TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 637–652

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12258

This article is part of the topic "More than Words: The Role of Multiword Sequences in Language Learning and Use," Morten H. Chrisiansen and Inbal Arnon (Topic Editors). For a full listing of topic papers, see: http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-2/issuetoc

Computational Investigations of Multiword Chunks in Language Learning

Stewart M. McCauley,[a] Morten H. Christiansen[b]

[a]*Department of Psychological Sciences, University of Liverpool*
[b]*Department of Psychology, Cornell University*

Received 22 October 2015; received in revised form 1 September 2016; accepted 26 September 2016

Abstract

Second-language learners rarely arrive at native proficiency in a number of linguistic domains, including morphological and syntactic processing. Previous approaches to understanding the different outcomes of first- versus second-language learning have focused on cognitive and neural factors. In contrast, we explore the possibility that children and adults may rely on different linguistic units throughout the course of language learning, with specific focus on the granularity of those units. Following recent psycholinguistic evidence for the role of multiword chunks in online language processing, we explore the hypothesis that children rely more heavily on multiword units in language learning than do adults learning a second language. To this end, we take an initial step toward using large-scale, corpus-based computational modeling as a tool for exploring the granularity of speakers' linguistic units. Employing a computational model of language learning, the Chunk-Based Learner, we compare the usefulness of chunk-based knowledge in accounting for the speech of second-language learners versus children and adults speaking their first language. Our findings suggest that while multiword units are likely to play a 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.

Keywords: Language learning; Chunking; L2; Computational modeling; Corpora

Correspondence should be sent to Stewart M. McCauley, Department of Psychological Sciences, University of Liverpool, Eleanor Rathbone Building, Bedford St South, Liverpool L69 7ZA, UK. E-mail: stewart.mccauley@liverpool.ac.uk

1. Introduction

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 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), 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 linguistic proposals, eschewing the grammar-lexicon distinction (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., Goldberg, 2006; Tomasello, 2003). 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 online comprehension and production (e.g., McCauley & Christiansen, 2011, 2014, unpublished data), 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' use of multiword sequences using large-scale, corpus-based modeling. We do this by employing a model of online language learning in which multiword sequences play a key role: the Chunk-Based

Learner (CBL) model (Chater, McCauley, & Christiansen, 2016; 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 online production of unseen utterances from the same learners. This modeling effort thereby aims to provide the kind of "rigorous computational evaluation" of the Traceback Method called for by Kol, Nir, and Wintner (2014).

In what follows, we first introduce the CBL model, including its 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. 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.

2. 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 backward transitional probabilities (BTPs),[1] to which both infants (Pelucchi, Hay, & Saffran, 2009) and adults (Perruchet & Desaulty, 2008) have been shown to be sensitive (see McCauley & Christiansen, 2011, for more about this choice of statistic, and for why the model represents a 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 (Chater et al., 2016).

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 the corpora of real learner and learner-directed speech taken from public databases.

2.1. 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.

2.1.1. Comprehension

The model processes utterances online, 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 online (incrementing the frequency of each word or word-pair by 1). This frequency information is then used online 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 processed, its count in the chunk inventory is incremented by 1, provided it already exists; otherwise, it is added to the inventory 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, online fashion as before. The model continues calculating BTPs while learning the same frequency information, but uses the chunk inventory to make online 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 online 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.

2.1.2. Production

While the model is exposed to a corpus incrementally, processing the utterances online 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.

3. Simulation 1: Modeling the role of multiword chunks in L1 versus 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 multiword 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. That is, CBL is trained on 90% of the utterances spoken by a given speaker and then tested on its ability produce the novel utterances from the remaining 10% of the corpus from that speaker. We compare the outcome of simulations performed using L2 learner speech ($L2 \rightarrow L2$) to two types of L1 simulation: production of child utterances based on learning from that child's own speech ($C \rightarrow C$) and production of adult caretaker utterances based on learning from the adult caretaker's own speech ($A \rightarrow A$). The $C \rightarrow C$ simulations provide a comparison to early learning in L1 versus L2 (as captured in the $L2 \rightarrow L2$ simulations), while the $A \rightarrow A$ simulations provide a comparison of adult L1 language to adult speech in an early L1 setting. A 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 ($A \rightarrow C$). Crucially, the $L2 \rightarrow L2$, $C \rightarrow C$, and $A \rightarrow A$ 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* (similar to the Traceback Method), while the $A \rightarrow C$ simulations provide a comparison to a more standard simulation of language development.

If L2 learners do rely less heavily on multiword 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.

3.1. Methods

3.1.1. 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 nonclassroom setting allows better comparison with child learners. The data were 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 the 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, the 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 ($t(6) = -1.3$, $p = .24$) nor the L1 caretaker utterances ($t(6) = 0.82$, $p = .45$) differed significantly from the L2 learner utterances in terms of number of words per utterance.

