Cognitive Science

Modeling the influence of language input statistics on children's speech production

--Manuscript Draft--

Manuscript Number:	19-129R1
Full Title:	Modeling the influence of language input statistics on children's speech production
Article Type:	Regular Article
Keywords:	statistical learning; language development; abstraction; developmental trajectory; age-invariance; CHILDES
Corresponding Author:	Marisa Alexa Casillas Max Planck Institute for Psycholinguistics Nijmegen, Gelderland NETHERLANDS
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Max Planck Institute for Psycholinguistics
Corresponding Author's Secondary Institution:	
First Author:	Ingeborg Roete
First Author Secondary Information:	
Order of Authors:	Ingeborg Roete
	Stefan L. Frank
	Paula Fikkert
	Marisa Casillas
Order of Authors Secondary Information:	
Abstract:	We trained a computational model (the Chunk Based Learner; CBL) on a longitudinal corpus of child-caregiver interactions to test whether one proposed statistical learning mechanism---backward transitional probability (BTP)---is able to predict children's speech productions with stable accuracy throughout the first few years of development. We predicted that the model less accurately generates children's speech productions as they grow older because children gradually begin to generate speech using abstracted forms rather than specific ``chunks'' from their speech environment. To test this idea, we trained the model on both recently encountered and cumulative speech input from a longitudinal child language corpus. We then assessed whether the model could accurately reconstruct children's speech. Controlling for utterance length and the presence of duplicate chunks, we found no evidence that the CBL becomes less accurate in its ability to reconstruct children's speech with age. Our findings suggest that BTP may be an age-invariant learning mechanism.

12 February 2020

Dear Dr. Cooper and Dr. Bonawitz,

We submit to you a revision of our manuscript entitled “Modeling the influence of language input statistics on children's speech production”, which tests the extent to which the Chunk Based Learner model (McCauley & Christiansen, 2011; 2014; 2019), which uses backwards transitional probability (Perruchet & Desaulty, 2008), can consistently and accurately account for children’s speech production across the first four years of life.

We have carefully reviewed and responded to the comments of the three reviewers and the action editor. We hope that they find this new version of the paper much improved. The paper is now reformatted as an Rmarkdown document and uploaded to the same anonymous OSF repository mentioned before (https://osf.io/ca8ts/), in case the reviewers or the editor would like to have a closer look at the data.

In-line with reviewer requests, we have added one additional figure, and a number of extra references. The main text of the manuscript is now 7027 words long (excluding the abstract, references, and the supplementary materials) and includes 7 figures, placed near where they are referred to in the text for ease of review. As before, we also submit a set of Supplementary Materials that include full statistical model output and model results using an alternative formulation of the model.

We greatly appreciate the chance to integrate reviewer comments and resubmit our manuscript. Please do not hesitate to contact us if there are any additional requests or questions.

Sincerely,

The Authors

########## Reviews and responses ##########

Reviewer #1:

- I really appreciate the detailed explanation on page 10 of the corrected score with the example calculations. I think this section could benefit from a figure to make the new score even more clear for readers. I would like to see a graph that plots, for both correct and incorrect utterances, the score as a function of the chance of getting the utterance right. I realize that this is a fairly simple plot of the logarithmic function for the domain [0,1], but I think it will help readers get a sense of the possible range of scores, and it would help illustrate the worked examples too.

> Thank you for this suggestion—we agree that a figure would clarify the importance of the corrected scoring over age. We have now added a figure related to what you describe, only plotting the score as a function of the number of chunks (see Figure 3).

- If I understand the paper correctly, the authors' version does create new chunks for unknown words when processing the corpus with caregiver utterances, but not when reconstructing the child utterances. That's what I had understood from the main text, but then the first paragraph of the supplemental materials made me doubt that, so I wonder if it could be made more explicit (that their own implementation does also accept unknown words during its training phase).

