Nelson, 1994). More relevant to the present study is work on phrase-level units. Though several studies suggest that phrases are not as reliably marked in the speech stream as are clauses (Beckman & Edwards, 1990; Fisher & Tokura, 1996), infants' sensitivity to these markers has been demonstrated (Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, & Piwoz, 1992). Moreover, it has been shown that infants can use these markers to segment larger prosodic units corresponding to clauses into smaller, phrase-level units (Soderstrom, Seidl, Kemler-Nelson, & Jusczyk, 2003). Results from the Soderstrom et al. (2003) study went beyond mere on-line recognition of prosodic ill-formedness, suggesting that infants formed representations based on the prosodic information in familiarization sequences, and used them to segment prosodically well-formed items into phrase-level units at test.

The incorporation of such prosodic information, however, represents a challenge for computational models of acquisition, given the limited availability of prosodic information in currently available corpora of child-directed speech. Fortunately, distributional information is also highly relevant to early chunk discovery. Some of children's earliest unanalyzed multiword chunks may stem from "errors" in word segmentation (as suggested by Bannard & Matthews, 2008). For instance, using mutual information between syllables to find word boundaries in an unsegmented corpus, Swingley (2005) found that 91% of bisyllablic false alarms were frequent word pairs, such as *come on,* while 68% of trisyllabic false alarms were frequently occurring multiword phrases. More recent models of word segmentation (e.g., Goldwater, Griffiths, & Johnson, 2009; Monaghan & Christiansen, 2010) have yielded similar results. Given that such models exploit some of the same distributional cues that infants have been shown to be sensitive to in experimental studies of artificial word segmentation, it would not be surprising if infants similarly undersegmented the speech stream to arrive at unanalyzed multiword chunks. Far from hindering the child's language development, such "mistakes" may actually impart an advantage, as predicted by usage-based theories (e.g., Arnon, 2009; Arnon & Christiansen, in press).

But what of chunks acquired after segmentation of the component words has taken place? Presumably, unanalyzed chunks, arrived at through under-segmentation and/or the use of prosodic information, are rare once the child reaches a certain level of experience. The usefulness of multiword chunks should be no less real for an experienced language user (and indeed, as shown in the studies reviewed above, older children and adults actively use multiword units). Thus, we should allow for the possibility that statistical information linking words can be used to arrive at multiword chunks by older children and adults.

How might learners chunk co-occurring words together as a unit after segmentation of the component parts has already taken place? The use of raw frequency of co-occurrence would lead to placing too much emphasis on co-occurring words that frequently occur adjacent to one another by mere virtue of being highly frequent words. Similarly, precise tracking of the frequencies of all encountered sequences would lead to a combinatorial explosion (cf. Baayen, Hendrix, & Ramscar, 2013).

Thus, while phrase-frequency effects are continuous rather than threshold-based (e.g., Caldwell-Harris et al., 2012), meaningful chunks cannot be identified on the basis of raw frequency alone[6], in much the same way as a word segmentation model based solely on raw frequency of co-occurrence would be largely ineffective. Consistent with the Now-or-Never perspective (Christiansen and Chater, 2016b), which forms part of the theoretical motivation for the present study, we explore the notion that many of the same cues and mechanisms involved in word segmentation may be involved in chunking at the level of multiword units. In what follows, we discuss previous computational accounts of chunking which, in the domain of language, have been primarily concerned with word segmentation. We then describe our own model, which extends chunk-based learning and processing to the sentence level.

Previous Computational Models of Chunking

[6] Indeed, it has been suggested that part of the problem experienced by second-language learners may be due to a suboptimal chunking strategy based on raw frequency (Ellis, Simpson-Vlach, & Maynard, 2008)—something that has been corroborated by simulations of second-language learning using the CBL model presented below (McCauley & Christiansen, in press).

Chunking has been regarded as a key learning and memory mechanism for over half a century (e.g., Feigenbaum & Simon, 1962; Miller, 1956), with many of the earliest computational implementations being concerned with expertise (e.g., Ellis, 1973; Simon & Gilmartin, 1973) or specific language-related phenomena such as spelling (Simon and Simon, 1973) and alphabet recitation (Klahr, Chase, & Lovelace, 1983). In recent decades, a number of chunking models related to implicit learning have emerged, and have been applied to word segmentation, particularly in the context of modeling data from artificial language learning experiments.

An early instance of one such model is the Competitive Chunking (CC) model of Servan-Schreiber and Anderson (1990), which views learning as the buildup of progressively larger chunks which are structured in a hierarchical network. Servan-Schreiber and Anderson argued that the implicit learning of artificial grammars (e.g., Reber, 1967) is primarily driven by chunking, and model the discrimination of grammatical vs. ungrammatical strings according to the number of stored chunks necessary to describe a sequence. The CC model operates according to activation of hierarchical chunks which match a current stimulus. Activated chunks which overlap with one another then "compete" to shape perception of the stimulus. Chunk creation and retrieval are determined by chunk strength, which is tied to free parameters involving decay and competition. In a Reber (1967) task analogue, CC was able to capture 87% of the variance in subject discrimination of grammatical vs. ungrammatical strings.

Perhaps the most influential chunking model devoted to implicit learning and word segmentation is PARSER (Perruchet & Vinter, 1998), which was directly inspired by the CC model of Servan-Schreiber and Anderson (1990). Unlike the CC model, however, PARSER does not build up a hierarchical network of chunks, being concerned primarily with identifying structurally relevant units, such as words (Perruchet et al., 2002). Like the CC model, PARSER operates through competition, or "interference," between overlapping chunks, and utilizes free parameters for setting decay and managing activation rates (chunk strength). PARSER has been used to successfully model some of the experimental data involving human word segmentation performance in artificial language learning contexts (e.g., Saffran, Newport, & Aslin, 1996; cf. Perruchet & Vinter, 1998), and has also been used to discover the syntactically relevant units in an artificial language generated by a finite-state grammar (Perruchet et al., 2002).

A more recent approach to chunking is the MDLChunker (Robinet, Lemaire, & Gordon, 2011), which operates according to the information-theoretic principle of minimum description length (MDL), following the notion that human cognition favors simpler representations as a general principle (Chater & Vitanyi, 2003). Like CC, MDLChunker involves hierarchies of chunks. Unlike CC, or PARSER, MDLChunker does not have free parameters. MDLChunker captures human chunking in a novel task involving meaningless visual symbols, as well as providing similar results to PARSER and CC on a well-known experiment by Miller (1958).

Another contemporary model of perhaps greater relevance is the TRACX model (French et al., 2011). A connectionist auto-associator, TRACX operates according to recognition-based processing rather than prediction, as in the case of prediction-based networks like simple recurrent networks (SRNs; Elman, 1990).

TRACX provides a better fit to human data than either PARSER or SRNs, across a range of sequence segmentation studies.

French et al. (2011) nicely exemplify the tensions in the implicit learning literature between recognition-based processing, through chunking, and statistically-based processing utilizing transitional probabilities (TPs). Despite the absence of predictive processing and lack of conditional probability calculation, TRACX is sensitive to TPs in both directions, as is the case with PARSER.

A number of studies have argued that recognition-based chunking provides a better fit to human performance than do TPs: PARSER offers a better fit to human data than learning based solely on TPs in a context in which the statistics of two consecutively learned artificial languages are pitted against one another (Perruchet, Poulin-Charronnat, Tillmann, & Peereman, 2014). Moreover, PARSER provides a better fit to adult learning of a semi-artificial language than SRNs (Hamrick, 2014). Poulin-Charronnat, Perruchet, and Tillmann (2016) developed a design which allowed them to dissociate the influence of familiarity and transitional probabilities using a pre-exposure phase in a standard artificial word segmentation task, in which recognition of familiar units appeared to override sensitivity to statistical cues, though findings were only partially captured by PARSER. These findings are compatible with a recent recognition-based model of word segmentation (Monaghan & Christiansen, 2010) which performs reasonably well on child-directed speech input in comparison to more computationally complex models.

Of the above computational approaches to chunking, PARSER has been the most widely explored in the context of human artificial-language data on chunking and segmentation performance. PARSER also best satisfies the memory constraints imposed by the Now-or-Never bottleneck, which forms part of the theoretical motivation for the present study. MDLChunker, for instance, has no memory limitations and is ill-suited to capturing on-line processing. Therefore, we chose PARSER for use as a baseline for comparison to our own model, alongside a standard, prediction-based model utilizing transitional probabilities over *n*-grams.

Integrating Recognition-based and Statistically-based Processing

While chunking models have been argued to offer a better fit than TPs to human performance in studies of artificial word segmentation, which involve brief periods of exposure, fewer studies have examined learning over longer periods of time, or the learning of higher-level chunks, such as would be useful in learning grammatical regularities.

Transitional probabilities have been found to be useful in segmenting out multiword phrases: Thompson and Newport (2007) found that peaks in forward transitional probabilities (FTPs) between form classes in an artificial language can be used by adult subjects to group artificial words together into multiword units, whereas dips in FTPs can be used to identify chunk boundaries. However, a number of instances arise in natural language in which sole reliance on forward transitional probabilities might prevent the segmentation of useful multiword chunks. For example, if learners were to compute statistics over individual words rather than form classes, the FTP between the words in an English phrase such as "*the dog*" will always be extremely low, given the sheer number of nouns that may follow a determiner.

Other sources of information, however, such as mutual information or backwards transitional probabilities (BTPs) provide a way around this issue: given the word “dog,” the probability that the determiner “the” immediately precedes it is quite high, considering the small number of determiners one might choose from. Thus, it makes sense that child learners might also make use of such sources of information to discover useful multiword units.

Along these lines, Saffran (2001, 2002) has shown that dependencies in the form of backwards transitional probabilities of between words in an artificial phrase-structure grammar not only facilitate learning, but aid in isolating specific phrases. That infants and adults are sensitive to backward transitional probabilities has been established in the chunking (French et al., 2011; Perruchet &Desaulty, 2008) and statistical learning (Pelucchi, Hay, & Saffran, 2009) literatures. Thus, the view of backward transitional probabilities as a potential cue to useful multiword phrases holds promise. That English speakers may be more sensitive to backwards than forward transitional probabilities between words during production is suggested by auditory corpus analyses showing that functors and content words are shorter when predictable given the following word, while the same effect in the forward direction appears only for the most frequent functors, and is absent for content words (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). Though much previous work examining adults' production of multiword units has been concerned with raw frequency counts rather than conditional probabilities, the Bell et al. (2009) findings might also be taken as initial support for idea that speaker sensitivity to backward transitional probabilities may help drive the representation and use of multi-word chunks.

The simple architecture of the model used in the present study—while inspired by the successes of recognition-based chunking in accounting for experimental data—also seeks to incorporate statistical learning, in line with the aforementioned evidence for the use of TPs in phrase-level chunking. Rather than rely on prediction-based processing, statistical information tied to BTPs is used as a cue for identifying chunks which are then stored as concrete units and used to support further processing through recognition. Forward prediction in the model is chunk-based rather than statistical. A purely recognition-based model which does not directly utilize statistical cues, PARSER (Perruchet &Vinter, 1998), serves as a baseline for comparison to our own model, alongside a purely prediction-based model utilizing transitional probabilities over *n*-grams.

The Chunk-Based Learner Model

In what follows, we present the Chunk-Based Learner (CBL) model of language learning. Following Christiansen and Chater (2016b), one of the primary aims of the CBL model is to provide a computational test of the idea that the discovery and on-line use of multiword units forms part of the backbone for children's early comprehension and production. To this end, the model gradually builds up an inventory of chunks consisting of one or more words—a "chunkatory"—which is used in both language comprehension and production. The model explicitly captures the shallow processing perspective outlined above—in which chunks of local information are used to process sentences—by learning to group words together into local chunks that are appropriate for arriving at an utterance's meaning (a key aspect of comprehension), while simultaneously learning to produce utterances in an incremental fashion using the same chunks of local information (a key aspect of production).

CBL was designed with several key psychological and computational features in mind:

1. **On-line processing**: During comprehension, input is processed word-by-word as it is encountered, reflecting the incremental nature of human sentence processing (e.g., Altmann & Steedman, 1988; Tanenhaus, Carlson, & Trueswell, 1989; Tyler & Marslen-Wilson, 1977); during production, utterances are constructed incrementally according to a chunk-to-chunk process rather than one of whole-sentence optimization. This approach is consistent with memory constraints deriving from the real-time nature of language processing (Christiansen & Chater, 2016b).
2. **Incremental learning**: At any given point in time, the model can only rely on what it has learned from the input encountered thus far (i.e., unlike the vast majority of computational approaches to acquisition, the current model does not rely on batch learning of any sort).
3. **Simple statistics**: For reasons detailed above, learning is based on the computation of backward transitional probabilities, which 8-month-old infants (Pelucchi, Hay, & Saffran, 2009) and adults (Perruchet &Desaulty, 2008) can track.
4. **Local information**: Learning is tied to local rather than global information; instead of storing entire utterances as sequences, the model learns about transitions between adjacent words and chunks.

5. **Item-based**: The model learns from concrete words, without recourse to abstract information such as that of syntactic categories (as is also the case with a number of other usage-based models; e.g., Freudenthal et al., 2006; Jones et al., 2000; Kolodny et al., 2015; Solan et al., 2005). This stands in stark contrast to most computational approaches emerging from the tradition of generative linguistics; rule-based processing in the "words-and-rules" framework operates over word classes rather than words themselves.
6. **Psychologically motivated knowledge representation**: In accordance with evidence for the role of multiword linguistic units in comprehension and production (reviewed above) as well as for the interconnectedness of comprehension and production processes more generally (Chater, McCauley, & Christiansen, 2016; Pickering & Garrod, 2007, 2013), aspects of both comprehension and production are performed by using the same inventory of single- and multiword linguistic units.
7. **Naturalistic input:** The model learns from corpora of child-directed speech taken from the CHILDES database (MacWhinney, 2000). As word segmentation itself lies outside the scope of the current model, the use of such pre-segmented corpora (which consist of words rather than phonemic transcriptions) enables us to expose the model to a far more diverse array of corpora than would be possible otherwise.
8. **Broad, cross-linguistic coverage**: The model is designed in such a way that it can be evaluated on corpora of child-directed speech in any language (following Chang, Lieven, & Tomasello, 2008). We therefore evaluate it using a typologically diverse set of 29 languages.

We begin by providing an initial glance at the general architecture of the model, before describing its inner workings in full detail. We then present results from simulations of the acquisition of English. Finally, we show that the model successfully extends to the acquisition of a wide array of typologically diverse languages.

Inner Workings of the Model

There is growing behavioral and neuroimaging evidence for the involvement of the same mechanisms in adult comprehension and production (for reviews, see Pickering & Garrod, 2007, 2013), which prompts us to extend the idea of a unified framework for comprehension and production to language development (Chater et al., 2016; McCauley & Christiansen, 2013). In light of this, we designed CBL to capture the idea that comprehension and production can be viewed as two sides of the same process. Comprehension is approximated by the segmentation of incoming speech into chunks relevant for determining the meaning of an utterance. These units are then stored in an inventory that makes no distinction between single- and multi-word chunks. Production is approximated by the model's ability to reproduce actual child utterances through retrieval and sequencing of units discovered during comprehension. Crucially, the very same distributional information underlying the model's comprehension-related processing form the basis for production, while production itself is taken to feed back into comprehension.

The model's inventory of single- and multi-word linguistic units—its chunk inventory, or "chunkatory"—is its core feature. Comprehension-related processing is used to build up the chunkatory while production-related processing both draws upon it and reinforces it. Through the chunkatory, CBL is able to approximate elements of comprehension and production within a unified framework.

The model begins by learning—in an incremental, on-line fashion—to segment incoming input into groups of related words (similar to phrasal units). These chunks are then stored in the chunkatory unless they have been encountered before, in which case the frequency of the relevant chunk is incremented by 1. In each simulation, the input consists of a corpus of speech directed to a single child (taken from the CHILDES database; MacWhinney, 2000). When the model encounters a multiword utterance produced by the target child, the production side of the model comes into play: the model's task is to produce an utterance which is identical to that produced by the child, using only statistics and chunks learned and used during comprehension-related processing up to that point in the simulation. Thus, we aimed to construct a fully incremental, on-line model of child language development that uses the same chunks and distributional statistics to perform aspects of both comprehension and production.

To summarize this initial snapshot of the model's inner workings: CBL approximates aspects of both *comprehension,* by learning an inventory of chunks and using them to segment child-directed speech into related groups of words (such as would be appropriate for arriving at an utterance's meaning via shallow processing), and *production,* by reproducing actual child utterances as they are encountered in a corpus, using the same chunks and statistics learned and used during comprehension. We hypothesized that both problems could, to a large extent, be solved by recognition-based processing tied to chunks which are discovered through sensitivity to *transitional probabilities* between linguistic units.

Transitional probability: The simple statistic at the heart of CBL: As reviewed above, transitional probability (TP) has been proposed as a cue to phrase structure in the statistical learning literature; peaks in TP can be used to group words together, whereas dips in TP can be used to find phrase boundaries (e.g., Thompson & Newport, 2007). The view put forth in such studies is that TP is useful for discovering phrase structure when computed over form classes rather than actual words. We hypothesized, instead, that distributional information tied to individual words provides richer source of information than has been assumed in such work. Because we adopted this purely item-based approach, and because of evidence for greater reliance on BTPs when chunking words together in English (Bell et al., 2009), we decided to focus initially on backward transitional probability.

The computation of transitional probabilities in the backward direction also has an unexpected advantage in the context of incremental, on-line calculation, in that the properties of the most recently encountered word attain the greatest importance (e.g., the BTP linking the sequence XY can be arrived at by normalizing P(X, Y) by P(Y) rather than involving X in the denominator when computing FTP). For the above-mentioned reasons, the current computational approach focuses initially on backward rather than forward transitional probabilities as a cue to multiword units—chunks of local information for use in processing—while comparing the types of transitional probability in a systematic way in Appendix B.

In what immediately follows, we provide an in-depth description of the inner workings of the model, showing how simple transitional probabilities support recognition-based chunk learning through comprehension- and production-related processes.

Comprehension

Though comprehension and production in the model represent two sides of the same coin, we describe them separately for the sake of simplicity. During comprehension, the model discovers its first chunks through simple sequential statistics. Processing utterances on a word-by-word basis, the model learns frequency information for words and word pairs, which is used on-line to track BTPs between words and maintain a running average BTP across previously encountered word pairs. When the model calculates a BTP between two words that is greater than expected, based on the running average BTP, it groups the word pair together such that it may form part 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 immediately preceding words, is added to the chunkatory. Then, the model moves on to process the next word in the utterance. The use of the running average BTP as a threshold allows the avoidance of a free parameter. This process is illustrated using a simple utterance in Figure 1.

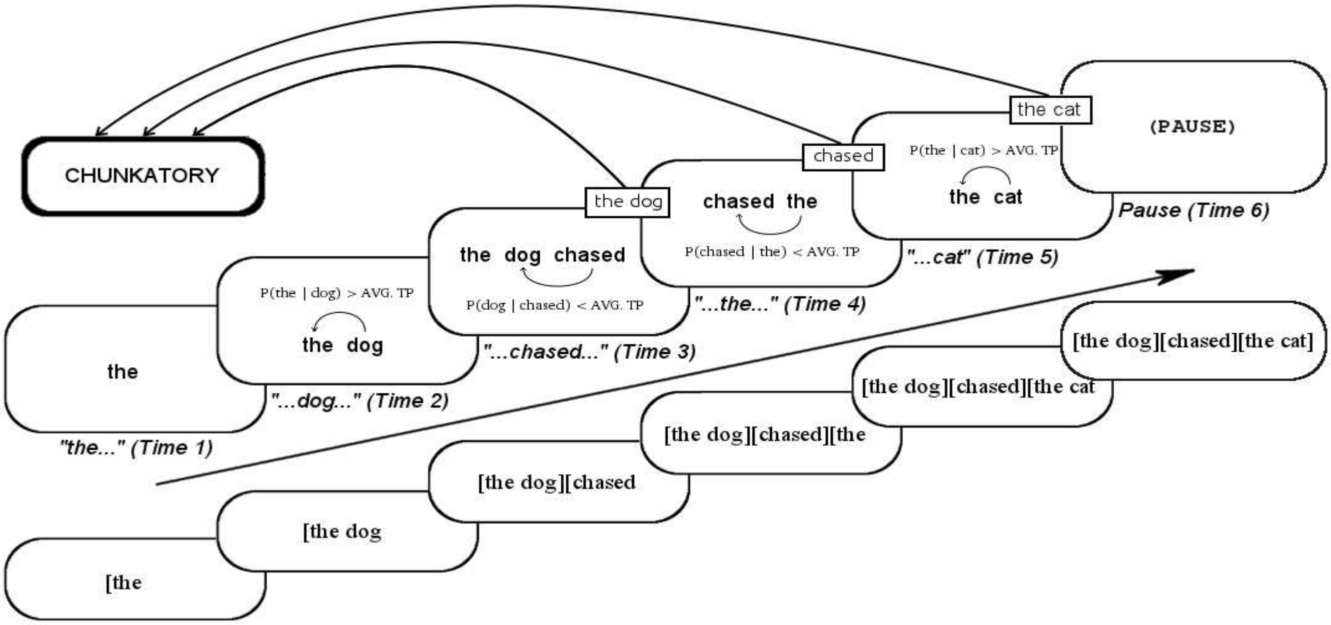

Fig. 1: Incremental, on-line processing of the simple utterance *"the dog chased the cat"*. Material above the diagonal arrow depicts the simple computations driving the model's on-line processing, while material below the arrow represents the resulting shallow parse (the model's interpretation of the sentence) as it unfolds over time. At Time 2, the model calculates the BTP between *the* and *dog*, which exceeds the average TP threshold (indicated by the backward arrow's position above the words), 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 chunkatory as a chunk. At Time 3, the BTP between *dog* and *chased* falls below the running average (indicated by the backward arrow's position below the words), so *chased* is not grouped together with the preceding material and *the dog* is then stored in the chunkatory. 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 chunkatory 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 chunked together and stored in the chunkatory.**

All newly-added chunks are initialized with a frequency count of 1. The frequency count of a chunk is incremented by 1 each time it is encountered subsequently.

Once the model has acquired its first chunk, it begins using its chunkatory in a recognition-based fashion to assist in processing the incoming input on the same incremental, word-to-word basis as before. The model continues learning the same low-level distributional information and calculating BTPs, but also uses the chunkatory to make on-line predictions as to which words should form a chunk, based on previously learned chunks. Crucially, these predictions are recognition-based rather than statistically-based. When a word pair is encountered, it is searched for in the chunkatory; if it has occurred more than once, either as a complete chunk or as part of a larger chunk, the words are automatically grouped together and the model moves on to the next word without placing a boundary. If the word pair has not occurred more than once in the chunks found in the chunkatory at that time step, the BTP is compared to the running average, with the same consequences as described above. Thus, there are no *a priori* limits on the number or size of chunks that can be learned.

As an example of how this can be understood as prediction, consider the following scenario in which the model encounters the phrase *the blue ball* for the first time and its chunkatory includes *the blue car* and *blue ball* (with frequency counts greater than 1). When processing *the* and *blue,* the model will not place a boundary between the two words because the word pair is already strongly represented in the chunkatory (as in *the blue car*). The model therefore predicts that this word-pair will form part of a chunk, even though the rest of the chunk has not yet been encountered. Next, when processing *blue* and *ball*, the model reacts similarly, as this word pair is also represented in the chunkatory. The model thereby combines its knowledge of two chunks to discover a new, third chunk, *the blue ball*, which is added to the chunkatory. As a consequence, the sequence *the blue* becomes even more strongly represented in the chunkatory, as there are now two chunks in which it appears.

Production

While the model makes its way through a corpus incrementally, segmenting and storing chunks during comprehension, it encounters utterances produced by the target child, at which point the production side of the model comes into play. The model's ability to generate the child's utterance, based on chunks learned from previous input, is then evaluated using a sentence production task inspired by the *bag-of-words incremental generation task* used by Chang et al. (2008), which offers a method for automatically evaluating syntactic learners on corpora in any language.

We loosely approximate the overall message that the child wants to convey by treating the utterance as an unordered set of words: a “bag-of-words,” corresponding to (again, very roughly) the set of concepts contributing to the semantics of the utterance to be produced. The task for the model, then, is to place these words in the correct order, as originally produced by the target child. Following evidence for the use of multiword sequences in child production, as well as usage-based approaches more generally, the model utilizes its chunkatory to generate the child's utterances. In order to model retrieval of stored chunks during production, the bag-of-words is filled by comparing parts of the child's utterance against the chunkatory. For instance, consider a scenario in which the model is to produce the child utterance *the dog chased a cat* and the largest chunk in the chunkatory consists of 3 words. To begin, the first 3 words are searched for storage as a single chunk. As this is not found in the chunkatory, *the dog* is searched for. This search succeeds, so the words are removed from the utterance and placed in the bag as a single chunk. Next, *chased a cat* is searched for, unsuccessfully, followed by *chaseda*, also without success. The word *chased* is placed in the bag as a single chunk. Then, *a cat* is searched for, and so on. Crucially, this procedure is only used to find chunks that the child already knows (i.e., that were in the chunkatory as a result of learning during comprehension) and thus would be likely to use as such (e.g., *the dog*). Once in the bag, the order of chunks is randomized.

During the second phase of production, the model attempts to reproduce the child's utterance using the unordered chunks in the bag-of-words. Again following Christiansen and Chater (2016b), we model this as an incremental, chunk-to-chunk process rather than one of whole-sentence optimization (e.g., calculating the probability of the entire utterance, etc.), in order to reflect the incremental nature of sentence processing (e.g., Altmann & Steedman, 1988; Christiansen & Chater, 2016b; Tanenhaus et al., 1989; Tyler & Marslen-Wilson, 1977). Thus, the model begins by removing from the bag-of-words the chunk with the highest BTP given the start-of-utterance marker (a hash tag representing the pause preceding the utterance in the corpus), and producing 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 previously produced chunk. In this manner, the model uses chunk-to-chunk BTPs to incrementally produce the utterance, adding chunks one-by-one until the bag is empty. The model's production of the child utterance *I'm gonna stop the train with my whistle* is depicted in Figure 2. In rare cases where two or more chunks in the bag-of-words are tied for the highest BTP, one of them is chosen at random.

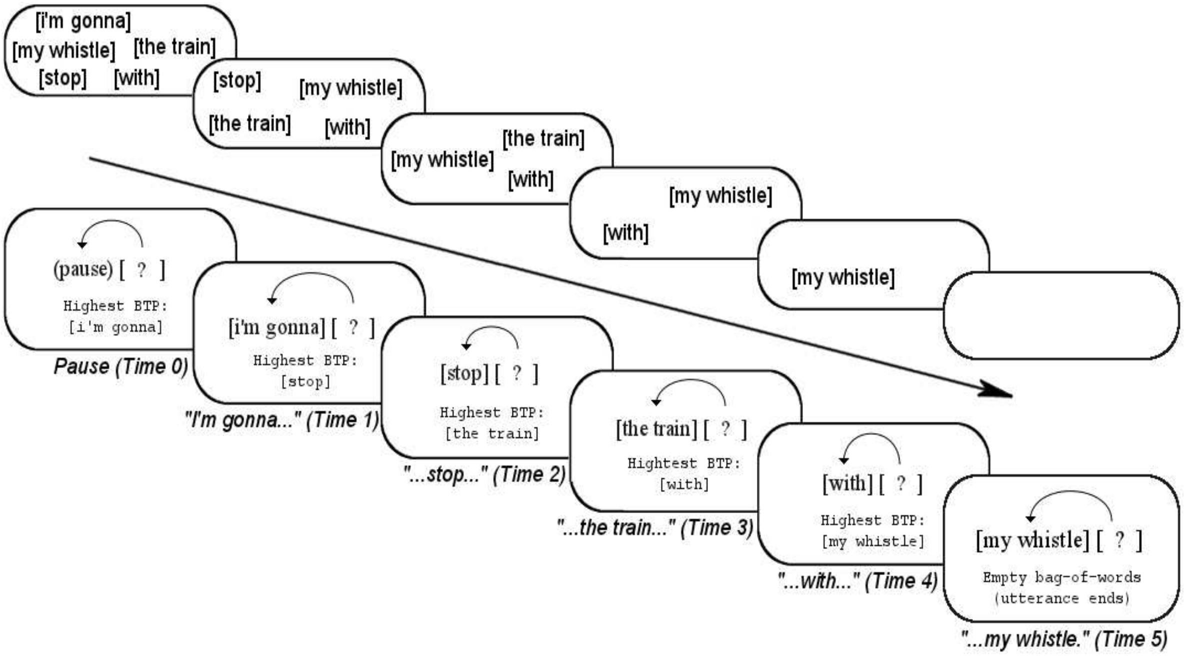

Fig. 2: Incremental, on-line production of the child utterance *"I'm gonna stop the train with my whistle."* Material above the diagonal arrow depicts the contents of the bag-of-words at each time step. Material below the arrow represents the

simple computations whereby the model selects the next item to be produced at each time step. At Time 0, the model selects its first chunk from the bag according to the highest BTP, given the pause preceding the utterance (which can be understood as a start-of-utterance marker); out of the chunks in the bag, [*i'mgonna*]has the highest BTP in this instance, so it is removed from the bag and produced at the next time step. At Time 1, the model calculates the BTP between [*i'mgonna*] and the remaining chunks in the bag; [*stop*] has the highest BTP and is therefore removed and produced at the next time step. This process continues, with the item possessing the highest BTP (given the previous item) being selected until the bag-of-words is empty, at which point the utterance ends.

Because comprehension and production are seen as two sides of the same process, a child's own productions are taken to reinforce statistics previously learned during comprehension. For this reason, immediately following the model's attempt to produce a given child utterance, the same utterance is used to reinforce the model's low-level sequential statistics as well as its chunkatory, through the performance of (incremental and on-line) comprehension on the utterance, in an identical manner to any other utterance of child-directed speech in the corpus. The child is taken to “hear” its own productions in a manner consistent with the position that no strong distinction can be drawn between the mechanisms and statistics underlying comprehension and production (as argued in Chater et al., 2016). Thus, the CBL model features some similarities to the “Traceback Method” of Lieven et al. (2003), while also providing the kind of "rigorous computational evaluation" of the general approach called for by Kol, Nir, and Wintner (2014).

Validity of the bag-of-words production task: In order to ensure that our chosen production task evaluates meaningful linguistic skills, we tested adult native speakers on a behavioral version of the task. Using the largest available child corpus of English (Maslen, Theakston, Lieven, & Tomasello, 2004), we extracted, at random, 50 grammatical child utterances and 20 child utterances which had been previously marked as ungrammatical, for a total of 70 test utterances. Twenty Cornell undergraduates (mean age 20.1 [SE 0.8], all native speakers of English) then received, for each utterance, an un-ordered set of chunks corresponding to the very same chunks the model used when attempting to produce the given utterance during a full simulation over the same corpus. The subjects' task, for each utterance, was to sequence the chunks to form a sentence. The mean accuracy rate was 95.6% across all subjects for the grammatical utterances, and 64% for the ungrammatical utterances (note that only a perfect match to the child's utterance was scored as accurate, just as with the model version of the task). Thus, we conclude that the bag-of-words task itself does provide a meaningful and valid test of linguistic skills which reflect knowledge of grammar.

Contrasting Recognition-based and Statistical Approaches to Chunking: Baseline Models

We compare CBL directly to two models: PARSER (Perruchet &Vinter, 1998) as well as a modified *n*-gram model. PARSER was chosen for comparison because it has been the most widely explored model in the context of human data on chunking and segmentation performance, while also best satisfying the memory constraints imposed by the Now-or-Never bottleneck (Christiansen & Chater, 2016b). As such, it provided an ideal instantiation of purely recognition-based processing for comparison to the alternative approach taken by the CBL model. As an additional baseline utilizing purely prediction-based processing, we implemented a variation on the standard *n*-gram model, focusing on Trigrams for reasons explained below. The contrasting unit types and processing styles of the model and its baselines are simplified in Table 1.

Table 1
Contrasting Unit Type and Processing Style

Model	**Stored,Variable-sized Chunks?**	**Recognition-based?**	**Prediction-based?**
CBL	Yes	Yes	Yes
PARSER	Yes	Yes	No
Trigram	No	No	Yes

Trigram baseline: To assess the usefulness of CBL and PARSER's variable-sized, recognition-based chunks as opposed to simpler sequential statistics tied to prediction, an additional alternate model was created which lacked a chunk inventory, relying instead on forward transitional probabilities computed over stored *n*-grams. Since trigram models (second-order Markov models) are commonly used in computational linguistics as well as the field of machine learning (Manning &Schütze, 1999), we chose to focus on three-word sequences. This decision was further motivated by findings that trigram models are quite robust as language models, comparing favorably even to probabilistic context-free grammars. Our trigram model acquired statistics in an incremental, on-line fashion, in the style of CBL, while simultaneously processing utterances through the placement of chunk boundaries.

If the FTP between the first bigram and the final unigram of a trigram fell below the running average for the same statistic, a chunk boundary was inserted. For instance, as the model encountered Z after seeing the bigram XY, it would calculate the FTP for the trigram by normalizing the frequency count of the trigram XYZ by the count of the bigram XY, and comparing the result to the running average FTP for previously encountered trigrams (inserting a chunk boundary if the running average was greater). The start-of-utterance marker made it possible for the Trigram model to place a boundary between the first and second words of an utterance.

During production attempts, which were also incremental and on-line in the style of CBL, the trigram model began constructing an utterance by choosing from the bag-of-words the word with the highest FTP, given the start-of-utterance marker (in other words, bigram statistics were used to select the first word). Each subsequent word was chosen according to trigram statistics, based on the two most recently placed words (or the initial word and the start-of-utterance marker, in the case of selecting the second word in an utterance). This meant the word with the highest FTP given the two preceding words was chosen at each time step.

PARSER: We implemented PARSER according to Perruchet and Vinter (1998) as well as personal communication with the first author. While a full description of PARSER is beyond the scope of the present article, it operates in much the same fashion as the Competitive Chunking model of Servan-Schreiber and Anderson (1990), but without building up a hierarchical network of chunks. Chunks are initially formed in PARSER through a stochastic process determining the size of percepts consisting of elementary units (in the present case, words). At each time step, chunks in the model's "lexicon" are affected by decay, with interference between partially overlapping chunks.

Perruchet (personal communication) advised us to retain the default values for the free parameters governing the threshold beyond which chunks shape percepts, the initial weight assigned to newly-segmented chunks, and weight added when existing chunks are reinforced by subsequent encounters(1.0, 1.0, and 0.5, respectively). Thus, the free parameters of primary interest for the present study were those governing decay and interference. We explored a range of values and adopted the one offering the best performance according to the gold standard for evaluating the models (described below).

While PARSER required no modifications to work with the shallow parsing task (merely the addition of a mechanism for recording its "percepts" as segmentations), Perruchet (personal communication) declined to offer suggestions for how it might be applied to sequencing in the bag-of-words task. PARSER, according to Perruchet and Vinter (1998), was designed merely to build up an inventory of chunks rather than capture any sort of on-line usage of those chunks: "The issue addressed by PARSER is quite different, insofar as it concerns the *creation* of the lexicon (p. 252; emphasis in the original)." As such, we chose to focus the use of PARSER on the comprehension-related shallow parsing task only. Moreover, substantial changes to the model would be necessary in order to adapt it for use with the bag-of-words task.

