## Title
Modeling Children’s Early Linguistic Productivity Through the Automatic Discovery and Use of Lexically-based Frames



Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 41(0)

Authors
McCauley, Stewart M.
Christiansen, Morten H.

Publication Date
2019

Peer reviewed

# Modeling Children's Early Linguistic Productivity Through the Automatic Discovery and Use of Lexically-based Frames

**Stewart M. McCauley (stewart-mccauley@uiowa.edu)**
Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242

**Morten H. Christiansen (christiansen@cornell.edu)**
Department of Psychology, Cornell University, Ithaca, NY 14853

**Abstract**

A central question for cognitive science is whether children's linguistic productivity can be captured by item-based learning, or whether the learner must be guided by abstract, system-wide principles governed by innate constraints. Here, we present a computational model of early language acquisition which learns to discover and use lexically-based frames in a fully incremental, on-line fashion. The model is rooted in simple prediction- and recognition-based processes, subject to the same memory limitations as language learners. When exposed to English corpora of child-directed speech, the model is able learn developmentally plausible frames and use them to capture over 70% of the utterances produced by target children aged 2 to 5. Across a typologically diverse range of 29 languages, the model is able to capture over 68% of child utterances. Together, these findings suggest that much of children's early linguistic productivity can be captured by item-based learning through computationally simple mechanisms.

**Keywords:** language learning; language acquisition; usage-based approaches; computational modeling; chunking

## Introduction

By four years of age, most children have mastered the basic grammatical structures of their native language, an achievement marking the transition to a seemingly unbounded capacity for communicating novel information. But how is such open-endedness possible, given the finite, noisy nature of the input? This is among the foundational questions of cognitive science. For over half a century, researchers have argued that children's learning is guided by system-wide, abstract principles and constrained by innate biases (e.g., Chomsky, 1965). In recent decades, an alternative perspective has emerged in the form of usage-based approaches, which hold that children's linguistic productivity emerges gradually as a process of storing and abstracting over the input (e.g., Tomasello, 2003). In this framework, children's earliest steps towards unbounded productivity come in the form of lexically-based frames: through knowledge of partially overlapping sequences, children form schemas with slots that are filled according to semantic, pragmatic, or phonological constraints (e.g., Braine, 1963).

Among the earliest quantitative studies offering evidence for lexical frames was that of Lieven, Behrens, Speares, and Tomasello (2003), who used a technique known as the "traceback method" to analyze the speech of a single child during its second year. Lieven et al. found that a high proportion of the child's linguistic productivity—utterances which went beyond frozen or recycled sequences to feature novel word combinations—could be explained in terms of lexically-specific frames, such as "*there's the* __ ." Subsequent work improved on the original traceback method and yielded similar findings (e.g., Lieven, Salomo, & Tomasello, 2003).

As highlighted by other researchers (e.g., Kol, Nir, & Wintner, 2014) the traceback method is not automated and is therefore severely limited in terms of the range of corpora and languages to which it can be applied. Moreover, the lack of a computationally explicit formulation means that the general approach does not make specific commitments to the types of learning mechanisms or representations that allow productivity to emerge from lexically-based representations.

This problem highlights a general lack of computational work examining item-based learning as a starting point for linguistic abstraction, which is reflected in the imprecise language with which usage-based theory is often discussed. For instance, researchers have appealed to complex psychological constructs such as analogical reasoning to explain lexically-based frames (e.g., Gentner & Namy, 2006; Tomasello, 2003). Even computational studies examining the transition from item-based learning to abstraction have appealed to analogy while remaining agnostic as to the lower-level mechanisms supporting it (e.g., Bod, 2009).

By contrast, we aim to provide an account of early abstraction which is rooted in basic processes of prediction and recognition. Moreover, we wish to capture such learning in a way that is consistent with the myriad sensory and memory limitations imposed on the learner (as discussed in Christiansen & Chater, 2016). This requires a fully incremental and on-line learning model, in line with memory constraints that force reliance on local rather than global syntactic information. It also means capturing learning in a way that is fully usage-based in the sense that all learning takes place in the context of specific processing events.