> Thank you for noticing this potential ambiguity. Indeed, during the training phase the model needs to create new chunks for unknown words. But what to do with previously unseen words when *testing* the trained model with the children's productions is another matter. We decided that the fairest treatment for unseen words during test was to throw them out, our reasoning being that there is no obvious way to give them valid default transition probabilities with other existing chunks. We have clarified this in the paper (subsection: "Child utterance reconstruction task") and in the Supplementary Materials (paragraph 1) in order to avoid confusion about how we implemented our version vs. the original one from McCauley and Mortensen.

- Throughout the paper whenever mixed effects regression models are mentioned, I think it would make the manuscript clearer to also provide the model formula. […] I also think a table with the model output for each regression analysis (maybe as a supplement) would be helpful. Right now the reader can rerun the R analysis to generate those tables but even a .Rmd document on OSF with the R output would already make accessing this information at a glance easier.

> We have now added the formulas to the main text as footnotes for each analysis, and have added each model output table to the Supplementary Materials. The reviewer will also find that the manuscript is now compiled from an Rmd document, as suggested.

- In section 3.3 I got confused whether or not the interaction between age and utterance length was also tested and if so, what it's results were, and if not, why not.

> This was somewhat ambiguous in the previous version; thanks for catching it! We did not have an *a priori* expectation that age and utterance length would interact in the likelihood of previously unseen words. Therefore they are not included in our analyses. Thanks to the comment above which resulted in us adding model formulas (main text) and table outputs (Supplementary Materials), the predictors used in each analysis should be much clearer now.

- I wonder if the authors could explain a bit more clearly in the manuscript why the corrected scores take reduplicated chunks rather than reduplicated words into account in their formula. I think it is because the ordering operation that the model performs happens at the chunk level, so that is the relevant unit for calculating probabilities of getting the order right. That being said, readers might benefit from a more explicit explanation about why reduplications of words are not relevant unless the whole chunk is reduplicated (e.g. the word 'the' might be in a sentence twice as part of two different chunks without affecting the chance of ordering it correctly).

> Your understanding (i.e., that the model only learns chunks, not words) is correct and we have clarified this in the paper (end of paragraph 1 under Methods subsection "Corrected reconstruction accuracy"), thanks!

- The first paragraph of the results section seems to refer to figure 3; if so, put that figure reference in earlier.

> We have improved figure placement more generally in the revised manuscript.

==========

Reviewer #2:

p. 3, the beginning of the second paragraph is misleading, "Change in SL behavior following further linguistic experience is also predicted in models that do not assume abstraction". When reading the paragraph, the conclusion is that the chunk-based models do not predict a change in SL behavior, the mechanism remaining the same but applying on larger chunks over time. The first sentence of the paragraph should be modified.

> Thanks very much for pointing this out! We have clarified and made more consistent our discussion about the extent to which these alternative models predict change in SL ability with age (among other changes, we now use "SL ability" instead of "SL behavior", which was itself ambiguous with respect to process vs. outcome).

p. 3, it is written was the period of interest for early speech production was 0;11-4;0, but in the Section 2.3 it is written 1;0-4;0.

> We have now corrected this error, thank you for catching it.

p. 3, in addition to McCauley & Christiansen (2011), Onnis & Thiessen (2013) and Pelucchi, Haye, & Saffran (2009), the authors should cite Perruchet & Desaulty (2008), who, as far as I know, are the first to evaluate the role of BTP.

> We have now incorporated this groundbreaking paper into our text and we apologize for overlooking it in the first submission.

p. 4, "As it sees more sentences, it would continue to add new chunks and track how often they co-occurred". I was wondering if the CBL continually updates the BTP. If some BTP were high at the beginning of the training and that the associated chunks were stored in the model memory, but overtime the BTP become low and under the running average BTP, what happens to the associated chunks and their BTP? Do they disappear from the model memory?

