Comparing Memory-based and Neural Network Models of Child Language Acquisition

Eva Portelance
Stanford University
portelan@stanford.edu

Abstract

This paper compares two models of early language development in children: the Chunk-based Learner of McCauley and Christiansen (2019), which memorizes frequently occurring chunks, and an LSTM recurrent neural network model in order to determine if the emergent abstractions provided by this model help us understand children's production behavior. The models are trained on 39 single-child corpora from the CHILDES database (MacWhinney, 2000) documenting caregiver and child interactions. A production task is used to determine how well each model reproduces the linguistic production of the children in these corpora. I find that the LSTM has better performance overall on this task, but that neither model can reliably reproduce longer child utterances, suggesting that even more abstraction then what is learned by the LSTM may be necessary to account for child production behavior.

1 Introduction

Children's lexicon and grammatical abilities grow in tandem, resulting in a tight correlation between vocabulary size and grammatical complexity (Bates et al., 1994; Brinchmann et al., 2019; Frank et al., 2019). This relationship is consistent with the hypothesis that children's early grammatical abilities are well-described by lexicalized models. In such models, grammar learning is equivalent to learning a distribution over lexical items, sometimes called signs, which encode both the function and form of units of meaning (e.g. words, morphemes). Learning a grammar, therefore, amounts to learning a Lexicon. Models can learn varying levels of abstraction over lexical items. Some learn no representational abstractions, for example n-gram language models which don't learn syntactic relations beyond collocation, while others have the capacity to learn some structural abstractions, for example LSTM recurrent neural networks learning long distance dependencies such as subject-verb number agreement (Linzen et al., 2016). The goal of this paper is to determine the degree to which the emergent abstractions provided by a lexicalized recurrent neural network model help in understanding children's production behavior.

I compare two lexicalized models based on their capacity to reproduce early child language production. The first is the Chunk-based Learner model (CBL) of McCauley and Christiansen (2019) described in §2; the second is a Long short-term memory (LSTM) recurrent neural network model described in §3. The CBL memorizes frequently-occurring chunks and does not learn any representational abstraction, similar to other pure-memory models (Perruchet and Vinter, 1998; Servan-Schreiber and Anderson, 1990). In contrast, the LSTM makes use of nested hidden layers to learn abstract representations that can predict sequential dependencies between words across a range of dependency lengths. The models are trained and tested on 39 different corpora from the CHILDES database (MacWhinney, 2000) of English speaking children and caregiver interactions. Analyses of the corpora are presented in §4. I evaluate the models using a production task which is a slightly modified version of the task originally used by McCauley and Christiansen (2019). The production task is described in §5. Performance on this task for both models is reported in §6 and then analysed in §7.

2 Memory-based Model

McCauley and Christiansen (2019) propose CBL as lexicalized model of early child language. CBL is a continuation of a long standing tradition of computational pure-memory based chunk-