## Modeling Children's Discovery and Use of Lexically-based Frames

Here, we seek to model children's discovery and use of lexical frames by modifying an existing usage-based computational framework, known as the Chunk-Based Learner (CBL; McCauley & Christiansen, 2014, 2019). Inspired by the aforementioned memory constraints, the CBL model aims to recreate individual children's utterances by learning from the linguistic input to which they have been exposed. The model offers strong performance across a typologically diverse range of languages (McCauley & Christiansen, 2019) while capturing psycholinguistic data from both children (McCauley & Christiansen, 2014) and adults (Grimm, Cassani, Gillis, & Daelemans, 2017). Importantly, previous research has only used CBL to model the discovery and use of concrete multiword units. In the present study, we implement this pre-existing model and modify it to support the incremental, on-line discovery and use of lexically-based frames.

In what follows, we describe the basic workings of the CBL model as well as the modifications we applied to enable the learning of lexical frames. Next, we examine qualitative and quantitative properties of the frames discovered by the model when exposed to corpora of English child-directed speech. We also evaluate the model's ability to use these frames in a sentence production task, exploring the extent to which they can support early linguistic productivity. Finally, we look at the model's ability to use frames in this sentence production task across a typologically diverse array of 29 different languages.

## Experiment 1: Modeling the Development of Lexically-based Frames in English

### The CBL Model

The model has been described in detail in previous work (e.g., McCauley & Christiansen, 2019). We therefore briefly provide sufficient information to implement the model. The model processes the input corpus on a word-by-word basis, tracking low-level frequency information for words and word pairs (bigrams). This information is used on-line to calculate the backward transition probability (BTP) between words. By maintaining a running average of BTP over previously seen word pairs and using it as a threshold, the model classifies BTPs linking words as either *high* or *low*. High BTPs are used to group words together to form part of a chunk, while low probabilities are used to define chunk boundaries. When a boundary is placed, the preceding word(s)—there is no *a priori* limit on the size of a chunk—are placed as a unit in the model's *chunk inventory*. When the model encounters a previously-discovered chunk in the input, its frequency count is incremented by 1. The resulting chunk inventory thus contains a mix of single-word and multiword units. The model maintains frequency counts for pairs of chunks occurring together, which supports the incremental construction of utterances during production.

The model also uses its chunk inventory on-line while processing the input. Through a combination of prediction- and recognition-based processing, knowledge of previously discovered chunks can assist in further discovery: when a word-pair is encountered, if it has occurred at least twice as part of an existing chunk, it is automatically grouped together (regardless of BTP). Otherwise, the BTP is evaluated against the running average threshold as described above.

A record of the model's on-line chunking of utterances is maintained for later evaluation against the output of a parser. CBL's ability to approximate the output of shallow parsers cross-linguistically has been suggested to capture key aspects of comprehension (cf. McCauley & Christiansen, 2019). The model also aims to capture key aspects of production: as the model makes its way through a corpus of child-directed speech, it encounters utterances produced by the target child of the corpus, at which point the production side of the model comes into play. The model must produce its own utterance by generalizing from the chunks and statistics it has learned up to that point in the simulation. This task is used to evaluate our version of the model and is described below in the subsection entitled Sentence Production Task.

### Modifications to the CBL Model

To enable the on-line discovery and use of lexical frames, we made some slight changes to the original CBL implementation. When the model has discovered 5 or more multiword chunks which overlap in all but one position, it creates a lexical frame—a chunk with an empty slot—and stores it in the chunk inventory. When chunks matching this frame are encountered, the frame's frequency count is incremented, as are the counts of matching chunks. The 5+ criterion was selected in light of previous corpus studies of evidence for lexical frames in child-directed speech (e.g., Cameron-Faulkner, Lieven, & Tomasello, 2003, who used a criterion of 4+ in their analyses). As the original version of CBL already uses its chunk inventory during on-line processing, we felt this change was in keeping with the model's intended psychological features.

As an example of frame creation, consider an instance in which the model has already discovered and used the chunks *in the box*, *in the tub, in the bag,* and *in the chair.* When the model discovers the chunk *in the cup*, it also discovers the frame *in the* __ as an automatic generalization over the previous multiword chunks. Both *in the cup* and *in the* __ are initialized in the chunk inventory with counts of 1, the starting frequency value for newly-discovered chunks. The frame's count is then incremented by 1 when the model later encounters *in the box*, a previously discovered chunk, as is the count for that chunk. The frame's count is also incremented by 1 again when the model discovers a new chunk, *in the sink*, and so forth.