> The Reviewer is correct that chunks are never removed from the inventory. So indeed some chunks that are added early in training would not have been stored as a chunk later on in training. However, this is more of a feature than a flaw of the model. Most of the early chunks are single words that come in handy when reconstructing child speech: in the reconstruction task, the utterance is broken up into the largest chunks possible from the inventory, so small chunks are used when a larger one can't be found. This feature of the model is fairly reasonable from an incremental learning perspective, although the Reviewer is right that a forgetting feature would be an interesting future addition to experiment with (see also Reviewer 3's comments below). We have made minor clarifications to the text quoted above to clarify this issue.

p. 7, is there a threshold for which two chunks provided from a child utterance could not be associated during reconstruction? I give a fictitious (and maybe impossible) example: in the utterance of the child we have three chunks A, B, and C. The BTP between # and A is .90, between A and B is .83, and between B and C is .10. Does the model reconstruct #ABC or only #AB because the BTP between B and C is two low?

> Thanks for this clarification question. Here's how it works: The first step of the reconstruction task involves decomposing the utterance into chunks that are already stored in the inventory. So in this case, assuming you've seen A, B, and C, you would end up with a list of three chunks and a start marker: {#, C, A, B}. Those would be stored as a 'bag of chunks', which is just to say that they are stored as unordered. After this first decomposition step, the utterance is re-composed, but this time using two sources of information: (1) the bag of chunks and (2) the transitional probability between chunks. So in this theoretical example, we would start with "#", then "A" (assuming 0.9 is the highest TP between # and another chunk), then "B" (assuming 0.83 is higher than the A->C TP), then "C"; we could get the original utterance '#ABC'. If the transitional probability between # and B were very high (e.g., 0.95) then we might get a different answer (e.g., "#BCA", "#BAC"), so really the whole matrix of TPs between each chunk pair matters for reconstruction. We have made some minor changes to the text (Methods subsection "Child utterance reconstruction task") and to the figure visually depicting this process (Figure 2) to try and make the reconstruction process more transparent.

p. 8, "With the local data sampling method we selected data within a two-month interval around each age point. For example, for age point 1;6 we selected transcripts in which the child was between 1;5.0 and 1;6.31". Maybe I missed something but how is it possible to go beyond 1;6, because this means that adult utterances the child has not already been exposed to are used in the training. Why not take, for example, one month up to the age point?

> We chose to sample input data as *proximal* to each age rather than up-to that age because we were trying to get a representative picture of the type of input each child was getting at the age points tested. As the reviewer suggests, under ideal circumstances we should have sampled input for the short period preceding each age to train the model to reconstruct utterances produced at that age. However, the corpus on which we based the analyses, while (relatively) quite densely sampled, did not provide sufficient data for this approach. We also reasoned that, because we are interested only in modeling the type of input the child is experiencing at that age, and because the recordings are incomplete, training the model on input *proximal* to the tested age was the best balance to getting a broad, but age-specific model of adult speech. As one can see from the results and discussion, even with our inclusive take on the age-appropriate input, there are still many words produced by the children that are unaccounted for in the adult speech. To us, this result suggests that much denser data are needed for future work, in which case the reviewer's suggestion would be very much worth trying. We have added a sentence on reasoning for our age-based sampling to the paper.

p. 9, "However, for the model of corrected reconstruction accuracy, […] in the dataset". Once again because of my weak expertise in modeling, I did not understand the rationale that led the authors to do that. (I have the same problem p. 13 at the end of the first paragraph of section 3.2).