Simulation 1: Modeling Aspects of Child Comprehension and Production of English

In this section, we describe CBL simulations of child language learning and processing using English language corpora which capture interactions between children and their caretakers. We begin by describing the criteria used in selecting these corpora, followed by a description of the automated procedure used to prepare each corpus prior to its use as input in a simulation. Following this, we report the results of simulations for each corpus, comparing the performance of CBL to that of the two baseline models. For the sake of simplicity, we report performance for comprehension- and production-related tasks separately.

Corpus Descriptions and Preparation Procedure

In keeping with the key psychological features of the model, we initially sought to assess what could be learned by CBL from the input available to individual children. We therefore selected developmental corpora involving single target children, rather than aggregating data across multiple corpora. From the English language sections of the CHILDES database (MacWhinney, 2000), we selected every corpus meeting the following criteria:

1) *Sufficient data* – In order to locate corpora that had sufficient diversity in terms of both vocabulary and syntactic constructions, we included only those corpora which contained at least 50,000 words.

2) *Dyadic* – Because we wished to model both comprehension and production for each child, we selected only corpora which featured a multiword child-to-adult utterance ratio of at least 1:10.

3) *Developmental* – As we sought to model the developmental progression of each child's language learning, we included only those corpora that spanned at least a 6-month period (in terms of the target child's age across the corpus).

The three criteria were met by corpora for 42 individual English-learning children (US: 25, UK: 17). For use in subsequent analyses, we collected, for each child, the age range (mean age of 1;11 at the beginnings of the corpora, 3;7 at the ends), number of months spanned by the corpus (mean: 20.6), total number of words in the corpus (mean: 183,388), number of child utterances (mean: 20,990), number of multiword child utterances (mean: 12,417), number of adult utterances (mean: 33,645), child mean length of utterance (MLU; the mean number of morphemes per utterance; mean: 3.17), and child mean number of words per utterance (mean: 2.6).

Corpus preparation: The corpora were submitted to an automated procedure whereby codes, tags, and punctuation marks were removed, leaving only speaker identifiers and the original sequence of words. To ensure that the input available to the model was representative of what children actually receive, apostrophes were also removed from the corpora along with the other punctuation symbols. Thus, the contraction *it's* and the word *its*, for instance, were both represented orthographically as *its*, reflecting their identical phonological forms. This offered a naturalistic approach, considering developmental work indicating that children treat contractions as single words (cf. Tomasello, 2003).

Lines spanning tagged prosodic breaks, such as pauses (indicated in CHILDES by the (.) code), were broken into separate utterances, following research indicating that infants are sensitive to the suprasegmental properties of utterances, such as the acoustic correlates of clause boundaries (e.g., Hirsh-Pasek et al., 1987). Pauses due to hesitation (as indicated by the [/] code) were dealt with in the same manner. Finally, hash marks were added to the beginning of each line to signal the pause preceding each utterance.

Dense UK English corpus: The corpora in the CHILDES database typically represent a small percentage of the input a typical child might receive during the months spanned by the recording sessions. To examine subtle developmental trends with the model, a denser sample may be necessary. For this reason, we also tested the model using a dense corpus of child-directed speech which contains an estimated 8-10% of the target child's total productions (the Thomas corpus, originally known as the Brian corpus, which is now part of CHILDES; Maslen et al., 2004).

The dense corpus was submitted to the same automated procedure used to prepare the other CHILDES corpora. The prepared corpus spanned 36 months from age 2;0 to 5;0, featured 2,437,964 words, 225,848 child utterances, 114,120 multiword child utterances, 466,484 adult utterances, and an overall child MLU (in morphemes, as above) of 2.84.

Form class corpora: A considerable amount of work in computational linguistics has assumed that statistics computed over form classes are superior to item-based approaches for learning about structure (hence the widespread use of tagged corpora). This assumption is also present throughout the statistical learning literature (e.g., Thompson & Newport, 2007; Saffran, 2002), but is at odds with the present model, which relies on statistics computed over concrete words and chunks rather than classes. To evaluate the usefulness of item-based chunking and statistics against those computed over word classes, we ran the model and its alternates on separate versions of each corpus, in which words were replaced by the names of their lexical categories. This process was automatically carried out by tagging each corpus using TreeTagger, a widely used, probabilistic part-of-speech tagger based on decision trees (Schmid, 1995). The tag set used by TreeTagger was reduced to the following 12 categories: noun, verb, adjective, numeral, adverb, determiner, pronoun, preposition, conjunction, interjection, infinitive marker, and proper name. Unknown words (e.g., transcribed babbling) were marked as such. As we removed the punctuation from each corpus as part of the preparation procedure, contractions were handled straightforwardly: contractions involving verbs were classed as verbs, while possessives were classed as nouns. Thus, contractions were classed according to the type of phrase they immediately appeared in (noun vs. verb phrases). This allowed us to avoid the use of a tokenizer (which would reflect an assumption that children represent contractions such as *don't* as two separate words), while being motivated by psychological considerations (e.g., a child may treat an utterance such as *that's the car* similarly to *see the car*; the verb-like aspect of the whole contraction takes precedence).

Evaluating Model Performance

Gold standard for testing comprehension performance of model and baselines:

Shallow parsing: As the model approximated comprehension by segmenting the incoming input into semantically related, phrase-like multiword units, we evaluated the model's comprehension performance—as well as that of the three baseline models—against the gold standard of a shallow parser. Shallow parsing is a widely used technique in the field of natural language processing, which aims to segment text into non-hierarchical (i.e., non-embedded) phrases which are labeled according to phrase type. As an example, take the sentence *the dog chased the cat.* A shallow parser would group the words together into noun and verb groups: *[NP the dog] [VP chased] [NP the cat]*. This choice of gold standard reflects the psychological motivation for the model; as observed by Sanford and Sturt (2002), shallow parsing identifies a subset of possible analyses for a sentence rather than giving the type of articulated analysis created by full syntactic parsers. This is in line with the previously discussed evidence for underspecification in sentence comprehension, as well as the shallow processing approach we adopt more generally, in which chunks of local information are used to arrive at a semantic interpretation of a sentence.

For each corpus, we generated a shallow parse for all utterances using the Illinois Chunker (Punyakanok& Roth, 2001), a widely used shallow parser based on constraint satisfaction with classifiers in a probabilistic framework. Phrase tags were then removed, leaving only the original sequence of words segmented via the phrase boundaries placed by the parser.

The model's on-line comprehension performance was scored according to two measures: *accuracy* and *completeness*, which are analogous to precision and recall, respectively. Each boundary marker placed by the model was scored as a *hit* if it corresponded to a boundary marker inserted by the shallow parser, and as a *false alarm* otherwise. Each boundary inserted by the shallow parser which was not placed by the model was scored as a *miss*. Thus, accuracy could be calculated as the proportion of *hits* out of all boundaries placed by the model, *hits / (hits + false alarms)*, and completeness as the proportion of *hits* out of all boundaries placed by the shallow parser, *hits / (hits + misses)*. To avoid score inflation due to trivial factors, the model was only scored on utterance-internal boundaries (i.e., no boundary placement decisions were made at the beginnings or ends of utterances). Single-word utterances were naturally excluded.

For purposes of scoring the model's comprehension performance on the form class corpora (in which individual items were replace by their lexical categories), the set of phrase boundaries placed by the shallow parser and used as the gold standard for scoring the original corpus was overlaid on the corresponding form class corpus. For instance, the utterance “[the dog] [chased] [the cat],” became “[DET N] [V] [DET N]” in the form class version, which therefore featured identical phrase boundary markers.

As an overall measure of comprehension performance for a given simulation, we relied on the F-score, which is widely used as a measure of performance in the fields of information retrieval and machine learning (e.g., van Rijsbergen, 1979). The F-measure combines both the precision (or *accuracy*, in the current case) and recall (*completeness*) of a test to compute a single score. We used the general F_β formula, which weights the completeness score according to β:

$$F_{\beta} = \left(\frac{accuracy * completeness}{(\beta^{2} * accuracy) + completeness}\right) \quad (1)$$

In other words, the F_β metric attaches β times as much importance to completeness as to accuracy.

In our case, β is the ratio of gold standard phrase boundaries to the total number of word pairs (the number of possible slots for boundary insertion) across a given corpus. The choice of the F_β metric reflects the need to control for score inflation stemming from trivial factors, such as over-segmentation (e.g., due to data sparseness). As an example of this, consider a toy corpus which features phrase boundaries between exactly half of its word pairs. A model which heavily over-segments, placing phrase boundaries in every possible position, would receive a completeness score of 100%, and an accuracy score of 50%. By simply taking the harmonic mean of accuracy and completeness (what is known as the F_1 score), the model would receive an F-score of 66.67, despite its heavy over-segmentation. The F_β score, on the other hand, uses the number of word pairs straddling gold standard phrase boundaries to appropriately weight completeness in the calculation. For the previous example, this would yield an F-score of 55.56 (as opposed to the score 66.67 yielded by F_1) thereby reducing the impact of the perfect completeness score, which was achieved through trivial means (segmenting the corpus to the maximum extent). Weighting accuracy more heavily than completeness in this way is also motivated by psychological considerations: phrases like *go to the shop* might be chunked as a single item by a child (as suggested by the results of Bannard & Matthews, 2008), or the model, whereas a shallow parser would segment it into three separate chunks: [*go*] [*to*] [*the shop*]. Therefore, the calculation also reflects the fact that accuracy, which reflects the model's ability to place boundaries that correspond to actual phrase boundaries (e.g., after *shop* or before *the* instead of between *the* and *shop*), may be more important than following the fine-grained chunking of a shallow parser (which penalizes the model through the completeness measure for not placing boundaries after *go* or *to* in a phrase like *go to the store*).

Gold standard for production: Child utterances. Each utterance produced by the model is evaluated against the corresponding child utterance in the original corpus, according to a simple all-or-nothing criterion: if the model's utterance matches the child utterance in its entirety, a score of 1 is assigned. In all other cases, a score of 0 is assigned, despite the degree of similarity between the model- and child-produced utterances. Thus, the overall percentage of correctly produced utterances provides the *sentence production performance* for a given child/corpus. This represents a fairly conservative measure, as the model may produce sentences that are grammatical but nevertheless fail to match the target utterance. For example, the model may produce a sentence such as *the cat chased the dog* when the target sentence is *the dog chased the cat*. In such instances, the model receives a score of 0, due to the lack of principled and efficient way of automatically evaluating mismatching utterances that are nevertheless grammatical.

Parameter selection for PARSER: Following communication with the model's creator (Perruchet, personal communication), we adjusted the interference and decay parameters along a wide range, maintaining their separation by a factor of ten (following the settings used by Perruchet &Vinter, 1998), and selected the one offering the best combination of accuracy and completeness. At higher settings (e.g., Decay: 0.001, Interference: 0.0001), the model heavily over-segmented, placing boundaries between 90% of words. At settings 0.0001 and 0.00001 (for decay and interference, respectively), the model saw substantial improvements in accuracy while segmenting at a rate comparable to the CBL model. Decreasing the parameters by a further factor of ten lead to slight drops in accuracy and completeness. Thus, for the natural language simulations, we adopted settings of 0.0001 and 0.00001 for decay and interference, respectively.

Results and Discussion: Simulating Aspects of Comprehension and Production of English

Shallow parsing performance. Across all 43 single-child English corpora, CBL attained a mean F-score of 75.4, while the PARSER model attained a mean F-score of 66.1. The Trigram model had a mean F-score of 65.9. Comprehension performance for each model is shown in Figure 3. As can be seen, the CBL model not only outperformed its baselines, but yielded a tighter, more uniform distribution of scores across the corpora.

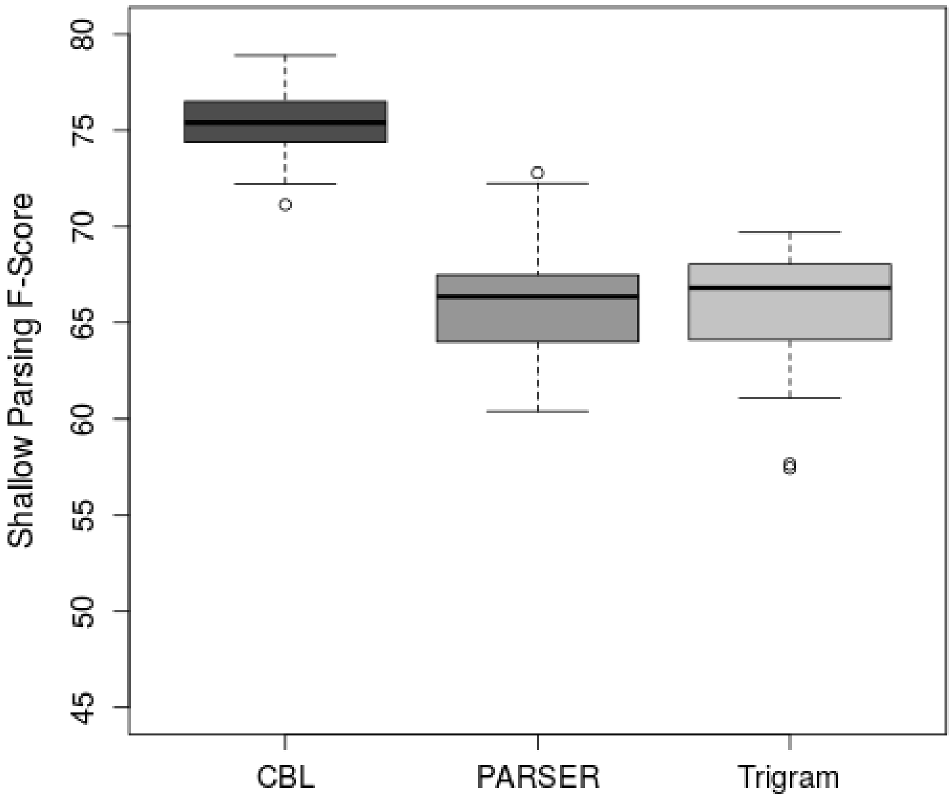

Fig. 3: Boxplots depicting shallow parsing F-scores for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges

representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

The F-scores for the model and its baselines were logit-transformed[7] and submitted to a repeated-measures ANOVA including the factor *Model* (3: CBL vs. PARSER vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant main effect of *Model* [F(2,84) = 643.3, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(42)=35.9, p<0.0001] and Trigram [t(42)=28.3, p<0.0001] models, with no significant difference in means between PARSER and the Trigram model [t(42)=0.49, p=0.63].

In line with the developmental motivation for the model, we also examined accuracy rates independently. Across the 43 child corpora, CBL achieved a mean accuracy rate of 76.4%, while PARSER attained a mean accuracy of 65.2% and the Trigram model reached a mean accuracy rate of 65.8%. The same general pattern was seen for completeness: CBL achieved a mean completeness of 73.8%, while the PARSER attained a mean completeness of 68.7% and the Trigram model reached a mean completeness rate of 66.5%. Accuracy and completeness scores are described more fully in Appendix A.

Thus, the best combination of accuracy and completeness (as measured by the F-score), as well as the best accuracy and completeness overall, was achieved by CBL's statistically-based chunking for the English child corpora. CBL was able to approximate the performance of a shallow parser through a combination of

[7] As the scores necessarily have both floors and ceilings, and represent proportional data, a logit transformation was applied prior to analysis in order to fit the assumptions of the test.

recognition- and statistically-based processing in an on-line, incremental fashion starting with a single distributional cue. This result is encouraging, as shallow parsing is regarded as a nontrivial problem in the field of natural language processing (e.g., Hammerton, Osborne, Armstrong, & Daelemans, 2002).

In addition to highlighting the wealth of distributional information in the input, these results suggest that purely item-based information may be far more useful to early learners than has been assumed previously; in addition to providing the basis for discovering useful multiword sequences (which may later be abstracted over, as proposed by usage-based approaches more generally; e.g., Tomasello, 2003), statistical information tied to concrete items can be used to uncover the chunks of local information necessary to interpret sentences ("phrase structure," in most approaches), as demonstrated by the present model.

CBL and the Trigram model may have outperformed PARSER in part because of the latter model's over-reliance on raw frequency of occurrence. For instance, CBL can identify high TPs between items which have occurred with very low frequency in the corpus: the relative, rather than absolute, frequency of the two items is stressed. While PARSER is indirectly sensitive to TPs, via its decay and interference parameters, the use of these parameters along with randomly determined percept sizes may requires more exposure.

CBL also has the additional advantage of being directly sensitive to *background rates* (cf. Ramscar, Dye, & McCauley, 2013). Words that occur extremely often in a variety of contexts have high background rates, which mean they are less informative about the items preceding them (or following them, in the case of the

Trigram model). Conditional probabilities directly reflect this. PARSER is only indirectly sensitive to background rates, through its interference feature: items that occur often as parts of larger chunks will lead to decreases in the strength of chunks featuring the same item. However, the impact of the decay parameter on less-frequent chunks may still lead to an overemphasis on items with high background rates.

For these reasons, PARSER may ultimately be best suited to working with small, artificial languages which involve fairly uniform frequency distributions over items, such as those featured in studies to which it has previously been applied (e.g., Perruchet et al., 2002; Saffran et al., 1996).

Development of the chunkatory. The development of the model's knowledge over the course of a simulation, independently of its performance, may offer potential predictions on which to base psycholinguistic work. To examine the development of the chunkatory, we tracked the percentage of stored chunks which consisted of multiple words, for chunk types as well as chunk types weighted by their frequency counts in the chunkatory (akin to chunk tokens). Figure 4a shows the percentage of multiword chunks in the chunkatory as it develops, for each of the 43 child corpora. As can be seen, the percentage of multiword chunk types increased logarithmically over the course of the simulations, leveling off below 90%. The first data point fell above 50% for all child corpora. When we weighted individual chunk types by their strength (frequency counts) in the chunkatory, however, we found higher percentages for single-word chunks, as shown in Figure 4b.

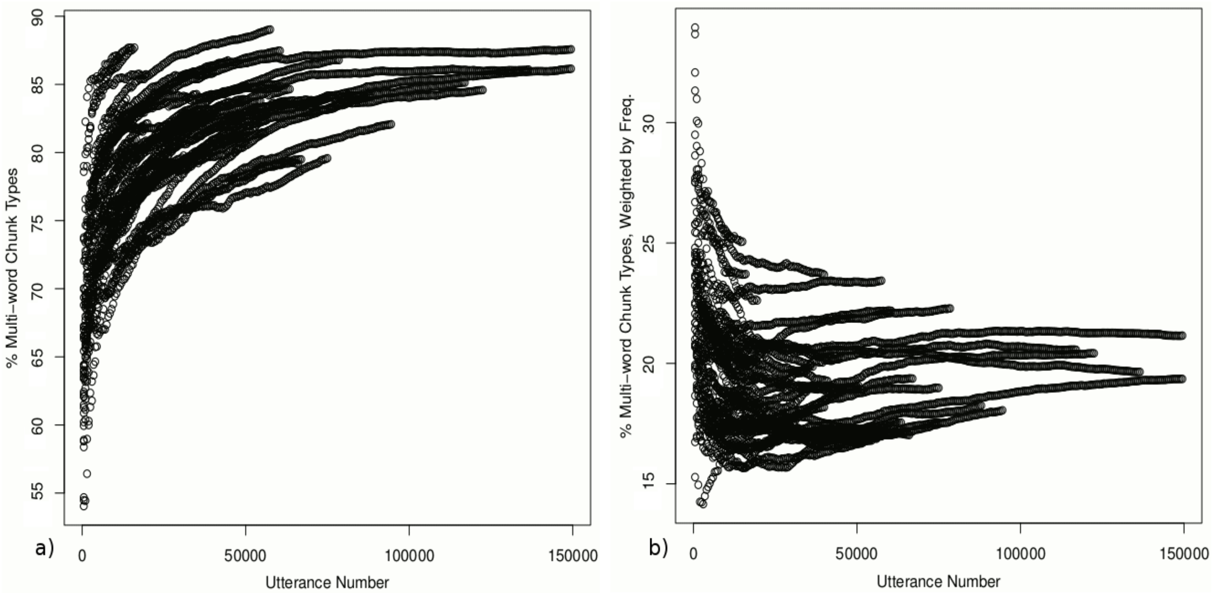

Fig. 4: a) Development of the chunkatory by percentage of multiword types; b) Development of the chunkatory by percentage of multiword types weighted by frequency of use.

As can be seen, the percentage of multiword chunk types, when weighted by frequency, began well below 50% at the first data point for all corpora, with percentages for many of the child simulations dipping within the first 20,000 utterances before rising sharply and then climbing more steadily.

The dense corpus (Thomas) exemplifies the same pattern. To look more closely at the makeup of chunks we might expect the child to actively use in production, we calculated the percentage of multiword chunk types, weighted by frequency, which were actively used by the model during the bag-of-words task. This is depicted in Figure 5 for the dense corpus (Thomas).

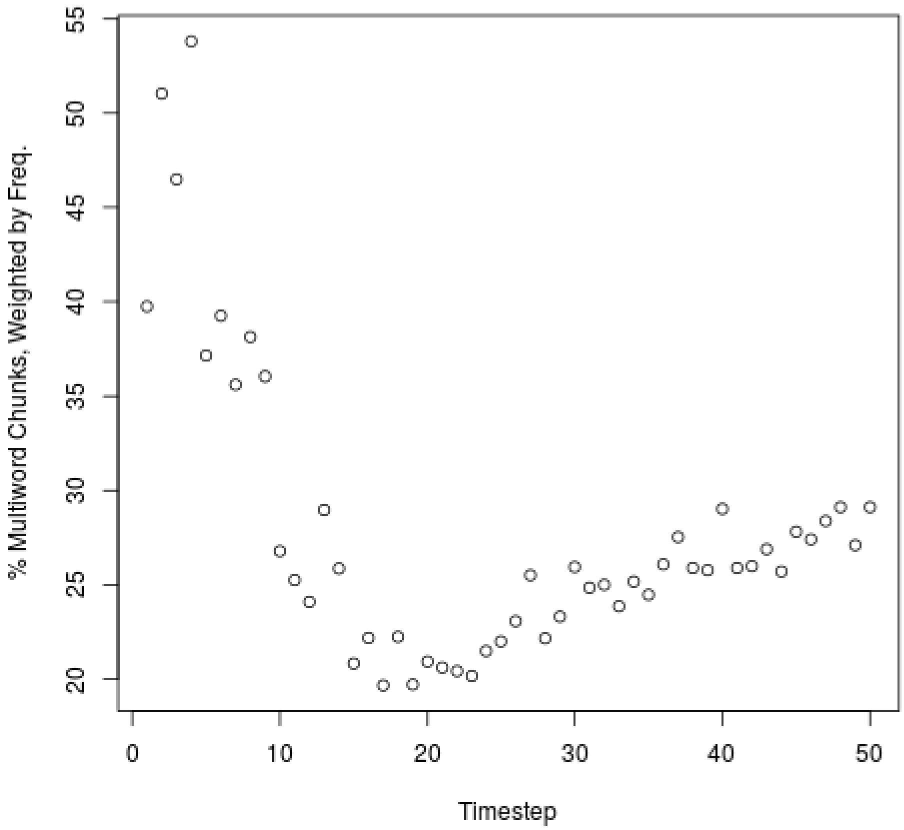

Fig. 5: Percentage of multiword types weighted by frequency of use in the production task for the dense corpus (Thomas). Each time step represents the mean percentage across 2,000 child utterances.

As can be seen, the prominence of multiword chunks in the chunkatory for the Thomas simulation mirrors the general pattern illustrated by Figure 4b, dipping early before rising once more. However, because the dense corpus extends well past the other corpora in terms of length, we were able to look at a more complete trajectory, which included a sharp dip followed by a more subtle increase which spanned the remainder of the simulation.

The U-shaped curve exhibited by the model mirrors a common developmental pattern which has been tied to several aspects of language learning, including phonological development (Stemberger, Bernhardt, & Johnson, 1999), morphological development (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992), relative clause comprehension (Gagliardi, Mease, &Lidz, submitted), and verb usage (Alishahi & Stevenson, 2008; Bowerman, 1982). Though the model's shifting reliance on chunks of differing granularity represents a small change numerically during the more stable latter half of the simulation, subtle changes in distributional statistics may have dramatic outcomes. The model's behavior leads us to consider that children's reliance on multiword chunks may shift in similar ways to that of the model, and that this may have some bearing on U-shaped trajectories in other areas, such as morphological development. For instance, Arnon and Clark (2011) found that over-regularization errors were less likely when irregular plurals were produced in the context of a lexically-specific frame; the facilitatory role played by chunks in this area (and others) may wax and wane with the "degree of chunkedness" of the child's linguistic representations, consistent with preliminary findings comparing children of different ages (Arnon, personal communication).

In summary, multiword chunks ultimately grow in importance to the model over the course of a simulation, both in terms of types and in terms of tokens. In light of psycholinguistic work with adults (Arnon & Snider, 2010; Bannard & Ramscar, 2007), this leads us to predict that children do not merely "start big" by relying on larger multiword sequences which break down over time, leaving single words; the child's memory-based processing is dynamic, and the degree to which representations of linguistic material are tied to multiword sequences ultimately grows in importance over time. The model's U-shaped reliance on weighted multiword chunks also leads us to propose that children may go through periods where new knowledge of the properties of single words may lead to a decreased reliance on multiword sequences, only to be followed by a renewed reliance on chunked representations.

Class- vs. item-based comprehension performance. As discussed above, most generative approaches as well as certain trends within the statistical learning literature (cf. Thompson & Newport, 2007) have assumed that language learning is tied to word classes. For this reason, we re-ran the 43 simulations reported above, using the form class corpora (see the corpus preparation above for a description of how words were converted to form classes).

Because PARSER is sensitive to overall number of unit types it is exposed to, we found that the highest parameter setting we tested for natural language (0.01 for decay, and 0.001 for interference) provided the best trade-off between accuracy and completeness when working with form classes.

Class- versus item-based performance for the model and its baselines is depicted in Figure 6. Performance for CBL was considerably worse when working with class-based statistics, with a sharp decrease in the mean F-score (from 75.4 to 39). For PARSER there was a far less drastic decrease in performance (from a mean F-score of 66.1 to 63.1). The Trigram baseline also fared worse under class-based statistics, though with less dramatic decreases in performance. The mean F-score dropped from 65.9 to 57.2.

In the case of the CBL model, the lower comprehension performance when working with class statistics was driven both by a drop in accuracy as well as a more drastic drop in completeness scores, the latter owing partly to the use of the chunkatory in phrase segmentation; the relatively small number of possible class combinations in a sequence lead to the automatic grouping of items together (based on the use of the chunkatory) with increasing frequency throughout the models' pass through a corpus. As more combinations exceeded the average TP threshold, the models placed progressively fewer phrase boundaries. PARSER, however, saw a slight increase in accuracy accompanied by a decrease in completeness. As PARSER was designed for use with small item sets, as discussed above, further experimentation with the parameter settings of PARSER may be necessary in order to improve performance on the form-class simulations.

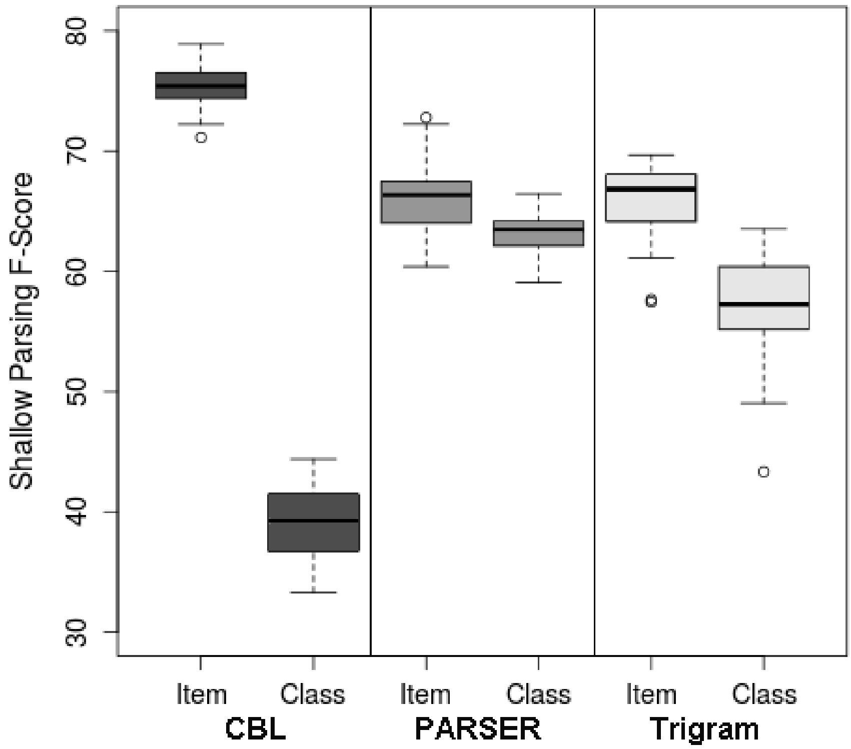

Fig. 6: Boxplots depicting comprehension performance (F-scores) for the CBL model and its baselines, comparing item-vs. class-based simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within the 1.5 IQR of the quartiles, while dots depict outliers.

We evaluated the effects of learning from class-based information using a two-way ANOVA with factors *Statistic Type* (2: item- vs. class-based) and *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor, over logit-transformed F-scores. This yielded main effects of *Statistic Type* [F(1,42) = 1562, p<0.0001], confirming stronger performance for item-based models, and of *Model* [F(2,84)=171.4, p<0.0001], with an interaction between *Statistic Type* and *Model* [F(2,84)=25.5, p<0.0001], due to a more drastic drop in performance for the CBL model relative to the baselines when working with classes.

Thus, a reliance on word classes did not improve the performance of the model or its baselines; instead, knowledge of classes lead to a decrease in performance, which was considerably more drastic for CBL. This result makes close contact with item-based approaches more generally (e.g., Tomasello, 2003), suggesting that successful language learning can take place without the sort of abstract syntactic categories that much previous psycholinguistic and computational work has focused upon. This also runs counter to claims made in the statistical learning literature that children and adults can use transitional probabilities to segment phrases by calculating statistics over word classes rather than concrete items (e.g., Saffran, 2002; Thompson & Newport, 2007). Indeed, we have shown elsewhere (McCauley & Christiansen, 2011) that comprehension through item-based learning in our model captures subject performance in one such study (Saffran, 2002) better than class-based learning.

Production performance. Across all 43 single-child corpora, CBL achieved a mean sentence production performance of 58.5%, while the Trigram model achieved a sentence production performance score of 45.0%. Recall that PARSER was not compatible with the production task (for reasons discussed in the Methods section above). The distributions of the scores for each model are depicted in Figure 7. As can be seen, the overall pattern of results was similar to that seen with comprehension, with CBL achieving the highest mean score.

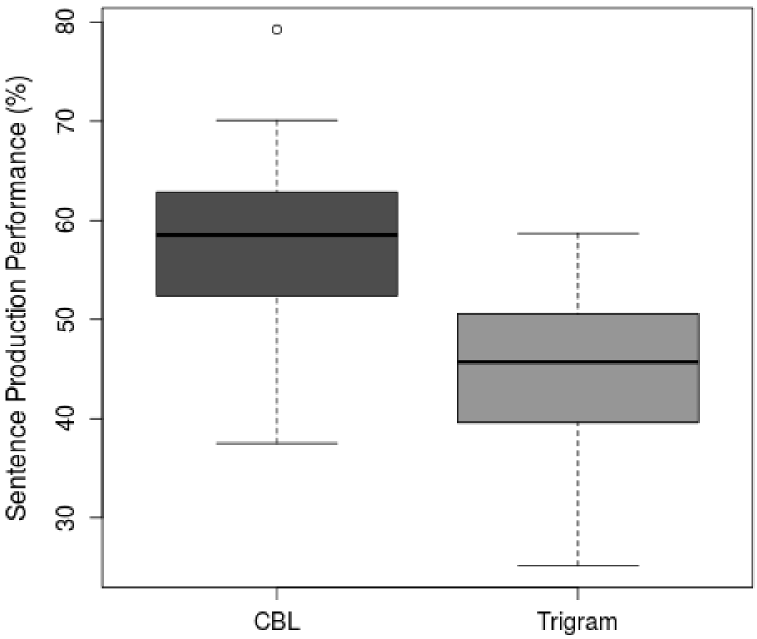

Fig. 7: Boxplots depicting Sentence Production Performance scores for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

A repeated-measures ANOVA including the factor *Model* (2: CBL vs. Trigram), with *Child Corpus* as a random factor, yielded a significant effect of *Model* [F(1,42) = 514.9, p < 0.0001], indicating better performance for CBL.

Thus, CBL exhibited clear advantages over the baseline. The advantage of stored-chunks (which do not have to be sequenced, in and of themselves) is clear in these results. What is less clear is that BTPs may offer an advantage over FTPs when there is a limited, specified set of possible items that can follow the most recently placed item in a sequence (such as a bag-of-words, in the present instance). The FTP-based Trigram model simply selects, at each time step, the item combination with the highest frequency, since the frequency of every possible sequence is normalized by the preceding item, which is fixed, as it has already been produced. The CBL model, however, through the use of BTPs, is sensitive to the background rate (discussed above; cf. Ramscar et al., 2013) of the candidate items: items that occur more often in contexts other than the present one will not be selected.

Production summary and discussion. CBL not only outperformed its baselines in reproducing child utterances, but was able to produce the majority of the target utterances it encountered, with a mean score of nearly 60% based on our conservative all-or-nothing measure of production performance. This not only underscores the usefulness of chunks and simple statistics such as BTPs, but serves to demonstrate that the same sources of information can be useful for learning about structure at multiple levels: a single distributional statistic (BTP) can be used to segment words when calculated over syllables (e.g., Pelucchi et al., 2009), to discover multiword sequences (or perhaps even "phrase structure") when calculated over words (demonstrated by the present model), and to construct utterances when calculated over stored multiword chunks themselves (demonstrated in the current section).

Despite this success, the model nevertheless failed to account for 40% of the child utterances, a significant proportion, and yielded a less dramatic advantage over its baseline than was the case with comprehension. This pattern of results, when considered alongside the idea of shallow processing as a central feature of a child's language comprehension, has immediate implications for the comprehension/production asymmetry that has been so puzzling to developmental psycholinguists. Through shallow processing of the sort captured by the model, a child can give the appearance of having utilized a construction (such as a transitive construction in canonical word order) in comprehension while still lacking the sequential knowledge to use it in production. This is especially true if one considers specific aspects of shallow processing in adults, as well as its ubiquitous nature in language comprehension more generally. Take, for instance, passive sentences. Ferreira (2003) found that when adults were exposed to anomalous sentences using passive constructions (*"The dog was bitten by the man"*) many readers utilized global pragmatic information rather than the passive construction to identify agents and patients of actions, and gave higher plausibility ratings than for the same content given in active voice (i.e., they interpreted the passive sentence to mean the dog bit the man). If even adults tend to interpret sentences according to local information, pragmatic expectations, and background world knowledge rather than the actual constructions used in the sentence, it seems likely that children also utilize such information, giving the appearance of having fully exploited a grammatical construction when in actuality they were able to interpret an utterance on a more superficial level.