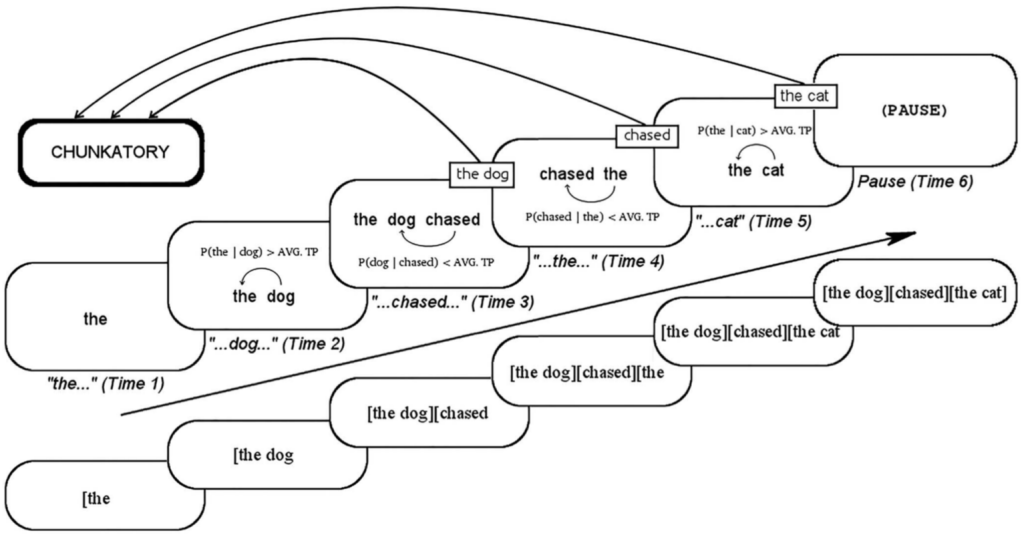

Figure 1: Processing of the utterance 'the dog chased the cat.' Material above the diagonal arrow depicts the computations driving the model processing; material below the arrow represents the resulting shallow parse as it unfolds over time. At Time 2, the model calculates the BTP between *the* and *dog*, which exceeds the average BTP threshold, resulting in the two words being grouped together. Because the next word has not yet been encountered, the two words are not yet stored in the chunkatory as a chunk. At Time 3, the BTP between dog and chased falls below the running average so chased is not grouped together with the preceding material and the dog is then stored in the chunkatory. (McCauley and Christiansen, 2019)

ing models for word and phrase segmentation. They learn frequent multi-word chunks, much like the CC model of Servan-Schreiber and Anderson (1990) or the PARSER model of Perruchet and Vinter (1998). McCauley and Christiansen (2019) motivate their model design by appealing to human cognitive behaviors, specifically, the sensitivity to transition probabilities during language learning (Bannard and Matthews, 2008; Arnon and Snider, 2010; Arnon and Clark, 2011) and the tendency to remember frequently co-occurring words as chunks (Saffran et al., 1996; Thompson and Newport, 2007; Pelucchi et al., 2009).

CBL uses simple heuristics to learn a lexicon - or as the authors call it, a chunkatory - of chunks, which are frequently co-occurring words. It does not learn any abstraction beyond collocation (i.e. no hierarchical structure or long distance dependencies). These chunks can be as short as one word to as long as a complete utterance. In order to learn these chunks, the model uses the following process; As CBL incrementally processes through utterances, it uses backward transition probabilities (BTP) between words to determine the boundaries between chunks. If the BTP between two words is above a running average, the two words are part of a same chunk, however, if it is below the running average than they are part of separate chunks. All words are also added to the chunkatory as uni-chunks to facilitate parsing in unseen contexts. This process is illustrated in figure 1, taken from McCauley and Christiansen (2019). The model also learns BTPs between chunk bigrams which can be used to predict new utterances.

3 LSTM Model

LSTMs are the current standard baseline for language model development in the natural language understanding literature. They process utterances incrementally and they have been shown to learn some structural abstraction (Linzen et al., 2016). For these reasons, LSTM language models as a neural network architecture lend themselves well to the question this paper explores: when comparing memory-based and neural network models of early syntactic development, do the emergent abstractions provided by the lexicalized recurrent neural network model help us understand children's production behavior?

I use a two layered LSTM recurrent neural network. The model uses randomly initialized 100-dimensional word embeddings as its input layer. The embeddings are updated during learning. At

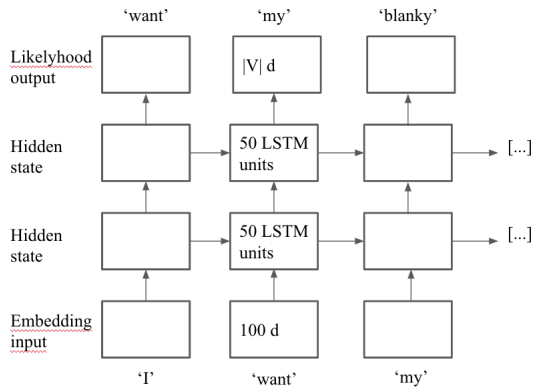

Figure 2: The LSTM model architecture incrementally processing the utterance 'I want my Blanky'.