> Thanks for this clarification question. We will try to give a thorough explanation here; apologies if some of this is already familiar! By fitting a linear mixed-effects model in this analysis we are asking the computer to (a) find a line that best matches our datapoints in a multidimensional space and then (b) tell us about how well the line fits the datapoints in each dimension. In the modeling package we use—lme4, one of the most popular in our field—the model output gives both (1) the estimated value of the dependent measure at a single reference point (i.e., one point along the line that it fitted) and (2) estimates of how the data is predicted to change from *that reference point* for each predictor (e.g., size of increase/decrease in accuracy for each unit of age, etc.). Conveniently the default model output for lme4 also tells us whether the intercept of the model (its reference point) differs significantly from zero. Going back to our current analyses, let's first think about modeling age numerically as 1, 2, 3, and 4. The lme4 software assumes the default reference value for numerical predictors is 0. Therefore the model output would give us accuracy estimates at the reference point of age = 0. We decided that estimates at age zero are not useful for the corrected accuracy score and that, instead, the middle-point in our age range—2;6—is much more indicative of the model's overall performance. All we do, then, is re-code the age predictor in the model so that the default value is in the middle of our age range—at 2;6 (ages 1;0, 1;6, 2;0, 2;6, 3;0, 3;6, and 4;0 are re-coded as -1;5, -1, -0.5, 0, 0.5, 1, and 1.5). That is, the model maps age from -1.5 to 1.5, and will give us estimates at zero. In sum, this age re-coding simply takes advantage of default behavior in lme4 to use age 2;6 as our reference point. All the while, the linear fit for the effect of age is identical to what it would be if we used 1–4. We give some extra clarification of this in the text (subsection: Analysis, paragraph 2), but do not go into detail, since this style of analysis is increasingly common.

p. 10-11, "Because there is no straightforward way to establish […] unseen words. This part also is not clear for me. Could the authors try to explain a little more the rationale underlying their choice.

> Thanks for pointing this out. In addressing this comment, we realized that we needed to adjust how we analyze unseen words, so we are very glad that you asked for clarification! Our previous analysis attempted to answer the question "what increases the likelihood of an utterance being unreconstructable?". However, we realized that we should instead be measuring the likelihood that words seen at test had been previously seen during training, since that is, by definition, what causes an utterance to be unreconstructable. In accordance with this update, we have changed the text under the subsections "Previously unseen words" and "Children's use of unseen words", along with the statistical models, their output, and the accompanying figure (Figure 7).

p. 14, "By taking a longer history of linguistic input into account (i.e., by using cumulative sampling), we expected to see a smaller increase in previously unseen words with age". I took some time to understand why the authors made this prediction. If I have well understood I think they made that prediction because an unseen word in the local sampling could be a seen word in the cumulative sampling. If this is the case, maybe the authors could make that point explicit to help the comprehension of the reader.

> We've added this clear wording suggestion in the new manuscript (subsection: "Children's use of unseen words"), thanks!

p. 14, the primary question recalled at the beginning of the discussion (stable accuracy prediction of CBL throughout development) is opposite to the predictions mentioned at the end of the introduction ("we expected to find that the CBL's ability to reconstruct children's speech decreases in-line with a concomitant increase in children's linguistic sophistication"). Maybe the authors could reformulate to be coherent.

> Thanks for pointing out this apparent inconsistency. We have edited two sentences in this first paragraph of the Discussion as well as a few wordings in the Predictions subsection to make sure this is more coherent.

The authors found that the corrected score they proposed lead to different results from the uncorrected measure initially proposed. Would they advise to use their corrected measures for the following uses of the CBL model to improve its performance?

> We would indeed suggest that our corrected measure is a better test of model performance. We have added a sentence to the end of the first paragraph of the Discussion along these lines.

==========

Reviewer #3:

There are three main ways in the SL community to produce chunking in streams of utterances. The first is Forward Transitional Probabilities (FTP), proposed by *Saffran et al. (1996)*, *Aslin et al. (1998)*, etc. The standard computational implementation (model) of FTPs was Cleereman et al.'s use of Elman's (1990) SRN. The second is Backward Transitional Probabilities, first suggested by *Perruchet and Desaulty (2008)* and applied to infants by *Pelucchi, Hay & Saffran in 2009*. The third is memory-based chunking, as implemented in PARSER (*Perruchet & Vinter, 1998, 2002*) and TRACX (French et al. 2011).

> We are grateful to Reviewer 3 for this brief but comprehensive overview of chunking research. The reviewer is absolutely correct that we missed some essential references, particularly regarding other established approaches to chunking. For that we sincerely apologize. We introduce these other lines of research immediately after introducing BTP and come back to them again briefly when discussing the limitations of the CBL (as suggested in a comment below).