Under such a view, children may understand specific utterances utilizing passive voice without having mastered the passive construction; to actively sequence the elements of its message into a passively voiced sentence, the child would necessarily need to have full knowledge of the passive construction schema (such as *PATIENT is ACTION by AGENT*) as well as knowledge of the pragmatic motivations for using it (for a model which learns to produce sentences using such semantic role information, see Chang, Dell, & Bock, 2006). Thus, shallow processing allows a child to make a great deal of progress towards understanding language input merely on the basis of an ability to chunk parts of utterances and form associations between those chunks and concrete parts of the world, as well as event schemas or scenarios; it is in the sequencing of those chunks that the problem of production becomes more difficult than that of comprehension. This idea is explored further using the CBL model by Chater et al. (2016).

Interim summary: Learning English. We have shown that the CBL model is able to approximate shallow parsing through the incremental, on-line discovery and processing of multiword chunks, using simple statistics computed over local information. The model is able to use the same chunks and statistics to produce utterances in an incremental fashion, capturing a considerable part of children's early linguistic behavior in the process. This is achieved through item-based learning, without recourse to abstract categorical information such as that of word classes. When the model learns class-based statistics, its ability to segment useful chunks is impaired. Furthermore, the development of the model's chunk inventory offers the novel prediction that subtle shifts in the “degree of chunkedness” of children's linguistic units may impact on other areas of language development. Finally, the model, which combines chunking with statistical cues, compares favorably to exclusively recognition-based and exclusively prediction-based baselines.

Simulation 2: Modeling the Development of Complex Sentence Processing Abilities

Whereas the previous simulations examined the ability of CBL to discover building blocks for language learning, in the present section we investigate the psychological validity of these building blocks. We report simulations of empirical data covering children's ability to process complex sentence types (Diessel & Tomasello, 2005). Usage-based approaches predict that stored chunks play an integral role in the development of complex grammatical abilities (e.g., Christiansen & Chater, 2016a; Tomasello, 2003), which have been argued to emerge from abstraction over multiword sequences (e.g., Goldberg, 2006; for models, see Kolodny et al., 2015; Solan et al., 2005).

Nevertheless, there is strong evidence of a role for concrete multiword chunks in adult processing of grammatically complex sentences, such those featuring embedded relative clauses (e.g., Reali & Christiansen, 2007), which in turn suggests that children's ability to comprehend and produce complex sentences should be influenced by the same type information. If this holds true, and if CBL provides a reasonable approximation of children's discovery and use of chunks, the model should be able to offer some insight into the development of complex grammatical abilities, despite its lack of abstract grammatical knowledge. In order to test this notion, we used CBL to model children's ability to produce different relative clause types (Diessel and Tomasello, 2005), as a great deal of previous developmental work on grammatically complex sentences has focused on relative clause constructions (see Christiansen & Chater, 2016a, for a review).

This particular study was chosen because its stimuli were designed to reflect the types of relative constructions children actually produce in spontaneous speech (specifically, those that attach to either the predicate nominal of a copular clause, or to an isolated head noun; Diessel, 2004; Diessel & Tomasello, 2000). Prior to this study, developmental work on relative clauses focused mainly on sentence types which children rarely produce spontaneously, and which therefore may not adequately reflect children's grammatical knowledge (e.g., Hamburger & Crain, 1982; Keenan & Hawkins, 1987; Tavakolian, 1977). Of further importance is the study's focus on children's production abilities as opposed to just comprehension; because the stimuli consisted of whole sentences, this allowed us to model child performance using the entire model architecture (comprehension as well as production).

Using a repetition paradigm, Diessel and Tomasello exposed a group of UK English-speaking children (mean age: 4;7) to sentences featuring one of six relative clause types (*subject relatives featuring intransitive verbs, subject relatives featuring transitive verbs, direct-object relatives, indirect-object relatives, oblique relatives, and genitive relatives*). An example of each relative clause type is shown in Table 2. Following exposure to a sentence, the child was prompted to repeat it to the experimenter. The authors found that children's ability to reproduce the relative clauses closely mirrored the similarity of each clause type to simple non-embedded sentences (with the greatest accuracy for subject relatives).

Table 2

Relative Clause Types from Diessel and Tomasello (2005)

Type	Example
S	There’s the boy who played in the garden yesterday.
A	That’s the man who saw Peter on the bus this morning.
P	That’s the girl who the boy teased at school this morning.
IO	There’s the girl who Peter borrowed a football from.
OBL	That’s the dog that the cat ran away from this morning.
GEN	That’s the woman whose cat caught a mouse yesterday.

Method. We began by exposing CBL to a corpus of UK English. As the original English study was conducted in Manchester, we focused on the dense Thomas corpus (which was recorded in Manchester; Maslen et al., 2004). Following exposure to the corpus, CBL was presented with the same test sentences heard by children in the original study. Immediately following comprehension on a given test sentence, the model simulated a repetition trial by attempting to produce the utterance (using the bag-of-words task in an identical manner to the child utterances in our original natural language simulations). If the utterance produced by the model matched the target utterance in its entirety, a score of 1 was assigned; otherwise, a score of 0 was assigned.

To order the test items, we used the same randomization procedure as was used in the original study: items were organized into four consecutive blocks of six randomly chosen sentences, with the constraint that each block included one sentence from each condition (Diessel, personal communication). This randomization allowed for small individual differences to arise between simulations (21 in total; a different randomization/simulation pair for each child in the original study).

Results and discussion. The children in the original study achieved the following correct response rates, as shown in Figure 8: 82.7% (S-Rel), 59.5% (A-Rel), 40.5% (P-Rel), 31% (IO-Rel), 31.5% (OBL-Rel), and 2.5% (Gen-Rel). As also shown in Figure 8, correct response rates for the model were 77.4% (S-Rel), 48.8% (A-Rel), 75% (P-Rel), 39.3% (IO-Rel), 34.5% (OBL-Rel), and 16.7% (GEN-Rel). As can be seen, the model followed the same general pattern as the children in the original study, with the exception of its performance on P-Relatives, which was almost as high as its performance on S-Relatives.

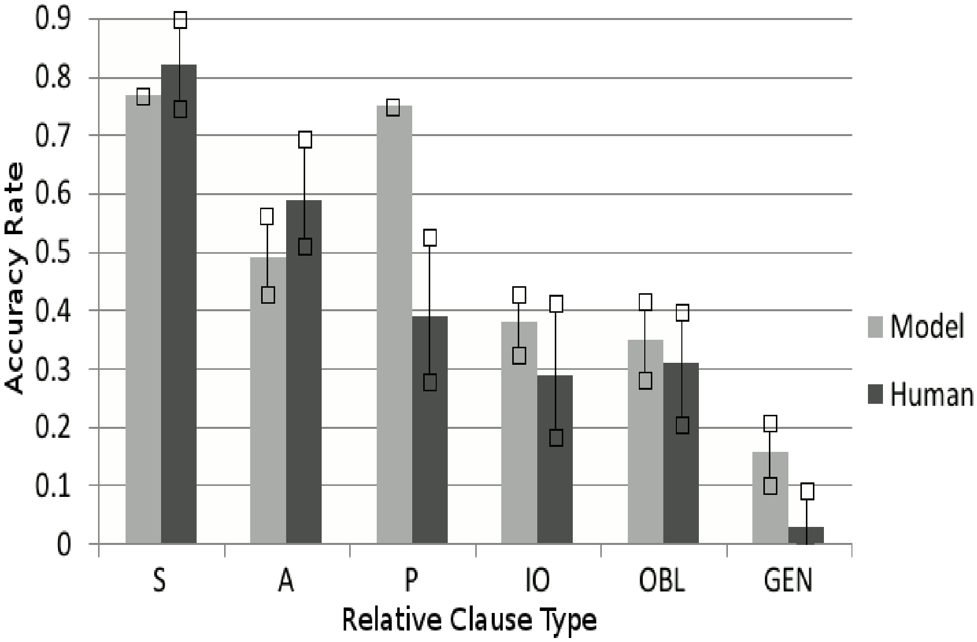

Fig. 8: Mean correct response rates for CBL model and the child subjects in Diessel & Tomasello (2005). Error bars denote 2x standard error.

That the model was able to mirror the child repetition performance for most of the clause types is unexpected, considering its complete lack of semantic/pragmatic information or distributional information spanning non-adjacent chunks (for a connectionist simulation of Diessel & Tomasello, 2005 which incorporates semantic information, see Fitz & Chang, 2008).

Despite the model's decent fit to the child data for 5 of the 6 relative clause types, its over-performance on the P-Relatives serves as a reminder of the limits of a purely distributional approach; semantics obviously plays a role not only in children's on-line chunking of test utterances upon first exposure (corresponding to the comprehension side of the model), but their incremental production during repetition attempts (corresponding to the sequencing stage of production in the model), and their recall of chunks throughout both (corresponding to recognition-based processing during comprehension and the retrieval stage of production in the model). As the model receives no input related to semantic roles, it received no information on the *patient* role of the main clause subject within the P-Relatives, and hence no interference from its deviation from the *agent-action-patient* sequence most commonly encountered in simple non-embedded sentences. Instead, the model relied on purely item-based similarity or dissimilarity to sentences in the child-directed speech it initially learned from. Moreover, the model's high performance on P-Rel trials might also stem in part from the nature of its input: P-Rels may be even more common in English child-directed speech than A- and S-Rels. In an analysis by Diessel (2004), 56.8% of all relative clauses produced by four English-speaking parents were P-relatives, while 35.6% were S- or A-relatives, 7.6% were OBL-relatives, and IO- and GEN-relatives did not occur.

Diessel and Tomasello (2005), in accord with usage-based approaches, propose that young children's ability to produce relative clauses depends on the degree of similarity between the type of relative clause and simple non-embedded sentences of the sort encountered most often in child-directed speech (an idea which has received considerable empirical support in recent years; e.g., Brandt, Diessel, & Tomasello, 2008; see Christiansen & Chater, 2016a, for a review). This stands in contrast to the hypothesis that the distance between filler and gap determines processing difficulty (Wanner& Maratsos, 1978), which initially sought to explain the well-documented phenomenon of greater processing difficulty for object relative as opposed to subject relative clauses (as demonstrated in Dutch, English, and French; Wanner& Maratsos; 1978, Frauenfelder, Segui, &Mehler, 1980; Holmes & O'Regan, 1981; Ford, 1983; Frazier, 1985; King & Just, 1991; Cohen &Mehler, 1996; though see also Brandt, Kidd, Lieven, & Tomasello, 2009). Under this view, object relatives cause more difficulties than subject relatives because they feature a greater distance between filler and gap, and the filler must be retained in working memory until the gap is encountered. The distance between filler and gap has also been hypothesized to play a role in the ease of acquisition of relative clauses, favoring relative clauses with a short distance between filler and gap (e.g., de Villiers et al., 1979; Clancy, Lee, &Zoh, 1986; Keenan & Hawkins, 1987).

In CBL, both comprehension and production rely on statistics computed over adjacent chunks; the model has no "working memory," and thus no sensitivity to the distance between filler and gap in relative clause constructions. Nevertheless, we observe better performance on the S-Rels (for which the distance between filler and gap is the smallest) than other relative clause types. At the same time, we fit the better S- than A-Rel pattern exhibited by the children in the original study; as Diessel and Tomasello note, the distance between filler and gap cannot account for this result, as the distance is the same in both relative clause types. Furthermore, we observed the worst performance with GEN-Rels (again, a pattern exhibited by the child subjects), despite the small distance between filler and gap for these sentences. Our results therefore have a direct bearing on the filler-gap hypothesis, beyond merely reinforcing Diessel and Tomasello's findings, suggesting that children's item-specific knowledge may play a greater role in relative clause processing than working memory constraints (see Christiansen & Chater, 2016a; MacDonald & Christiansen, 2002; McCauley & Christiansen, 2015, for extensions of this perspective to individual differences in adult relative clause processing).

The results of this simulation also allow us to derive a novel prediction from the model: the very factor driving the model's performance, item-based statistics computed over adjacent chunks, may well be a factor in the apparent ease with which children learn to produce certain kinds of sentences while encountering difficulties in learning to produce others. Indeed, previous evidence from children's elicited question formation indicates a role for such surface distributional statistics (Ambridge, Rowland, & Pine, 2008). This prediction can be tested further with a simple repetition paradigm such as that used by Diessel and Tomasello, using stimuli which systematically pit adjacent chunk statistics against statistics derived from large corpora of child and child-directed speech, in such a way that local information conflicts with the global properties of coherent target utterances.

Modeling Child Comprehension and Production across a Typologically Diverse Array of Natural Languages

We have shown that CBL can capture a considerable part of children's early linguistic behavior, in addition to making close contact with developmental psycholinguistic data from a key study on children's item-based distributional learning. Nevertheless, these findings—like most of the psycholinguistic findings forming the basis for the model—are based entirely on the use of English data; the computational approach we have adopted may not actually characterize aspects of learning held in common by learners of typologically different languages. In the next series of simulations, we explore the question of whether the model can extend beyond English to cover a typologically diverse set of languages.

The goal of attaining broad, cross-linguistic coverage extends beyond merely building support for the model; we aim to address certain limitations of the psycholinguistic literature. For instance, a potential problem with the view of multiword chunks as an important feature of language use is that most of the directly supporting psycholinguistic evidence has been gathered from English-speaking subjects. Importantly, English is an *analytic* language; it has a low ratio of words to morphemes, relative to *synthetic* languages, which have higher ratios due to the many ways in which morphemes can be combined into words. What may apply to arguments about unit size in the learning of analytic languages (such as Mandarin or English) may not apply to the learning of synthetic languages (such as Tamil or Polish), and vice versa. It is therefore essential to test the predictions of both CBL and previous empirical work with English-speaking subjects by modeling chunk-based learning across a typologically diverse set of languages. The breadth of material in the CHILDES database (MacWhinney, 2000) makes it possible to test the model on a typologically diverse array of languages.

Following a description of the corpora used to simulate learning cross-linguistically, we report comprehension performance for the languages for which an automated scoring method was available. We then report sentence production performance for 28 additional languages.

Corpus Selection and Preparation

Corpora were selected from the CHILDES database (MacWhinney, 2000), and covered a typologically diverse set of languages, representing 15 genera from 9 different language families (Haspelmath, Dryer, Gil, & Comrie, 2005). As with the English simulations, we sought to assess what could be learned by CBL from the input available to individual children. We therefore selected, once more, developmental corpora involving single target children, rather than aggregating data across multiple corpora. However, due to the limited availability and size of corpora representing several of the languages in the CHILDES database (MacWhinney, 2000), we relaxed our criteria somewhat. Thus, we used corpora that met the following criteria:

1) *Sufficient data* – As we sought to use corpora of a sufficient density to offer input of representative diversity in terms of both vocabulary and sentence types, we included only those corpora which contained at least 10,000 words.

2) *Dyadic* – Because we wished to model production for each child, we selected only corpora which featured at least 1000 multiword child utterances, with a multiword child-to-adult utterance ratio of no less than 1:20.

These criteria were met by corpora for 160 individual children (Afrikaans: 2, Cantonese: 8, Catalan: 4, Croatian: 3, Danish: 2, Dutch: 12, Estonian: 3, Farsi: 2, French: 15, German: 22, Greek: 1, Hebrew: 6, Hungarian: 4, Indonesian: 8, Irish: 1, Italian: 8, Japanese: 10, Korean: 1, Mandarin: 7, Polish: 11, Portuguese: 2, Romanian: 1, Russian: 2, Sesotho: 3, Spanish: 11, Swedish: 5, Tamil: 1, Welsh: 6). We recorded, for each child, the age range (mean age of 1;11 at the beginnings of the corpora, 3;10 at the ends), the number of months spanned by the corpus (mean: 23), the total number of words in the corpus (mean: 103,555), the number of child utterances (mean: 14,248), the number of multiword child utterances (mean: 7,552.7), the number of adult utterances (mean: 23,207), the multiword child-to-adult utterance ratio (mean: 0.49), and the mean words per child utterance (overall mean: 2.3). Since a method for automated morpheme segmentation was not available for all 28 languages, we do not include MLU calculations.

The final set of 28 languages (including English, 29) differed typologically from one another in a number of important ways. Four dominant word orders were represented: SVO (18), VSO (2), SOV (4), and no dominant order (5; Haspelmath et al. 2005). The languages varied widely in their morphological complexity, ranging from languages with no morphological case marking (e.g., Sesotho; Demuth, 1992) to languages with 10 or more cases (e.g., Estonian; Haspelmath et al., 2005). Table 3 shows the family, genus, dominant word order, and number of cases for each of the 28 languages, in addition to English.

Table 3

Typological Properties of the 29 Languages

Language	Family	Genus	Word Order	# Cases
Irish	Indo-European	*Celtic*	VSO	2
Welsh	Indo-European	*Celtic*	VSO	0
English	Indo-European	*Germanic*	SVO	2
German	Indo-European	*Germanic*	N.D.	4
Afrikaans	Indo-European	*Germanic*	N.D.	0
Dutch	Indo-European	*Germanic*	N.D.	0
Danish	Indo-European	*Germanic*	SVO	2
Swedish	Indo-European	*Germanic*	SVO	2
Greek	Indo-European	*Greek*	N.D.	3
Farsi	Indo-European	*Iranian*	SOV	2
Romanian	Indo-European	*Romance*	SVO	2
Portuguese	Indo-European	*Romance*	SVO	0
Catalan	Indo-European	*Romance*	SVO	0
French	Indo-European	*Romance*	SVO	0
Spanish	Indo-European	*Romance*	SVO	0
Italian	Indo-European	*Romance*	SVO	0
Croatian	Indo-European	*Slavic*	SVO	5
Russian	Indo-European	*Slavic*	SVO	7
Polish	Indo-European	*Slavic*	SVO	7
Estonian	Uralic	*Finnic*	SVO	10+
Hungarian	Uralic	*Ugric*	N.D.	10+
Sesotho	Niger-Congo	*Bantoid*	SVO	0*
Hebrew	Afro-Asiatic	*Semitic*	SVO	0
Tamil	Dravidian	*S. Dravidian*	SOV	7 or 8**
Indonesian	Austronesian	*Malayic*	SVO	0

Cantonese	Sino-Tibetan	*Chinese*	SVO	0
Mandarin	Sino-Tibetan	*Chinese*	SVO	0
Korean	Korean	*Korean*	SOV	7
Japanese	Japanese	*Japanese*	SOV	9

Note: Information from Haspelmath et al. (2005), except where noted otherwise

**Demuth, 1992*

***Schiffman, 1999*

We sought to gauge the morphological complexity of the languages quantitatively, thereby placing them on an analytic/synthetic spectrum (Greenberg, 1960). Analytic languages such as Mandarin or English have a low morpheme-to-word ratio, whereas synthetic languages like Polish or Hungarian have a high morpheme-to-word ratio. We therefore carried out an analysis of the type/token ratio for each language (following Chang et al., 2008). This allowed us to approximate the morpheme-to-word ratio differences between languages without the aid of an automated method for morpheme segmentation: Morphological richness mirrors type/token ratio in the sense that morphologically complex languages yield a greater number of unique morpheme combinations, and thus a higher number of unique word types, relative to the number of tokens, whereas analytic languages rely on a smaller number of unique morpheme combinations.

Thus, type/token ratio was used to compute a *Morphological Complexity* score for each language. For the type/token ratio calculation, we used only the adult utterances in the included corpora. Because type/token ratios are highly sensitive to the size of speech samples, we controlled for the lengths of individual corpora by calculating the mean type/token ratio per 2,000 words across all corpora representing a given language. The results of these calculations are depicted on an analytic/synthetic spectrum in Figure 9, and demonstrate the wide variation of the 29 languages, including English, in terms of morphological complexity. While some languages had relatively low *Morphological Complexity* scores (e.g., Cantonese), others had much higher scores (e.g., Tamil), and others had scores falling between the two (e.g., Sesotho).

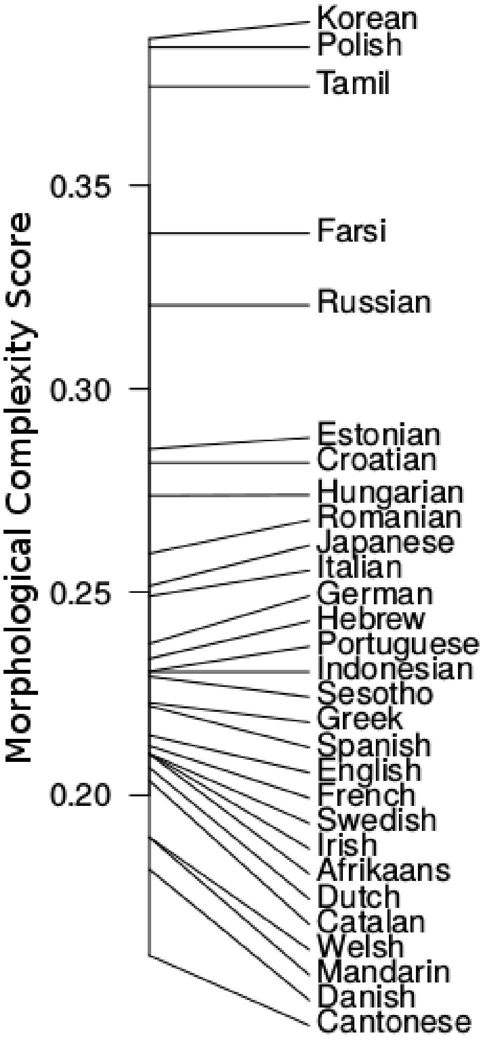

Fig. 9: Morphological complexity scores for each of the 29 languages.

Simulation 3: Modeling Child Comprehension of French and German

Shallow parsers (providing a gold standard for comprehension performance) were only available to us for two of the additional languages: French and German. TreeTagger (Schmid, 1995) was used to evaluate comprehension-related performance (through shallow parsing) for both languages. In this section, we report shallow parsing performance for French and German CBL simulations. Fifteen French and 22 German child corpora in the CHILDES database met our selection criteria and were used to simulate aspects of comprehension and production in exactly the same model architecture as used in the English simulations (as was also the case for the baseline models).

French: Comprehension performance. Across all 15 French single-child corpora, CBL achieved a mean F-score of 71.6, while the PARSER model reached a mean F-score of 64.4. The Trigram model attained a mean F-score of 59.0. Comprehension performance for each model is shown in Figure 10. As with the English simulation, the model not only outperformed its baselines, but yielded a tighter, more uniform distribution of scores.

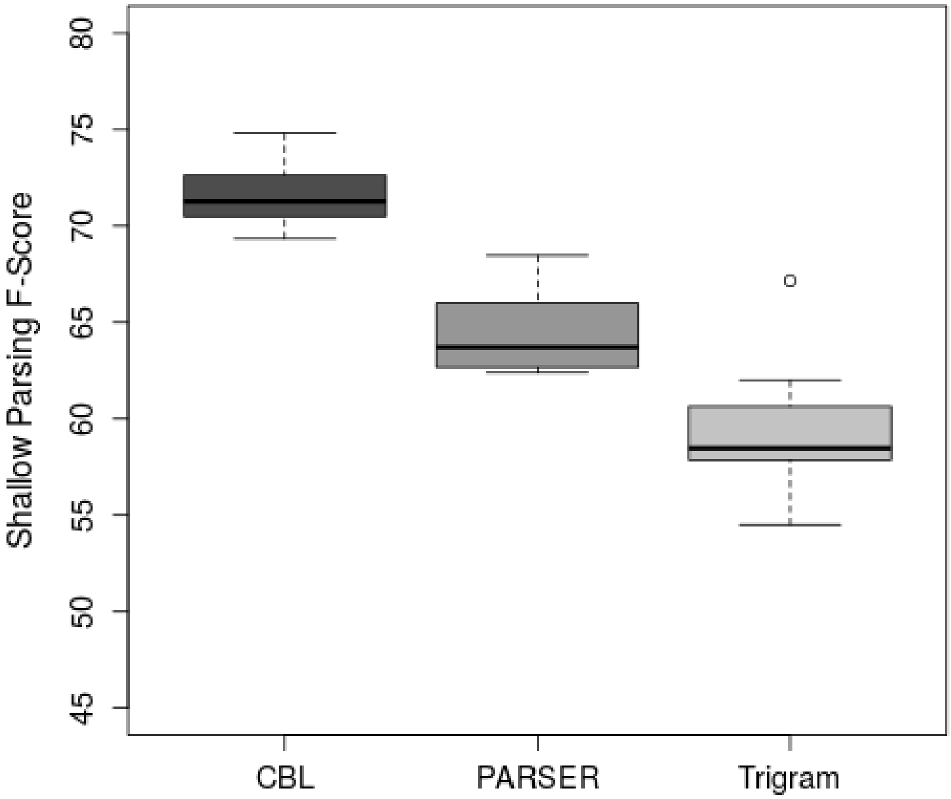

Fig. 10: Boxplots depicting shallow parsing F-score (%) for the CBL model and its baselines for the French simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

The F-scores for the model and its baselines were logit-transformed and submitted to a repeated-measures ANOVA including the factor *Model* (2: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,26) = 214.6, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(14)=14.33, p<0.0001] and Trigram [t(14)=15.72, p<0.0001] models, as well as for PARSER compared to the Trigram model [t(14)=11.99, p=0.0001].

As with the English simulations, we followed up this analysis by examining accuracy separately. Across the 15 child corpora, CBL attained a mean accuracy rate of 72.0%, while the PARSER model attained a mean accuracy rate of 61.8%. The Trigram model attained a mean accuracy rate of 57.0%. For completeness, CBL attained a mean score of 70.8%, while the PARSER model attained a mean completeness rate of 73.5%. The Trigram model attained a mean accuracy rate of 66.1%. Detailed analysis of accuracy and completeness scores are provided in Appendix A.

Therefore, similar to our English simulations, the best combination of accuracy and completeness (as measured by the F-score), as well as the best accuracy specifically, was achieved by CBLfor the French child corpora.

German: Comprehension performance. Across all 22 single-child corpora, CBL attained a mean F-score of 75.7, while the PARSER model attained a mean F-score of 73.4. The Trigram model reached a mean F-score of 67.4. Though CBL once more attained the highest scores, its performance advantage over the PARSER model was markedly smaller than in the English and French simulations. The distributions of scores for the model and its baselines are shown in Figure 11.

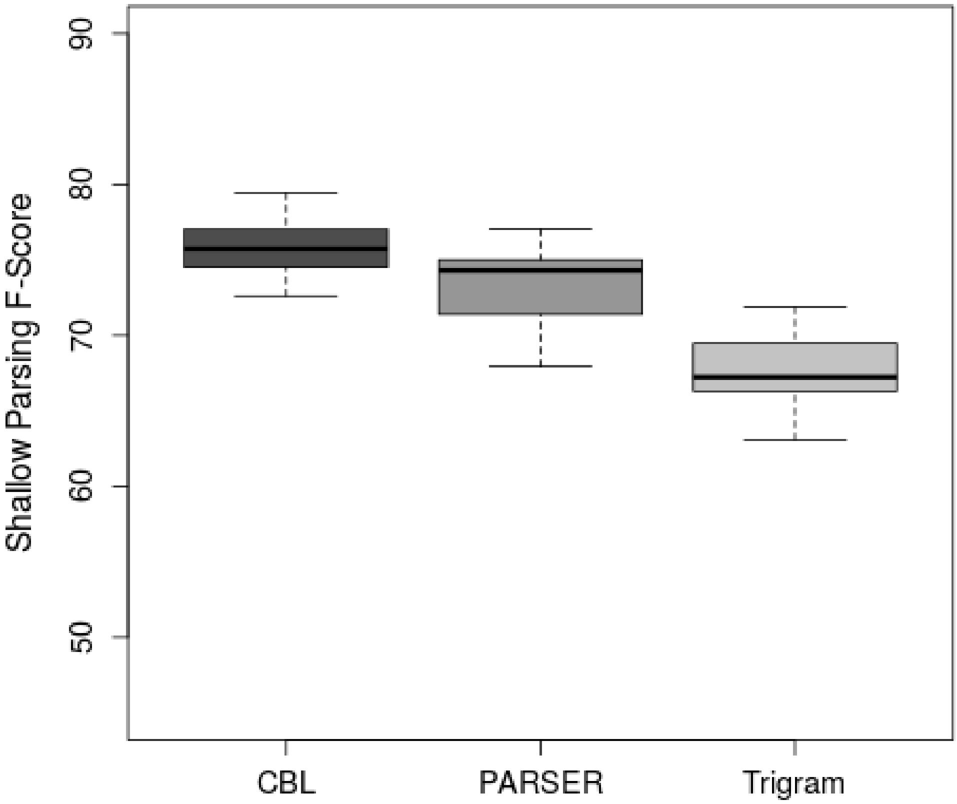

Fig. 11: Boxplots depicting shallow parsing F-score (%) for the CBL model and its baselines for the German simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

The F-scores for the model and its baselines were once more logit-transformed and submitted to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,40) = 69.43, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(21)=2.78, p<0.05] and Trigram [t(21)=17.67, p<0.001] models, as well as for PARSER compared to the Trigram model [t(21)=6.8, p=0.0001].

As with the English and French simulations, we followed up this analysis by examining accuracy separately. Across the 22 child corpora, CBL attained a mean accuracy rate of 78.0%, while PARSER attained a mean accuracy rate of 69.4%. The Trigram model attained an accuracy of 70.5%. For completeness, CBL attained a mean score of 72.2%, while PARSER attained a mean completeness of 83.5%. The Trigram model attained a completeness of 62.9%. Accuracy and completeness scores are analyzed in Appendix A.

Thus, as with our English and French simulations, the best accuracy, as well as the best combination of accuracy and completeness (as measured by the F-score), was achieved by CBL across the German child corpora. PARSER tended to segment more heavily than with English and French, leading to a drop in accuracy, relative to baselines, and a boost in completeness.

Comparing CBL's French and German shallow parsing performance to English performance: The CBL model's performance was highly similar across English, French, and German. Beyond outperforming baseline models on all three languages (both in terms of Accuracy and in terms of the overall F-score), the model yielded mean scores which were highly similar across languages: mean F-scores of 75.4 (English), 71.6 (French), and 75.7 (German) were achieved, alongside mean Accuracy rates of 76.4 (English), 72 (French), and 78 (German) and mean completeness scores of 73.8 (English), 70.8 (French), and 72.2 (German).

Thus, the model's ability to group related words together in an incremental, on-line fashion was remarkably stable across the three languages, despite important differences along a number of dimensions such as morphological complexity (French and German are morphologically richer than English) and word order (while English and French have an SVO word order, German has no dominant word order). This result offers important cross-linguistic support not only for the importance of multiword sequences, but for memory-based (as opposed to purely predictive) on-line learning as well as the plausibility of shallow linguistic processing based on local information.

Class-based simulations. We sought to test our item-based approach cross-linguistically by once more creating form class corpora from the single-child corpora used in the French and German simulations (using the same corpus preparation procedure described for English). We then tested the comprehension performance of the model and its baselines when learning from class-based statistics.

Figures 12 and 13 compare class- and item-based F-scores for comprehension across the French and German corpora. For CBL, performance was considerably worse when working with class-based statistics, with a sharp decrease in the mean F-score for both French (from 71.6 to 53.5) and German (from 75.7 to 39.1). For PARSER there was a similar decrease in performance for German (from 73.4 to 67.5), but an increase in score for French, from a mean F-score of 64.4 to 68.6. In the case of CBL, the lower comprehension performance when working with class statistics was driven by a drastic drop in completeness scores (French: from 70.8 to 29.9; German: from 72.2 to 21.4), owing partly to the use of the chunkatory in phrase segmentation;

the relatively small number of possible class combinations in a sequence lead to the automatic grouping of items together (based on chunkatory searches) with increasing frequency throughout the models' pass through a corpus. As more combinations exceeded the average TP threshold, the models placed progressively fewer phrase boundaries. For CBL there were less drastic changes in accuracy (French: from 72 to 75.4; German: from 78 to 68.9).

PARSER actually increased French F-scores under class-based information, owing to an increase in accuracy (from 61.8 to 71.9), though with a drop in completeness (from 73.5 to 60.5), while German scores decreased due to a drop in completeness (from 83.5 to 56.9), while accuracy scores increased (from 69.4 to 74.8).

The Trigram baselines showed less drastic changes in performance: the mean F-score rose for French (from 59 to 66.2) while decreasing slightly for German (67.4 to 66). Though accuracy scores increased for sharply French (from 57 to 73.9), completeness dropped (from 66.1 to 51.5). German accuracy increased (70.5 to 76.9) while completeness also dropped (from 62.9 to 53.1).

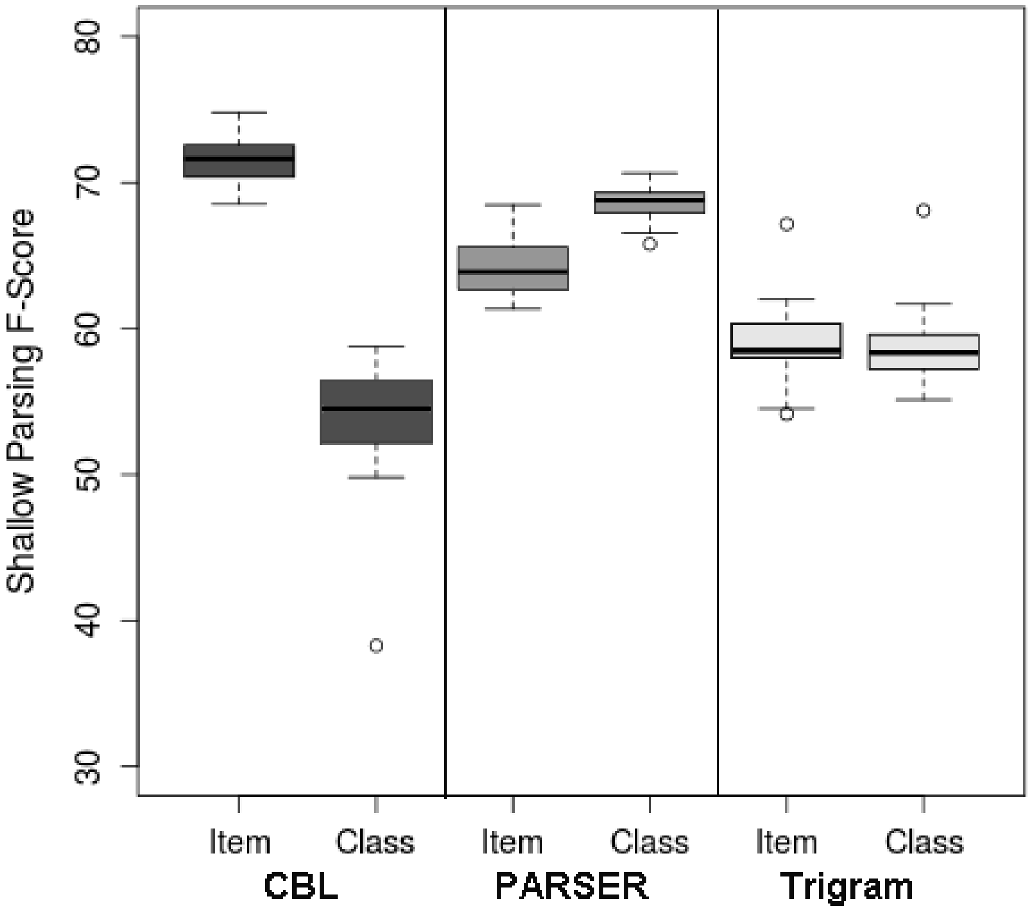

Fig. 12: Boxplots depicting class- vs. item-based comprehension performance (F-scores) across French simulations for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

We evaluated the effects of learning from French class-based information using a two-way ANOVA with factors *Statistic Type* (2: item- vs. class-based) and *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor, over logit-transformed F-scores. This yielded main effects of *Statistic Type* [F(1,14) = 6.09, p<0.05], confirming stronger performance for item-based models, and of *Model* [F(2,28)=19.91, p<0.0001], with an interaction between *Statistic Type* and *Model* [F(2,28)=348.4, p<0.0001], due to a more drastic drop in performance for the CBL model relative to the baselines when working with classes.