While limitations on the number of available corpora made it impossible to match the corpora along every relevant linguistic dimension, we controlled for additional relevant 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), which has previously been shown to reflect syntactic development (e.g., Brown, 1973; de Villiers & de Villiers, 1973). Additionally, type-token ratio (TTR) served as a measure of vocabulary range, as the corpora were matched for size. Because the corpora are matched for length (number of word tokens), TTR allows us to factor the number of unique word types used into an overall measure of vocabulary breadth. Details for each corpus and speaker are presented in Table 1.

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.

3.1.2. Simulations

For each simulation, we ran 10 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 a particular L2 learner and must subsequently attempt to produce the withheld subset of 10% of this L2 learner's utterances ($L2 \rightarrow L2$), 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

Table 1
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

other utterances produced by this child [$C \rightarrow C$], one in which the model must attempt to produce the withheld L1 caretaker utterances after exposure to the other L1 utterances produced by the same adult/caretaker [$A \rightarrow A$], and one in which the model must attempt to produce a random 10% of the child utterances after exposure to the adult/caretaker utterances [$A \rightarrow C$]). 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 set of unseen utterance types.

3.2. Results and discussion

As can be seen in Fig. 1, the model achieved stronger mean sentence production performance for all three sets of L1 simulations than for the L2 simulations (*L2* → *L2*: 36.3%, *SE*: 0.6%; *Child* → *Child*: 49.6%, *SE*: 0.8%; *Adult* → *Adult*: 42.1%, *SE*: 0.7%; *Adult* → *Child*: 47.5%, *SE*: 0.9%). 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 ($B = -5.67$, $t = -1.98$, $p < .05$), reflecting a significant difference between the L2

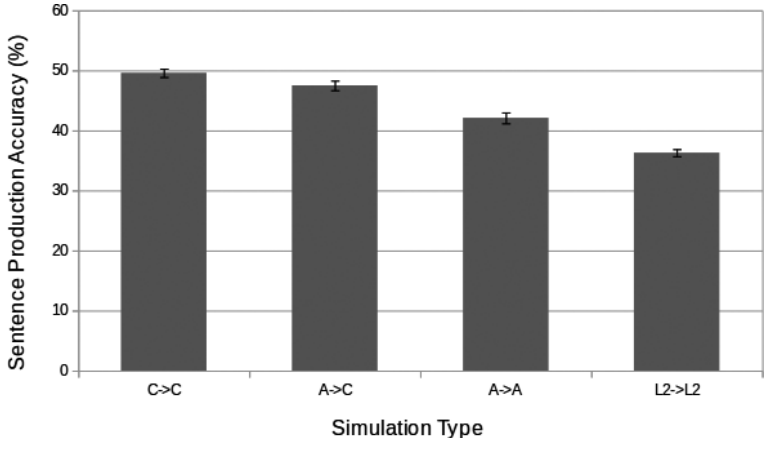

Fig. 1. Graph depicting the mean sentence production accuracy scores on the leave-10%-out task for each of the four simulation types.

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 ($B = -0.7$, $t = -1.7$, $p = .08$), none of the other variables or interactions reached significance. The model had an adjusted R^2 value of 0.83.

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 language use (see Arnon & Christiansen, 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 versus 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, McCauley, & Christiansen, 2017; Arnon & Snider, 2010; Jolsvai et al., 2013).

To compare the structure of the chunk inventories built by models for each learner type, we calculated the overall percentage of chunks falling within each of four size groupings: one-word, two-word, three-word, and four-or-more-word chunks. The results of this comparison are depicted in Fig. 2. As can be seen, there were close similarities in terms of the size of the chunks extracted from the input across learner types, despite clear differences in the ability of these units to account for unseen learner speech. In