McCauley and Christiansen's CBL model applied to child language learning has become something of cottage industry: papers on similar, if not identical topics, in Psych Review paper, in Topics in Cognitive Science, and several papers in the Proceedings of the Cognitive Science Society. There is considerable overlap with, at least, papers from 2014 and 2019, and I would like to see the authors more clearly delineate their work from other published work.

> We see the unique contribution of our study as: (1) diving into the longitudinal predictions and performance of the CBL and (2) establishing a better (corrected) measure of the model's output. Our study offers some additional value in that it replicates and tests the CBL with a team of authors completely independent from McCauley and Christiansen, though that's probably not worth mentioning in the paper itself. We have highlighted point 1 (longitudinal predictions) in the Introduction ("We extend this work by testing how the model performs with longitudinal data; it is not yet known how well it functions as a predictor of what children can say as they become more linguistically sophisticated.") and Conclusion ("This work extended previous CBL studies by testing the robustness of utterance reconstruction across an age range featuring substantial grammatical development and by also introducing a new controlled accuracy measure for reconstruction."). We feel that point 2 (i.e., the new measure) is already well highlighted from the methods onward.

I would like the notion of BTPs as a driving mechanism to be justified more thoroughly. Since Perruchet et al. (2008) and Pelucchi et al. (2009), pretty much everyone acknowledges that BTPs are able to play a role in chunking. This kind of "backward prediction" is fine, but what is the role of forward prediction in this model? It is obvious that one could rig an SRN to do "backward prediction" (Maskara & Noetzel, 1993, did something related). Now, of course, an SRN doesn't form chunks, as the CBL does, but would the predictions of the BTP-SRN model be as good as the CBL? And what about models like PARSER or TRACX that have been shown to be sensitive to both FTPs and BTPs? One of the major problems of this paper is that none of this other modeling work is even cited, let alone used to compare the performance of CBL on the data presented.

> We now cite these other lines of research immediately after introducing the CBL and BTP. We agree that the lack of citations was a major problem with our prior draft. We have also added a short paragraph to the end of the paper noting (1) again why we chose the CBL (in short, because the age ranges it's been tested on fit well with our target longitudinal range and it had not yet been tested for longitudinal robustness) and (2) that the CBL has some drawbacks that might be better addressed in future work with some other SL models. We hope that, with these changes, it is now clear that the use of the CBL in the present study is little more than one way to test the idea of age-invariance in SL.

Further, a discussion is needed of the mechanisms required for BTPs (and FTPs) to work as chunking cues. These issues have been raised in a number of places, but they are not part of the discussion in justifying the use of CBL. The point is this: any model relying of FTPs or BTPs as chunking cues, must REMEMBER TP information in order to compare it to the current TP. How is this accomplished? In presentations of SRNs doing FTP-based chunking, this issue is simply glossed over. Somehow the system "just knows" that the current TP is lower than the TPs that preceded it and therefore, the spot of the current TP must be a word boundary. How, exactly, does the system know this? (This issue does not come up with models like PARSER or TRACX.) So, let's assume that there must be a

mechanism for storage of TP information and the appropriate comparison of this stored information with the current TP (whether backward or forward). But presumably this memory-and-comparison mechanism, part of executive-control functions, improves (very) significantly between the ages of 1 and 4. So, why is this not reflected in the CBL results? Perhaps the strongest claim of the paper is the BTPs represent an "age-invariant" mechanism of chunking. In light of the above remark, this claim needs considerably more justification as to why it might be.

> Thanks for pointing out this interesting prediction regarding the interaction of memory, executive control, and expected CBL performance. We agree that it's intriguing that, even though the CBL does not at all model maturational changes in, e.g., memory, between ages 1;0 and 4;0, it still manages to reconstruct child utterances better than chance over this age range. One thing to keep in mind is that, while the model is performing, on average, above chance, it still gets many utterances wrong. That is, there is room for improvement. It would be interesting in future work to see how other models would fit the data, as now discussed toward the end of the paper (e.g., in the final paragraph before the Conclusions, where we come back to other work on SL mechanisms).