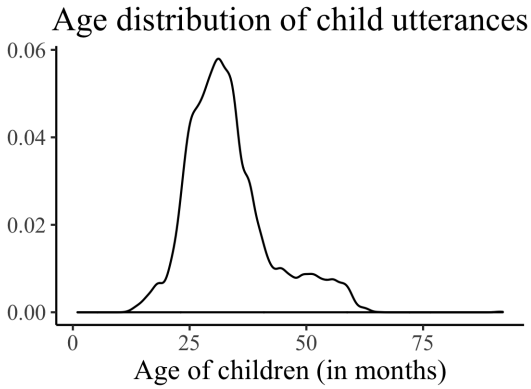

Figure 3: The overall all distribution of children's age at the time of utterance production for all child produced speech across the 30 corpora.

Child	Example utterance
Abe	I wanna go outside okay Mom ?
Becky	we going over the bridge .
Dominic	want picking up Mummy .
Jimmy	wheres this go wheres the other one go ?
Roman	does do you wanna play the um game again the store game ?

Table 1: Examples of longer utterances produced by children in the corpora.

each time-step a the current word embedding is combined with a hidden state in the first LSTM layer. After cross validation, I found that that ideal size for the each LSTM layer was 50 hidden units. The output of this first layer is fed to a second LSTM layer with its own hidden state. The model uses both L1 and L2 regularization when calculating the output layer. The output at each time-step is a distribution over the whole vocabulary, representing the next word. This model architecture is illustrated in figure 2.

Though recurrent neural networks are capable of learning longer distance dependencies, it is still unclear how much and what kinds of abstractions the LSTM learns and furthermore, whether it learns enough abstraction to properly account for the linguistic production of children. I will pursue these questions further in the discussion of §7.

4 Data

The data for this study comes from 39 different corpora of single child and caregiver interactions. These corpora are all available through the CHILDES database (MacWhinney, 2000) which was accessed through the childes-db API (Sanchez et al., 2018). The complete list of corpora is available in appendix A. All utterances were stripped of punctuation and transformed into lowercase.

These corpora include 39 of the 42 originally used for CBL by McCauley and Christiansen (2019). I was unable to locate 3 of the corpora they used. The criteria for selecting these corpora were as follows:

1. they have at least 20 000 words;

2. they have at least a 1:20 ration of child produced to child-directed speech.

I adopt a 'model as participant' paradigm where I trained a separate model for each corpus, resulting in 39 instantiations of both the CBL and the LSTM models. For this reason, it is important to analyze how similar the different corpora are prior to reporting averaged performance results.

All of the corpora are longitudinal, collecting data across multiple months usually in the form of spread out 2 or so hour sessions. Thus, for each child there are utterances produced at a variety of developmental stages. The ages of children at the time of utterance production range from 9 months to about 5 years of age, with a mean of 33 months. Most of the data was produced between the ages of 2 to 4 years. The overall distribution over the age range is available in figure 3.

The corpora vary significantly in size, the smallest corpus having 23 428 tokens while the largest has 1 981 183 tokens. However, the ratio of vocabulary size, or unique tokens, to corpus size is relatively proportional across corpora as figure 4 shows.

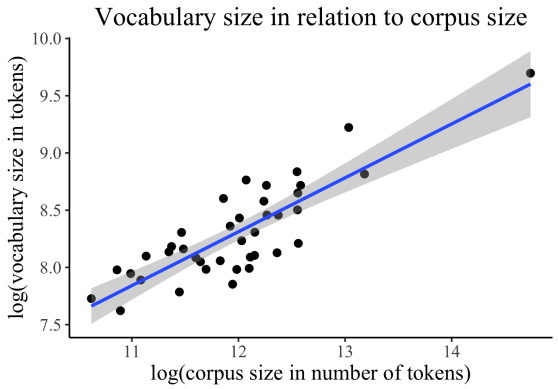

Figure 4: The relation between corpus size and vocabulary size. Each point is one corpus and the linear relation uses the formula, `log(vocabulary size)` ∼ `log(corpus size)`.