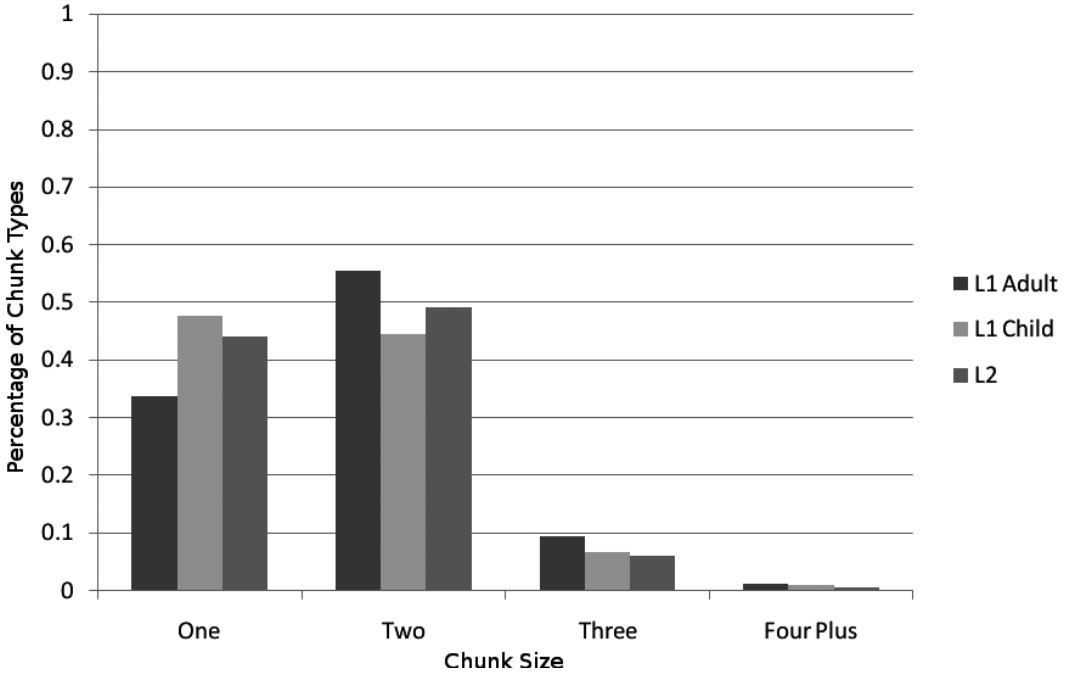

Fig. 2. Percentage of chunk types by size for each learner type.

Appendix A, we list the top ten most frequent chunks across L1 child and L2 learners for the English language corpora.

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 versus 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 preexisting linguistic knowledge from a learner's L1 (cf. Arnon, 2010; Arnon & Christiansen) and the role of overall L2 exposure (e.g., Matusevych, Alishahi, & Backus, 2015).

While these additional factors may be key sources of the differences between L1 and L2 learning outcomes, 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. Nevertheless, L2 and child L1 inventories exhibited similarities in terms of structure: McCauley (2016) shows, using a series of network analyses, that the chunk inventories constructed by the model for L2 and L1 child simulations exhibit similar patterns of connectivity (between chunks) while differing significantly from chunk inventories constructed for L1 adult simulations.

One intriguing possibility for explaining lower performance on the L2 simulations is that L2 learners are less sensitive to coherence-related information (such as transition

probability, in the present case), and may rely more on raw frequency of exposure. If so, 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 the second simulation.

4. Simulation 2: Evaluating the role of raw frequency versus 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 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 performance across the L1 simulations.

4.1. Method

4.1.1. Corpora

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

4.1.2. 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.

4.1.3. 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.

4.2. 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 compared 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 ($L2 \rightarrow L2$: +1%, $A \rightarrow A$: −1%, $C \rightarrow C$: −2%), there were considerable individual differences across simulations (*SD* of 4%, with change sizes ranging from 0% to 11%).

These differences in performance across learner types were further underscored by an interaction between learner type and model version: a four-way ANCOVA with Simulation Type (Original CBL versus Raw Frequency), Learner Type (Adult L1, Child L1, and Adult L2), MLU, and TTR as factors confirmed a significant interaction between Simulation Type and Learner Type ($F(1, 412) = 4.31$, $p < .05$), in addition to main effects of Learner Type ($F(1, 412) = 393.7$, $p < .001$), TTR ($F(1, 412) = 68.2$, $p < .001$), and MLU ($F(1, 412) = 769.7$, $p < .001$), with post hoc tests confirming greater improvement for L2 simulations over Adult L1 ($t(129.6) = 2.39$, $p < .05$) and Child L1 simulations ($t(119.6) = 4.39$, $p < .001$).

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. Preexisting

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, Simulation 1 revealed remarkable similarities in the L1 Child versus L2 learner chunk inventories in terms of chunk structure (see also McCauley, 2016), 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.

5. General discussion

This study represents an initial step toward the use of large-scale, corpus-based computational modeling to uncover similarities and differences in the linguistic building blocks used by different learner types (in this case, L1 and L2 learners). We have sought to answer the call of Kol et al. (2014) and have provided a computationally explicit approach to the Traceback Method (Lieven et al., 2003), relying on CBL as a psychologically motivated model of language acquisition. 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, unpublished data), 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. This finding complements previous work showing that L2 learners do not use multiword sequences to support grammatical development to the same extent as children do (e.g., Wray, 1999).

Secondly, 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 improved scores on L2 simulations to a statistically significant extent, while models of L1 child and adult speech were 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 versus 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 versus 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), further work using longitudinal corpora of L2 learner speech will be necessary 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—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.

Acknowledgments

This work was supported in part by BSF grant number 2011107 awarded to MHC (and Inbal Arnon). Thanks to Erin Isbilen and Inbal Arnon for helpful comments and suggestions. We thank two anonymous reviewers for suggestions which improved the paper.

Note

1. 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.

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., 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., & 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.

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.

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

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

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 & Brain Sciences*, *39*, e62.

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

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.

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

Felser, C., & Clahsen, H. (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.

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.

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* (pp. 692–697). Austin, TX: Cognitive Science Society.

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, CA: 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. (2016). Language learning as language use: Statistically-based chunking in development. Doctoral dissertation, Cornell University.

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

Neville, H. J., & Bavelier, D. (2001). Variability of developmental plasticity. In J. McClelland & R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 271–301). 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 old 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*, *102*, 11629–11634.

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

de Villiers, J. G., & de Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research*, *2*, 267–278.

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.

Appendix A: Top 10 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