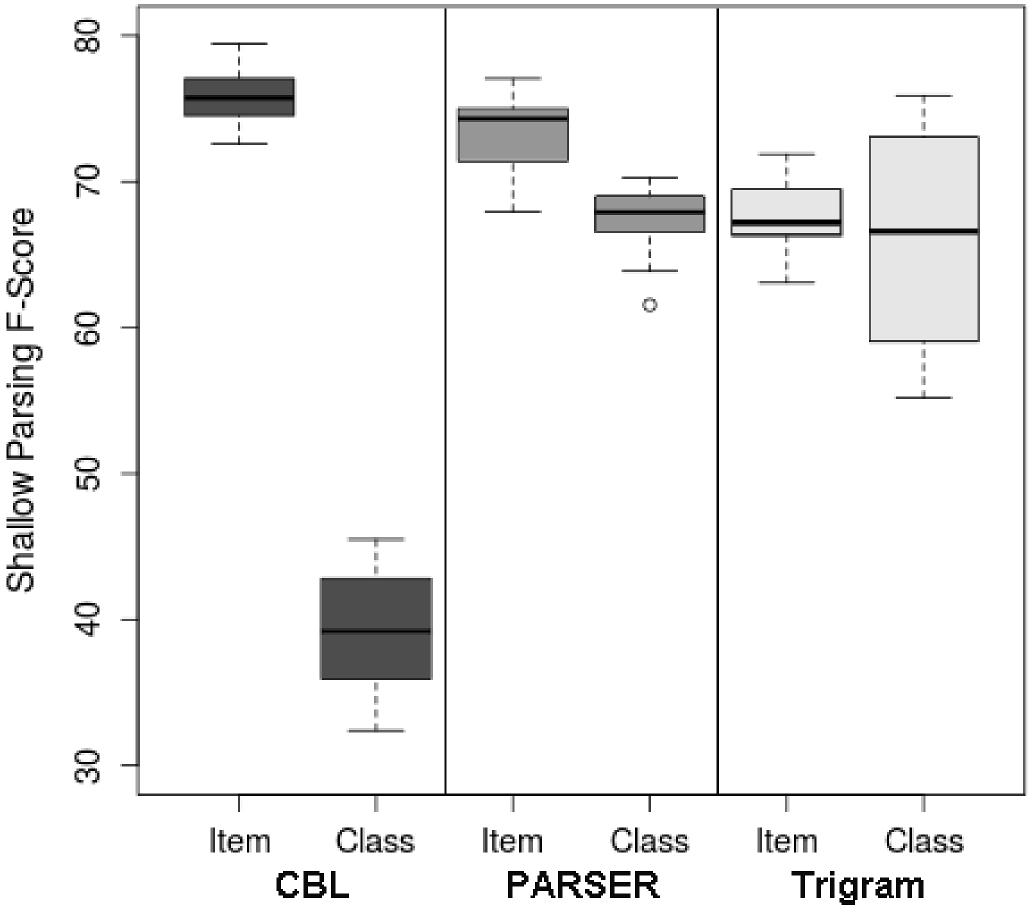

Fig. 13: Boxplots depicting class- vs. item-based comprehension performance (F-scores) across German simulations for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

We evaluated the effects of learning from German class-based information using a two-way ANOVA with factors *Statistic Type* (2: item- vs. class-based) and *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor, over logit-transformed F-scores. This yielded main effects of *Statistic Type* [F(1,21) = 243.4, p<0.0001], confirming stronger performance for item-based models, and of *Model* [F(2,42)=324, p<0.0001], with an interaction between *Statistic Type* and *Model* [F(2,42)=301.3, p<0.0001], due to a more drastic drop in performance for the CBL model relative to the baselines when working with classes.

As was the case with the English simulations, a reliance on word classes did not improve the performance of the model; instead, the use of classes lead to a decrease in performance. Though performance did increase for the baseline models when using French class-based information, CBL still yielded the strongest performance out of all simulations in its original item-based form. This result reaffirms the item-based approach, as well as the broader notion that initial language learning can take place without abstract syntactic categories. This also resonates with our previous simulation of child artificial grammar learning (McCauley & Christiansen, 2011), casting further doubt on to claims that children and adults can use transitional probabilities to segment phrases by calculating statistics over word classes (e.g., Saffran, 2002; Thompson & Newport, 2007) rather than concrete items.

Case study: Learning grammatical gender: Cross-linguistically, children master grammatical gender quite early in development (e.g., Slobin, 1986), and rarely make the sort of gender agreement errors often made by second language learners (e.g., Rogers, 1987; Holmes & de la Bâtie, 1999). Such findings resonate with the proposal that children treat article-noun pairs as single units (e.g., MacWhinney, 1978; Carroll, 1989), an idea which receives support from item-based patterns observed in children's use of articles (e.g., Mariscal, 2008; Pine & Lieven, 1997). More recently, Arnon and Ramscar (2012) used an artificial language learning paradigm to test the idea that learning article-noun pairings as chunks imparts an advantage in the learning of grammatical gender. They found that subjects receiving initial exposure to unsegmented article-noun sequences, which was only later followed by exposure to the noun labels in isolation, exhibited better mastery of grammatical gender in the artificial language at test than did those subjects who underwent the very same exposure phases in reverse order.

The findings of Arnon & Ramscar (2012), as well as children's item-based patterns in article usage and the apparent ease with which they master grammatical gender more generally, lead us to examine the model's ability to construct the right article-noun pairings during production. While CBL does not possess chunks arrived at via under-segmentation (the model initially recognizes articles and nouns as separate entities by virtue of the fact that the input corpora are in the form of words), the model may nevertheless learn to chunk articles and nouns together, leading to an item-based mastery of grammatical gender. To explore the model's learning of grammatical gender and determine whether its production of article-noun pairs exhibits the early mastery demonstrated by children, we analyzed the model's productions from the simulation involving the largest German corpus from CHILDES (Leo). We found that the model's article-noun pairings were correct over 95% of the time. Out of those article-noun pairs produced by the target child in the corpus, the model correctly captured 11,703, pairing the wrong article with a noun in only 557 cases. When we considered only those 513 utterances which featured multiple articles (as in the sentence ***die*** *Katzejagte* ***den*** *Hund*), rather than two or more instances of the same article being paired with different nouns (as in ***die*** *Katzejagte* ***die*** *Maus*), we found that the model paired nouns with the wrong gender marker in only 14 cases (an error rate of 2.7%). Thus, consistent with the findings of Arnon & Ramscar (2012), the distributional learning of article-noun sequences as chunks leads the model to mirror both children's early mastery of grammatical gender as well as the item-based nature of children's early article usage.

Simulation 4: Modeling Child Production Performance across a Typologically Diverse Set of 29 Languages

We conducted full simulations for each corpus from the additional languages, using the same model architecture and baseline models as used in each of the previous simulations. In the overall analysis of Sentence Production Performance, we include the scores for the English corpora for a total of 204 individual child simulations. CBL achieved a mean sentence production accuracy of 55.3%, while the Trigram model achieved a mean sentence production accuracy of 46%. The results for each language are depicted in Figure 14.

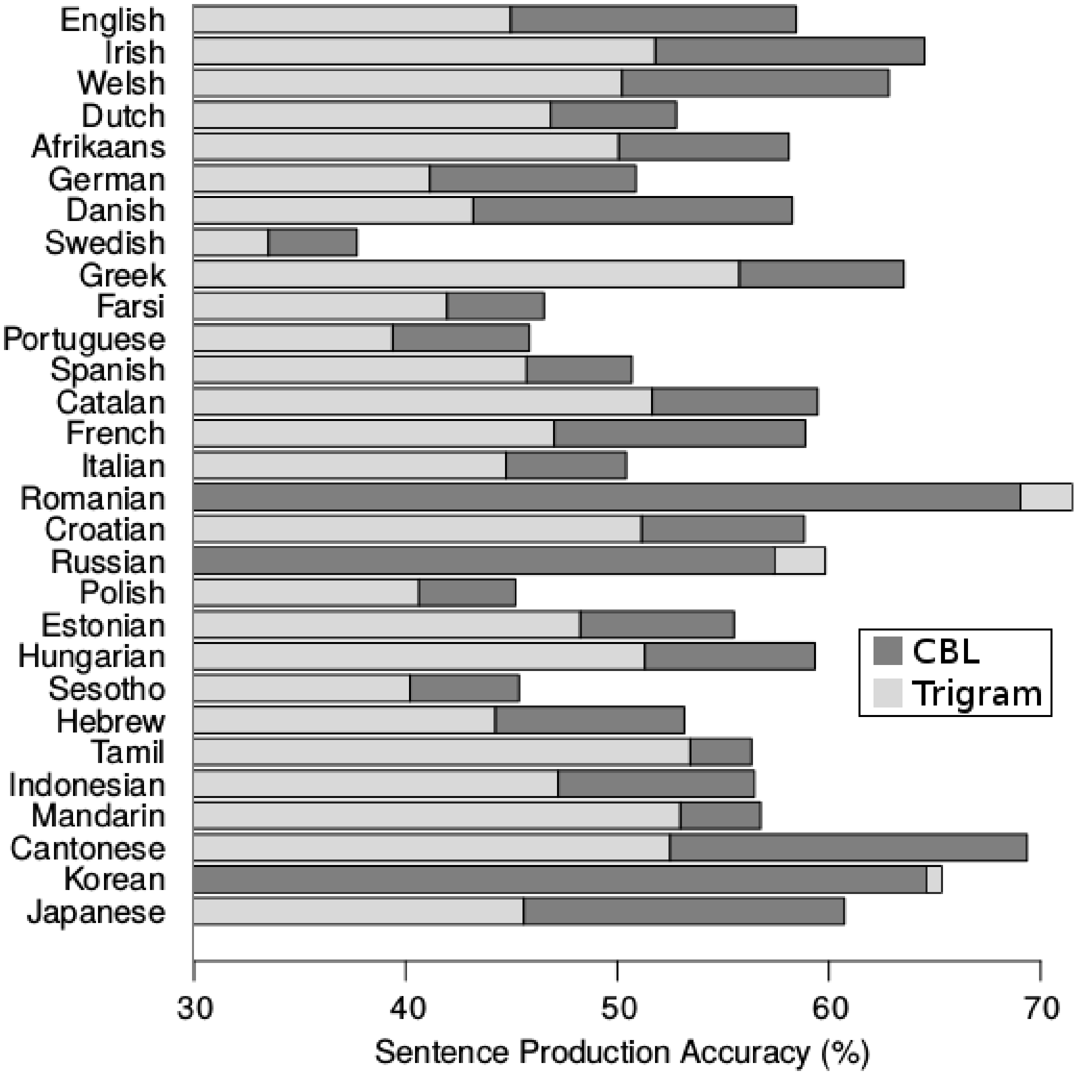

Fig. 14: Mean Sentence Production Accuracy scores for the CBL model and its trigram baseline across all 29 languages, including English (shown at top). Bars are non-cumulative (e.g., the Japanese CBL score was just over 60%, while the Trigram score was near 45%).

As with the English-only production simulations, we submitted the production scores to a repeated-measures ANOVA including, once more, the factor *Model* (2: CBL vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(1,203) = 575.2, $p < 0.0001$], indicating better performance for CBL.

When all utterances across the simulations were considered together, CBL was able to produce the majority of those utterances. The same pattern held for all but five of the individual languages, with CBL failing reach the 50% mark for Sesotho, Polish, Farsi, Portuguese, and Swedish.

As can be seen, CBL outperformed its baseline for 26 of the 29 languages; the exceptions were Russian, Romanian, and Korean for which the Trigram scored highest. It should be noted that for two of the exceptions (Romanian and Korean), there was only one child corpus; for Russian, there were only two available corpora. Moreover, all three of these languages fall towards the extreme synthetic end of the analytic/synthetic spectrum estimated by our morphological analyses. Below, we explore the notion that the CBL model performed worse as a function of morphological complexity.

Effects of morphological complexity and word order: To assess the effect of morphological complexity on the model's performance, we fit a linear regression model to the Sentence Production Accuracy scores across the 204 simulations using the morphological complexity measure calculated for each language previously. This yielded a significant negative value for morphological complexity [β=-0.17, t(202)=-4.35, $p<0.0001$], indicating that the model's sentence production tended to be less accurate when learning morphologically rich languages, although the amount of variance explained by the linear model was moderate to low (Adjusted R-squared:

0.08). Figure15 depicts the Sentence Production Accuracy for each simulation according to the morphological complexity score of the 29 languages.

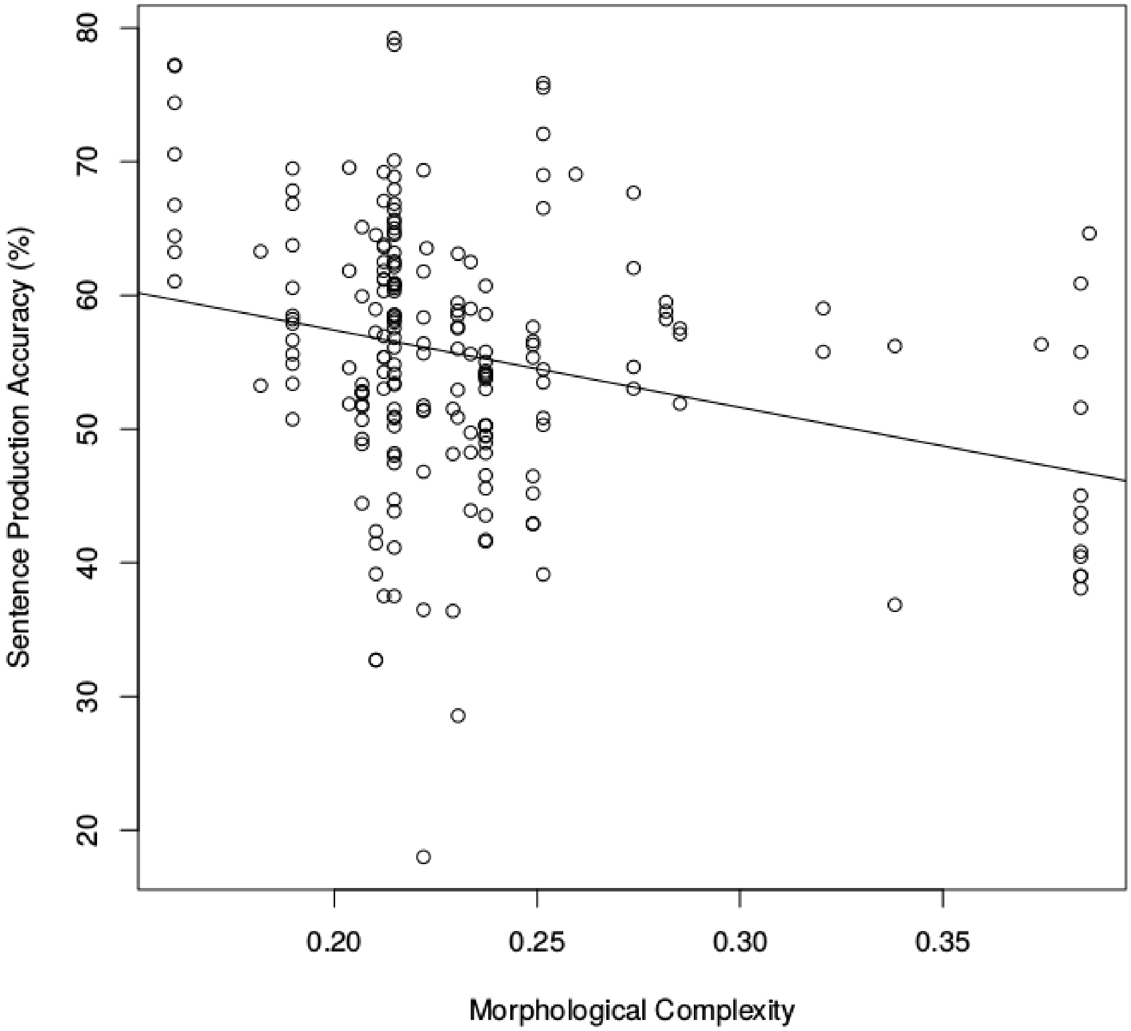

Fig. 15: Scatterplot depicting Sentence Production Accuracy (%) for each simulation by the Morphological Complexity score for the corresponding language.

Because the three languages on which the word-based Trigram model outperformed CBL were on the extreme synthetic end of the analytic/synthetic spectrum, we tested whether the model's sentence production accuracy advantage over the Trigram baseline decreased as a function of morphological richness. A linear model with *Morphological Complexity* as a predictor of the difference between CBL and Trigram scores for each simulation confirmed this to be the case [β=-0.4, t(201)=-5.6, p<0.0001], with the overall model explaining a significant amount of variance in the difference scores (Adjusted R-squared: 0.13).

That the advantage of a chunk-based model such as CBL would decrease as a function of morphological complexity is perhaps unsurprising. However, segmenting corpora into their component morphemes may better accommodate a chunk-based model while helping to deal with data sparseness tied to high type/token ratios. Since an automated method for morpheme segmentation was not available for many of the languages, we were unable to test this intuition cross-linguistically. However, the CHILDES segmentation system for one of the more synthetic languages, Japanese, treats verb morphemes as separate words (Chang et al., 2008; Miyata, 2000; Miyata & Naka, 1998). Interestingly, model performance, on average, is far stronger for Japanese corpora than for languages of comparable morphological complexity, especially relative to the Trigram baseline. Additionally, the greatest difference in scores between the model and its baselines was seen for Japanese. Future work will focus on comparing model performance on synthetic languages with morphologically segmented vs. standard corpora.

Despite the effect of morphological complexity on model production performance, there was no effect of word order (mean scores: SOV, 58.7; SVO, 55.5; VSO, 63; no dominant order: 52.9). We used a linear model to test for a potential effect of word order on model performance, while controlling for morphological richness. As the effect of *Morphological Complexity* on model performance was significant, we included it as a predictor alongside *Word Order* in the model. However, no significant effect of word order emerged. Figure 16 depicts the mean Sentence Production Accuracy score across each of the four word orders.

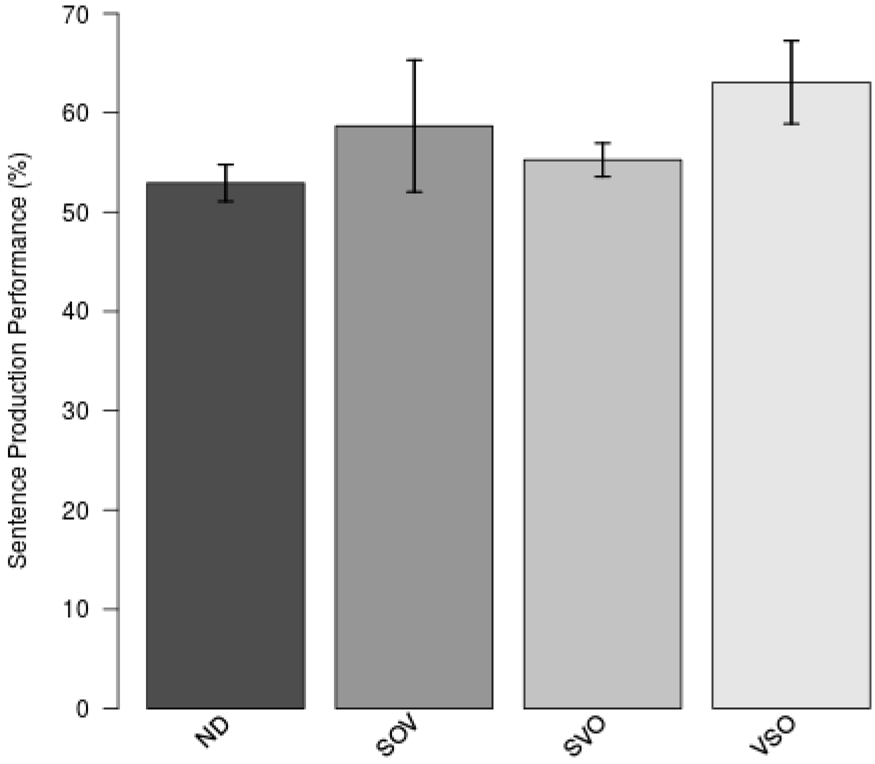

Fig. 16: Barplot depicting the mean Sentence Production Accuracy (%) for each of the four word orders represented across the 29 languages. Error bars depict standard errors. While there were only two corpora tested with VSO word order (precluding a statistical test), visual inspection of the error bars indicates that the model's performance was highly similar across the four word orders represented.

Production summary: The model outperformed its baseline for 26 of the 29 languages, correctly producing the majority of child utterances for 24 languages. The usefulness of BTP-based chunking across so wide an array of languages is somewhat surprising, given previous work demonstrating that the usefulness of forward vs. backward probability calculations in word segmentation is highly sensitive to cross-linguistic differences in distributional features (e.g., heavy suffixing in case of Hungarian, phrase-initial function words in the case of Italian; Gervain& Guevara Erra, 2012). While our model was somewhat sensitive to differences in morphological complexity, tending to perform slightly better on morphologically simple languages, it did not appear to be sensitive to differences in dominant word order. The fact that multi-word units were useful even for the learning of morphologically rich languages is of particular interest, considering the difficulties inherent in working with morphologically rich languages in the field of computational linguistics (cf. Tsarfaty, Seddah, Kübler, & Nivre, 2012). Taken together with previous findings of item-based patterns in children's learning of morphologically rich languages (e.g., MacWhinney, 1978), this result is quite encouraging in the context of future cross-linguistic item-based modeling work.

These results offer substantial cross-linguistic support for CBL, and, more broadly, for the view that simple learning mechanisms underlie a large part of early linguistic behavior. The outcome of our simulations strengthens previous psycholinguistic evidence for chunk-based learning, which has been gained primarily from English speakers (e.g., Bannard & Matthews, 2008), suggesting that multiword sequences play an important role cross-linguistically, in analytic and synthetic languages alike.

General Discussion

We have shown that the CBL model can approximate children's comprehension and production of language by learning in a purely incremental fashion through on-line processing. The model gradually builds up an inventory of chunks consisting of one or more words, which unites aspects of comprehension and production within a single framework. On the comprehension side, the model chunks incoming words together to incrementally build an item-based “shallow parse” of each utterance as it is encountered. Chunks discovered in this fashion are used to make predictions about upcoming words in subsequent input, facilitating the model's on-line processing. When the model encounters an utterance produced by the target child of the input corpus, it attempts to generate an identical utterance using the same chunks and statistics used in shallow parsing. Importantly, production is modeled as an incremental, chunk-to-chunk process rather than one of whole-sentence optimization (as would be the case by, e.g., choosing among candidate sentences based on whole-string probabilities).

The model achieves strong performance across English single-child corpora from the CHILDES database, approximating the performance of a shallow parser with high accuracy and completeness. In line with expectations derived from usage-based theories, item-based information is shown to be more useful than statistics or chunks computed over form classes. The model is also able to fit much of children's early linguistic behavior, correctly generating the majority of the child utterances encountered across the English corpora. The model exhibits similar shallow parsing performance for 15 French and 22 German corpora alongside similar sentence production performance for nearly 200 additional child corpora drawn from a typologically diverse set of 28 languages (also from CHILDES). In each case, the model outperforms baseline models.In addition to its strong cross-linguistic performance in approximating aspects of comprehension and production, the model provides close quantitative fits to children's production of complex sentences featuring six different relative clause types (Diessel & Tomasello, 2005).

Together, these findings suggest that a fair amount of children's early language use can be accounted for by incremental, on-line learning of item-based information using simple distributional cues. In what follows, we discuss the limitations of the model and directions for future modeling work addressing them. We then place the model in the larger context of usage-based computational approaches to acquisition. Finally, we highlight novel predictions made by the model which will be tested in future psycholinguistic work.

Limitations of the Model

As an initial step towards a comprehensive computational account of language learning, CBL is not without limitations. Perhaps most immediately obvious is that the model learns from segmented input. It does not confront one of the early challenges facing language learners: that of segmenting an uninterrupted speech stream. The problem of segmenting the speech stream (traditionally thought of as word segmentation) and discovering useful multiword sequences are likely to impact one another: Children do not learn about phonology, words, multiword sequences, and meanings in discrete, separable stages, but instead learn about multiple aspects of linguistic structure simultaneously and their interaction with each other (see also Christiansen &Chater, 2016b). Indeed, many of children's earliest, unanalyzed chunks are likely to stem from under-segmentation “errors,” which may offer insights into a number of phenomena tied to early language learning (Arnon, 2009). Future work will focus on using the model to learn from unsegmented corpora in ways that maintain a fluid rather than rigid relationship between individual words, unanalyzed chunks, and chunks which the model is capable of breaking down into its component words.

A further limitation stems from what may also arguably be one of the model's greatest strengths: reliance on a single source of distributional information. CBL was designed to be as simple as possible, in order to demonstrate that a model can approximate aspects of comprehension and production through incremental, on-line processing based on simple statistics. Though the model, which relies upon BTPs, is evaluated against a baseline which uses FTPs, it is clear from the modeling results that both information sources are potentially useful. Infants, children, and adults have been shown to be sensitive to TPs calculated in both directions (e.g., French et al., 2011; Pelucchi et al., 2009; Perruchet &Desaulty, 2008). Future work should be based on a principled, parameter-free method for seamlessly integrating TP calculations in both directions (in addition to other potentially useful distributional and non-distributional cues).

A limitation which is perhaps more difficult to address lies in the model's lack of semantic information; the model never learns “meanings” corresponding to the chunks it discovers, and is never called to interpret the meanings of utterances. We have adopted the perspective that semantic/conceptual information—such as that tied to event schemas, situational settings, and background world knowledge—is a key factor in generalizing to unbounded productivity of the sort exhibited by mature language users, superseding the importance of abstract “syntactic” knowledge, such as that of form classes. We are currently expanding the model to incorporate such information in idealized forms (for a review of the prospects and challenges of incorporating semantic information into computational models of language acquisition, see McCauley & Christiansen, 2014).

A further, related limitation of the model stems from its inability to produce utterances “from scratch:” The randomly ordered bag-of-words which the model attempts to sequence during production is always populated by words from one of the target child's actual utterances. This means that the model cannot be used to capture children's systematic errors (the model can only commit errors which are made by the target child). Previous models capable of producing novel utterances have successfully captured such developmental trends, such as optional infinitive errors (e.g., MOSAIC; Freudenthal et al., 2006, 2007). Future work will therefore seek to incorporate "free production," both on the basis of distributional input of the sort modeled here, as well as semantic input (similar to production in the model of Chang et al., 2006).

Finally, work with adult subjects suggests that there is no frequency "threshold" beyond which a multiword sequence is stored as a chunk, but rather that the extent to which sequences cohere as multiword units is graded in nature (cf. Caldwell-Harris et al., 2012). While CBL does not make use of raw whole-sequence frequency information in chunk discovery, it does rely on the use of a running average BTP as a threshold. Future work will emphasize the graded nature of "chunk" status for multiword units, while also seeking to make predictions about part/whole interactions (reflecting findings that stored multiword sequences both prime and are primed by their component words; e.g., Sprenger et al., 2006).

Relationship to other Usage-Based Modeling Approaches

Despite its current limitations, the CBL model may be viewed as a possible foundation for a comprehensive computational account of language acquisition. While previous computational models within the usage-based tradition have boasted great success, CBL possesses a number of features that have been largely absent from language modeling, several of which represent desiderata for a fully comprehensive approach to acquisition.

Firstly, and perhaps most importantly, CBL takes usage-based theory to its natural conclusion in making no distinction between language learning and language use; the model learns solely by attempting to comprehend and produce language. That is, the very processes by which input is interpreted and output is constructed are the same processes involved in learning; at no point does the model engage in a separate "grammar induction" process. This sets the present model apart from a number of extant usage-based models that have focused on grammar induction (e.g., Bannard et al., 2009) or conceived of learning and processing separately.

Also of great importance is that CBL learns incrementally, without batch learning of the sort used by most existing computational approaches (e.g., Bannard et al., 2009; Jones et al., 2004). While more sophisticated models of grammatical development have captured incremental learning (e.g., Bod, 2009; Kolodny et al., 2015), CBL is unique in offering an account of the on-line processes leading to linguistic knowledge over time; the model learns incrementally not only from utterance-to-utterance, but within individual utterances themselves as input is received, on a word-by-word basis. Thus, its design reflects the constraints imposed by the Now-or-Never bottleneck (Christiansen & Chater, 2016b).

While a number of previous usage-based models have captured the generation of novel utterances (e.g., Jones et al., 2004; Solan et al., 2005), none have simultaneously sought to approximate aspects of comprehension in an explicit fashion. A further contribution of CBL is that it not only captures aspects of both comprehension and production, but also in unites them within a single framework. Pickering and Garrod (2007; 2013) argue that comprehension and production should not be seen as separate processes, a view compatible with usage-based approaches more generally (cf. Chater et al., 2016; McCauley & Christiansen, 2013). While connectionist models have utilized the same network of nodes to simulate comprehension and production (e.g., Chang et al., 2006; see also MacKay, 1982), ours is the first full-scale (taking full corpora as input) model to offer a unified framework.

Finally, CBL was designed to reflect psychologically parsimonious processing and knowledge representation. Outside the realm of word segmentation, the model is unique in its reliance solely on simple recognition-based processing and simple statistics of a sort that infants, children, and adults have been shown to be sensitive to (BTPs; French et al., 2012; Pelucchi et al., 2009; Perruchet &Desaulty, 2008). While a number of more complex computational approaches have made use of transitional probabilities (e.g., Kolodny et al., 2015; Solan et al., 2005), CBL relies solely on transitional probabilities computed in an incremental, on-line fashion, and is not supplemented by more complex processes. Furthermore, the model relies on local information; rather than automatically storing entire utterances, the model shallow parses and produces utterances in an incremental, chunk-to-chunk fashion rather than relying on whole-sentence representation or optimization.

Predictions Derived from the Model

Beyond CBL's unique features, its ability to capture much of children's early linguistic behavior cross-linguistically, and its success in accounting for key psycholinguistic findings, the model makes several explicit predictions that may be tested through psycholinguistic research:

1) **Simple distributional cues are useful at every level of language learning.** The model was able to use a simple distributional cue previously shown to be useful in word segmentation (BTP; Pelucchi et al., 2009; Perruchet &Desaulty, 2008) in order to segment speech into useful multiword units, as well as to combine them to create sentences. Though based on this simple statistic, the model was able to make close contact with psycholinguistic results on children's production of complex sentence types (Diessel & Tomasello, 2005).

2) **Previous artificial grammar learning results may reflect item-based rather than class-based computations.** The decision to focus on learning through purely item-based statistics stands in contrast to several threads of argument within the statistical learning literature, which hold that learners discover phrase structure by computing statistics over form classes rather than individual words (e.g., Saffran, 2002; Thompson & Newport, 2007). As discussed above, we have found that the discovery of useful chunks of local information (which, being as our model was scored against a shallow parser, is analogous to phrase segmentation of the sort discussed by Thompson & Newport, 2007) was actually enhanced as a consequence of a reliance on item-based statistics, as was also often the case with the model's baselines. The model performed worse when exposed to class-based information – a pattern which was replicated in our simulation of Saffran's (2002) child artificial

language learning experiment (McCauley and Christiansen, 2011).

We suggest, then, that artificial language learning results which have previously been taken to reflect learners' sensitivity to the phrase-like structure of the stimuli be reassessed to determine whether item-based calculations might be sufficient to capture the learning of both children and adults. Previous modeling work on chunking has been shown to better account for segmentation performance in artificial language studies than more complex learning mechanisms (e.g., Perruchet et al., 2002; French et al., 2011). This general approach may be extended to full-blown language development.

3) **Most of the difficulty faced by the learner lies outside the distributional realm.** The difficulty of learning from distributional information may be compounded by the problem combining multiple probabilistic cues (which CBL, relying on a single distributional cue, does not attempt to capture). However, given the rapidity with which the model was able to learn how to identify useful chunks of local information, as well as to sequence those chunks to create new utterances, we suggest that the greatest difficulties children face in learning to process sentences may have less to do with distributional information or even “linguistic structure,” but instead derive from conceptual/semantic dimensions of the problem, such as learning event schemas and scenarios to map chunks onto.

4) **Multiword sequences remain important throughout development.** The model ultimately relied more heavily on multiword units over time, as shown in various analyses of the model's chunk inventory. This leads us to suggest that instead

of “starting big” by merely relying upon multiword units during the early stages of learning (e.g., Arnon, 2009), learners continue to rely on chunks throughout development; representations may actually become *more* chunk-like instead of less. At the same time, subtle shifts in the “degree of chunkedness” of children's representations, as suggested by the U-shaped development of the model's chunk inventory, may interact with U-shaped trajectories observed in seemingly disconnected areas, such as the learning of irregular forms.

5) Learners rely on multiword units even in morphologically rich languages. The model benefited from the use of chunks in learning analytic and synthetic languages alike. At the same time, chunk-to-chunk sequential information of the type learned by the model clearly matters less in synthetic languages, where there may be stronger pragmatic constraints on ordering. CBL lacks such information and this may partly explain the lower performance of the model when learning morphologically rich languages.

Conclusion

We have presented the foundations of a new approach to modeling language learning in the form of the CBL model, which provides a computational framework based on incremental, on-line learning from simple chunks and statistics. The model makes close contact with psycholinguistic evidence for both multiword unit storage and shallow, underspecified language processing; rather than attempting to induce a target grammar, the model learns chunks of local information which are used to simulate aspects of comprehension and production. CBL approximates the performance of a shallow parser by segmenting utterances into chunks of related words on-line, and simultaneously uses the same chunks to incrementally produce new utterances. The model's production abilities can account for a considerable part of children's early linguistic behavior. The model offers broad, cross-linguistic coverage and successfully accounts for key developmental psycholinguistic findings, in addition to making several predictions on which to base subsequent psycholinguistic work. But, perhaps most importantly, CBL demonstrates how language learning can modeled as language use.