Another point: Mareschal & French (2017) used TRACX2 to model developmental chunking data in infants (see section 3: Modelling infant statistical learning) by varying learning rates in TRACX2 from 0.0005 for newborns, 0.0015 for two-month olds, and 0.005 for eight-month olds. Now, granted this is not the age range for the present paper, but surely there are some parameter differences that vary between ages 1 and 4?

> This comment is in-line with the one immediately above it, so we have cited this paper along with the changes made in response to the prior comments.

So, one key question that needs to be answered is: to what extent are BTPs the whole picture? Clearly, as semantics enters the picture (referred to as "future work") in the article, other mechanisms will come into play. Can this model do without the memory mechanisms for FTPs. Would an SRN that was retrofitted to do BTPs work as well? Could a recursive auto-encoder model like TRACX or TRACX2 produce the same results?

> We hope that, with the changes described above, it is now very clear to readers that BTPs in general (and the CBL specifically) are far from being the whole picture when it comes to segmenting and learning meaningful chunks from the input. The reviewer has made a compelling case that the current model we focus on is limited; we hope that our agreement with this sentiment is more apparent in the new version of the paper.

Another (mildly) irritating feature is the existence of an explicit "chunk inventory". One of the main criticisms of PARSER was that it, too, kept an explicit "chunk memory". The authors write that "The only information that the model tracks and stores are the discovered chunks, the BTPs between words, and the BTPs between discovered chunks". That is A LOT of explicitly stored information! And how does this work, exactly? Is there an explicit rule like: if the BTP at this point is 0.5 of the previous two (or average...) BTPs, then this must be a chunk boundary." But this would lead to all-or-nothing chunks. Is that reasonable? Chunks, in real life, are *graded*: chunks like "cupboard" or "football" are far more chunked than, say, "sunburn", which in turn is more chunked than "smartphone" or "petshop". In highly chunked items, you are unaware of the components; in chunks that are nascent, you still hear them. So, does CBL also store how strong a chunk is, along with all of the other information? Also, in order to keep track of the "BTPs between words", this is problematic. The chunk "dog" can be preceded by a whole lot of words ("a", "the", "big", "mean", "my", "your", "little", "yappy", "gentle", "and", "or", etc., etc.) does CBL really keep track of all of these BTPs. And surely, children from 1 to 4 would differ in their ability to do so. Problems like this need explaining.

> We think these criticisms of the CBL are both insightful and fair. The model depends on a highly simplified and idealized version of reality. Anyone who is putting the CBL forth as the premiere model for chunking/segmentation (that is, not us) should address these points. In the revised manuscript we have done what we can to bring attention to some of its limitations in the added text in the Discussion.

Also, it would seem like reconstruction above the word level requires some form a co-occurrence memory. The logic of "N unique chunks can be reconstructed in N! different orders", while mathematically accurate, is mostly a straw-man. The point is that if chunk-based models are ever going to do grammar, even elementary grammar, forward prediction and co-occurrence would seem to necessarily enter the picture. I would like to see this point discussed in some detail. In short, what are the limits of the CBL approach? This is not dealt with at present.

> We have made sure that the added Discussion text mentions some of these limitations of the CBL approach and also highlights other SL approaches that could feasibly be used to assess reconstruction accuracy over age.

This paper, while technically accurate, is too narrow. Yes, the results and simulations are reasonable, but they are far too limited in scope. This work absolutely needs to be fit into the bigger picture of models of SL in children. In other words, the fact is that other models exist and have been applied to child-language acquisition, but virtually no mention is made of any of them in this paper. If one were to read this paper naively, one would think that CBL is the only show in town, which is far from the case. Further, major issues need to be discussed in some detail, as discussed above.

> We again thank Reviewer 3 exposing us to other work in this domain. Their expert perspective has helped us gain a much broader *and* deeper view of current issues in the field. We have done what we can to include references to this other work at the start and finish of the paper and to discuss some limitations of the CBL approach. It should now be clear to readers that the CBL is not "the only show in town", simply the place where we started in doing this work. Model comparison is not our objective in the present paper; we leave it to future work to create parallel tests of age-invariance using longitudinal data with other models.