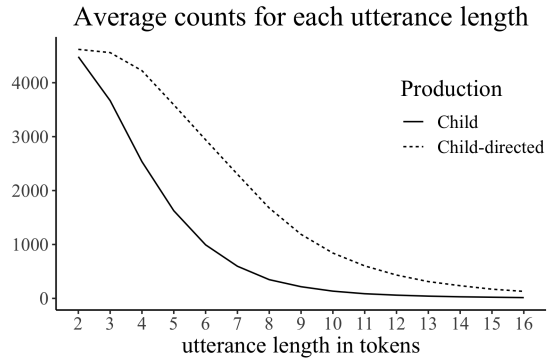

Figure 5: The mean number of utterances produced by a child or caregivers (child-directed) for each utterance length.

The models are trained on all of the child-directed speech in a given corpus and 60% of the child produced speech; the remaining 40% were used for test. These were randomly selected throughout the corpus. Utterances were seen by the models in the order in which they appeared in the transcripts.

Children tend to produce shorter utterances; the overall mean utterance length across corpora for child-produced sentences is about 2.6 words. That said, they still produce many longer utterances as well, some examples taken from the corpora are provided in table 1. Figure 5 shows the mean number of utterances produced in the corpora by utterance length. It compares the numbers produced by children to those produced by caregivers. In §6, I report the performance of the models on the production task described in the next section as a function of utterance length. Longer utterances are interesting because they cannot be as easily memorized. Thus, I assume that they require generalizing forms of abstraction over linguistic input in order to then be able to produce them.

5 Production Task

In order to compare the models' capacity to reproduce a child's production behavior, I will use a slightly modified version of the production task which was used by McCauley and Christiansen (2019) to evaluate their CBL model. The original task called for incrementally evaluating the model when test utterances (child-produced utterances) were encountered during training. This is not a standard evaluation method and it does not facilitate performance comparison with other models. Instead, I evaluate model performance at the end of training for both the CBL and LSTM models.

For every test utterance, a bag-of-words of lexical items is created; for the CBL, this is a bag of chunks and for the LSTM, this is a bag of word vectors. From the bag-of-words, the model's task is to predict the order in with the lexical items where produced. If the predicted order matches the original order in the child produced utterance, then this is counted as a correct prediction. The production score is the proportion of correct predictions over test examples.

In order to retrieve predictions from the models a decoder is necessary. I report performance results using both a greedy decoder and a beam search decoder. In the case of the beam search decoder, a beam of size 5 was used and a correct prediction was counted if the original order appeared in the 5 resulting beams from the last state of the decoder algorithm.

It should be noted that this task gives a certain advantage to a CBL type model where the lexical items in the model are chunks. The reason for this is that for any given utterance with n tokens, CBL will have $\leq n$ lexical items to reorder, while the LSTM will always have n tokens to reorder. Regardless of this disadvantage, however, the LSTM still gets better overall performance on the task.