Appendix A: Shallow parsing accuracy and completeness statistics for English, French, and German

English

In line with the developmental motivation for the model, we examined accuracy rates independently. Across the 43 child corpora, CBL achieved a mean accuracy rate of 76.4%, while PARSER attained a mean accuracy of 65.2% and the Trigram model reached a mean accuracy rate of 65.8%. The distributions are shown in box plots in Figure A1. As can be seen, CBL not only outperformed its baselines, but once more yielded a tighter, more uniform distribution of scores.

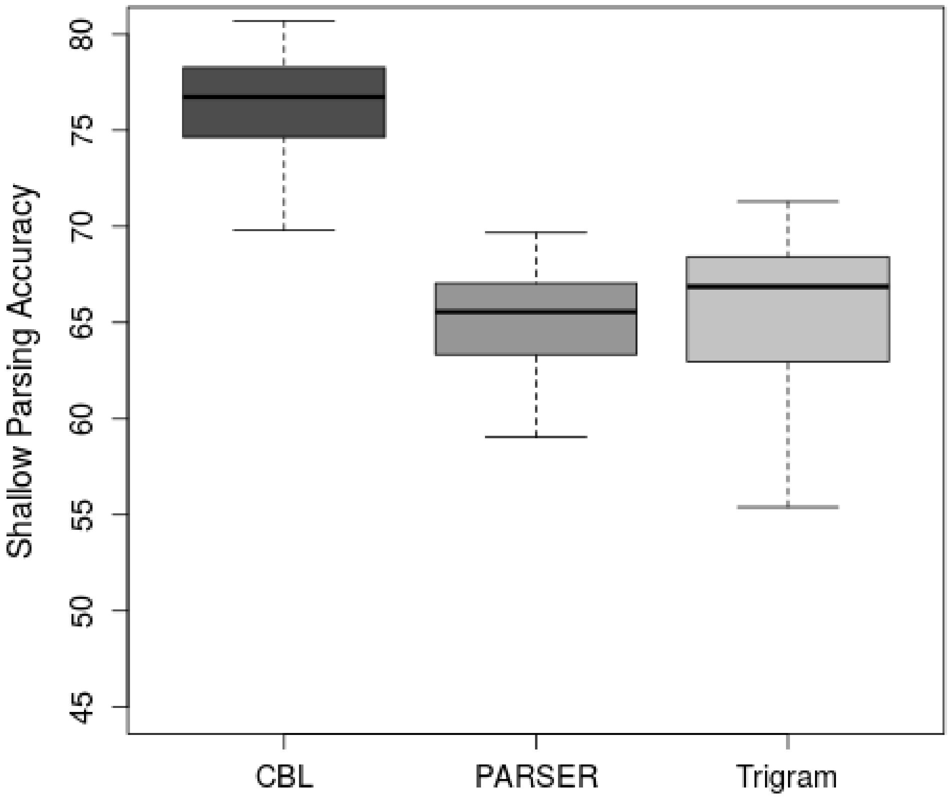

Fig. 17: Boxplots depicting shallow parsing accuracy (%) for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

As in our analysis of the overall comprehension performance scores, we submitted the logit-transformed accuracy scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant main effect of *Model* [F(2,84) = 1877, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(42)=65.17, p<0.0001] and Trigram [t(42)=39, p<0.0001] models, as well as stronger performance for the Trigram model compared to PARSER [t(42)=2.63, p<0.05].

Finally, we looked at completeness scores. Across the 43 child corpora, CBL achieved a mean completeness of 73.8%, while the PARSER attained a mean completeness of 68.7% and the Trigram model reached a mean completeness rate of 66.5%. The distributions are shown in box plots in Figure A2.

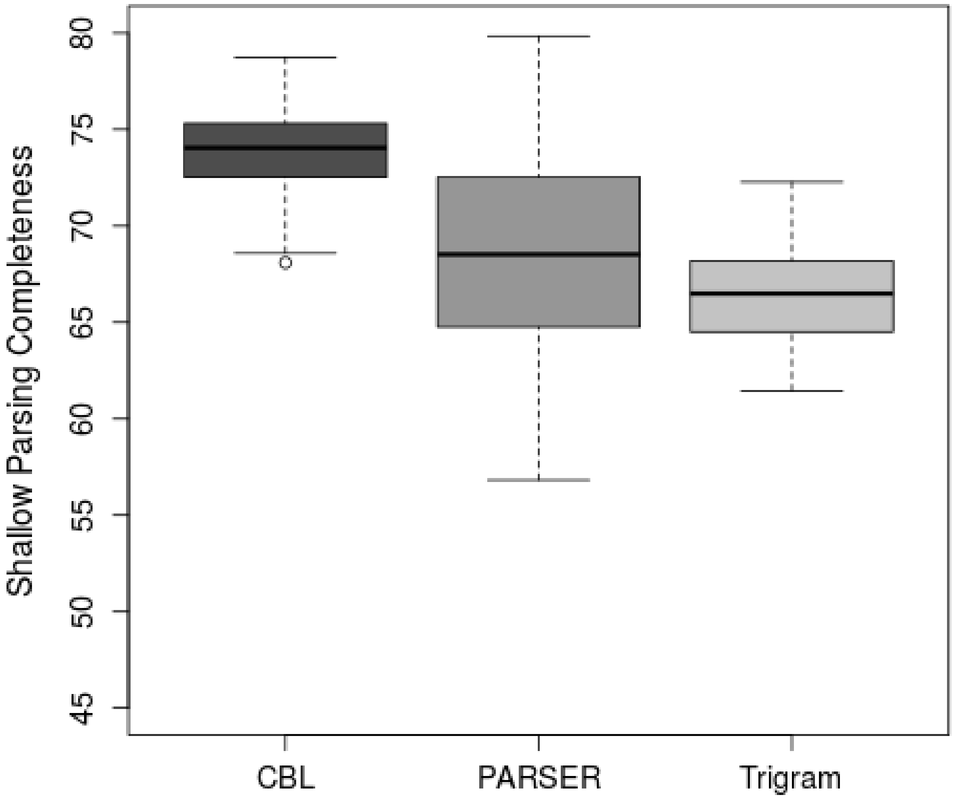

Fig. 18: Boxplots depicting shallow parsing completeness (%) for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

As with accuracy, we submitted the logit-transformed completeness scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram) with *Child Corpus* as a random factor. This yielded a significant main effect of *Model* [F(2,84) = 42.14, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(42)=7.77, p<0.001] and Trigram [t(42)=11.9, p<0.0001] models, with no significant difference in means for PARSER relative to the Trigram model [t(42)=1.94, p=0.06].

French

As with the English simulations, we examined accuracy separately. Across the 15 child corpora, CBL attained a mean accuracy rate of 72.0%, while the PARSER model attained a mean accuracy rate of 61.8%. The Trigram model attained a mean accuracy rate of 57.0%. The distributions are shown in box plots in Figure A3.

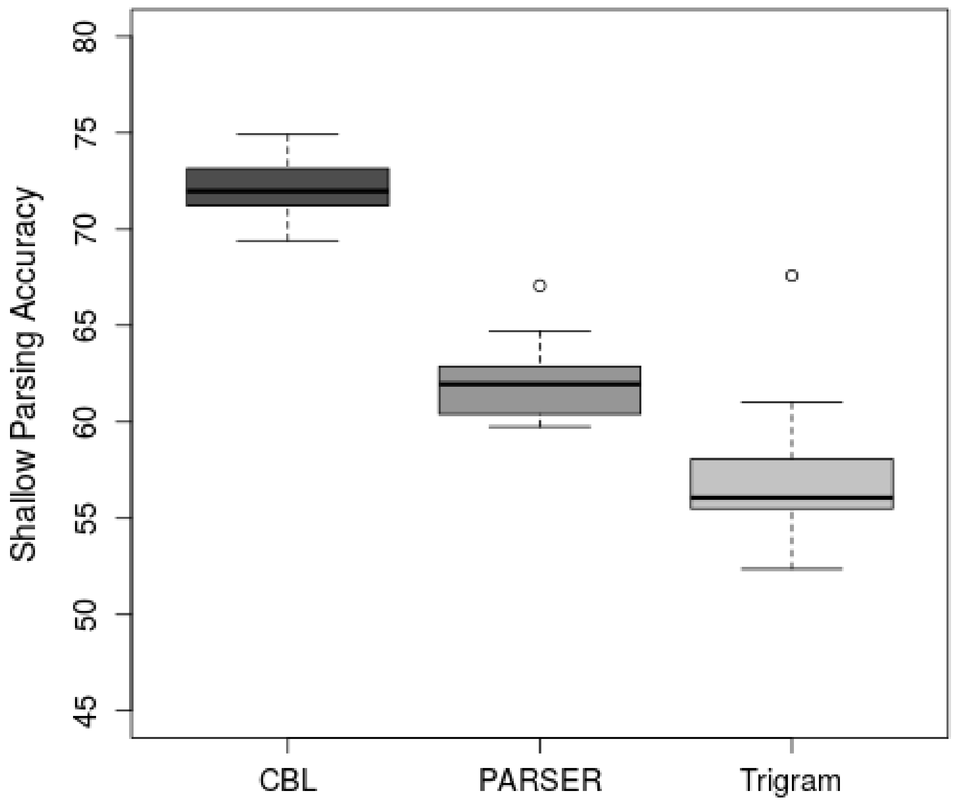

Fig. 19: Boxplots depicting shallow parsing accuracy (%) for the CBL model and its baselines for the French simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

As with the previous analyses, we submitted the logit-transformed accuracy scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,26) = 342.3, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(14)=24.54, p<0.0001] and Trigram [t(14)=18.69, p<0.0001] models, as well as for PARSER compared to the Trigram model [t(14)=9.7, p=0.0001].

We also analyzed completeness: across the 15 child corpora, CBL attained a mean completeness score of 70.8%, while the PARSER model attained a mean completeness rate of 73.5%. The Trigram model attained a mean accuracy rate of 66.1%. The distributions are shown in box plots in Figure A4. As with accuracy, we submitted the logit-transformed completeness scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,26) = 21.96, p < 0.0001], with post-hoc analyses confirming stronger performance for PARSER compared to the CBL [t(14)=2.54, p<0.05] and Trigram [t(14)=5.29, p<0.001] models, as well as for CBL compared to the Trigram model [t(14)=6.35, p=0.0001].

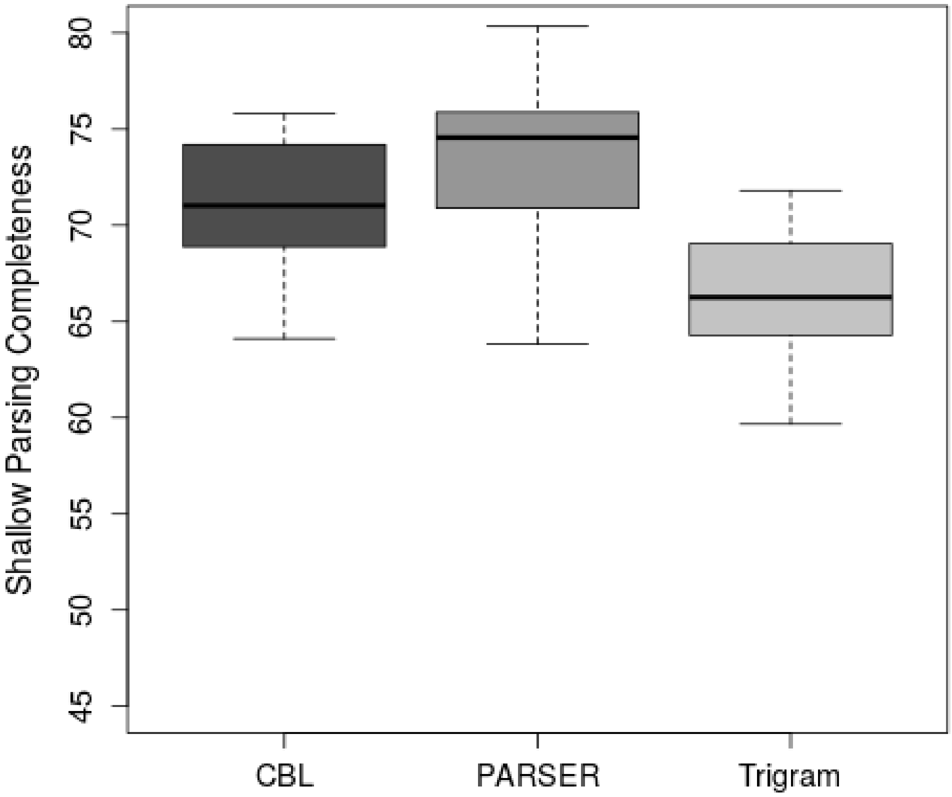

Fig. 20: Boxplots depicting shallow parsing completeness (%) for the CBL model and its baselines for the French simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

German

As with the English and French simulations, we examined German accuracy separately. Across the 22 child corpora, CBL attained a mean accuracy rate of 78.0%, while PARSER attained a mean accuracy rate of 69.4%. The Trigram model attained an accuracy of 70.5%. The distributions are shown in box plots in Figure A5.

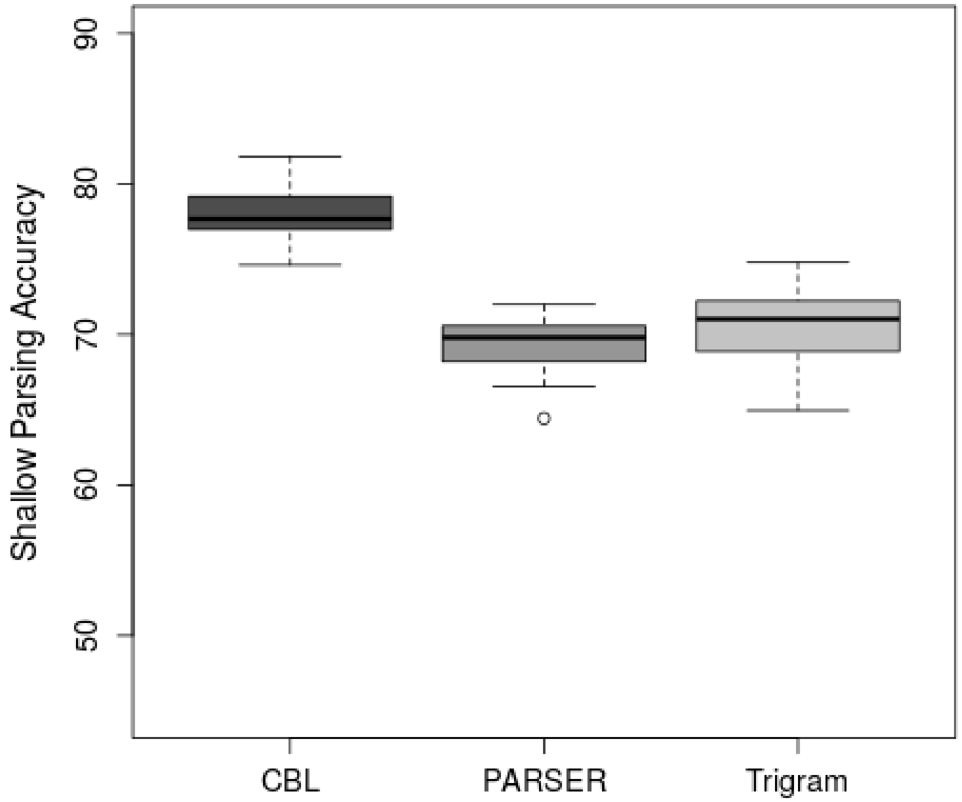

Fig. 21: Boxplots depicting shallow parsing accuracy (%) for the CBL model and its baselines for the German simulations. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores falling within 1.5 IQR of the quartiles, while dots depict outliers.

We once more submitted the logit-transformed accuracy scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,40) = 475.2, p < 0.0001], with post-hoc analyses confirming stronger performance for CBL compared to the PARSER [t(21)=29.4, p<0.0001] and Trigram [t(21)=20.05, p<0.0001] models, as well as for the Trigram model compared to PARSER [t(21)=4.23, p=0.001].

As with the English and French simulations, we also examined completeness separately. Across the 22 child corpora, CBL attained a mean completeness of 72.2%, while PARSER attained a mean completeness of 83.5%. The Trigram model attained a completeness of 62.9%. The distributions are shown in box plots in Figure A6.

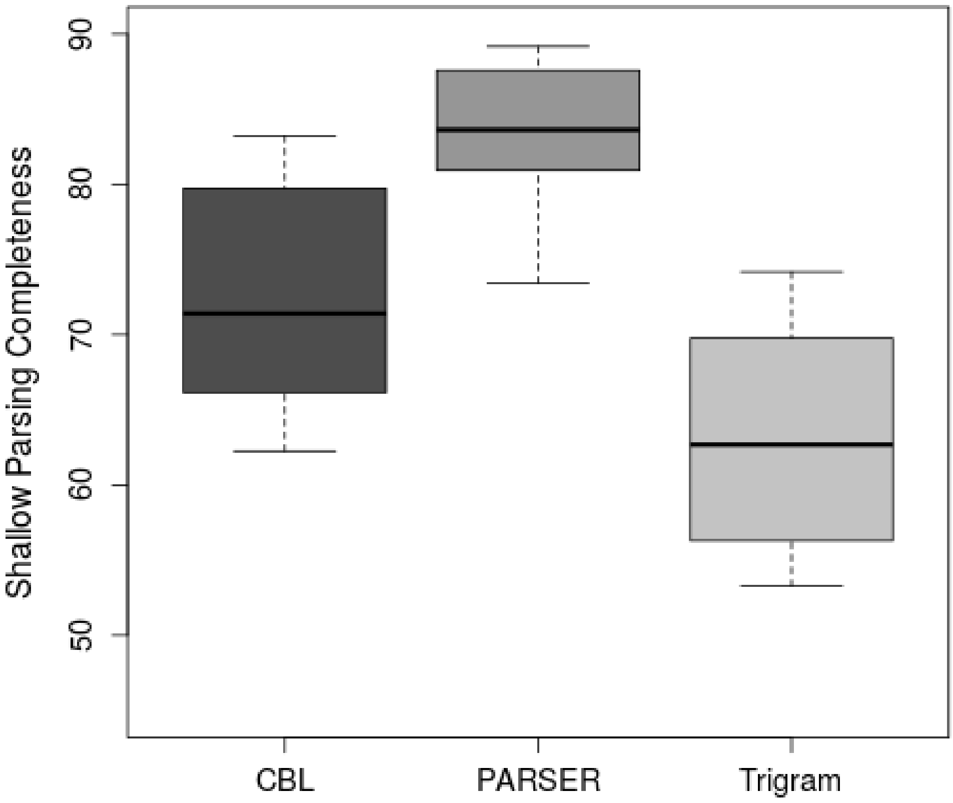

Fig. 22: Boxplots depicting shallow parsing completeness (%) for the CBL model and its baselines. Boxes depict the median (thick line), with upper and lower edges representing the respective quartiles. Whiskers depict the range of scores

falling within 1.5 IQR of the quartiles, while dots depict outliers for the German simulations.

We once more submitted the logit-transformed completeness scores to a repeated-measures ANOVA, including the factor *Model* (3: CBL vs. PARSER vs. Trigram), with *Child Corpus* as a random factor. This yielded a significant effect of *Model* [F(2,40) = 61.7, p < 0.0001], with post-hoc analyses confirming stronger performance for PARSER compared to the CBL [t(21)=5.49, p<0.0001] and Trigram [t(21)=8.59, p<0.0001] models, as well as for CBL compared to the Trigram model [t(14)=11.3, p=0.0001].

Appendix B: Evaluating the Effects of Forwards vs. Backwards Transitional Probability

Baseline Models

We created three baseline models in order to explore a 2 x 2 design, depicted in Table B1, including the factors *unit type* (chunks vs. *n*-grams) and *direction* (backward vs. forward transitional probability).

Table 4
Contrasting Direction and Unit Type

	Chunks	**N-grams**
BTP	CBL	BTP3G
FTP	FTP-Chunk	FTP3G

As previous work in the statistical learning literature has focused on FTP as a cue to phrase structure (e.g., Thompson & Newport, 2007), an alternate model was created to compare the usefulness of this cue against the BTPs used by CBL. Thus, the first baseline model, hereafter referred to as the FTP-Chunk model, was identical to CBL, with the sole exception that all BTP calculations were replaced by FTP calculations.

As the Trigram model described in the main paper relied on FTPs, we created an otherwise identical baseline model which relied on BTP rather than FTP calculations. Both models learned trigram statistics in an incremental, on-line fashion, in the style of CBL, while simultaneously processing utterances through the placement of chunk boundaries. In the present appendix we refer to the Trigram baseline as the FTP3G baseline, and the backwards transitional probability version as the BTP3G baseline.

In the case of the FTP3G baseline, if the FTP between the first bigram and the final unigram of a trigram fell below the running average for the same statistic, a chunk boundary was inserted. For instance, as the model encountered Z after seeing the bigram XY, it would calculate the FTP for the trigram by normalizing the frequency count of the trigram XYZ by the count of the bigram XY, and comparing the result to the running average FTP for previously encountered trigrams (inserting a chunk boundary if the running average was greater). In the case of the BTP3G baseline, a chunk boundary was placed if the BTP between the first unigram and the final bigram of the trigram fell below the running average. The start-of-utterance marker made it possible for the 3G baselines to place a boundary between the first and second words of an utterance. During production attempts, which were also incremental and on-line in the style of CBL, both trigram models began constructing an utterance by choosing from the bag-of-words the word with the highest TP (FTP for the FTP3G model, and BTP for the BTP3G model), given the start-of-utterance marker (in other words, bigram statistics were used to select the first word). Each subsequent word was chosen according to trigram statistics, based on the two most recently placed words (or the initial word and the start-of-utterance marker, in the case of selecting the second word in an utterance). For the FTP3G model, this meant the word with the highest FTP given the two preceding words was chosen; for the BTP3G model, the word resulting in the highest BTP between the final bigram and the first unigram of the resulting trigram was chosen. Thus, like CBL and its FTP-based counterpart, both trigram baseline models relied on identical statistics during comprehension and production (either BTPs or FTPs, computed over trigrams).

Shallow Parsing Results

Shallow parsing results for the same English child corpora are shown for the model and its baselines in Table B2.

Table 5: English Shallow Parsing F-Scores

	Chunks	**N-grams**
BTP	75.4	61.3
FTP	67.5	65.9

We submitted the shallow parsing F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams)and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [F(1,42) = 1184, p < 0.0001] and *Direction* [F(1,42) = 78.01, p < 0.0001], indicating better performance for chunk-based models and BTPs, respectively, and a significant *Unit Type* x *Direction* interaction [F(1,42) = 792.5, p < 0.0001], indicating better performance for the CBL model’s combination of BTPs and chunks.

The French shallow parsing scores, depicted in Table B3, followed the same qualitative pattern as the English data.

Table 6: French Shallow Parsing F-Scores

	Chunks	**N-grams**
BTP	71.6	51.6
FTP	61.7	59.0

We submitted the French shallow parsing F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams)and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [F(1,14) = 573.5, p < 0.0001] and *Direction* [F(1,14) = 14.7, p < 0.01], indicating better performance for chunk-based models and BTPs, respectively, and a significant *Unit Type* x *Direction* interaction [F(1,14) = 234.2, p < 0.0001], indicating better performance for the CBL model’s combination of BTPs and chunks.

The German shallow parsing scores, shown in Table B4, followed once more the same general pattern.

Table 7: German Shallow Parsing F-Scores

	Chunks	**N-grams**
BTP	75.7	53.2
FTP	71.0	67.4

We submitted the German shallow parsing F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams)and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [F(1,203) = 890.2, p < 0.0001] and *Direction* [F(1,203) = 47.5, p < 0.0001], indicating better performance for chunk-based models and BTPs, respectively, and a significant *Unit Type* x *Direction* interaction [F(1,203) = 389.8, p < 0.0001], indicating better performance for the CBL model's combination of BTPs and chunks.

Sentence Production Performance Results

The sentence production performance results for the CBL model, the FTP-Chunk model, and the two 3G baselines are shown in Figure B1.

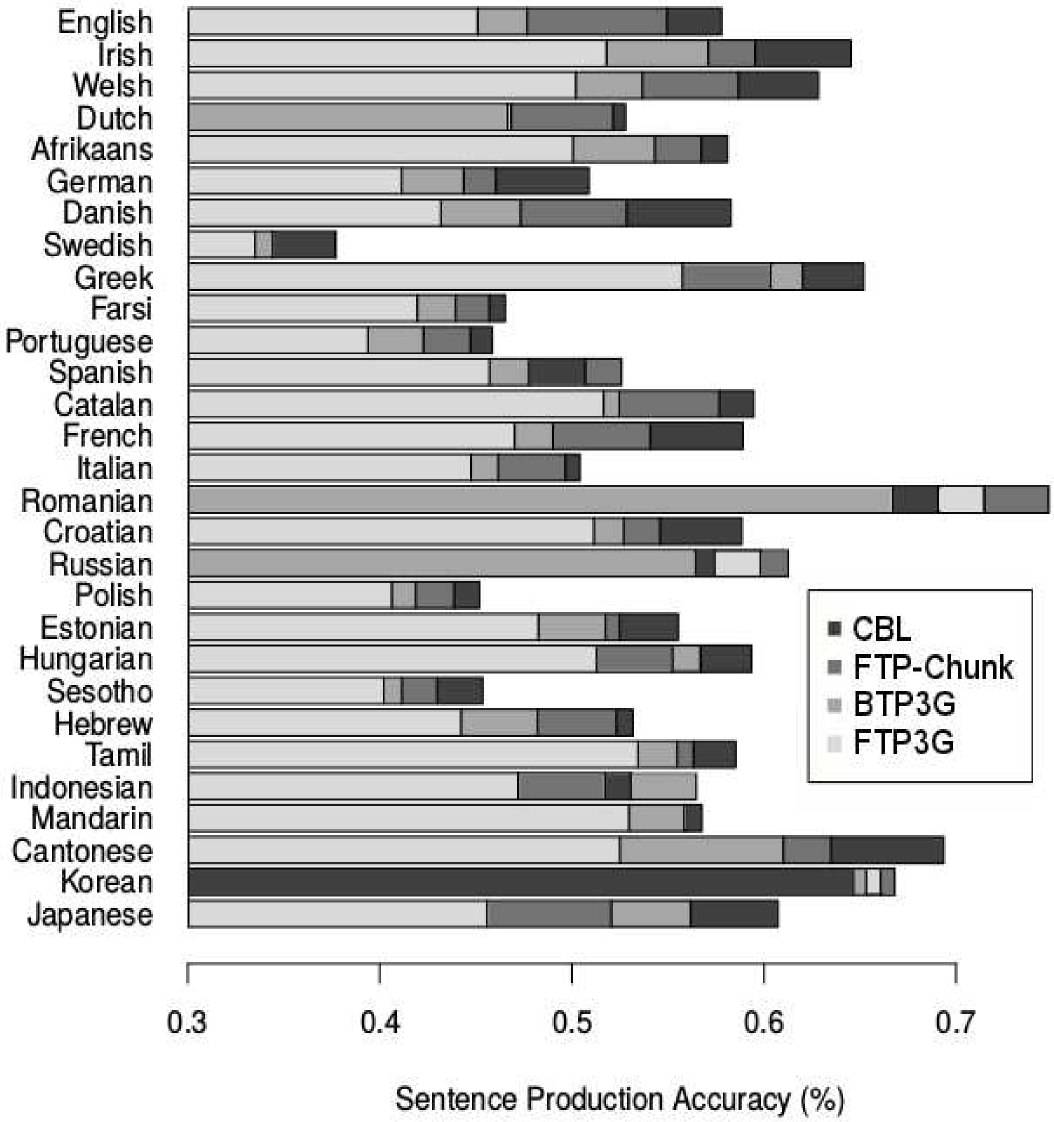

Fig. 23: Mean Sentence Production Accuracy scores for the CBL model and its trigram baseline across all 29 languages, including English (shown at top). Bars are non-cumulative

We submitted the sentence production performance F-scores (logit-transformed) to a repeated-measures ANOVA with the factors *Unit Type* (2: Chunks vs. *n*-grams)and *Direction* (2: BTP vs. FTP), with *Child Corpus* as a random factor. This yielded main effects of *Unit Type* [F(1,21) = 801.4, p < 0.0001] and *Direction* [F(1,21) = 128.4, p <

0.0001], indicating better performance for chunk-based models and BTPs, respectively.

REFERENCES

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language, 26,* 261–278.

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*, 789–834.

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

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science, 32*, 222-255.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191-238.

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

Arnon, I. & Christiansen, M.H. (in press). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science.*

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., & Cohen Priva, U. C. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*, 349-371.

Arnon, I., McCauley, S.M. & Christiansen, M.H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language, 92*, 265-280.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of- acquisition affects what gets learned. *Cognition, 122*, 292-305.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language*, *62*, 67–82. Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*, 329-347.

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.

Bannard, C., & Ramscar, M. (2007). Reading time evidence for storage of frequent multiword sequences. Abstract in *Proceedings of the Architectures and*

Mechanism of Language Processing Conference (AMLAP-2007), Turku, Finland.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*, 477–487.

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (Eds.), *Between the grammar and physics of speech: Papers in laboratory phonology I* (pp. 152–178). Cambridge, UK: Cambridge University Press.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111.

Berman, R. (1982). Verb-pattern alternation: The interface of morphology, syntax, and semantics in Hebrew child language. *Journal of Child Language, 9*, 169–191

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*, 752–793.

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.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner, & L. Gleitman (Eds.)*, Language acquisition: The state of the art* (pp. 319-346). New York: Academic Press.

Brandt, S., Diessel, H., & Tomasello, M. (2008). The acquisition of German relative clauses: A case study. *Journal of Child Language, 35*, 325.

Brandt, S., Kidd, E., Lieven, E., & Tomasello, M. (2009). The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics, 20*, 539-570.

Caldwell-Harris, C.L., Berant, J.B., & Edelman, S. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In S. T. Gries& D. Divjak (Eds.), *Frequency effects in cognitive linguistics (Vol. 1): Statistical effects in learnability, processing and change.* The Hague, The Netherlands: De Gruyter Mouton.

Carroll, S. (1989). Second-language acquisition and the computational paradigm. *Language Learning, 39*, 535-594.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

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.

Chater, N. & Christiansen, M.H. (2016). Squeezing through the Now-or-Never bottleneck: Reconnecting language processing, acquisition, change and structure. *Behavioral & Brain Sciences, 39,* e62.

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244-254.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19-22.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Christiansen, M.H. & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*, 126–161.

Clancy, P. M., Lee, H., &Zoh, M. H. (1986). Processing strategies in the acquisition of relative clauses: Universal principles and language-specific realizations. *Cognition*, *24*, 225–262.

Cohen, L., &Mehler, J. (1996). Click monitoring revisited: An on-line study of sentence comprehension. *Memory & Cognition*, *24*, 94–102.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. New York: Oxford University Press.

Dabrowska, E. (2000). From formula to schema: The acquisition of English questions. *Cognitive Linguistics, 11*, 83-102.

Demuth, K. 1992. *Acquisition of Sesotho*. In D. Slobin (ed.), *The cross-linguistic study of language acquisition, Vol. 3* (pp. 557-638). Hillsdale, N.J.: Lawrence Erlbaum Associates.

de Villiers, J. G., TagerFlusberg, H. B., Hakuta, K., & Cohen, M. (1979). Children's comprehension of relative clauses. *Journal of Psycholinguistic Research*, *8*, 499–518.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge, UK: Cambridge University Press.

Diessel, H., & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, *11*, 131–152.

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

Ellis S. H. (1973). Structure and Experience in the Matching and Reproduction of Chess Patterns. Doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second- language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 41, 375-396.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Edelman, S. (2008). *Computing the mind: how the mind really works*. Oxford University Press.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*, 540–551.

Feigenbaum, E. A., & Simon, H. A. (1962). A theory of the serial position effect. *British Journal of Psychology*, *53*, 307-320.

Ferreira, F. (2003). The misinterpretation of non-canonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11.

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

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*, 574-578.

Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development*, *67*, 3192–3218.

Fitz, H., & Chang, F. (2008). The role of the input in a connectionist model of the accessibility hierarchy in development. In *Proceedings of the 32nd Boston University Conference on Language Development* (pp. 120-131).

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 203–218.

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.

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

Frauenfelder, U., Segui, J., &Mehler, J. (1980). Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, *19*, 328–337.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.) *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge, UK: Cambridge University Press.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*, 181–200.

French, R. M., Addyman, C., &Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614.

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.

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1366-1383.

Gagliardi, A., Mease, T., &Lidz, J. (submitted). *U-shaped development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month olds*.

Gathercole, V., Sebastian, E., & Soto, P. (1999). The early acquisition of Spanish verbal morphology: Across-the-board or piecemeal knowledge? *International Journal of Bilingualism, 3*, 138–182

Gertner, Y., & Fisher, C. (2012). Predicted errors in children’s early sentence comprehension. *Cognition, 124,* 85-94.

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

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Actapsychologica*, *115*, 43-67.

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).

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447-474.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

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

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American linguistics*, *26*, 178–194.

Hamburger, H., & Crain, S. (1982). Relative acquisition. *Language development*, *1*, 245–274.

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.

Hamrick, P. (2014). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*, 247-278.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of linguistic structures*. Oxford, UK: Oxford University Press.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*, 269–286.

Holmes, V. M., & de la Bâtie, B. D. (1999). Assignment of grammatical gender by native speakers and foreign learners of French. *Applied Psycholinguistics, 20*, 479-506.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*, 417–430.

Jackendoff, R. (1995). The boundaries of the lexicon. *Idioms: Structural and Psychological Perspectives*, 133–165.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar,evolution.* Oxford, UK: Oxford University Press.

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

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*, 789-845.

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological bulletin*, *131*, 684.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., &Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, *24*, 252–293.

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (in press). Meaning overrides frequency in idiomatic and compositional multiword chunks. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. *Frontiers in Psychology, 3*, 167. doi: 10.3389/fpsyg.2012.00167.

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.

Keenan, E. L., & Hawkins, S. (1987). The psychological validity of the accessibility hierarchy. *Universal Grammar*, *15*, 60–85.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Kirjavainen, M., Theakston, A., & Lieven, E. (2009). Can input explain children's me-for-I errors? *Journal of Child Language, 36*, 1091-1114.

Klahr, D., Chase, W. G., & Lovelace, E. A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. 9, 462–477

Koh, S., Sanford, A. J., Clifton, C., &Dawydiak, E. J. (2008). Good-enough representation in plural and singular pronominal reference: Modulating the conjunction cost. In J. Gundel& N. Hedberg (Eds.) *Reference: Interdisciplinary Perspectives* (pp. 123-139). Oxford: Oxford University Press.

Kol, S., Nir, B., &Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language*, *41*, 176-199.

