Acquiring formulaic language

A computational model

Stewart M. McCauley and Morten H. Christiansen
Cornell University

In recent years, psycholinguistic studies have built support for the notion that formulaic language is more widespread and pervasive in adult sentence processing than previously assumed. These findings are mirrored in a number of developmental studies, suggesting that children's item-based units do not diminish, but persist into adulthood, in keeping with a number of approaches emerging from cognitive linguistics. In the present paper, we describe a simple, psychologically motivated computational model of language acquisition in which the learning and use of formulaic expressions represents the foundation for comprehension and production processes. The model is shown to capture key psycholinguistic findings on children's sensitivity to the properties of multiword strings and use of lexically specific multiword frames in morphological development. The results of these simulations, we argue, stress the importance of adopting a developmental perspective to better understand *how* formulaic expressions come to play an important role in adult language use.

Keywords: language acquisition, formulaic expressions, computational modeling, chunking, statistical learning, cognitive linguistics

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).[1] 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

[1] 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.*

The Mental Lexicon 9:3 (2014), **419–436.** DOI 10.1075/ml.9.3.03mcc
ISSN 1871–1340 / E-ISSN 1871–1375

(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, implying 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 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,[2] before evaluating the model's ability

[2] 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.

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, in press; 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 (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.[3] 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 multiword units. Aside from the aforementioned role of the chunk inventory in

[3] 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).

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. (in press) 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 shows CBL's ability to capture child sensitivity to multiword sequence frequency (Bannard & Matthews, 2008) while the second concerns the learning of formulaic sequences and their role in morphological development (Arnon & Clark, 2011).

Simulation 1: 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 et al., 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.[4] 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. This is depicted in Figure 1.

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

[4] 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.

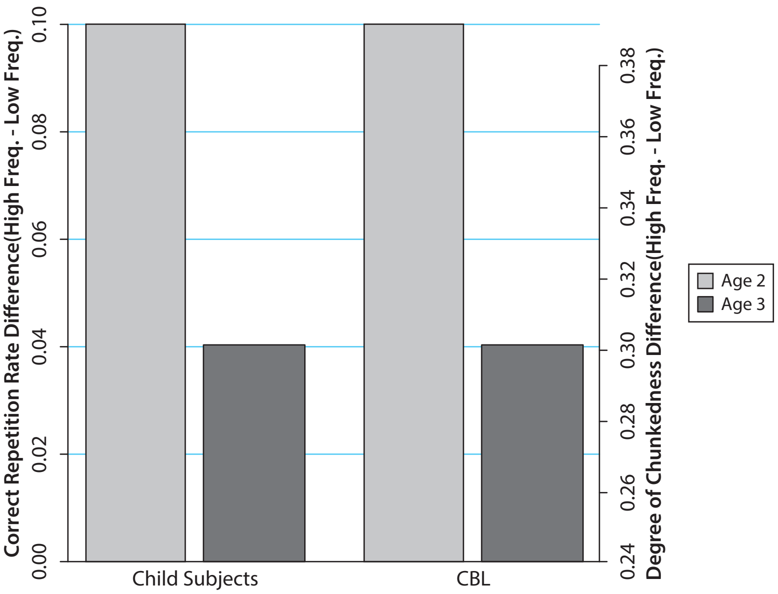

Figure 1. The difference in correct repetition rates between high- and low-frequency phrase conditions for both age groups in Bannard & Matthews (2008) (at left), and the difference in the mean degree of chunkedness (BTP) of the stimuli in high- and low-frequency conditions for the two- and three-year-old CBL results from Simulation 1 (at right).

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 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*.[5] 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 multiword sequences grows, rather than diminishes, over time. Figure 2 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).

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

[5] 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.