6 Results

Overall, the LSTM has better performance on the production task than the CBL model. Using the greedy decoder, the average performance for the LSTM is .62, 95%CI[.58-.66] and for the CBL, .57 95%CI[.53-.60]. In the case of the beam seach decoder the performance of the LSTM is on av-

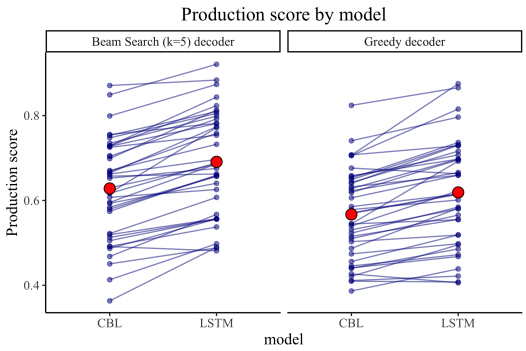

Figure 6: Overall performance of both the CBL and LSTM models for each corpus using the beam search decoder and the greedy decoder. The use of the LSTM increases performance in most cases.

erage .69 95%CI[.65-.73] and the performance of the CBL is .63 95%CI[.59-.66]. Figure 6 shows the individual differences in overall performance of the models for each corpus.

The size of the individual corpora does not seem to affect performance as shown in figure 7.

Given that the goal of this paper is to weigh the benefits of the LSTM's capacity to learn some structural abstraction over the data, I am most interested in the models performance on utterances which are longer than 2 or 3 words. Figure 8 shows the models' mean production score by utterance length. Both the performance of the CBL and the LSTM drop for utterances longer than 3 words. Neither model seems to learn enough abstraction to truly account for a child's production behavior.

7 Discussion

Though there is some improvement in performance when using the LSTM, neither model seems to do well on longer utterances. There are a few possible explanations for this which I will explore in this section.

First, the production task is difficult. Reordering 10 words, may be difficult even for humans. I hope to run an experiment with human participants to get a sense of what the performance ceiling is on this task. In a future iteration of this task, I could also provide preceding context to the models to aid with prediction. This may be a slightly more *realistic* objective.

Second, the models are only being exposed to a fraction of the linguistic input that children receive in the time frame during which they participated in these recorded sessions. It is possible that with

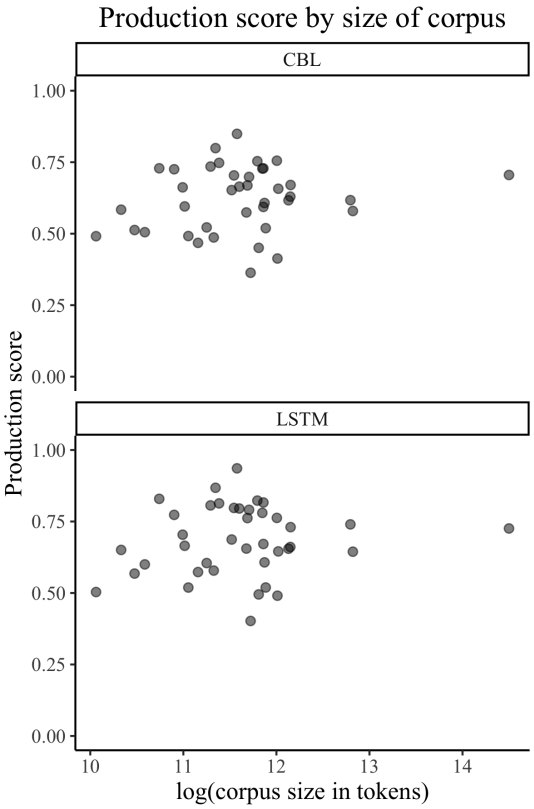

Figure 7: Performance using the beam search decoder for both the CBL and LSTM models as a function of corpus size. Each point represents a different corpus of the 39 corpora.