As described in the below section entitled Sentence Production Task, the model can rely on its knowledge of lexical frames during production.

## Input Corpora

Rather than aggregate across multiple corpora, each of our simulations involved exposing the model to a single corpus of child-directed speech. We selected, from the English language portion of the CHILDES database (MacWhinney, 2000) all corpora meeting the following three criteria: i) contained at least 50,000 words; ii) featured a multiword child-to-adult utterance ratio of at least 1:10; and iii) spanned at least a 6-month period in terms of the target child's age across the corpus. These criteria were met by individual corpora for 43 English-learning children (US: 25; UK: 18). Tags and punctuation were removed from the corpora, leaving, for each utterance, only speaker identifiers and the original sequence of words.

## Learning Lexical Frames in English

Across the entire set of 43 simulations, the model discovered a mean of 14 lexical frames per 10,000 words of input. Rather than leading to a combinatorial explosion of units—as might suggest psychological implausibility, or coverage due to trivial factors in subsequent evaluation tasks--frames made up just 5% of the total chunk inventory for the simulation involving the largest corpus (*Thomas*; Maslen, Theakston, Lieven, & Tomasello, 2004), with smaller percentages for smaller corpora (3% on average).

To offer a sense of the qualitative nature of the model's lexical frames, we show, for the largest corpus (Thomas), a range of frequent frames as well as less-frequent but developmentally interesting frames.

**Table 1:** Frequent and Developmentally Interesting Frames Learned from the Thomas (Dense) Corpus and Corresponding Counts in the Chunk Inventory

| **Frequent Frames** | | **Developmentally Interesting Frames** | |
|---|---|---|---|
| *the* __ | 56117 | *a little* __ | 2131 |
| *a* __ | 42937 | *what's* __ | 2122 |
| *your* __ | 8366 | *a big* __ | 1401 |
| *in the* __ | 7718 | *are you going to* __ | 1196 |
| *on the* __ | 6950 | *what do you* __ | 945 |
| *this* __ | 6742 | *more* __ | 837 |
| *that* __ | 6343 | *I want to* __ | 427 |
| *very* __ | 4911 | *on* __ *own* | 228 |
| *I don't* __ | 3386 | *the red* __ | 120 |
| *going to* __ | 3348 | *more* __ | 103 |

As can be seen in Table 1, even though slots are allowed anywhere in a chunk, the vast majority of lexical frames featured a slot in the final position. Across all the English corpora, slot-final frames accounted for a large percentage of overall frames utilized, ranging from 85% to 98%.

There is good overlap between the frames appearing in Table 1 and frames postulated by other researchers on the basis of corpus analyses, including some of the earliest to advance the notion of lexical frames: for instance, *more* __ is one of the first frames identified in Braine (1963).

Next, we turn to the question of whether the lexical frames discovered by the model can offer insights into the nature children's early productivity. To this end, we evaluate the model according to its ability to capture children's actual utterances in these corpora, and measure the extent to which the model's lexical frames can support production above and beyond concrete multiword chunks.

## Sentence Production Task

The sentence production task was based on the bag-of-words incremental generation task first described by Chang, Lieven, and Tomasello (2008). The task rests on the simplifying assumption that the overall message the child wishes to convey can be—very roughly—approximated by treating the utterance as an unordered bag-of-words. When the model encounters a multiword utterance produced by the target child of a corpus, its task is to sequence the items in the bag to produce its own utterance, using only the words and statistics it has discovered prior to that point.

We used a nearly identical version of the task to that described by McCauley and Christiansen (2019): following psycholinguistic evidence for children's use of multiword units (see above), the model was allowed to draw upon previously discovered chunks to populate the bag-of-words. To produce an utterance, the model begins by selecting from the bag the word or chunk with the highest transition probability given the start-of-utterance marker (a marker preceding every line in the corpus). At each subsequent time step the model removes and produces the word or chunk with the highest probability *given the most recently placed chunk*. This process continues until the bag is empty.

Thus, production is implemented as fully incremental, chunk-to-chunk process, relying entirely on local information. In other words, there is no global whole-sentence optimization. In this sense, the model captures the sorts of memory limitations described in the introduction.