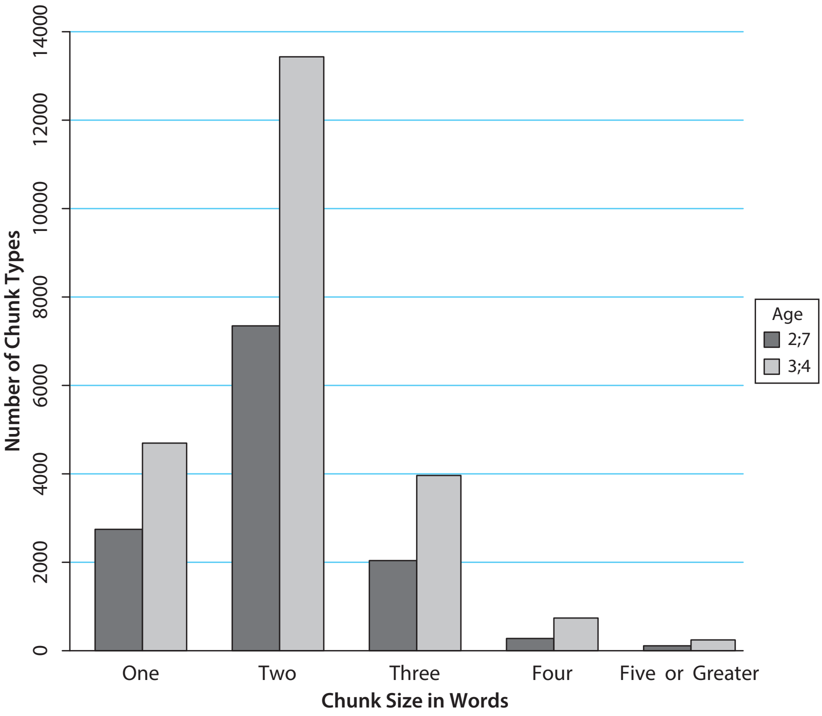

Figure 2. Number and size (in words) of chunk types in the chunk inventory at the early and late stages of Simulation 1. Each stage corresponds to the average age of the children in each age group in Bannard and Matthews (2008).

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 2: 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 1.

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 3. 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 3; in order to depict the results in an intuitive format, we divided –1 by the mean log BTP for each condition.[6]

As can be seen in Figure 3, 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

[6] 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.

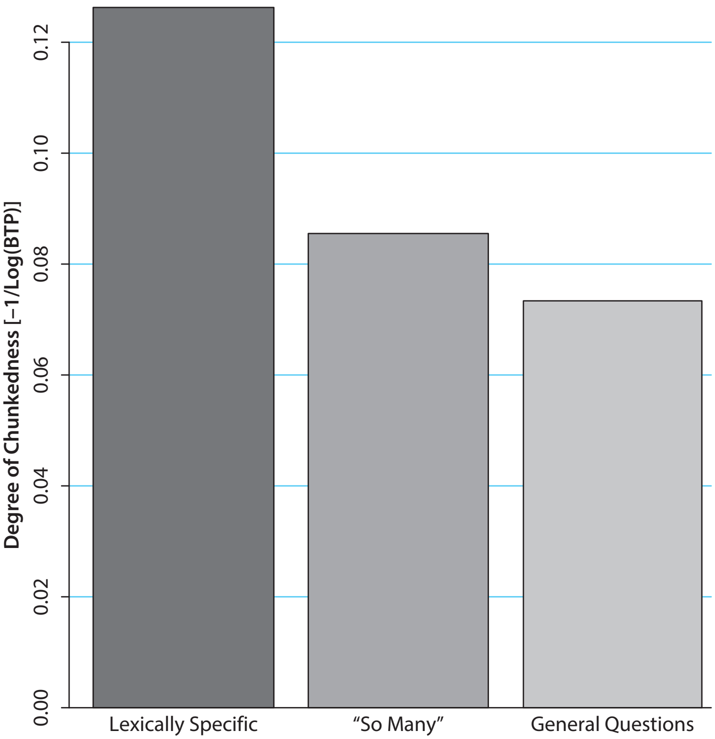

Figure 3. Mean degree of chunkedness (−1/mean log(BTP)) for the each of the three conditions in Simulation 2.

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., in press). 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. DOI: 10.1016/0010-0277(88)90020-0

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. DOI: 10.1080/15475441.2010.505489

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. DOI: 10.1177/0023830913484891

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122, 292–305. DOI: 10.1016/j.cognition.2011.10.009

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language*, 62, 67–82. DOI: 10.1016/j.jml.2009.09.005

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, 19, 241. DOI: 10.1111/j.1467-9280.2008.02075.x

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. DOI: 10.1016/j.jecp.2012.01.005

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, 9, 198–213. DOI: 10.1016/j.cogsys.2007.10.002

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

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780198299554.001.0001

Ferreira, F., & Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, 1, 71–83. DOI: 10.1111/j.1749-818X.2007.00007.x

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22, 829. DOI: 10.1177/0956797611409589

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PloS one*, 7, e33202. DOI: 10.1371/journal.pone.0033202

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. DOI: 10.1044/1092-4388(2004/099)

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). DOI: 10.1017/S0140525X12002658

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. DOI: 10.1002/wcs.1295

McCauley, S. M., Monaghan, P., & Christiansen, M. H. (in press). 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. DOI: 10.1016/j.cognition.2009.07.011

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, 36, 1299–1305. DOI: 10.3758/MC.36.7.1299

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39, 246–263. DOI: 10.1006/jmla.1998.2576

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. DOI: 10.1017/S0140525X12001495

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. DOI: 10.1353/lan.2013.0068

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31, 927–960. DOI: 10.1080/03640210701703576

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. DOI: 10.1080/17470210600971469

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, 6, 382–386. DOI: 10.1016/S1364-6613(02)01958-7

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. DOI: 10.1111/j.1467-9922.2010.00622.x

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 32, 213–231. DOI: 10.1017/S0261444800014154

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511519772

Corresponding Address

Stewart M. McCauley
Department of Psychology
Uris Hall, Cornell University
Ithaca, NY 14850, USA

smm424@cornell.edu