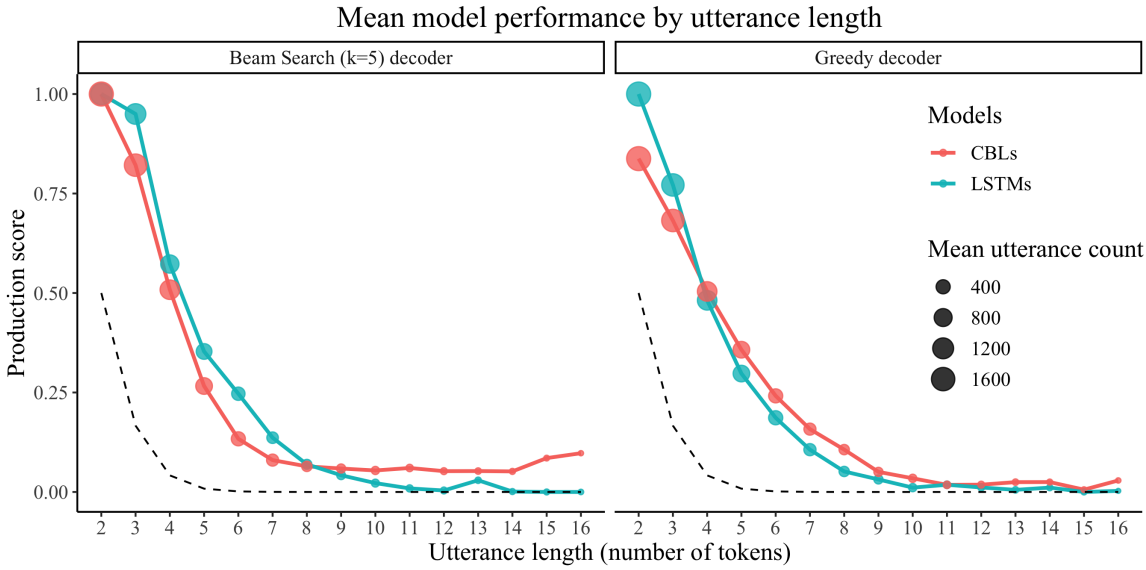

Figure 8: The mean production score of models by utterance length. The points represent the mean number of exemplars per corpus. The dotted line represents chance performance if the number of lexical items to reorder for a n long utterance is also equal to n - i.e. chance performance for the LSTM, but not necessarily for the CBL.

more data exposure, the LSTM would do better. In future work, I hope to test the performance of GPT-2, the current state-of-the-art language model (Radford et al., 2019), which has been trained on many times more linguistic input than a three year old child has heard or seen in their lifetime. If this model still cannot predict the productive behavior of children, then I may conclude that the scarcity of training data is not at fault.

Third, though LSTM are known to have the capacity to learn some abstraction, it may still not be enough to account for the generalization capabilities of young children. Models which learn explicit types of structural abstractions (e.g. syntactic and semantic categories) which are believed to be informative may have a better chance at performing well on longer utterance production.

8 Conclusion

In this paper, I compared the performance of the Chunk-based Learner model (CBL) of McCauley and Christiansen (2019) to an LSTM recurrent neural network (LSTM) on a production task designed to determine how well these models reproduce the production behavior of young children. I trained models on 39 different English corpora of single child and caregiver interactions. I found that the LSTM had better overall performance than the CBL on this task, but that neither model was able to reliably predict longer utterances produced by children of around 2 to 4 years of age. In the post-experiment analysis, I explored possible explanations for the LSTM's poor performance on longer utterances and I proposed research steps that I will engage with in future work in order to distinguish between these explanations.

Acknowledgments

I would like to thank George Kachergis and the members of the Language and Cognition Lab for the initial discussions which sparked this project. I would also like to thank the people who came to my CogSci Seminar presentation in May 2019 whose valuable comments have helped me ameliorate this project.

Authorship Statement

This project serves as a subsection of my Qualifying Paper (A Linguistics department requirement for the Ph.D.) and as such, I have received advising from faculty members beyond this course, primarily Professor Mike C. Frank, as well as Professors Judith Degen and Boris Harizanov. However, the work on developing and writing this project is solely my own.

References

Inbal Arnon and Eve V Clark. 2011. Why brush your teeth is better than teeth: Children's word production

is facilitated in familiar sentence-frames. *Language Learning and Development*, 7(2):107–129.

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82.

Colin Bannard and Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological science*, 19(3):241–248.