Where our version of the task differed from that described by McCauley and Christiansen (2019) was in the additional use of lexical frames: if the model lacked experience of a given sequence in the child's utterance, but had learned a lexical frame capable of fitting that sequence, it was allowed to utilize the frame in the bag-of-words task. Consider the model's attempt to produce the child utterance: "*red one stuck in the jam.*" In a case in which the model has discovered the lexical frame *in the* __ but has never encountered the sequence *in the jam* in the input, the model is allowed to use the lexical frame to complete this pattern. Statistics are then calculated over the frame itself, as if it were a fully concrete chunk.

**Gold Standard for Sentence Production Task** Following each production attempt, the model's utterance is scored against the child's original utterance according to an all-or-nothing scoring metric: if the two utterances do not match completely, a score of 0 is assigned. Otherwise, a score of 1 is given. Thus, the overall accuracy of the model across a corpus can be calculated as a percentage of correctly produced multiword utterances (single-word utterances are excluded to avoid inflating performance). We call this the *Sentence Production Accuracy* (SPA) score.

**Alternate Distributional Models** We evaluate the model against two baseline models: the first is the basic version of CBL used as a starting point for the present study (described above; cf. McCauley & Christiansen, 2019). The second is a standard trigram model; this approach was selected as a baseline due to its widespread use and generally robust performance as a probabilistic language model across a range of genres (Manning & Schütze, 1999).

**Results and Discussion** Across all 43 English simulations, the lexical frames version of the CBL model (CBL+LF) achieved a median Sentence Production Accuracy of 71.3% (mean: 69.5%). This is compared to a median score of 58.5% (mean: 57.8%) for the original CBL model and a median score of just 45.7% (mean: 45.1%) for the trigram (3G) baseline. The distribution of scores for each model are shown in Figure 1.

A linear mixed-effects model fit using logit-transformed SPA scores, with child as a random factor, confirmed that both the CBL+LF model ($\beta$=0.53, $t$=22.8, $p < 0.001$) and the 3G model ($\beta$=-0.53, $t$=-22.6, $p < 0.001$) differed significantly from the original CBL model, in opposite directions.[1]

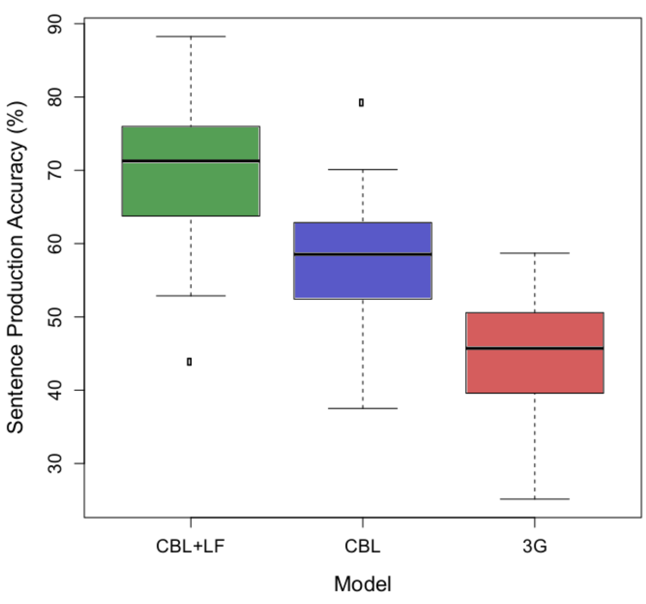


**Fig. 1:** Box and whisker plots depicting English Sentence Production Accuracy (%) for the model and its baselines.

Thus, in addition to discovering developmentally and psychologically plausible lexical frames, the CBL+LF model was able to use these units to improve upon the CBL model's production performance by nearly 12 percentage-points, surpassing the performance of a standard trigram model by nearly 25 percentage-points.

[1]All $p$-values computed via Satterthwaite approximation.

## Experiment 2: Modeling the Development of Lexically-based Frames Across Typologically Diverse Languages

The vast majority of computational modeling work in the study of language acquisition has focused on English. It is crucial, however, to determine whether the types of linguistic representations and learning mechanisms we ascribe to children can plausibly accommodate languages with typological features that differ greatly from those of English. In the case of the present model, which uses multiword units as much of the basis for learning and processing, morphological features are of particular interest.