Modeling the influence of language input statistics on children's speech production

Abstract

We trained a computational model (the Chunk Based Learner; CBL) on a longitudinal corpus of child-caregiver interactions to test whether one proposed statistical learning mechanism—backward transitional probability (BTP)—is able to predict children's speech productions with stable accuracy throughout the first few years of development. We predicted that the model less accurately generates children's speech productions as they grow older because children gradually begin to generate speech using abstracted forms rather than specific "chunks" from their speech environment. To test this idea, we trained the model on both recently encountered and cumulative speech input from a longitudinal child language corpus. We then assessed whether the model could accurately reconstruct children's speech. Controlling for utterance length and the presence of duplicate chunks, we found no evidence that the CBL becomes less accurate in its ability to reconstruct children's speech with age. Our findings suggest that BTP **may be** an age-invariant learning mechanism.

Keywords: statistical learning, language learning, abstraction, developmental trajectory, age-invariance, CHILDES, children

Word count: **8726 (7027, excluding references and abstract)**

Modeling the influence of language input statistics on children's speech production

During the first few years of life children learn the basic building blocks of the language(s) around them. One way they do so is via statistical learning (SL), the process of extracting regularities present in the language environment. Over the past few decades, SL has become a major topic in the field of first language acquisition, ranging in application from speech segmentation (Jusczyk & Aslin, 1995; Saffran, Aslin, & Newport, 1996) and phonotactic learning (Chambers, Onishi, & Fisher, 2003) to producing irregulars (Arnon & Clark, 2011), discovering multi-word structures (Bannard, Lieven, & Tomasello, 2009; Chang, Lieven, & Tomasello, 2006; Frost, Monaghan, & Christiansen, 2019), and much more (see Saffran and Kirkham (2018) for a recent review). By its nature, work in this domain is heavily concerned with at least two major topics: (1) the information available in children's language environments (the 'input') from which they can pick up on patterns, and (2) the precise mechanisms by which children convert these 'raw' environmental statistics into internalized knowledge about language. A third issue is whether and how children's SL behavior changes as they develop (Shufaniya & Arnon, 2018). The current paper taps into each of these three issues: we train a computational model on a longitudinal corpus of child-caregiver interactions to test whether one proposed SL mechanism—backward transitional probability (BTP**; Perruchet & Desaulty, 2008**)—is able to predict children's speech productions with stable accuracy as they get older.

SL over development

The ability to detect and store patterns in the environment begins in infancy (e.g., Johnson et al., 2009; Kidd, Junge, Spokes, Morrison, & Cutler, 2018; Saffran et al., 1996; Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009), continues into adulthood (e.g., Conway,

Bauernschmidt, Huang, & Pisoni, 2010; Frost & Monaghan, 2016; Saffran, Johnson, Aslin, & Newport, 1999), and crosses a range of modalities (Conway & Christiansen, 2005; Emberson, Conway, & Christiansen, 2011; Monroy, Gerson, & Hunnius, 2017). However, it is still a matter of debate whether SL is an age-invariant skill or not (Arciuli & Simpson, 2011; Raviv & Arnon, 2018; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Shufaniya & Arnon, 2018). Recent work that investigates SL abilities in 5–12-year-old children suggests that, while both visual and auditory SL improve with age for non-linguistic stimuli, performance stays the same across childhood for linguistic stimuli (Raviv & Arnon, 2018; Shufaniya & Arnon, 2018). From this finding, the authors conclude that SL for language might be age-invariant. On the other hand, infant SL abilities do appear to shift within the first year, both for linguistic (Kidd et al., 2018) and non-linguistic (Johnson et al., 2009) stimuli. For example, while 11-month-olds can detect and generalize over regularities in a sequence, 8-month-olds are only capable of detecting the regularities, and neither group succeeds yet at learning visual non-adjacent dependencies (Johnson et al., (2009); see also Bulf, Johnson, & Valenza, 2011, and Slone & Johnson, 2015).