Kolodny, O., Lotem, A., & Edelman, S. (2015). Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. *Cognitive Science*, *39*, 227-267.Reali, F. & Christiansen, M.H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60,* 161-170.

Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*, 68–101.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333–370.

MacDonald, M.C. & Christiansen, M.H. (2002). Reassessing working memory: A comment on Just &Carpenter (1992) and Waters & Caplan (1996). *Psychological Review, 109,* 35-54.

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review, 89,* 483-506.

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.

Mandel, D. R., Jusczyk, P. W., &Kemler Nelson, D. G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, *53*, 155–180.

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

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., &Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development, 57 (4, Serial No. 228).*

Mariscal, S. (2008). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *Journal of Child Language, 35,* 1-29.

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.

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. (2013). Toward a unified account of comprehension and production in language development. *Behavioral and Brain Sciences*, *36*, 366-367.

McCauley, S.M. & Christiansen, M.H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9,* 419-436.

McCauley, S.M. & Christiansen, M.H. (in press). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science.*

McCauley, S.M. & Christiansen, M.H. (2015). Individual differences in chunking ability predict on-line sentence processing. In D.C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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.

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

Miller, G. A. (1958). Free recall of redundant strings of letters. Journal of Experimental Psychology, 56, 485–491.

Miller, G. A., & Taylor, W. G. (1948). The perception of repeated bursts of noise. *The Journal of the Acoustical Society of America*, *20*, 171-182.

Miyata, S. (2000). The TAI corpus: Longitudinal speech data of a Japanese boy aged 1;5.20–3;1.1. *Bulletin of Shukutoku Junior College, 39*, 77–85.

Miyata, S., & Naka, N. (1998). Wakachigaki Guideline for Japanese: WAKACHI98 v.1.1. The Japanese Society for Educational Psychology Forum Report No. FR-98-003, The Japanese Association of Educational Psychology.

Monaghan, P. & Christiansen, M.H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data(TILAR Series)* (pp. 139-163). Amsterdam: John Benjamins.

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

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., &Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*, 1299-1305.

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., &Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Actapsychologica*, *149*, 1-8.

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

Perruchet, P., Vinter, A., Pacteau, C., &Gallego, J. (2002). The formation of structurally relevant units in artificial grammar learning. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(2), 485-503.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge

University Press.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105–110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329-347.

Pine, J. M., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123-138.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Harper Collins.

Pizutto, E. and Caselli, C. (1992). The acquisition of Italian morphology. *Journal of Child Language, 19*, 491–557.

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

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*, 760-793.

Reali, F. & Christiansen, M.H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60,* 161-170.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855-863.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

Remez, R. E., Ferro, D. F., Dubowski, K. R., Meer, J., Broder, R. S., &Davids, M. L. (2010). Is desynchrony tolerance adaptable in the perceptual organization of speech?.*Attention, Perception, & Psychophysics*, *72*, 2054-2058.

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive science*, *35*, 1352-1389.

Rogers, M. (1987). Learners difficulties with grammatical gender in German as a foreign language. *Applied Linguistics, 8*, 48-74.

Rubino, R. and Pine, J. (1998) Subject–verb agreement in Brazilian Portuguese: What low error rates hide. *Journal of Child Language, 25*, 35–60.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

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

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.

Sanford, A. J. (2002). Context, attention, and depth of processing during interpretation. *Mind and Language*, *17*, 188–206.

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations of comprehension beyond the sentence*. New York: Wiley.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes*, *26*, 159–190.

Sanford, A. J. S., Sanford, A. J., Filik, R., &Molle, J. (2005). Depth of lexical-semantic processing and sentential load. *Journal of Memory and Language*, *53*, 378–396.

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

Schiffman, H. F. (1999). *A reference grammar of spoken Tamil.* Cambridge, UK: Cambridge University Press.

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

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592.

Simon, H. A. (1974). How big is a chunk?.*Science*, *183*, 482-488.

Simon H. A. &Gilmartin K. J. (1973). A simulation of memory for chess positions. Cognitive Psychology, 5, 29-46.

Simon D. P. & Simon H. A. (1973). Alternative uses of phonemic information in spelling. Rev. Educ. Res. 43, 115–137.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*, 644–649.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*, 251-272.

Slobin, D. I. (1986). *The crosslinguistic study of language acquisition: The data (Vol. 2).* London: Psychology Press.

Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*, 249–267.

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

Sprenger, S. A., Levelt, W. J., &Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*, 161–184.

Stemberger, J.P., & Bernhardt, B.H., & Johnson, C.E. (1999). "Regressions" ("u"-shaped learning) in the acquisition of prosodic structure. Poster presented at the 6th International Child Language Congress, July 1999.

Studdert-Kennedy, M. (1986). Some developments in research on language behavior. *Behavioral and Social Science: 50 Years of Discovery*, 208.

Sturt, P., Sanford, A. J., Stewart, A., &Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, *11*, 882–888.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, *36*, 201-216.

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

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*, 355-370.

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes, 4*, 211-234.

Tavakolian, S. L. (1977). *Structural principles in the acquisition of complex sentences*. Unpublished doctoral dissertation. University of Massachusetts, Amherst.

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.

Tomasello, M. and Brooks, P. (1998) Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics, 9,* 379–395.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.) *Perspectives on formulaic language: Acquisition and communication* (pp. 151-173). London: Continuum International Publishing Group.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing Advantages of Lexical Bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*, 569-613.

Tsarfaty, R., Nivre, J., &Andersson, E. (2012). Joint evaluation of morphological segmentation and syntactic parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2 (pp. 6-10)*. Association for Computational Linguistics.

Tunstall, S. L. (1998). *The interpretation of quantifiers: Semantics and processing*. Unpublished doctoral dissertation. University of Massachusetts, Amherst.

Tyler, L. K., & Marslen-Wilson, W. D. (1977). The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 683-692.

Wanner, E. & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.) *Linguistic theory and psychological reality* (pp. 119-161). Boston: MIT Press.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science*, *164*, 586-587.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*, 591–597.

Wray, A. (2005). *Formulaic language and the lexicon.* Cambridge, UK: Cambridge University Press.

CHAPTER 3

ACQUIRING FORMULAIC LANGUAGE[8]

Formulaic expressions have long been held to be a key component of language use within cognitive linguistics (e.g., Croft, 2001; Langacker, 1987; Wray, 2002).[9] Lending support to this perspective, a number of psycholinguistic studies have demonstrated that adults are sensitive to the frequency of multiword sequences. These include reaction time studies (Arnon & Snider, 2010; Jolsvai, McCauley, & Christiansen, 2013), as well as studies of complex sentence comprehension (Reali & Christiansen, 2007), self-paced reading and sentence recall (Tremblay, Derwing, Libben, & Westbury, 2011), and event-related brain potentials (Tremblay & Baayen, 2010). Similar findings have been shown for production, with naming latencies decreasing as a function of phrase frequency (Janssen & Barber, 2012) and reduced phonetic duration for frequent multiword strings in spontaneous and elicited speech (Arnon & Cohen-Priva, 2013). Together, these studies suggest the active use of fixed multiword sequences as linguistic units in their own right, which implies a far greater role for formulaic language processing than has previously been assumed.

Importantly, such results have been mirrored in psycholinguistic studies with young children (Arnon & Clark, 2011; Bannard & Matthews, 2008). In addition to lending support to usage-based approaches (which hold that linguistic productivity

[8] An abridged version of this paper was published as McCauley, S.M. & Christiansen, M.H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9,* 419-436.

[9] For the purposes of the present paper, we define “formulaic expression” according to Wray (1999): *a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.*

emerges from abstraction over multiword sequences; e.g., Tomasello, 2003), such findings suggest that children's item-based linguistic units—and their active use during processing—do not diminish, but persist throughout development and into adulthood. If this is indeed the case, it holds that researchers can better understand the role of formulaic sequences in adult language by studying the processes and mechanisms whereby children discover and use multiword units during the acquisition process.

The aim of the present paper is to take the first steps toward establishing the computational foundations of a developmental approach to adult formulaic language use. To this end, we describe two simulations performed using a computational model of acquisition which instantiates the view that the discovery and on-line use of concrete multiword units forms the backbone for children's early language processing. The model tests explicit mechanisms for the acquisition of formulaic language and is used to evaluate the extent to which children's linguistic behavior can be accounted for using concrete multiword units. Importantly, the role of multiword sequences in the model grows rather than diminishes over time, in keeping with the perspective that children's linguistic units persist throughout development and into adulthood. Moreover, the model takes usage-based theory to its natural conclusion; the model learns by attempting to comprehend and produce utterances, such that no distinction is made between language learning and language use. By avoiding a separate process of grammar induction, the model captures the usage-based notion that linguistic knowledge arises gradually through what is learned during concrete usage events (the notion of *learning by doing*).

In what follows, we first discuss the psychological and computational features of the model, as well as its inner workings[10], before evaluating the model's ability to account for key psycholinguistic findings on young children's formulaic language use.

The Chunk-Based Learner (CBL) Model

As our model is primarily concerned with the learning and use of concrete multiword linguistic units, or “chunks,” we refer to it as the Chunk-Based Learner (CBL; McCauley & Christiansen, in preparation; McCauley, Monaghan, & Christiansen, 2015; see also McCauley & Christiansen, 2011). We designed CBL with a number of key psychological and computational features in mind:

1) **Incremental, on-line processing:** In the model, all input and output is processed in a purely incremental, on-line, word-by-word fashion, as opposed to involving batch learning or whole-utterance optimization, reflecting the incremental nature of human sentence processing (e.g., Altmann & Steedman, 1988; Borovsky, Elman, & Fernald, 2012). At any given point in time, the model can only rely on what has been learned from the input encountered thus far.
2) **Psychologically inspired learning mechanisms and knowledge representation:** The model learns by calculating simple statistics tied to backward transitional probabilities, to which both infants (Pelucchi, Hay, & Saffran, 2009) and adults (Perruchet & Desaulty, 2008) have been shown to be sensitive. Moreover, the model learns from local linguistic information as opposed to storing entire utterances, in accordance with evidence for the primacy of local information in sentence processing

[10] All source code for the model and simulations will be made publicly available in the near future. Interested parties can contact the authors for model-specific code.

(e.g., Ferreira & Patson, 2007). In keeping with evidence for the unified nature of comprehension and production (Pickering & Garrod, 2013), comprehension and production are two sides of the same coin in the model, relying on the same statistics and linguistic knowledge.

3) **Usage-based learning:** In the model, the problem facing the learner is characterized as one of learning to process language. All learning takes place during individual usage events; that is, specific attempts to comprehend and produce utterances.
4) **Naturalistic linguistic input:** To ensure representative, naturalistic input, the model is trained and evaluated using corpora of child and child-directed speech taken from the CHILDES database (MacWhinney, 2000).

This combination of features makes CBL unique among computational models of language development, in terms of psychological plausibility. Language development in the CBL model involves learning—in an unsupervised manner—to perform two tasks: 1) “comprehension,” which is approximated by the segmentation of incoming utterances into phrase-like units useful for arriving at the utterances' meanings, and 2) “production,” which involves the incremental generation of utterances using the same multiword units discovered during comprehension. Importantly, comprehension and production in the model form a unified framework, as they rely on the same sets of chunks and statistics (cf. McCauley & Christiansen, 2013).

Architecture of the Model

Comprehension. The model processes input word-by-word as it is encountered, from the very beginning of the input corpus. At each time step, the model updates frequency information for words and word-pairs, which is used on-line to track the backward transitional probability (BTP) between words[11]. While processing each utterance incrementally, the model maintains a running average of the mean BTP calculated over the words encountered in the corpus so far. Peaks are defined as those BTPs which match or rise above this average threshold, while dips are defined as those which fall below it (allowing the avoidance of a free parameter). When a peak in BTP is encountered between two words, the word-pair is chunked together such that it forms part (or all) of a chunk. When a dip in BTP is encountered, a “boundary” is placed and the resulting chunk (which consists of the one or more words preceding the inserted boundary) is placed in the model's *chunkatory*, an inventory of chunks consisting of one or more words.

Importantly, the model uses its chunk inventory to assist in segmenting input and discovering further chunks as it processes the input on-line. As each word-pair is encountered, it is checked against the chunk inventory. If the sequence has occurred before as either a complete chunk or part of a larger chunk, the words are automatically chunked together regardless of their transitional probability. Otherwise, the BTP is compared to the running average threshold with the same consequences as usual (see McCauley & Christiansen, 2011, for further detail).

Because there are no fixed limits on the number or size of chunks that the model can learn, the resulting chunk inventory contains a mixture of words and

[11] BTPs were chosen over forward transitional probabilities because BTPs involve evaluating the probability of a sequence based on the most recently encountered item, as opposed to moving back one step in time (as is necessary when calculating forward transitional probabilities).

multiword units. Aside from the aforementioned role of the chunk inventory in processing input, chunks stored in the model's inventory are treated as separate and distinct units; chunks may contain overlapping sequences without interference. Moreover, chunks do not weaken or decay due to overlap or disuse. These representational properties allow the model to function without free parameters (in contrast to other well-known computational models of distributional learning, such as PARSER; Perruchet & Vinter, 1998).

The model's comprehension performance can be evaluated against the performance of shallow parsers (sophisticated tools widely used in natural language processing), which segment texts into series of non-overlapping, non-embedded phrases. We chose to focus on shallow parsing in evaluating the model in accordance with a number of recent psycholinguistic findings suggesting that human sentence processing is often shallow and underspecified (e.g., Ferreira & Patson, 2007; Frank & Bod, 2011; Sanford & Sturt, 2002), as well as the item-based manner in which children are hypothesized to process sentences in usage-based approaches (e.g., Tomasello, 2003).

Production. As the model makes its way through a corpus, segmenting utterances and discovering chunks in the service of comprehension, it encounters utterances made by the target child of the corpus, which are the focus of the production task. The production task begins with the idea that the overall message the child wishes to convey can be roughly approximated by treating the utterance as an unordered bag-of-words (cf. Chang, Lieven, & Tomasello, 2008). The model's task, then, is to reproduce the child's utterance by outputting the items from the bag in a sequence that matches that of the original utterance. Importantly, the model can only rely on the chunks and statistics it has previously learned during comprehension to achieve this.

Following evidence for children's use of multiword units in production, the model utilizes its chunk inventory when constructing utterances. To allow this, the bag-of-words is populated by comparing parts of the child's utterance to the model's chunk inventory; word combinations from the utterance that are represented as multiword chunks in the model's chunk inventory are placed in the bag-of-words. The model then begins producing a new utterance by selecting the chunk in the bag which has the highest BTP, given the start-of-utterance marker (which marks the beginning of each utterance in the corpus). The selected chunk is then removed from the bag and placed at the beginning of the utterance. At each subsequent time step, the chunk with the highest BTP given the most recently placed chunk is removed from the bag and produced as the next part of the utterance. This process continues until the bag is empty. Thus, the model's production attempts are based on incremental, chunk-to-chunk processing, as opposed to whole-sentence optimization.

Each utterance produced by the model is scored against the child's original utterance. Regardless of grammaticality, the model's utterance receives a score of 1 for a given utterance if (and only if) it matches the child utterance in its entirety; in all other cases, a score of 0 is received. The model's production abilities can then be evaluated on any child corpus in any language, according to the overall percentage of correctly produced utterances.

Previous Results Using the CBL Model

While the focus of the present paper is on simulations that directly capture psycholinguistic data, we note here that previous work using CBL has underscored the robustness and scalability of the model more generally. Thus, McCauley et al. (2015) described the results of over 40 simulations of individual children from the CHILDES database (MacWhinney, 2000). On the comprehension task, the model was shown to learn useful multiword units, approximating the performance of a shallow parser (e.g., Punyakanoth & Roth, 2001) with high accuracy and completeness. In production, the model was able to produce the majority of the child utterances encountered in each corpus. Furthermore, McCauley & Christiansen (in preparation; see also McCauley & Christiansen, 2011) demonstrated that the model is capable of producing the majority of child utterances across a typologically diverse array of 28 additional languages (also from the CHILDES database). Importantly, the CBL model outperformed more traditional bigram and trigram models (cf. Manning & Schütze, 1999) cross-lingustically in both comprehension and production.

In what follows, we evaluate the model according to its ability to account for key psycholinguistic findings on children's distributional learning of multiword units, as well as their use in early comprehension and production.

Modeling Developmental Psycholinguistic Data

Whereas previous simulations have examined the ability of CBL to discover building blocks for language learning, in the current paper we investigate the psychological validity of these building blocks. We report simulations of empirical data covering two key developmental psycholinguistic findings regarding children's distributional and item-based learning. The first simulation demonstrates CBL's ability to capture child artificial grammar learning (Saffran, 2002). The second simulation shows CBL's ability to capture child sensitivity to multiword sequence frequency (Bannard & Matthews, 2008) while the third concerns the learning of formulaic sequences and their role in morphological development (Arnon & Clark, 2011).

Simulation 1: Modeling Comprehension in Child Artificial Grammar Learning

If CBL offers a reasonable developmental account of learning from the distributional information available in language input, it should make contact with available developmental psycholinguistic data involving the learning of distributional information from language-like artificial stimuli. We therefore attempted to model data from an artificial language learning experiment (Saffran, 2002) using the CBL model. This particular study was chosen because it was purported to involve the use of predictive dependencies on the part of learners to group words into phrase-like units; as such, it provided the means to test the model's ability to learn phrase-level units based on BTPs. We focused solely on Experiment 2 because it involved child subjects (the other experiments described in Saffran, 2002 involved adult subjects). Children ranging in age from 7;6 to 9;8 were exposed to items from one of two artificial languages; one contained predictive dependencies between words within phrases (Language P), while the other lacked this cue to phrase structure (Language N). As illustrated by Table 8, sentences in the two languages were generated by a set of rewrite rules, with A, C, D, F, and G corresponding to separate categories of nonwords. In Language P the items from categories D and G were always preceded by items from categories A and C, respectively (though the same relationship did not hold in reverse; A and C items were not dependent on a subsequent appearance by D and G items). Language N featured no such predictive dependencies; items from either category making up a phrase could appear either together, or independently. On a forced-choice task between pairs of grammatical and ungrammatical items, children exposed to Language P outperformed those exposed to Language N.

Table 8

Rewrite rules underlying Language P and Language N

Language P	**Language N**
S → AP + BP + (CP)	**S** → AP + BP
AP → A + (D)	**AP** → [(A) + (D)]
BP → CP + F	**BP** → CP + F
CP → C + (G)	**CP** → [(C) + (G)]

Note: Parentheses denote optional elements.

Method. The model architecture was identical to that used with natural languages. Importantly, this meant that the model continued learning during exposure to test items. For each language, 15 simulations were performed, corresponding to the 30 child subjects from the original study. The model received the same amount of exposure to the original stimuli as the subjects did (for each language, 50 sentences repeated 8 times for a total of 400 training items, followed by 24 test item pairs). As the original presentation order was no longer available (Saffran, personal communication), 10 random orderings of the training items were generated (5 for each language), and 3 simulations run on each. Each test item pair consisted of one sentence which was grammatical, and one which was ungrammatical. The languages from the original study were structured such that the same set of test items could be used for both groups. Five random presentation orders of the test item pairs were generated for each language, and were counterbalanced such that the grammatical item came first in exactly half of the item pairs. The model was evaluated against a version of the test items which contained the correct phrase boundaries, as defined by the rewrite rules used to generate the sentences in the original study. Boundaries were placed between phrases in a non-embedded fashion that emulated the shallow parsing technique used to evaluate the model's performance on natural languages.

To further contrast the usefulness of item-specific vs. class-based distributional information, a separate set of simulations was performed after each word had been replaced by the corresponding class symbol from the rewrite rules in the original study.

Scoring. We introduced a new scoring method in order to model the two-alternative forced choice (2AFC) task used in the original study. Each item in a given test pair was scored according to the number of correctly placed phrase boundaries, and the item with the higher score was selected as the response. If the model scored the same number of hits (including zero) for both items, a choice was made at random, allowing individual differences across simulations.

Results and discussion. The child subjects in Saffran (2002) had overall correct response rates of 71.8% for Language P and 58.3% for Language N. The model provided a close quantitative fit, with overall correct response rates of 70.5% for Language P, and 57.5% for Language N (as shown in Figure 24). In neither case did the model's correct response rate differ significantly from that of the children in the original study (Language P, t[14] = -0.94, p = 0.36; for Language N, t[14] = -0.612, p = 0.55)[12].

[12] Because the standard deviations and individual subjects scores were not reported in the original study and were no longer available (Saffran, personal communication), we report t-tests (two-tailed) against the mean child correct response rate for each language. We use this particular test because Saffran used it in her analyses of the original results.

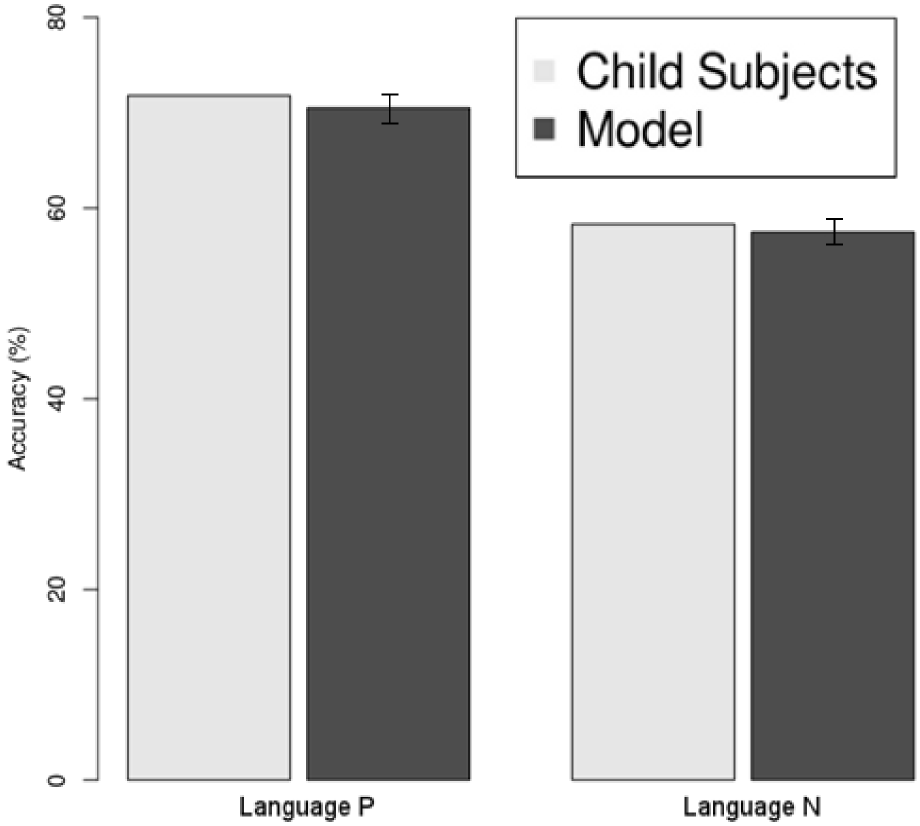

Fig. 24: Child subject and CBL accuracy rates for the forced-choice task in Saffran (2002). Error bars denote standard error.

However, when the model learned information on word classes instead of concrete items, it provided a poor fit to the child data, with 81.1% accuracy for Language P and 39.7% accuracy for Language N (as shown in Figure 25). In both cases, the overall correct response rates differed significantly from those of the child subjects [Language P, $t(14) = 5.75$, $p < 0.001$; Language N, $t(14) = -10.81$, $p < 0.001$]. This runs counter to a common assumption in the statistical learning literature: that subjects in AGL studies calculate statistics over form classes as opposed to concrete items (cf. Saffran, 2002; Thompson & Newport, 2007).

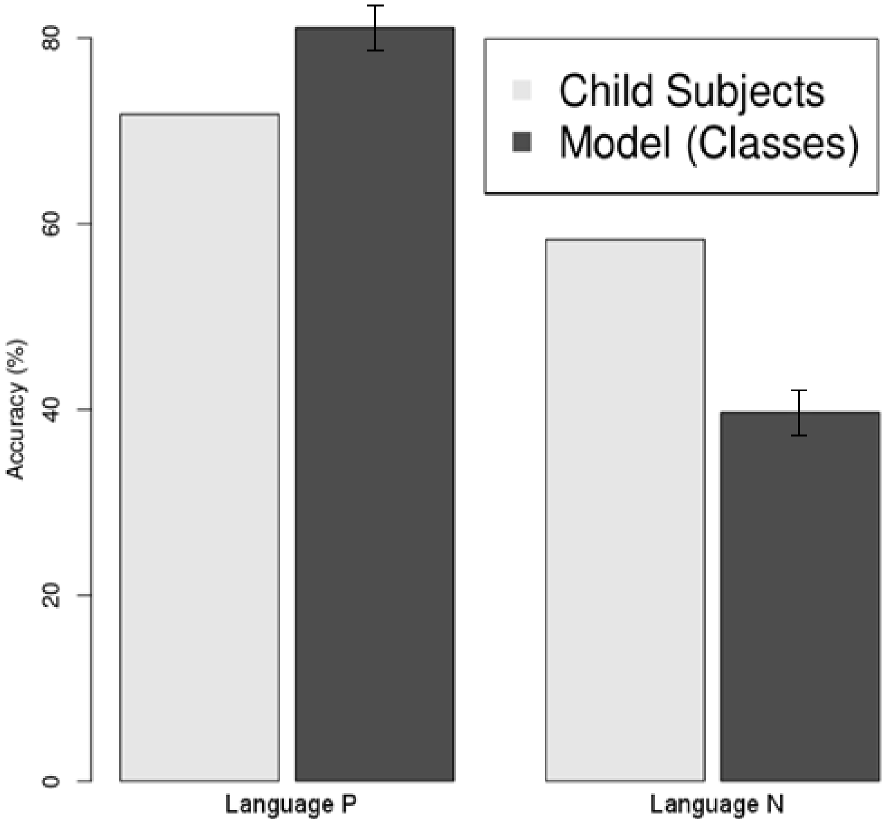

Fig. 25: Child subject and class-based model accuracy rates for the forced-choice task in Saffran (2002).

To further explore the model's sensitivity to the phrase-like structure of the stimuli, we conducted an analysis of the model's responses to specific properties of the grammar, in accordance with an analysis reported by Saffran in the original study. These properties, along with examples of test item pairs designed to test subject sensitivity to them, are shown in Table 9.

Table 9

Rules Tested in Saffran (2002)

Rule		
1	Sentences must contain an A phrase.	
	BIFF KLOR SIG PILK JUX	[A-D-C-G-F]
	*SIG PILK JUX	[C-G-F]
2	D words follow A words, while G words follow C words.	
	HEP PELL LUM PILK JUX	[A-D-C-G-F]
	*HEP PILK LUM PELL JUX	[A-D-C-D-F]
3	Sentences must contain an F word.	
	MIB LUM PILK VOT	[A-C-G-F]
	*MIB LUM PILK	[A-C-G]
4	C phrases must precede F words.	
	RUD PELL NEB DUPP	[A-D-C-F]
	*RUD PEL DUPP	[A-D-F]

In Saffran's study, children's correct responses after exposure to Language P were numerically above chance for items exhibiting each these properties, though the difference only reached significance for properties 1, 3, and 4. In a similar fashion, CBL's overall correct response rate for each of the four item types was numerically above chance, though the difference reached significance in all four cases (significance tests reported in Table 10). For Language N, children's correct response rates were once more numerically above chance for all four properties, though this was only significant for properties 3 and 4. Similarly, CBL's correct response rates did not differ significantly from chance for the first two properties, and significantly exceeded chance for property 3. However, the model's correct response rate was at chance for property 4 (see Table 10).

Table 10

Mean Scores and Significance Tests (Two-tailed) against Chance (Three of Six Possible), for Language P and Language N

Language	**Subjects/Model**	**Rule 1**	**Rule 2**	**Rule 3**	**Rule 4**
P	Children	4.60**	3.20	5.20**	4.27**
	CBL	3.60*	4.60**	4.67**	4.07**
N	Children	3.07	3.33	3.80*	3.80*
	CBL	3.26	2.67	4.93**	2.93

*p < 0.05, **p < 0.01

Thus, when trained on the exact same items encountered by subjects, the model not only provided a close numerical fit to the data but was able to capture differences in children's sensitivity to specific properties of the grammar across the two languages. Nevertheless, the model's performance did not match that of the child subjects perfectly; it scored above chance on Rule 2 for Language P (children were numerically above chance but statistically at chance level), and was statistically at chance level for Rule 4 for Language N (whereas children were statistically above chance). These differences may stem in part from various features of the stimuli which the model is not sensitive to, such as sequence length.

The model's close fit to child performance expands CBL's coverage of distributional learning, and suggests that an ability to group words into larger units can indeed account for subject performance in the original study. While predictive dependencies between word classes were a potentially useful cue, the calculation of statistics over classes in the present model could not account for subject performance as well as could the item-based approach. This resonates with the superior performance of the model on natural languages when working with words as opposed to lexical categories.

Simulation 2: Modeling Children's Sensitivity to Phrase Frequency

Bannard & Matthews (2008) provide some of the first direct evidence that children store frequent multiword sequences and that such sequences may be processed differently than similar, less frequent sequences. Their study contrasted children's repetition of four-word compositional phrases of varying frequency (based on analysis of a corpus of child-directed speech; Maslen, Theakston, Lieven, & Tomasello, 2004). For instance, *go to the shop* formed a high-frequency phrase which was contrasted with a low-frequency phrase, *go to the top*. Two and 3-year-olds were more likely to repeat an item correctly when its fourth word combined with the preceding trigram to form a frequent chunk, and 3-year-olds were significantly faster to repeat the first three words. As the stimuli were matched for the frequency of the final word and final bigram, only the frequencies of the final trigram and entire four-word phrase differed across conditions, suggesting that children do, in some sense, store multiword sequences as units.

If CBL provides a reasonable account of children's multiword chunk formation, it should show similar phrase frequency effects to those found in the Bannard and Matthews study, despite the fact that it is not directly sensitive to raw whole-string frequency information (the frequency of a sequence is only maintained if it has first been discovered as a chunk). To test this prediction, we exposed CBL to a corpus of child-directed speech and computed the “chunkedness” of the test items' representations in the model's chunkatory.

Method. The model architecture was identical to that used in prior simulations (e.g., McCauley & Christiansen, 2011). We began by exposing the model to the dense corpus of child-directed speech that was previously used in our natural language simulations (Maslen, Theakston, Lieven, & Tomasello, 2004). This corpus was chosen not only because of its density, but also because it was recorded in Manchester, UK, where the Bannard and Matthews study was carried out. To capture the difference between the 2- and 3-year-old subject groups in the original study, we tested the model twice: once after exposure to the corpus up to the point at which the target child's age matched the mean age of the first subject group (2;7), and once after exposure up to the point at which the target child's age matched that of the second group (3;4).

Following exposure, the chunkedness of each test item's representation in the model's chunkatory was determined.

Scoring. Our previous analyses of the chunkatories built by CBL during exposure to various corpora in previous natural language simulations showed that most of the model's multiword chunks involved 2- or 3-word sequences. As the stimuli in Bannard and Matthews all consisted of 4-word phrases, we focused on the chunk-to-chunk statistics that would be used by the model to construct each phrase during production, thereby offering a simulation of children's production attempts. A phrase's score was calculated as the product of the BTPs linking each chunk in the sequence, yielding the *degree of chunkedness* for that sequence. If a sequence happened to be stored as a 4-word chunk in the chunkatory, the model received a chunkedness score of 1, indicating a BTP of 1 (as no chunk-to-chunk probability calculation was necessary). In the case of an item represented as two separate chunks, the degree of chunkedness for the test item was calculated as the chunk-to-chunk BTP between the two chunks.

Results and discussion. Two-year-olds in the original study were 10% more likely to repeat a high-frequency phrase correctly than a phrase from the low-frequency condition, while 3-year-olds were 4% more likely (both differences were significant). There was also a duration effect found for the 3-year-olds, who were significantly faster to repeat the first three words on high-frequency trials. CBL exhibited phrase frequency effects that were graded appropriately across the three frequency bins used in the original study.[13] In the 2-year-old simulation, the mean degree of chunkedness (BTP) scores were: 0.4 (high-frequency), 0.2 (intermediate-frequency), and 0.008 (low-frequency). In the 3-year-old simulation, the mean BTP scores were: 0.38 (high-frequency), 0.21 (intermediate-frequency), and 0.08 (low-frequency). Thus, CBL was able to capture the general developmental trajectory exhibited across subject groups: the difference in performance between high- and low-frequency conditions was lower in our 3-year-old simulation, just as in Bannard and Matthew's child subject group.

Thus, the model not only captured the graded phrase frequency effect exhibited by the child subjects, but also fit the overall pattern of a less dramatic difference in performance between high- and low-frequency conditions for the 3-year-old subject group. As the stimuli in the original study were matched for unigram and bigram substring frequencies, a simple bigram model could *not* produce a phrase frequency effect like the one exhibited by the model; the result necessarily stems from CBL's ability to discover multiword chunks. This is despite the fact that many of the test items, even in the high-frequency group, were stored as two separate chunks in the model's chunkatory. The chunk-to-chunk BTPs linking two-word chunks like *a drink* and *of milk* (chunks forming a high-frequency phrase) were higher than the BTPs linking chunks like *a drink* and *of tea* (chunks forming a low-frequency phrase), despite the fact that *of milk* and *of tea* had nearly identical token counts in the chunkatory. This is not a trivial consequence of overall phrase frequency in the

[13] Note that while items in the Intermediate condition were listed by Bannard and Matthews, they reported no results or analyses for children's repetition of them, beyond inclusion in a regression analysis. We report CBL's performance for these items to emphasize the graded nature of the phrase frequency effect exhibited by the model.

corpus; because the model relies on backward rather than forward transitional probabilities, the raw frequency count of the entire sequence was not the only important factor (and was never utilized by the model). Of greater importance was the number of different chunks that could immediately precede the non-initial chunks in the sequence. For instance, because the bigrams *of milk* and *of tea* are matched for frequency, and the sequence *a drink* immediately precedes *of milk* with greater frequency than *of tea*, there are necessarily a greater number (in terms of token rather than type frequency) of different two-word sequences that precede *of tea* which are not *a drink*, resulting in a lower chunk-to-chunk BTP linking stored chunks like *a drink* and *of tea* than *a drink* and *of milk*[14]. Importantly, this difference in the statistical properties of the sequences suggests that the overall *cohesiveness* of the sequence (as captured by BTPs in the current instance) may be as important as overall phrase frequency when it comes to the representation of multiword sequences. Future behavioral work with children (and adults) should target this issue.