A previous study using the CBL model has demonstrated that multiword units do indeed facilitate production for typologically diverse languages, including morphologically rich languages (McCauley & Christiansen, 2019). Here, we ask the question of to what extent limited productivity based on lexical frames can improve the ability of CBL to capture the utterances of children learning a typologically diverse set of languages, above and beyond what can be captured through learning tied to concrete chunks.

### Corpora

We selected from the CHILDES database (MacWhinney, 2000) corpora involving single target children, rather than aggregating data across multiple corpora. Due to limitations on the number of corpora for several of the languages in CHILDES, these were selected according to slightly relaxed criteria: each corpus contained at least 10,000 words, at least 1,000 multiword child utterances, and a child-to-adult utterance ratio of no less than 1:20.

These criteria were met by corpora for 160 additional target children from 28 different languages (Afrikaans: 2, Cantonese: 8, Catalan: 4, Croatian: 3, Danish: 2, Dutch: 12, Estonian: 3, Farsi: 2, French: 15, German: 22, Greek: 1, Hebrew: 6, Hungarian: 4, Indonesian: 8, Irish: 1, Italian: 8, Japanese: 10, Korean: 1, Mandarin: 7, Polish: 11, Portuguese: 2, Romanian: 1, Russian: 2, Sesotho: 3, Spanish: 11, Swedish: 5, Tamil: 1, Welsh: 6). Table 2 lists some basic typological properties of these languages.

To get a rough quantitative measure of morphological complexity for child-directed speech in each language, we calculated word type/token ratios (following the reasoning and methods of Chang et al., 2008). We refer to this as the *Morphological Complexity Score*.

### Sentence Production Task

We used the same sentence production task as in Exp. 1.

### Results and Discussion

Across all 29 languages and 200+ corpora, the lexical frames version of CBL achieved a mean SPA score of 68.4%, compared to 55.3% for the original CBL model and just 45.9% for the trigram model. Means for each language are shown in Figure 2. By discovering and utilizing lexical frames, the model was able to reproduce the majority of the child utterances in every language, with mean scores ranging from 55% (Swedish) to 81% (Romanian).

A linear mixed-effects model was fit to logit-transformed SPA scores with language and child as random effects, and a by-language random slope of model. This confirmed that the CBL+LF model ($\beta$=0.58, $t$=26.8, $p$ < 0.001) and the trigram model ($\beta$=-0.32, $t$=-8.6, $p$ < 0.001) differed significantly from the original CBL, in opposite directions.

Because previous work with the original version of CBL demonstrated that the model's performance decreased as a function of morphological richness (McCauley & Christiansen, 2019), we compared CBL performance to CBL+LF in order to determine whether this effect was reduced by the use of lexical frames. Figure 3 depicts the relationship between the CBL+LF model and Morphological Complexity Score.

**Table 2:** Typological Properties of the 29 Languages

| Language | Family | Genus | Word Order | # Cases |
|---|---|---|---|---|
| Irish | Indo-European | *Celtic* | VSO | 2 |
| Welsh | Indo-European | *Celtic* | VSO | 0 |
| English | Indo-European | *Germanic* | SVO | 2 |
| German | Indo-European | *Germanic* | N.D. | 4 |
| Afrikaans | Indo-European | *Germanic* | N.D. | 0 |
| Dutch | Indo-European | *Germanic* | N.D. | 0 |
| Danish | Indo-European | *Germanic* | SVO | 2 |
| Swedish | Indo-European | *Germanic* | SVO | 2 |
| Greek | Indo-European | *Greek* | N.D. | 3 |
| Farsi | Indo-European | *Iranian* | SOV | 2 |
| Romanian | Indo-European | *Romance* | SVO | 2 |
| Portuguese | Indo-European | *Romance* | SVO | 0 |
| Catalan | Indo-European | *Romance* | SVO | 0 |
| French | Indo-European | *Romance* | SVO | 0 |
| Spanish | Indo-European | *Romance* | SVO | 0 |
| Italian | Indo-European | *Romance* | SVO | 0 |
| Croatian | Indo-European | *Slavic* | SVO | 5 |
| Russian | Indo-European | *Slavic* | SVO | 7 |
| Polish | Indo-European | *Slavic* | SVO | 7 |
| Estonian | Uralic | *Finnic* | SVO | 10+ |
| Hungarian | Uralic | *Ugric* | N.D. | 10+ |
| Sesotho | Niger-Congo | *Bantoid* | SVO | 0 |
| Hebrew | Afro-Asiatic | *Semitic* | SVO | 0 |
| Tamil | Dravidian | *S. Dravidian* | SOV | 7 or 8 |
| Indonesian | Austronesian | *Malayic* | SVO | 0 |
| Cantonese | Sino-Tibetan | *Chinese* | SVO | 0 |
| Mandarin | Sino-Tibetan | *Chinese* | SVO | 0 |
| Korean | Korean | *Korean* | SOV | 7 |
| Japanese | Japanese | *Japanese* | SOV | 9 |