Elizabeth Bates, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language*, 21(1):85–123.

Ellen Irén Brinchmann, Johan Braeken, and Solveig-Alma Halaas Lyster. 2019. Is there a direct relation between the development of vocabulary and grammar? *Developmental science*, 22(1):e12709.

Michael Frank, Mika Braginsky, Virginia Marchman, and Daniel Yurovsky. 2019. Variability and Consistency in Early Language Learning: The Wordbank Project. Https://langcog.github.io/wordbank-book/index.html.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

B. MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stewart M McCauley and Morten H Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological review*, 126(1):1.

Bruna Pelucchi, Jessica F Hay, and Jenny R Saffran. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2):244–247.

Pierre Perruchet and Annie Vinter. 1998. PARSER: A model for word segmentation. *Journal of memory and language*, 39(2):246–263.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jenny R Saffran, Elissa L Newport, and Richard N Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4):606–621.

Alessandro Sanchez, Stephan Meylan, Mika Braginsky, Kyle MacDonald, Daniel Yurovsky, and Michael C Frank. 2018. childes-db: a flexible and reproducible interface to the child language data exchange system.

Emile Servan-Schreiber and John R Anderson. 1990. Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):592.

Susan P Thompson and Elissa L Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language learning and development*, 3(1):1–42.

A List of corpora

On next page.

Child	Citation
Abe	Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16, 589600.
Adam	Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
Alex	Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of childrens early English articles. *Journal of Child Language*, 36, 173200.
Anne	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Aran	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Barbara	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York, NY: Oxford University Press.
Becky	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Carl	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Conor	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York, NY: Oxford University Press.
David	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York, NY: Oxford University Press.
Dominic	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Emily	Weist, R. M., & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development*, 4, 291308.
Emma	Weist, R. M., & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development*, 4, 291308.
Ethan	Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of childrens early English articles. *Journal of Child Language*, 36, 173200.
Eve	Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
Gail	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Jimmy	Demetras, M. (1989). *Changes in parents conversational responses: A function of grammatical development*. Paper presented at ASHA, St. Louis, MO.
Joel	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.

Table 2: Citations for all 39 English speaker child language corpora used for this study. (Continues...)

Child	Citation
John	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Lara	Rowland, C. F., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language*, 33, 859877.
Lily	Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of childrens early English articles. *Journal of Child Language*, 36, 173200.
Liz	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Matt	Weist, R. M., & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development*, 4, 291308.
Michelle	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York, NY: Oxford University Press.
Naomi	Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parentchild discourse. In K. E. Nelson (Ed.), *Childrens language*, Vol. 4 (pp. 128), Hillsdale, NJ: Lawrence Erlbaum Associates.
Nathaniel	MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, 17, 457472.
Nina	Suppes, P. (1974). The semantics of childrens language. *American Psychologist*, 29, 103114.
Peter	Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology*, 6, 380420.
Roman	Weist, R. M., & Zevenbergen, A. (2008). Autobiographical memory and past time reference. *Language Learning and Development*, 4, 291308.
Ross	MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk.* Hillsdale, NJ: Erlbaum.
Ruth	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.
Sarah	Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
Seth	Peters, A. (1987). The role of imitation in the developing syntax of a blind child. *Text*, 7, 289311.
Shem	Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In R. J. A. Sinclair & W. Levelt (Eds.), *The child's conception of language* (pp. 1743). Berlin, Germany: Springer Verlag.
Thomas	Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics,* 20, 481508.
Tow	Demetras, M., Post, K., & Snow, C. (1986). Feedback to first-language learners. *Journal of Child Language*, 13, 275292.
Trevor	Demetras, M. (1989). *Working parents conversational responses to their two-year-old sons.* Working paper. University of Arizona.
Violet	Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of childrens early English articles. *Journal of Child Language*, 36, 173200.
Warren	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127152.

Table 3: (Continued) Citations for all 39 English speaker child language corpora used for this study.