As noted above, the model captured the general finding that the high- and low-frequency stimuli were processed more similarly by the older children. That this counter-intuitive pattern was exhibited by the model supports the view that CBL does indeed offer a psychologically plausible and informative account of children's discovery and use of multiword chunks. Moreover, this result also resonates with the developmental trajectory of the model's chunk inventory, in which the importance of

[14] Because the stimuli were not matched for trigram substring frequency (the final trigram in high-frequency phrases being of higher frequency that that of low-frequency phrases), the same pattern would hold even if *a* and *drink*, in the previous example, were not represented as a single chunk by the model; the BTP between *drink* and *of milk* would still be higher than that between *drink* and *of tea*, for the same reasons discussed above.

multiword sequences grows, rather than diminishes, over time. Figure 26 depicts the number and size of the chunk types learned by the model during training up to 2;7 and 3;4 on the dense corpus. Importantly, while the number of types grows consistently across chunks of various sizes, chunks of size four and greater are discovered at an increased rate during the period between 2;7 and 3;4. Thus, while Bannard and Matthews' (2008) finding of a less dramatic difference between conditions for the older children might appear to suggest a decreased reliance on multiword sequences, the model's ability to capture this pattern is actually driven by an *increased* reliance on chunks. Because the model already has strong coverage of the items in the high-frequency condition at 2;7, the discovery of new chunks between 2;7 and 3;4 primarily increases the model's coverage of the test sequences in the low-frequency condition. This leads us to reaffirm our prediction that the importance of multiword units may actually grow, rather than diminish, throughout development. In this context, the chunking mechanism made explicit in the model could help explain the apparent pervasiveness of multiword units in adult language processing (as reviewed in the introduction).

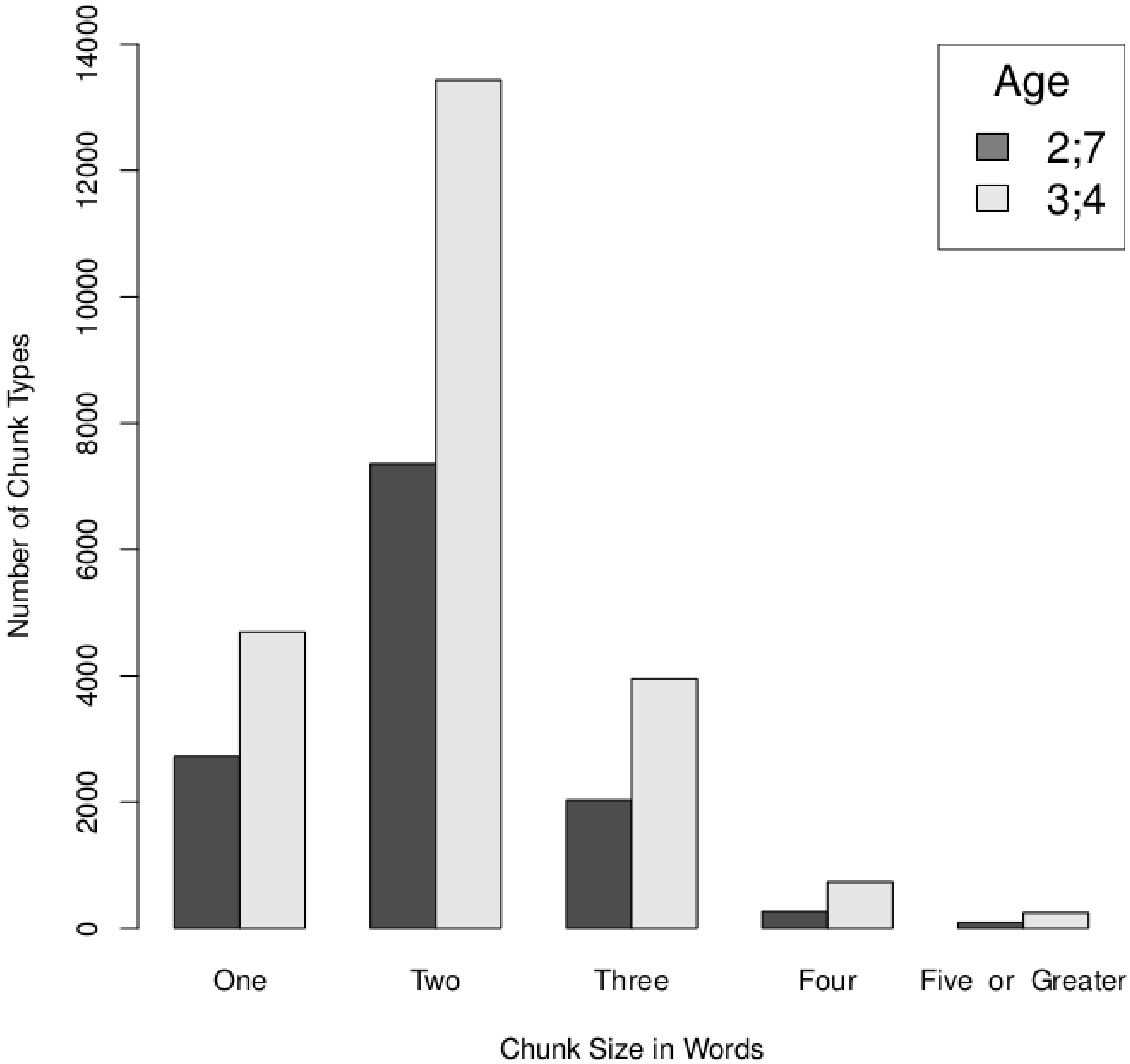

Fig 26: Number and size (in words) of chunk types in the chunk inventory at the early and late stages of Simulation 2. Each stage corresponds to the average age of the children in each age group in Bannard and Matthews (2008).

Our results also have implications for approaches to multiword chunk storage more generally: the fact that the model was able to capture phrase-frequency effects by learning to form chunks over pre-segmented input underscores the idea that not all of children's stored chunks need stem from initial “under-segmentation” of the speech stream (see also Arnon & Christiansen, in preparation, for discussion).

Simulation 3: Modeling the Role of Multiword Units in Children's Morphological Development

Arnon and Clark (2011) examined the impact of multiword sequences on children's morphological development, specifically with respect to patterns of over-regularization when producing irregular noun plurals. American English speaking children (mean age: 4;6) were tested in an elicitation paradigm in which images depicting items corresponding to irregular plurals (e.g., a group of mice) were displayed, followed by either a spoken lexically specific frame (e.g., *three blind* __) or non-specific frame (e.g., *so many* __), which the child was then asked to complete (e.g., by saying *mice*, or *mouses* in the case of an over-regularization error). A third condition involved a general question (*What are all these?*). Children produced irregular plurals more accurately, and with fewer over-regularization errors, when prompted with lexically specific frames, though the *so many* frame did provide some advantage over the general question. The facilitatory role of lexically specific frames demonstrated by this study implies not only that children store multiword units, consistent with the findings of Bannard and Matthews (2008), but also that such units may play an active role in morphological development.

As this study demonstrates child sensitivity to frame+plural chunks of much lower frequency than the sequences used by Bannard and Matthews (2008), with several of the noun plurals being relatively infrequent in child-directed speech to begin with (such as *mice*, which occurs only 8 times in the largest single corpus of American

English in CHILDES), the ability to fit the child data from this study stands as a strong challenge for a computational account of multiword unit discovery. A current limitation of CBL is that it cannot over-regularize independently of the utterances it attempts to produce (i.e., during production, the model is simply faced with the task of retrieving and sequencing chunks from a random collection of words corresponding to the words in the child's utterances, only some of which include over-regularized plurals). We were nevertheless able to model Arnon and Clark's results by looking at the pattern of chunk-to-chunk BTPs linking together the stimuli used in the study, using the exact same method as employed in Simulation 2.

Method. To model the Arnon and Clark results, we first constructed an aggregated corpus from the entire US English portion of CHILDES (we focused on American English because the original study was conducted in the US). The aggregated US corpus was used instead of a single corpus because of the infrequency of irregular noun plurals. The aggregated corpus was constructed by interweaving the individual recording files chronologically by the age of the target child at the start of each individual recording session, with the aim of approximating a naturalistic developmental trajectory. Files featuring multiple target children of different ages were excluded (to preserve a realistic developmental trajectory). The resulting aggregated corpus was stripped of tags and punctuation, leaving only the original sequence of words in each utterance (cf. McCauley & Christiansen, 2011). Proper names (including the names of individual target children) were preserved.

We then exposed CBL to the aggregated corpus, stopping at a point that met the corpus target child age corresponding to the mean subject age in the original study (4;6). To simulate the test, we treated each frame+plural combination as a sequence (e.g., *brush your teeth* in the case of a lexically-specific frame sequence, and *so many teeth* in the case of the corresponding general plural frame sequence) and examined its representation in the model's chunkatory. As the target sequences consisted of 3 words, we focused on the chunk-to-chunk statistics which would be used by the model to construct each sequence during production, thereby offering a simulation of children's production attempts which relied on a probabilistic rather than all-or-nothing measure. An item's score was calculated as the product of the BTPs linking each chunk in the sequence (in the case of an item represented as two separate chunks, the score for the test item was calculated as the chunk-to-chunk BTP between the two chunks). If a sequence happened to be stored as a 3-word chunk in the chunkatory, the model received a score of 1, indicating a BTP of 1 (as no chunk-to-chunk probability calculation was necessary). In order to simulate the general question trials, which featured no frame, we simply normalized the target irregular plural's count in the chunkatory by the total number of chunk tokens represented by the model (this would correspond to the model's likelihood of selecting the target irregular in the absence of distributional/frame or semantic information).

Results and discussion. The children in the Arnon and Clark study attained accuracy rates of 72% for the lexically-specific frame condition, 53% for the *so many* frame condition, and 32% for the general question condition (all differences significant, with accuracy defined as the proportion of trials in which irregular plurals were named correctly). The mean CBL BTP scores for the items in each condition are shown in Figure 26. Because *so many* did not immediately precede several of the plurals as a chunk, the path from *so* to *many* to the irregular plural was necessarily relied upon in certain instances (thus, the mean BTP for the *so many* condition was quite low). For this reason, log BTP scores are given in Figure 27; in order to depict the results in an intuitive format, we divided -1 by the mean log BTP for each condition.[15]

[15] As the materials across conditions in Arnon & Clark (2012) were not controlled for substring frequency, we carried out a series of bigram analyses to ensure that a comparable effect could not be gained with simple word-to-word transitional probabilities. As the stimuli in Bannard & Matthews (2008) were controlled for substring frequencies, we did not perform a bigram analysis of the materials used in Simulation 1.

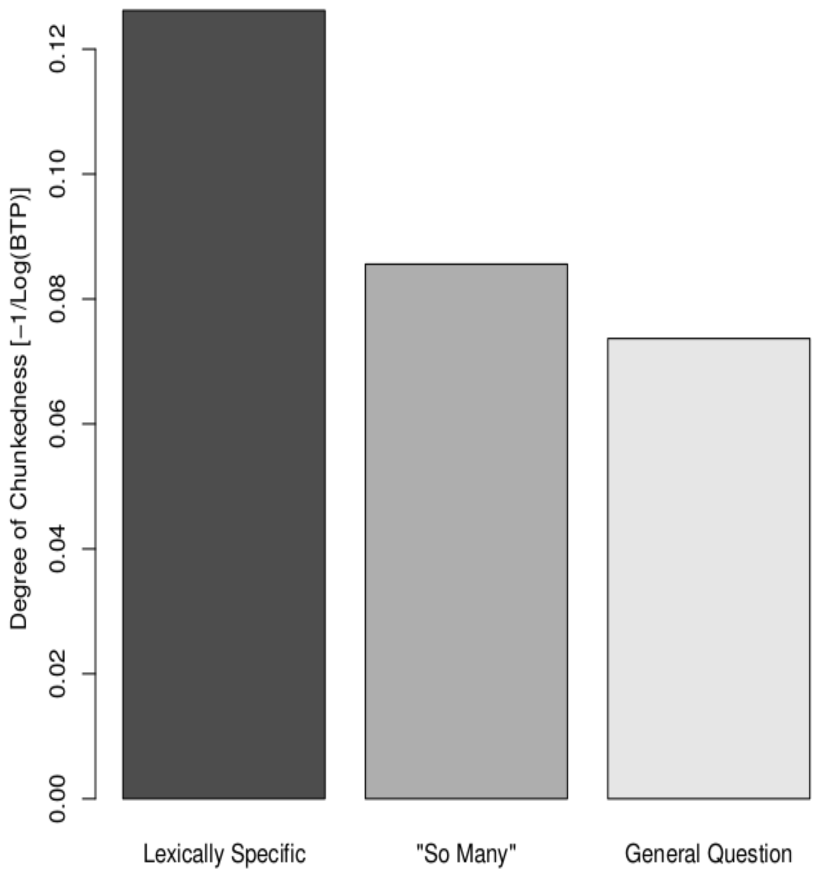

Fig. 27: Mean degree of chunkedness (-1/mean log(BTP)) for the each of the three conditions in Simulation 3.

As can be seen in Figure 27, the model was able to capture the facilitatory effect of lexically-specific frames on irregular production through its chunking of frame+plural sequences, despite the relatively low occurrence of such sequences in the corpus. Similarly to our simulation of the Bannard and Matthews (2008) study, this implies that the overall *cohesiveness* of a sequence is no less important than frequency when it comes to chunk discovery. In other words, whether something is chunked together with the material preceding it depends as much on how likely the preceding material is, given the item being considered (as predicted by CBL's reliance on BTPs), as on how strongly it is predicted by the preceding material (as would be predicted by a reliance on FTPs). It is important to note that transitional probabilities offer a measure of how likely a sequence is *given the frequency of its component parts.* Thus, CBL does not rely on raw frequency of co-occurrence. This can be related to the learning-theoretic notion of *background rate*: the more frequently an event occurs, the less informative it is about the events it sometimes co-occurs with (for a thorough discussion of background rates in language acquisition, see Ramscar, Dye, & McCauley, 2013).

Despite this promising result, the model nevertheless exaggerated the difference between the lexically-specific and general frame conditions. As noted above, this stems from the lack of a path from the *so many* chunk to specific irregulars throughout the corpus (mostly due to a lack of the relevant trigram's appearance in the aggregated corpus to begin with), forcing the model to rely, instead, on a path across three single-word chunks in such instances. Enhanced child performance in the *so many* frame condition may have more to do with semantic than purely distributional information at the word level. For instance, retrieval of the correct irregular form may be easier in a context that clearly requires a plural (cf. Ramscar et al., 2013; Ramscar & Yarlett, 2007). Nevertheless, distributional information of the sort learned by the model clearly played a role in children's performance. For instance, in the original study, there was no facilitatory effect of the *so many* frame for only one of the plurals (*geese*); in a similar fashion, *so many geese* was the only sequence for which no path through the chunkatory was found in the simulation (both findings clearly reflect the relatively low frequency of *geese* in child-directed speech and, therefore, the corpora available in CHILDES).

Further modeling work incorporating information about the *nonlinguistic* cues available to children (in the case of the above study, the images depicting irregular plurals) and semantic information (such as the notion of plurality), in concert with the types of linguistic distributional information used by CBL, will be necessary in order to fully capture children's performance on this study (for a discussion of the prospects and challenges of incorporating semantic information into models of language development, see McCauley & Christiansen, 2014).

Conclusion

Our previous simulations have shown that the CBL model can account for a considerable part of children's early linguistic behavior by using multiword units as building blocks for learning to comprehend and produce sentences (McCauley and Christiansen, 2011; McCauley et al., 2015). In the present paper, we have demonstrated the psychological validity of these building blocks, showing that CBL can also capture key psycholinguistic findings on children's discovery of multiword units in the speech stream as well as their use in language processing (Arnon & Clark, 2011; Bannard & Matthews, 2008). Importantly, CBL functions by computing simple statistics—to which infants and adults have been shown to be sensitive—in a purely incremental, on-line fashion. That so much of children's distributional learning can be accounted for by such a simple architecture is encouraging for the prospect of developing more comprehensive computational accounts of formulaic language learning and use, as well as of language development and sentence processing more generally.

The developmental findings we model here are mirrored in a number of recent studies on adult comprehension (e.g., Arnon & Snider, 2010), production (e.g., Janssen & Barber, 2012), and artificial language learning (e.g., Arnon & Ramscar, 2012). This points to an intriguing hypothesis: the pervasive role of formulaic language in adult processing may reflect the prior importance of multiword sequences for language acquisition—especially when considering the difficulties adult second language learners experience with formulaic language (e.g., Wray, 2002). That is, multiword units may be so crucial for acquisition that they become key building blocks of the emerging language system (see also Arnon & Christiansen, in preparation). This idea contrasts with traditional approaches to language, which incorporate sharp distinctions between lexicon and grammar (e.g., Chomsky, 1957), but fits quite naturally with theoretical frameworks emerging from cognitive linguistics, such as cognitive grammar (e.g., Langacker, 1987) and construction grammar (e.g., Croft, 2001), which eschew the distinction between lexicon and grammar. The parallels between psycholinguistic findings on children and adults' multiword linguistic units suggests that we can reach a fuller understanding of formulaic language by adopting a developmental perspective. In the present paper, we have sought to provide the initial steps towards a developmental approach to studying adults' formulaic language use, one which has its basis in explicit computational mechanisms that are psychologically plausible and can account for developmental psycholinguistic data.

REFERENCES

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191-238.

Arnon, I. & Christiansen, M.H. (in preparation). *Building blocks of language learning*.

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., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*, 349-371.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292-305.

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

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

Borovsky, A., Elman, J.L. & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*, 417-436.

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

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

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

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

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

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Langacker, R. (1987). *The foundations of cognitive grammar: Theoretical prerequisites (Vol. 1)*. Palo Alto: Stanford University Press.

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

Manning, C. & Scütze, H. (1999). *Foundations of Statistical Natural Language*

Processing. Cambridge, MA: 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.

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

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. (2013). Toward a unified account of comprehension and production in language development. *Behavioral and Brain Sciences*, *36*, 366-367 (commentary on Pickering & Garrod).

McCauley, S. M. & Christiansen, M. H. (2014). Prospects for usage-based computational models of grammatical development: Argument structure and semantic roles. *Wiley Interdisciplinary Reviews: Cognitive Science, 5,* 489-499.

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.* Hoboken, NJ: Wiley-Blackwell.

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., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*, 1299-1305.

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).

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329-347.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Expectation and error distribution in learning: The curious absence of "mouses" in adult speech. *Language, 89*, 760-793.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acvcquisition. *Cognitive Science*, *31*, 927–960.

Reali, F. & Christiansen, M. H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60,* 161-170.

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

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

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.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.) *Perspectives on formulaic language: Acquisition and communication* (pp. 151-173). London: Continuum International Publishing Group.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*, 569-613.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching, 32,* 213-231.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

CHAPTER 4

COMPUTATIONAL INVESTIGATIONS OF MULTIWORD CHUNKS IN LANGUAGE LEARNING[16]

Despite clear advantages over children in a wide variety of cognitive domains, adult language learners rarely attain native proficiency in pronunciation (e.g., Moyer, 1999), morphological and syntactic processing (e.g., Felser & Clahsen, 2009; Johnson & Newport, 1989), or the use of formulaic expressions (e.g., Wray, 1999). Even highly proficient second-language users appear to struggle with basic grammatical relations, such as the use of articles, classifiers, and grammatical gender (DeKeyser, 2005; Johnson & Newport, 1989; Liu & Gleason, 2002), including L2 speakers who are classified as near-native (Birdsong, 1992).

Previous approaches to explaining the differences between first-language (L1) and second-language (L2) learning have often focused on neural and cognitive differences between adults and children. Changes in neural plasticity (e.g., Kuhl, 2000; Neville & Bavelier, 2001) and the effects of neural commitment on subsequent learning (e.g., Werker & Tees, 1984) have been argued to hinder L2 learning, while limitations on children's memory and cognitive control have been argued to help guide the trajectory of L1 learning (Newport, 1990; Ramscar & Gitcho, 2007).

While these approaches may help to explain the different outcomes of L1 and L2 learning, we explore an additional possible contributing factor: that children and

[16] An abridged version of this paper has been resubmitted after an initial round of reviews: McCauley & Christiansen (under review). Computational investigations of multiword chunks in language learning.

adults differ with respect to the concrete linguistic units, or *building blocks*, used in language learning. Specifically, we seek to evaluate whether L2-learning adults may rely less heavily on stored multiword sequences than L1-learning children, following the “starting big” hypothesis of Arnon (2010; see also Arnon & Christiansen, this issue), which states that multiword units play a lesser role in L2, creating difficulties for mastering certain grammatical relations. Driving this perspective on L2 learning are usage-based approaches to language development (e.g., Lieven, Pine, & Baldwin, 1997; Tomasello, 2003), which build upon earlier lexically-oriented theories of grammatical development (e.g., Braine, 1976) and are largely consistent with previous theoretical alternatives to dual-system models proposed by linguists (e.g., Langacker, 1987). Within usage-based approaches to language acquisition, linguistic productivity is taken to emerge gradually as a process of storing and abstracting over multiword sequences (e.g., Tomasello, 2003; Goldberg, 2006). Such perspectives enjoy mounting empirical support from psycholinguistic evidence that both children (e.g., Arnon & Clark, 2011; Bannard & Matthews, 2008) and adults (e.g., Arnon & Snider, 2010; Jolsvai, McCauley, & Christiansen, 2013) in some way store multiword sequences and use them during comprehension and production. Computational modeling has served to bolster this perspective, demonstrating that knowledge of multiword sequences can account for children's on-line comprehension and production (e.g., McCauley & Christiansen, 2011, 2014, 2016), as well as give rise to abstract grammatical knowledge (e.g., Solan, Horn, Ruppin, & Edelman, 2005).

In the present paper, we compare L1 and L2 learners’ discovery and use of multiword sequences using large-scale, corpus-based modeling. We do this by using a model of on-line language learning in which multiword sequences play a key role: the Chunk-Based Learner model (CBL; McCauley & Christiansen, 2011, 2014, 2016). Our approach can be viewed as a computational model-based variation on the “Traceback Method” of Lieven, Behrens, Speares, and Tomasello (2003). Using matched corpora of L1 and L2 learner speech as input to the CBL model, we compare the model's ability to discover multiword chunks from the utterances of each learner type, as well as its ability to use these chunks to generalize to the on-line production of unseen utterances from the same learners. This modeling effort thus aims to provide the kind of “rigorous computational evaluation” of the Traceback Method called for by Kol, Nir, & Wintner (2014).

In what follows, we first introduce the CBL model, including its the key computational and psychological features. We then report results from two sets of computational simulations using CBL. The first set applies the model to matched sets of L1 and L2 learner corpora in an attempt to gain insight into the question of whether there exist important differences between learner types in the role played by multiword units in learning and processing. Following these initial simulations, we apply network-theoretic analyses to the resulting multiword chunk inventories in order to better understand the extent to which they might support processes not yet captured by CBL, such as abstraction to item-based schemas (e.g., Tomasello, 2003) or constructions (e.g., Goldberg, 2006). In the second set of simulations, we use a slightly modified version of the model, which learns from raw frequency of occurrence rather than transition probabilities, in order to test a hypothesis based on a previous finding (Ellis, Simpson-Vlach, & Maynard, 2008) suggesting that while L2 learners may employ multiword units, they rely more on sequence frequency as opposed to sequence coherence (as captured by mutual-information, transition probabilities, etc.). We conclude by considering the broader implications of our simulation results.

The Chunk-based Learner Model

The CBL model is designed to reflect constraints deriving from the real-time nature of language learning (cf. Christiansen & Chater, 2016). Firstly, processing is incremental and online. In the model, all processing takes place item-by-item, as each new word is encountered, consistent with the incremental nature of human sentence processing (e.g., Altmann & Steedman, 1988). At any given time-point, the model can rely only upon what has been learned from the input encountered thus far. This stands in stark contrast to models which involve batch learning, or which function by extracting regularities from veridical representations of multiple utterances. Importantly, these constraints apply to the model during both comprehension-related and production-related processing.

Secondly, CBL employs psychologically-inspired learning mechanisms and knowledge representation: the model's primary learning mechanism is tied to simple frequency-based statistics, in the form of backwards transitional probabilities (BTPs)[17], which both infants (Pelucchi, Hay, & Saffran, 2009) and adults (Perruchet & Desaulty, 2008) have been shown to be sensitive to (cf. McCauley & Christiansen, 2011 for more about this choice of statistic, and for why the model represents a

[17] We compute backward transition probability as $P(X|Y) = F(XY) / F(Y)$, where $F(XY)$ is the frequency of an entire sequence and $F(Y)$ is the frequency of the most recently encountered item in that sequence.

departure from standard *n*-gram approaches, despite the use of transitional probabilities). Using this simple source of statistical information, the model learns purely local linguistic information rather than storing or learning from entire utterances, consistent with evidence suggesting a primary role for local information in human sentence processing (e.g., Ferreira & Patson, 2007). Following evidence for the unified nature of comprehension and production processes (e.g., Pickering & Garrod, 2013), comprehension- and production-related processes rely on the same statistics and linguistic knowledge.

Thirdly, CBL implements usage-based learning. All learning arises from individual usage events in the form of attempts to perform comprehension- and production-related processes over utterances. In other words, language learning is characterized as a problem of learning to process, and involves no separate element of grammar induction.

Finally, CBL is exposed to naturalistic linguistic input. It is trained and evaluated using corpora of real learner and learner-directed speech taken from public databases.

CBL Model Architecture

The CBL model has been described thoroughly as part of previous work (e.g., McCauley & Christiansen, 2011, 2016). Here, we offer an account of its inner workings sufficient to understand and evaluate the simulations reported below. While comprehension and production represent two sides of the same coin in the model, as noted above, we describe the relevant processes and tasks separately, for the sake of simplicity.

Comprehension. The model processes utterances on-line, word by word as they are encountered. At each time step, the model is exposed to a new word. For each new word and word-pair (bigram) encountered, the model updates low-level distributional information on-line (incrementing the frequency of each word or word-pair by 1). This frequency information is then used on-line to calculate the BTP between words. CBL also maintains a running average BTP reflecting the history of encountered word pairs, which serves as a “threshold” for inserting chunk boundaries. When the BTP between words rises above this running average, CBL groups the words together such that they will form part (or all) of a multiword chunk. If the BTP between two words falls below this threshold, a “boundary” is created and the word(s) to the left are stored as a chunk in the model's chunk inventory. The chunk inventory also maintains frequency information for the chunks themselves (i.e., each time a chunk is computed, its count in the chunk inventory is incremented by 1, provided it already exists; otherwise, it is initialized with a count of 1).

Once the model has discovered at least one chunk, it begins to actively rely upon the chunk inventory while processing the input in the same incremental, on-line fashion as before. The model continues calculating BTPs while learning the same frequency information, but uses the chunk inventory to make on-line predictions about which words should form a chunk, based on existing chunks in the inventory. When a word pair is processed, any matching sub-sequences in the inventory's existing chunks are activated: if more than one instance is activated (either an entire chunk, or part of a larger one), the words are automatically grouped together (even if the BTP connecting them falls below the running-average threshold) and the model begins to process the next word. Thus, knowledge of multiple chunks can be combined to discover further chunks, in a fully incremental and on-line manner. If less than two chunks in the chunk inventory are active, however, the BTP is still compared to the running average threshold, with the same consequences as before. Importantly, there are no *a priori* limits on the size of the chunks that can be learned by the model.

Production. While the model is exposed to a corpus incrementally, processing the utterances on-line and discovering/strengthening chunks in the service of comprehension, it encounters utterances produced by the target child of the corpus (or, in the present study, target *learner*, which is not necessarily a child) – this is when the production side of the model comes into play. Specifically, we assess the model's ability to produce an identical utterance to that of the target learner, using only the chunks and statistics learned up to that point in the corpus. We evaluate this ability using a modified version of the *bag-of-words incremental generation task* proposed by Chang, Lieven, and Tomasello (2008), which offers a method for automatically evaluating a syntactic learner on a corpus in any language.

As a very rough approximation of sequencing in language production, we assume that the overall message the learner wishes to convey can be modeled as an unordered bag-of-words, which would correspond to some form of conceptual representation. The model's task, then, is to produce these words, incrementally, in the correct sequence, as originally produced by the learner. Following evidence for the role of multiword sequences in child production (e.g., Bannard & Matthews, 2008), and usage-based approaches more generally, the model utilizes its chunk inventory during this production process. The bag-of-words is thus filled by modeling the retrieval of stored chunks by comparing the learner's utterance against the chunk inventory, favoring the longest string which already exists as a chunk for the model, starting from the beginning of the utterance. If no matches are found, the isolated word at the beginning of the utterance (or remaining utterance) is removed and placed into the bag. This process continues until the original utterance has been completely randomized as chunks/words in the bag.

During the sequencing phase of production, the model attempts to reproduce the learner's actual utterance using this unordered bag-of-words. This is captured as an incremental, chunk-to-chunk process, reflecting the incremental nature of sentence processing (e.g., Altmann & Steedman, 1988; see Christiansen & Chater, 2016, for discussion). To begin, the model removes from the bag-of-words the chunk with the highest BTP given a start-of-utterance marker (a simple hash symbol, marking the beginning of each new utterance in the prepared corpus). At each subsequent time-step, the model selects from the bag the chunk with the highest BTP given the most recently placed chunk. This process continues until the bag is empty, at which point the model's utterance is compared to the original utterance of the target child.

We use a conservative measure of sentence production performance: the model's utterance must be identical to that of the target child, regardless of grammaticality. Thus, all production attempts are scored as either a 1 or a 0, allowing us to calculate the percentage of correctly-produced utterances as an overall measure of production performance.

SIMULATION 1: MODELING THE ROLE OF MULTIWORD CHUNKS IN L1 VS. L2 LEARNING

In Simulation 1, we assess the extent to which CBL, after processing the speech of a given learner type, can "generalize" to the production of unseen utterances. Importantly, we do not use CBL to simulate language development, as in previous studies, but instead as a psychologically-motivated approach to extracting multi-word units from learner speech. The aim is to evaluate the extent to which the sequencing of such units can account for unseen utterances from the *same speaker*, akin to the Traceback Method of Lieven et al. (2003).

To achieve this, we use a leave-10%-out method, whereby we test the model's ability to produce a randomly-selected set of utterances using chunk-based knowledge and statistics learned from the remainder of the corpus. We compare the outcome of simulations performed using L2 learner speech (L2 → L2) to two types of L1 simulation: production of child utterances based on learning from that child's own speech (C → C) and production of adult caretaker utterances based on learning from the adult caretaker's own speech (A → A). The C → C simulations provide a comparison to early learning in L1 vs. L2 (as captured in the L2 → L2 simulations), while the A → A simulations provide a comparison of adult L1 language to adult speech in an early L2 setting. The third type of L1 simulation is included as a control, allowing comparison to model performance in a more typical context: production of child utterances after learning from adult caretaker speech (A → C). Crucially, the L2 → L2, C → C, and A → A simulations provide an opportunity to gauge how well chunk-based units derived from a particular speaker's corpus generalize to unseen utterances from the *same speaker*, while the A → C simulations provide a comparison to a more standard simulation of language development.

If L2 learners do rely less heavily on multi-word units, as predicted, we would expect for the chunks and statistics extracted from the speech of L2 learners to be less useful in predicting unseen utterances than for L1 learners, even after controlling for factors tied to vocabulary and linguistic complexity.

Methods

Corpora: For the present simulations, we rely on a subset of the European Science Foundation (ESF) Second Language Database (Feldweg, 1991), which features transcribed recordings of L2 learners over a period of 30 months following their arrival in a new language environment. We employ this particular corpus because its non-classroom setting allows better comparison with child learners. The data was transcribed for the L2 learners in interaction with native-speaker conversation partners while engaging in such activities as free conversation, role play, picture description, and accompanied outings. Thus, the situational context of the recorded speech often mirrors the child-caretaker interactions found in corpora of child-directed speech.

For child and L1 data, we rely on the CHILDES database (MacWhinney, 2000). We selected the two languages most heavily represented in CHILDES (German and English), which allowed for comparison with L2 learners of these languages (from the ESF corpus), while holding the native language of the L2 learners constant (Italian). We then used an automated procedure to select, from the large number of available CHILDES material, corpora which best matched each of the available L2 learner corpora in terms of size (when comparing learner utterances) for a given language. Thus, we matched one L1 learner corpus to each L2 learner corpus in our ESF subset. The final set of L2 corpora included: Andrea, Lavinia, Santo, and Vito (Italians learning English); Angelina, Marcello, and Tino (Italians learning German). The final set of matched CHILDES corpora included: Conor and Michelle (English, Belfast corpus); Emma (English, Weist corpus); Emily (English, Nelson corpus); Laura, Leo, and Nancy (German; Szagun corpus). Because utterance length is an important factor, we ran tests to confirm that neither the L1 child utterances [t(6) = -1.3, p = 0.24] nor the L1 caretaker utterances [t(6) = 0.82, p = 0.45] differed significantly from the L2 learner utterances in terms of number of words per utterance. Thus, we included all utterances from the original corpora, which remained unfiltered with respect to utterance length.

While limitations on the number of available corpora made it impossible to match the corpora along every relevant linguistic dimension, we controlled for these factors in our statistical analyses of the simulation results. In particular, we were interested in controlling for linguistic complexity and vocabulary range: as a proxy for linguistic complexity, we used mean number of morphemes per utterance (MLU), while type-token ratio (TTR) served as a measure of vocabulary range, as the corpora were matched for size. Details for each corpus and speaker are presented in Table 11.

Table 11: Details of Corpora and Speaker Types

Corpus	Speaker Type	Language	MLU	TTR
Conor	Child	English	3.10	0.10
Emily	Child	English	4.13	0.09
Emma	Child	English	3.32	0.07
Michelle	Child	English	4.60	0.09
Laura	Child	German	2.51	0.13
Leo	Child	German	2.41	0.10
Nancy	Child	German	1.92	0.08
Conor	Adult	English	5.26	0.05
Emily	Adult	English	7.20	0.10
Emma	Adult	English	3.67	0.06
Michelle	Adult	English	6.07	0.05
Laura	Adult	German	3.90	0.09
Leo	Adult	German	3.78	0.10
Nancy	Adult	German	3.33	0.09
Andrea	L2	English	3.98	0.10
Lavinia	L2	English	5.32	0.08
Santo	L2	English	4.60	0.10
Vito	L2	English	3.05	0.12
Angelina	L2	German	3.70	0.13

Marcello	L2	German	4.72	0.15
Tino	L2	German	5.10	0.12

Each corpus was submitted to an automated procedure whereby tags and punctuation were stripped away, leaving only the speaker identifier and original sequence of words for each utterance. Importantly, words tagged as being spoken by L2 learners in their native language (Italian in all cases) were also removed by this automated procedure. Long pauses within utterances were treated as utterance boundaries.

Simulations: For each simulation, we ran ten separate versions, each using a different randomly-selected test group consisting of 10% of the available utterances. In each case, the model must attempt to produce the randomly withheld 10% of utterances after processing the remaining 90%. For each L1-L2 pair of corpora, we conduct four separate simulation sets: one in which the model is exposed to the speech of the L2 learner and must subsequently attempt to produce the withheld subset of 10% of the L2 learner utterances (L2 → L2), and three simulations involving the L1 corpus (one in which the model is tasked with producing the left-out 10% of the child utterances after exposure to the other child utterances [C → C], one in which the model must attempt to produce L1 caretaker utterances after exposure to the other L1 adult/caretaker utterances [A → A], and one in which the model must attempt to produce a random 10% of the child utterances after exposure to the adult/caretaker utterances [A → C]). Thus, we seek to determine how well a chunk inventory built on the basis of a learner's speech (or input) helps the model generalize to a group of unseen utterance types.