*Note: Information from Haspelmath et al. (2005)*

Though Morphological Complexity Score was indeed a predictor of CBL+LF performance ($\beta$=-2.01, $t$=-3.5, $p$ < 0.001, $r$=0.23), we found that the presence of lexical frames reduced this effect in comparison to that observed for the original CBL model ($\beta$=-2.4, $t$=-4.1, $p$ < 0.001, $r$=0.27), as confirmed by a significant interaction between model and Morphological Complexity Score ($\beta$=0.15, $t$=3.03, $p$ < 0.01) in a linear mixed model which included model as a categorical factor.

A close inspection of the lexical frames discovered by the model when exposed to English revealed that they were both psychologically and developmentally plausible, but we currently lack the cross-linguistic expertise to offer a detailed analysis of lexical frames for the 28 additional languages. Nevertheless, these simulations offer clear evidence that, in principle, the same types of representations and mechanisms can support the discovery of lexical frames across a typologically diverse range of languages. Indeed, for all the 29 languages, lexical frames capture early linguistic productivity above and beyond what can be achieved through concrete words and chunks: CBL+LF lead to a 13 percentage-point improvement over mean CBL performance and a 23 percentage-point improvement over trigram models.

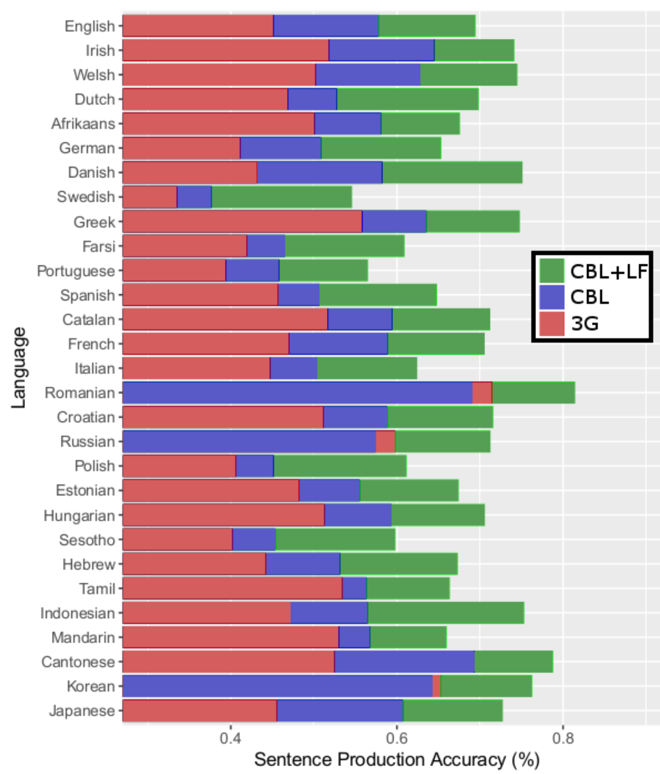


**Fig. 2:** Sentence Production Accuracy (%) for the model and its baselines across 29 languages. Bars are overlapping.

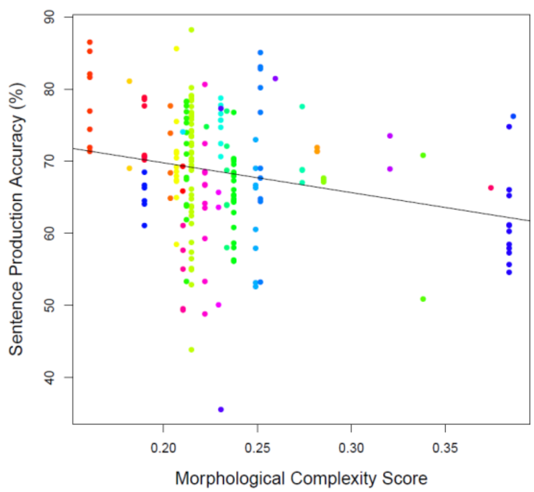


**Fig. 3:** CBL+LF SPA scores across all children and languages plotted against Morphological Complexity Score. Different colors denote distinct languages. Trendline from simple linear regression.