Results and Discussion

As can be seen in Figure 28, the model achieved stronger mean sentence production performance for all three sets of L1 simulations than for the L2 simulations (Child → Child: 49.6%, SE: 0.8%; Adult → Child: 47.5%, SE: 0.9%; Adult → Adult: 42.1, SE: 0.7%; L2 → L2: 36.3%, SE: 0.6%). To examine more closely the differences between the speaker types across simulations while controlling for linguistic complexity and vocabulary breadth, we submitted these results to a linear regression model with the following predictors: Learner Type (L1 Adult vs. L1 Child vs. L2 Adult, with L1 Adult as the base case), MLU, and TTR. The model yielded a significant main effect of L2 Adult Type [$B=-5.67$, $t=-1.98$, $p<0.05$], reflecting a significant difference between the L2 performance scores and the base case (L1 Adult). The Child L1 Type did not differ significantly from the Adult L1 Type. While there was a marginal effect of TTR [$B=-0.7$, $t=-1.7$, $p=0.08$], none of the other variables or interactions reached significance. The model had an adjusted R-squared value of 0.83.

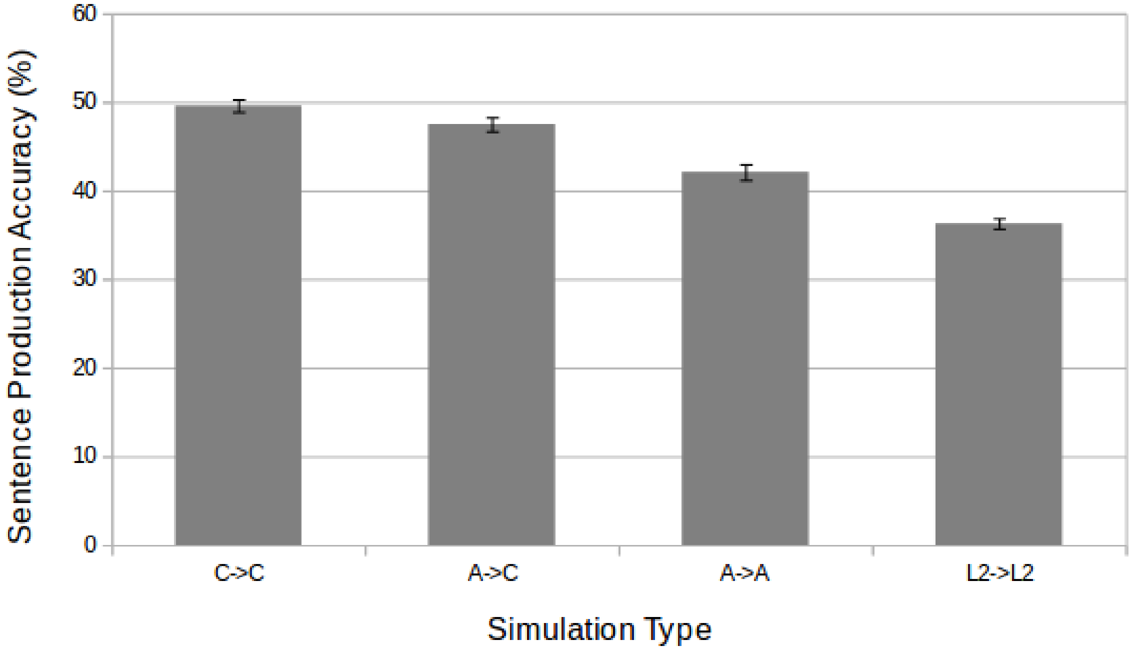

Fig. 28: Graph depicting the mean sentence production accuracy scores on the leave-10%-out task for each of the four simulation types.

Thus, CBL's ability to generalize to the production of unseen utterances was significantly greater for L1 children and adults, relative to L2 learners. This suggests that the type of chunking performed by the model may better reflect the patterns of L1 speech than those of L2 speech. This notion is consistent with previous hypotheses suggesting that adults rely less heavily than children on multiword chunks in learning, and that this can negatively impact mastery over certain aspects of grammar (see Arnon & Christiansen, this issue, for discussion). It also fits quite naturally alongside findings of differences in L2 learner use of formulaic language and idioms (e.g., Wray, 1999).

In addition, CBL exhibited no significant difference in its ability to capture L1 adult vs. child speech, once linguistic factors tied to MLU and TTR were controlled for. This is consistent with previous work using the CBL model, which suggests that multiword chunks discovered during early language development do not diminish, but may actually grow in importance over time (McCauley & Christiansen, 2014), reflecting recent psycholinguistic evidence for the use of multiword chunks in adults (e.g., Arnon & Snider, 2010; Jolsvai et al., 2013).

It is important to reiterate that the aims of Simulation 1 are to compare the extent to which multiword units extracted from the speech of L1 vs. L2 learners can generalize to unseen utterances from the same speakers; though CBL could theoretically be used to do so, the present simulations are not intended to provide an account of L2 acquisition. For such an endeavor, it would be necessary to account for a variety of factors, such as the influence of pre-existing linguistic knowledge from a learner's L1 (cf. Arnon, 2010; Arnon & Christiansen, this issue) and the role of overall L2 exposure (e.g., Matusevych, Alishahi, & Backus, 2015).

In what follows, we use network analyses to explore the actual structure of the chunk inventories discovered by the model during these simulations, following the notion that greater overlap between chunks can better support the gradual process of abstraction from which linguistic productivity is taken to emerge in usage-based approaches (e.g., Tomasello, 2003; Goldberg, 2006; Solan et al., 2005). If L2 learners' knowledge of multiword chunks—as estimated by our simulations—differs structurally from that of L1 learners, this could be an important factor in some of the qualitative and quantitative differences observed for actual L2 vs. L1 learning outcomes.

NETWORK ANALYSIS OF MODEL CHUNK INVENTORIES

In usage-based approaches to language, knowledge of multiword sequences is taken to be an important starting point for the emergence of linguistic productivity (e.g., Arnon, 2010; Tomasello, 2003). While the version of CBL used in the present study does not capture processes of abstraction, being limited solely to the discovery and use of *concrete units*, the structure of this knowledge provides a window onto the potential utility of such in scaffolding the acquisition of more abstract and productive units, such as grammatical constructions (e.g., Goldberg, 2006).

Given that the underlying motivation for this study was to determine whether corpus-based modeling can provide clues about the role of multiword units in L2 vs. L1 speech, and given that differences in such a role have been hypothesized to contribute to diverging learning outcomes observed for L2 vs. L1 speakers (e.g., Arnon, 2010; Arnon & Christiansen, this volume), the present section seeks to further explore the structure of the chunk inventories extracted from the speech of each learner type in Simulation 1.

In accordance with usage-based approaches to language development, we seek to determine whether the L2 learners' chunk inventories—as estimated by our simulations—differ in structural properties such as connectivity and local structure, which could support early steps towards abstract knowledge, such as the development of lexical frames or item-based schemas (e.g., Tomasello, 2003; Theakston & Lieven, this issue). Thus, we are primarily focused on *overlap* between chunks, from which a considerable amount of productivity may emerge.

As an example of such a process, consider the learning of a schema with a single slot: *I want to* [process] *it.* Through storage of overlapping multi-word sequences matching this pattern (e.g., *I want to get it; I want to change it; I want to smash it*) the *[process]* slot may begin to emerge, while at the same time reinforcing a semantic relationship between *process* verbs (for a computational approach to abstract grammar emerging from processes of "alignment and comparison," see Solan et al., 2005). Once the learner has established the *I want to* [process] *it* schema, novel utterances can be produced through the insertion of a *process* verb.

In what follows, we analyze the chunk inventories of each learner from Simulation 1 by conceiving of the inventories as networks of partially overlapping chunks. We analyze the inventories of each learner type according to three graph-theoretic measures of network structure, which are seen to measure the connectivity of the chunk inventory, and therefore the extent to which it may support—or reflect—productivity in the form of item-based schemas and constructions. If the types of multi-word units evident in L2 learner speech exhibit less overlap, this may be an important source of differences in learning outcomes.

Methods

We built 21 different networks: one for each L2 learner (L2 → L2), L1 child (C → C), and L1 caretaker (A → A) corpus (7 of each type). Following exposure of the CBL model to a given corpus, we treated each chunk in the model's chunk inventory as a vertex. Edges were drawn between chunks in the case of *multiword overlap* between chunks, with multiword overlap defined as sharing a sub-sequence of at least two words. For example, the chunks *in the box* and *in the bag* would be connected by an edge because of the overlapping subsequence *in the,* while *open the bag* and *empty the bag* would be connected because of the overlapping subsequence *the bag*.

For each of the 21 simulations, we calculated three network statistics which were averaged over all nodes in the network. These were: path length, clustering coefficient, and degree (for an overview of these statistics, see Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). Path length describes the average shortest path length between any two nodes in the network. The clustering coefficient of a node is the overall proportion of connections between its immediate neighbors out of all those possible. Degree simply describes the number of edges connecting a node to other nodes. Together, these network statistics provide a sense of the interconnectedness and structure of a graph: average degree provides a measure of overall connectivity, while path length offers a measure of global access across the network, and clustering coefficient offers a measure of local interconnectedness and structure.

For each network, we constructed 100 undirected Erdös-Rényi (ER) random graphs (Erdös & Rényi, 1960), which were equivalent in the number of nodes and edges, but with connections made at random, thereby removing any structure inherent in the chunk inventories themselves. This provided a baseline against which to compare each network's actual statistics: we rely upon the difference between each average statistic and the average for that statistic across all 100 matching ER graphs as a measure of how much the statistic deviates from what would be expected based merely on network size (note, however, that average degree is necessarily identical for the chunk-based and random networks, since they are matched for number of nodes and edges; for this reason we report the actual number rather than the difference in what follows).

Results and Discussion

The difference between each average network statistic and that of the matched set of ER random graphs is shown in Table 12, averaged across each learner/corpus type. As can be seen, there were remarkable similarities between the L1 child and L2 adult networks, while the L1 adult networks displayed higher degree and clustering coefficient, but with much less marked difference from the ER random graphs in average path length relative to the other two network types.

Table 12:

Mean network statistics for each simulation / learner type

Learner Type	Path Length	Clustering Coefficient	Degree
L1 Adult	-0.62	0.14	7.76
L1 Child	-5.99	0.12	2.83
L2 Adult	-5.75	0.12	2.99

Note: Path Length and Clustering Coefficient reflect differences between the average network statistics and the average for the matched ER random graphs, while degree describes the actual degree.

The overall closeness of the network statistics for the L1 child and L2 adult networks, as well as their differences compared with the L1 adult networks, were statistically reliable: for degree, a one-way ANOVA confirmed significant differences across learner types: [F(1,18) = 9.67, p<0.01], with pairwise comparisons confirming a higher degree for adults relative to child learners [t(12) = 3.28, p < 0.01] and L2 learners [t(12) = 3.29, p < 0.01], but no difference between child and L2 learners [t(12) = 0.22, p = 0.83]. For clustering coefficient, a one-way ANOVA also yielded significant differences across learner types: [F(1,18) = 5.34, p<0.05], with pairwise comparisons confirming a higher clustering coefficient (relative to the paired ER graphs) for adults relative to child learners [t(12) = 2.94, p < 0.05] and L2 learners [t(12) = 3.1, p < 0.01], but no difference between child and L2 learners [t(12) = 0.24, p = 0.81]. For path-length, a one-way ANOVA once-more confirmed significant differences across learner types: [F(1,18) = 9.1, p<0.01], with pairwise comparisons confirming a higher path length (relative to the paired ER graphs) for adults relative to child learners [t(12) = 3.24, p < 0.01] and L2 learners [t(12) = 3.05, p < 0.05], but no difference between child and L2 learners [t(12) = 0.22, p = 0.83].

Thus, there were no significant differences between the L1 child and L2 adult networks in terms of connectivity, global access, or local structure. However, there were marked differences when comparing either network type to those of the L1 adults, which displayed greater connectivity (as reflected by degree) and local structure (reflected by clustering coefficient; adult networks more closely approached a small-world structure), but with less global access (as reflected by path length). The difference in path length may be due in part to the heavier tail of the distribution of words (and therefore chunks) for the L1 adults: chunk types containing infrequent words may be less likely to overlap with other chunks.

We would expect for L1 children's chunk inventories to come to more closely resemble those of the L1 adults as language experience increases; from the perspective of usage-based approaches, in which the importance of multiword linguistic units is paramount, connectivity and local structure would be necessary to support abstraction over chunks. This could support greater linguistic productivity through the formation of lexical frames or schemas (e.g., Tomasello, 2003), an important stepping stone towards more fully abstract grammatical constructions (e.g., Goldberg, 2006). Such network properties may also support the “alignment and comparison” of sequences, which could give rise to a considerable amount of grammatical knowledge (e.g., as implemented in Solan et al., 2005).

However, there may be reason to expect that L2 learner chunk inventories drawn from more experienced L2 speakers would not grow to resemble the L1 adults, as we might expect of older L1 children's networks. While previous work has shown that adults can learn fixed “phrases” in an artificial language (Ellis, 1996), consistent with the present analyses, other evidence suggests that adult learners do not use multiword sequences to support grammatical development to the same extent as children do (e.g., Wray, 1999). Future work of the sort described in the present paper is called for, following access to more longitudinal corpora of L2 learner speech which can be matched to larger corpora in the CHILDES database.

Despite the much greater connectivity and structure of the L1 adult networks, the L1 child and L2 adult networks nevertheless displayed a considerable amount of connectivity, structure, and access. While the lack of a statistically significant difference between the two network types lends support to the notion that the amount of overall language exposure may be a key source of the differences between L1 and L2 learning outcomes (cf. Matusevych et al., 2015), the results of Simulation 1 support the idea that L1 and L2 learners learn different types of chunk-based information or use that information differently. In our simulations, L2 chunk inventories were less useful in generalizing to unseen utterances, despite their similarity to child L1 inventories in terms of structure.

One intriguing possibility is that L2 learners are less sensitive to coherence-related information (such as transition probability, in this instance), and may rely more on raw frequency of exposure. Thus, a model based on raw co-occurrence frequencies may provide a better account of L2 learner chunking than the CBL model. It is to this possibility that we turn our attention in what immediately follows.

SIMULATION 2: EVALUATING THE ROLE OF RAW FREQUENCY VS. COHERENCE

The chunk inventories learned by the CBL model for L2 learners are structurally quite similar to those learned for L1 children. Nevertheless, the results of Simulation 1 strongly suggest that there may be important differences in the utility of these chunks, as well as the extent to which they are relied on by the two types of learner. Here, we turn our attention to exploring a possible difference in the means by which the two learner types arrive at chunk-based linguistic units: in a study conducted by Ellis et al. (2008), L2 learners were shown to rely more heavily on raw sequence frequency, while L1 adult subjects displayed a sensitivity to sequence coherence (as reflected by mutual information). Following this finding, we explore the hypothesis that raw frequency-based chunks may provide a better fit to the speech of L2 learners than those discovered through transition probabilities (as in the CBL model), while yielding the opposite result for L1 child and adult speakers. Thus, the purpose of Simulation 2 is to determine the extent to which the move to a purely raw frequency-based style of chunking affects performance when compared to the transition probability-based chunking of CBL.

To this end, we conduct a second round of production simulations, identical to those of Simulation 1, but with a modified version of the model in which chunks are acquired through the use of raw frequency rather than transitional probabilities. If the findings of Ellis et al. (2008) do indeed correspond to a greater reliance on raw frequency—as opposed to overall sequence coherence—in L2 learners, we would expect that a raw frequency-based version of the model would improve production performance in the L2 simulations while lowering across the L1 simulations.

Method

Corpora: The corpora and corpus preparation procedures were identical to those described for Simulation 1.

Model Architecture: We implemented a version of the model which was identical in all respects, save one: all BTP calculations were replaced by the raw frequency of the sequence in question (i.e., Frequency[XY] as opposed to

Frequency[XY]/Frequency[Y]). Thus, boundaries were inserted between two words when their raw bigram frequency fell below a running average bigram frequency, while they were grouped together as part of a chunk if their raw bigram frequency was above this running average. During production, incremental chunk selection took place according to raw frequency of two chunks' co-occurrence in sequence (as opposed to using the BTPs linking them): at each time step, the chunk in the bag which formed the highest-frequency sequence when combined with the preceding chunk was chosen.

Simulations: Using the modified, raw frequency-based version of the model, we ran a parallel series of simulations (one simulation corresponding to each simulation in Simulation 1). The outcomes of both sets of simulations were then compared to assess whether the switch to raw frequency-based chunking affected the outcomes of L1 and L2 learner simulations differently.

Results and Discussion

The aim of Simulation 2 was to determine how much the switch to raw frequency-based chunking affected performance for a parallel version of each original CBL simulation (as before, 10 simulations for each corpus and simulation type). We compare the two model/simulation sets directly by calculating the difference in performance scores between Simulations 2 and 1. As predicted, this switch tended to improve L2 performance scores while decreasing L1 adult and child scores. While the differences in overall means calculated across learner types were small (L2 → L2:

+1%, A → A: -1%, C → C: -2%), there were considerable individual differences across simulations (Std. Dev of 4%, with change sizes ranging from 0% to 11%). The mean differences across learner types were statistically reliable: a three-way ANCOVA with Learner Type (Adult L1, Child L1, and Adult L2), MLU, and TTR as factors confirmed a significant main effect of Type [$F(1,198)=12.78$, $p<0.001$], with post-hoc tests confirming greater improvement for L2 simulations over Adult L1 [$t(129.6)=2.39$,$p<0.05$] and Child L1 simulations [$t(119.6)=4.39$,$p<0.001$].

Therefore, as hypothesized, the raw frequency-based chunking model was better able to capture the speech of the L2 adult learners, while the transition probability-based chunking of the CBL model provided a better fit to the L1 child and L1 adult learners alike. Why might this be the case? It is clear that both types of information are highly complementary: for instance, McCauley and Christiansen (2014) show that developmental psycholinguistic results which appear to stem from overall sequence frequency (e.g., Bannard & Matthews, 2008) can also be accounted for using transition probability-based chunking of the sort performed by CBL. If amount of exposure to the target language was the primary driving factor, we would expect the L1 child speech to behave more similarly to that of the L2 adults in this context. This supports the notion that L2 learning adults learn from the input differently, in ways that go beyond mere exposure. Pre-existing knowledge of words and word-classes may lead L2 learners to employ different strategies than those used in L1—and while such knowledge is not factored into our simulations explicitly, it is implicitly reflected in the nature of the L2 speech being chunked and sequenced by the model. Nevertheless, the network analyses reported above bear out remarkable similarities in the L1 Child vs. L2 learner chunk inventories, suggesting that knowledge of multiword sequences could still play an important role in the speech of our L2 sample. It may merely be that these sequences are discovered and used in ways that are less closely captured by the CBL model.

GENERAL DISCUSSION

This study represents an initial step towards the use of large-scale, corpus-based computational modeling to uncover similarities and differences in the linguistic building blocks used by L1 and L2 learners. Together, our findings suggest that multiword sequences play a role in L1 and L2 learning alike, but that these units may be arrived at through different means and employed to different extents by each type of learner.

The first set of simulations shows that a chunk-based model of acquisition, CBL (McCauley & Christiansen, 2011, 2014, 2016), better generalizes to the production of unseen utterances when exposed to corpora of children and adults speaking their L1 than when exposed to corpora of L2 learners. Our subsequent network-theoretic analyses of chunk inventories learned by the model for each corpus showed considerably more connectivity and local structure in the networks of L1 adults than those of the L1 children and L2 adults, which were statistically indistinguishable from one another along the considered dimensions. We interpret the greater connectivity and local structure of the adult L1 networks according to usage-based approaches which underscore the importance of stored multiword sequences in abstraction: greater overlap supports the learning of item-based schemas, lexically specific frames, and construction-like units (e.g., Tomasello, 2003; Goldberg, 2006). Because the chunks in these networks were based on the actual speech of each learner type (as opposed to passive input), the higher overlap can also be interpreted as reflecting greater use of schemas or frames in the L1 adults. If we follow previous evidence that L2 learners do not use multiword sequences to support grammatical development to the same extent as children do (e.g., Wray, 1999), we might expect networks built from the speech of more experienced L2 adults to resemble the adult L1 networks to a lesser extent than those of more experienced L1 children.

Finally, we tested the notion, derived from the findings of Ellis et al. (2008), that L2 learners may arrive at knowledge of multiword chunks through different means than L1 learners. The study of Ellis et al. (2008) showed that L2 learners were sensitive to raw sequence frequency but not the overall coherence of a sequence (such as would be reflected by mutual information, transition probabilities, etc.), in contrast to L1 adults. As expected, we found that the switch to a raw frequency-based version of the CBL model lead improved scores on L2 simulations to a statistically significant extent, while models of L1 child and adult speech were significantly stronger when using the transition probabilities (as in the original set of simulations).

Thus, taken together, our findings support the notion that there may be important differences in the building blocks typically involved in L1 vs. L2 learning, and that these differences cannot be explained away merely on the basis of amount of exposure: despite similarities in the structure of the chunk inventories learned by CBL when exposed to L1 child and L2 adult speech, those chunks were more useful for production of the child utterances, with further simulations supporting the notion that multiword units may be arrived at through different means in L1 vs. L2. While these findings can be taken to support the hypothesis that multiword units play a lesser role in L2, creating difficulties for mastering certain grammatical relations (e.g., Arnon, 2010; Arnon & Christiansen, this issue), further work using longitudinal corpora of L2 learner speech will be necessary in order to gain a clearer picture of the development of multiword units in L2. Another potential contributing factor to the differences observed in the present study is that knowledge of semantic and/or syntactic categories tied to words in a learner's L1 may shape the types of units drawn upon in their L2 learning. Implicit attempts to overlay L2 words upon L1 categories and constructions may lead to sensitivity to statics over abstract categories—for instance, statistics computed over word classes (e.g., Thompson & Newport, 2007)—which could account for L2 learners' lesser sensitivity to item-based coherence, as observed by Ellis et al. (2008) and in the present simulations. The present study serves to demonstrate the promise of large-scale, corpus-based modeling for exploring these and other questions related to the differences between L1 and L2 learning.

APPENDIX

Table 13

Top Ten Most Frequent Chunks for English Learners

LEARNER	TYPE	TOP TEN CHUNKS
Emma	**Child L1**	**this is; you can; I think; I want; and then; I don’t; in there; how about; I’m gonna; you tell**

Conor	**Child L1**	**I’ve got; that one; I have; in there; Jurassic Park; this is; look at; I don’t; you see; a big**
Michelle	**Child L1**	**I don’t know; that one; I don’t; I have; and then; this one; in there; my mummie; this here; my bed room**
Emily	**Child L1**	**and then; I want; I need; go to; I think; I don’t know; I don’t; my daddy; my back; my bed**
Andrea	**Italian Speaker L2**	**I don’t know; I think; there are; I have; the door; in Italy; we have; the bag; in front; the table**
Lavinia	**Italian Speaker L2**	**you know; I don’t know; I think; I don’t; my husband; I have; this one; you have; to find; to go**
Santo	**Italian Speaker L2**	**I think; in Italy; for me; this is; I don’t know; I see; I mean; I don’t; I go; you know what**
Vito	**Italian Speaker L2**	**I dunno; I don’t know; I think; the house; too much; come back; in Italy; in the; the left; this girl**

REFERENCES

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191-238.

Arnon, I. (2010). Starting Big: The role of multiword phrases in language learning and use. Doctoral dissertation, Stanford University.

Arnon, I. & Clark, E. V. (2011). When '*on your feet*' is better than '*feet*': Children's word production is facilitated in familiar sentence-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., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science, 19,* 241-248.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N. & Christiansen, M.H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences, 17,* 348-360.

Birdsong, D. (1992). Ultimate attainment in second language acquisition, *Language*, 68, 706-755.

Braine, M. D. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development, 41,* 104.

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

Christiansen, M.H. & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences*.

DeKeyser, R. M. (2005). What makes learning second language grammar difficult? A review of issues. *Language Learning*, *55*, 1-25.

Ellis, N. C. (1996). Sequencing in SLA: Phonological Memory, Chunking and Points of Order. *Studies in Second Language Acquisition*, 18, 91-126.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic Language in Native and Second-Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly* 41, 375-396.

Erdös, P. & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5,* 7–61.

Feldweg, H. (1991). *The European Science Foundation Second Language Database*. Nijmegen: Max-Planck-Institute for Psycholinguistics.

Felser, C. & H. Clahsen (2009). Grammatical processing of spoken language in child and adult language learners. *Journal of Psycholinguistic Research 38*, 305-319

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

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science

Society.

Johnson, J. S. & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60-99.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language*, *41*, 176-199.

Kuhl, P.K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Science*, *97*, 11850-11857.

Langacker, R. (1987). *The Foundations of Cognitive Grammar: Theoretical Prerequisites* (Vol. 1). Palo Alto: Stanford University Press.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333-370.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187-219.

Liu, D. & Gleason, J. L. (2002). Acquisition of the article the by nonnative speakers of English. *Studies in Second Language Acquisition*, *24*, 1-26.

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

Matusevych, Y., Alishahi, A., & Backus, A. (2015). Distributional determinants of learning argument structure constructions in first and second language. In D.C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 1547-1552)*. Austin, TX: Cognitive Science Society.

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*. Austin, TX: Cognitive Science Society.

McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *The Mental Lexicon*, *9*, 419-436.

McCauley, S.M. & Christiansen, M.H. (2016). *Language learning as language use: A cross-linguistic model of child language development*. Manuscript in preparation.

Moyer, Alene (1999). Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition*, 21, 81–108.

Neville, H. J. and Bavelier, D. (2001). Variability of developmental plasticity. In J. McClelland and R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives*. Riverton, NJ: Foris.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11-28.

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

Perruchet, P. & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*, 1299-1305.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329-347.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Science, 11*, 274-279.

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

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

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

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49-63.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, *32*, 213–231.

CHAPTER 5

TESTING PREDICTIONS, ADDRESSING LIMITATIONS, AND CHARTING DIRECTIONS FOR FUTURE WORK

While the computational modeling work presented in this dissertation succeeds along a number of lines, its limitations are readily apparent. First and foremost is the fact that it never moves beyond purely item-based processing to capture the sort of productivity linguistic theories ultimately aim to explain. While Chapters 2 and 4 discuss, at length, the ways in which the chunk-based linguistic knowledge built up by the model can serve as an ideal starting point for abstracting to partially productive frames, schemas, and constructions, the transition to a more comprehensive account of grammatical development is an area of ongoing research and will remain a priority for future work. Computationally more complex models such as the ADIOS model of Solan, Horn, Ruppin, and Edelman (2005) have served to demonstrate the power of item-based learning in abstracting over sequences to arrive at grammatical constructions.

The second greatest limitation of the CBL approach is that it cannot be said to actually model comprehension or production on the level of *meaning*. Rather, it captures specific aspects of comprehension and production, but—absent extra-linguistic information—does not actually form semantic representations during comprehension, or use semantic information in the incremental production of its utterances. This, of course, is a limitation of most usage-based computational accounts of grammatical development, which have primarily focused on what can be learned from distributional information. While the distributional approach has met with considerable success, as outlined in the introduction of Chapter 2, it cannot provide a complete account of children's language use: the child must learn to compute meanings based on previously unencountered utterances, and to generate novel utterances conveying meanings they themselves wish to communicate.

The relative lack of semantic information in computational accounts of grammatical development stems in part from the difficult challenge of simulating naturalistic semantic representations that children may use. Moreover, the disciplinary segregation within developmental psycholinguistics further exacerbates the problem: separate sub-fields have typically focused on largely distinct areas, along traditional boundaries, such as those dividing phonology from word learning, word learning from grammatical development, and grammatical development from semantic development. As a result, much of the computational work on grammatical development has focused on structural considerations.

Because the linguistic theories which form the basis for usage-based approaches to language acquisition hold that grammatical development is fundamentally tied to form-meaning mappings (e.g., e.g., Croft, 2001; Fillmore, 1985; Goldberg, 1995, 2006; Langacker, 1987, 2008), the relative dearth of computational cognitive modeling work on grammatical development to incorporate semantics presents a serious challenge for usage-based approaches to overcome more broadly. Initial progress, however, has been made in two key areas tied to early grammatical development: verb-argument structure and semantic role assignment (e.g., Alishahi & Stevenson, 2010; Chang, 2008; Chang, Dell, & Bock, 2006; Dominey, Hoen, & Inui, 2006; Perfors, Tenenbaum, & Wonnacott, 2010; Shayan, 2008). The success of these models is encouraging with respect to future work in modeling semantics-driven grammatical development more broadly. For a review of computational models of language development which incorporate semantic representations, as well as the prospects and challenges of extending these approaches, see McCauley and Christiansen (2014).

Testing Key Predictions of Chunk-based Learning and Processing

Future experimental work with human subjects will aim to test predictions drawn from the modeling work presented in this dissertation as well as the theoretical approach underpinning it. Initial work along these lines has met with promising results. McCauley and Christiansen (2015) tested a key prediction of the theoretical perspective embodied by the CBL model: that chunking ability shapes individual differences in on-line sentence processing abilities. In the initial part of the study, we tested a novel twist on a paradigm previously used to study chunking: the serial recall task. The stimuli consisted of triplets and pairs of letter consonants which appear in naturally occurring text at varying frequencies, and were designed to avoid resemblance to syllables or words (as there were no vowels). The results revealed considerable variation in participants' ability to successfully generalize previous knowledge of these sub-lexical chunks of letter consonants to novel contexts. In the second half of the experiment, subjects processed complex sentences, featuring two relative clause types (object- and subject-relative clauses), in a self-paced reading task. Chunking performance from the initial part of the study was then used to predict reaction times at the critical main verb for both relative clause types. Chunking ability successfully predicted processing times for both complex sentence types.

These findings confirm the modeling approach taken in this dissertation, demonstrating that chunking is relevant for understanding language processing, in line with the notion that chunking takes place at multiple levels: low-level chunking of sub-lexical letter sequences successfully predicted complex sentence processing abilities, consistent with the idea that chunking may reduce the computational burden imposed by long-distance dependencies during sentence processing.

This work is also of relevance to understanding language acquisition more broadly: as described in Chapters 1 and 2, the Now-or-Never bottleneck (Christiansen & Chater, 2016) requires that language learning take place in an incremental, on-line fashion, suggesting an integral role for chunking. This is consistent both with the modeling work described in this dissertation, and with previous computational modeling work showing that chunking can account for key findings relevant to children's phonological knowledge and word learning abilities (e.g., Jones, Gobet, Freudenthal, Watson, & Pine, 2014). Future behavioral work will examine individual differences in chunking ability in a developmental context, attempting to trace the impact of chunking on specific aspects of acquisition, including the early development of complex sentence processing.

The need for further individual differences work with adults is underscored by the finding that good chunkers had fewer difficulties in relative clause processing, while poor chunkers were shown to have greater difficulties in object-relative clause processing relative to subject-relative clause processing, consistent with previous findings from individual differences studies on statistical learning (Misyak et al., 2010) and verbal working memory (King & Just, 1991). This raises the intriguing possibility that experience-based chunking may partly mediate the relationship between those more nebulous constructs and sentence processing, consistent with the finding that individual differences in language experience are tied to similar subject-relative vs. object-relative effects (Wells et al., 2009; cf. Christiansen & Chater, 2016a). Future work will seek to gauge the relative importance of chunking for language processing in individual differences studies which examine chunking ability alongside measures of working memory and statistical learning.

Conclusion

The computational modeling work described in this dissertation has served to demonstrate the power simple learning and memory mechanisms, such as chunking, in accounting for key facets of children's linguistic behavior. In **Chapter 2,** I presented the CBL model. In this chapter, I compared CBL's simple framework for statistically-based chunking to other on-line models, including a purely recognition-based chunking model (PARSER; Perruchet & Vintner, 1998) as well as a purely statistical model. CBL not only outperformed both of these baselines, but achieved strong, stable shallow parsing performance while accounting for the majority of child utterances across corpora drawn from a typologically diverse set of 29 Old World languages.

Having shown, in Chapter 1, that the CBL model is capable of discovering and using chunk-based building blocks to perform aspects of comprehension and production, in **Chapter 3** I test the psychological validity of these building blocks. Studies covering several key developmental psycholinguistic findings regarding children's distributional and item-based learning are simulated. It is shown that CBL can capture developmental data from a range of findings spanning from child artificial language learning (Saffran, 2002) to child sensitivity to multiword sequence frequency (Bannard & Matthews, 2008) to morphological development (Arnon & Clark, 2011). These results are discussed in the context of their significance for understanding formulaic language use over the course of development, spanning into adulthood.

Chapter 4 explores the notion that children rely more heavily on multiword units in language learning than do adults learning a second language. To this end, I take an initial step towards using large-scale, corpus-based computational modeling as a tool for exploring differences between the linguistic units of different learner types. The CBL model is used to compare the usefulness of chunk-based knowledge in accounting for the speech of second-language learners vs. children and adults speaking their first language. In the same vein as the TraceBack method (Lieven, Behrens, Speares, & Tomasello, 2003), I compare the CBL model's ability to use chunks discovered in the speech of single learners to generalize to the on-line production of unseen utterances from the same learners. This modeling effort thus aims to provide the kind of “rigorous computational evaluation” of the Traceback Method called for by Kol, Nir, & Wintner (2014). Moreover, I explore the nature of the chunk inventories acquired by the model for each learner type, using a network-theoretic approach. Together, these findings suggest that while multiword units are likely to play an important role in second-language learning, adults may learn less useful chunks, rely on them to a lesser extent, and arrive at them through different means than children learning a first language.

Finally, **Chapter 5** explored the limitations and future directions for the modeling work presented in this dissertation, highlighting the need to incorporate semantic representations in future extensions. Ongoing behavioral research testing key predictions of the theoretical foundations of the CBL model was described.

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., & 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.

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

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

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

Christiansen, M.H. & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Dominey, P. F., Hoen, M., & Inui, T. (2006). A neurolinguistic model of grammatical construction processing. Journal of Cognitive Neuroscience, 18(12), 2088-2107.

Fillmore, C. (1985). Syntactic intrusions and the notion of grammatical construction. In Mary Niepokuj et al. (Eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society* (pp. 73-86). Berkeley: Berkeley Linguistics Society.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: The case of nonword repetition. *Developmental science*, *17*, 298-310.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, *30*, 580-602.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language*, *41*, 176-199.

Langacker, R. (1987). *The foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Palo Alto: Stanford University Press.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333–370.

McCauley, S.M. & Christiansen, M.H. (2014). Prospects for usage-based computational models of grammatical development: Argument structure and semantic roles. *Wiley Interdisciplinary Reviews: Cognitive Science, 5,* 489-499.

McCauley, S.M. & Christiansen, M.H. (2015). Individual differences in chunking ability predict on-line sentence processing. In D.C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, *1*, 31.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of child language*, *37*, 607-642.

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

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

Shayan, S. (2008). *Emergence of roles in English canonical transitive construction*. Unpublished doctoral dissertation. University of Indiana.

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