## General Discussion

In this paper, we have demonstrated that a simple, developmentally-motivated model rooted in concrete prediction- and recognition-based processes can discover lexically-based frames that are not only psychologically plausible but also can capture a significant amount of children's early linguistic productivity. In 200+ simulations of individual children across a typologically diverse array of languages, the CBL+LF model was able to capture a significantly higher proportion of child utterances than a version of the model relying solely on concrete words and chunks, offering an even larger improvement over trigram models. Moreover, this was achieved while accommodating the sorts of memory limitations that drive children (and adults) to rely on local information during comprehension and production (e.g., Christiansen & Chater, 2016).

In contrast to previous quantitative studies examining evidence for lexical frames in child speech (e.g., those using the traceback method of Lieven et al., 2003), we 1) capture the actual learning of frames during comprehension, as well as their use in production, and 2) do this for children beyond their second year, with corpora covering child productions during the third and fourth year.

Nonetheless, the CBL+LF approach is not without limitations. Firstly, frames operate on the level of words appearing within chunks; chunks themselves are not yet able to appear in slots. By overcoming this limitation in a principled way, a wider variety of linguistic phenomena could be captured. For instance, non-adjacent dependencies can be learned in the current version of the model: frames like *this* __ *one* and *those* __ *ones* capture a number dependency. Extending the model to allow entire chunks in slots will be a necessary subsequent step towards capturing more abstract processing of long-distance dependencies.

A more serious limitation of the present work is that it does not incorporate the learning or use of semantic information. The semantic properties of words and frames are needed to provide constraints on which items that can appear in lexical frames. The learning of such information is crucial for moving towards a framework capable of producing utterances based on meaning representations (and forming meaning representations during comprehension). To this end, ongoing work aims to simulate the learning of lexical semantics, semantic roles, and argument structures through the use of automatically generated, idealized "visual scenes" which are paired with utterances in corpora.

More generally, the promise of item-based computational approaches for tracing a path to more sophisticated forms of linguistic abstraction is great: previous work has shown that the systematic use of pseudographs to align and compare the sentences in a text can give rise to complex context-free grammars (Solan, Horn, Ruppin, & Edelman, 2005). Bayesian induction of item-based grammars from the speech of single target children has also yielded good coverage of those children's increasing productivity (in a manner akin to the traceback method; Bannard, Lieven, & Tomasello, 2009). However, these models are not subject to memory limitations, and involve computations beyond what children are capable of. A motivation for the current approach, therefore, was to take initial steps towards modeling increasingly productive linguistic representations in a way that is psychologically motivated, incremental, and on-line.

## Acknowledgments

Thanks to Erin Isbilen for helpful comments and feedback. Thanks to Nick Chater, Padraic Monaghan, Colin Bannard, and Ben Ambridge for helpful discussion.

## References

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

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

Braine, M. D. S. (1963). The ontogeny of English phrase structure: The first phrase. *Language, 39*, 1-13.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27*, 843-873.

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

Chomsky, N. (1965). *Aspects of a theory of syntax.* Cambridge: MIT Press.

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

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*, 297-301.

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical

processing and word learning: A computational investigation. *Frontiers in Psychology, 8:555*.
Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of linguistic structures.* Oxford, UK: Oxford University Press.
Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language, 41*, 176-199.
Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*, 481-507.
Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language, 30*, 333–370.
Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge: MIT.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database.* Mahwah, NJ: Lawrence Erlbaum Associates.
Maslen, Theakston, Lieven, & Tomasello (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research, 47*, 1319-1333.
McCauley, S.M. & Christiansen, M.H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9*, 419-436.
McCauley, S.M. & Christiansen, M.H. (2019). Language learning as language use: A cross-linguistic model of language development. *Psychological Review, 126*, 1-51.
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.* Cambridge: Harvard University Press.