These changes in SL **ability** during infancy and early childhood may relate to changes in other fundamental cognitive skills. For example, SL-relevant brain regions, such as the pre-frontal cortex, continue maturing through childhood (Casey, Giedd, & Thomas, 2000; Diamond, 2002; Rodríguez-Fornells, Cunillera, Mestres-Missé, & Diego-Balaguer, 2009; Uylings, 2006), which may change how children attend to the linguistic information around them as they get older. Similarly, infants' long-term memory continuously improves between ages 0;2 and 1;6 (Bauer, 2005; Wojcik, 2013). Therefore, the manner in which they store linguistic regularities in long-term memory may also shift during this period. Relatedly, working memory and speed of processing change continuously throughout early childhood (Gathercole, Pickering, Ambridge, &

Wearing, 2004; Kail, 1991), implying that there could be a developmental change in the rate and scale at which children can process chunks of information from the unfolding speech signal.

Continued exposure to linguistic input itself can also be an impetus for change in SL **ability**—a view supported by multiple, theoretically distinct, approaches to early syntactic learning. For example, Yang (2016) proposes that children gather detailed, exemplar-based statistical evidence until it is more cognitively efficient for them to make a categorical abstract generalization. He proposes that, at that point, the learner instantiates a rule to account for patterns in the data. Usage-based theories of early language development **alternately** propose that children first learn small concrete linguistic sequences from their input that are made up of specific words or word combinations (e.g., 'dog' and 'I wanna'; or multi-word combinations, 'where's the …'; Tomasello, 2008). Then, over time, children are proposed to form abstract schemas centered on lexical items (see also Bannard et al., 2009, and Chang et al., 2006). This representational shift, from probabilistic and lexical to abstract and syntactic, is used to account for how children can eventually create utterances that they have never heard before. Crucially, the representational shift implies a change in the way children apply the original SL mechanism(s) to incoming linguistic information (see also Lany & Gómez, 2008).

Change in SL **ability** following further linguistic experience is also predicted in models that do not assume abstraction. In chunk-based models of language learning (Arnon, McCauley, & Christiansen, 2017; Christiansen & Arnon, 2017; Christiansen & Chater, 2016; Misyak, Goldstein, & Christiansen, 2012; StClair, Monaghan, & Christiansen, 2010), children use statistical dependencies in the language input (e.g., between words or syllables) to store chunks of co-occurring forms. Dependencies between the chunks themselves can also be tracked **with** continued exposure and chunk storage **(see, e.g., Jost & Christiansen, 2016). In this case, the**

development of a detailed chunk inventory can gradually change overt SL performance. Fundamentally, however, this apparent change in SL still comes through the use of the original underlying mechanisms (Misyak et al., 2012); there is no qualitative change in how the system processes data, and the mechanisms for processing, storing, and deploying information stay the same.

We investigated the possibility of developmental change in SL using computational modeling, which enables us to define and test the goodness-of-fit for any given learning mechanism on a dataset of natural speech. We chose to use a longitudinal child language dataset, in which the same children were tracked across the developmental period of interest for early speech production (1;0–4;0). By choosing data in this age range, we could test whether use of a learning mechanism changed for each child across the studied developmental time points. We tested for developmental change in the use of a single proposed statistical learning mechanism: backward transitional probability (McCauley & Christiansen, 2011; Onnis & Thiessen, 2013; Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008).

BTP and the Chunk-Based Learner

Our model is based on McCauley and Christiansen's (2011, 2014a) Chunk-Based Learner (CBL) model, which uses one measure—backward transitional probability (BTP; Perruchet & Desaulty, 2008)—to detect statistical dependencies in the speech stream. Backward transitional probability is one of multiple approaches for dividing streams of continuous speech into meaningful units; other approaches include, for example, forward transitional probability and memory-based chunking (Aslin, Saffran, & Newport, 1998; Cleeremans & Elman, 1993; French, Addyman, & Mareschal, 2011; Mareschal & French, 2017; Onnis & Thiessen, 2013; Pelucchi et al., 2009; Perruchet & Desaulty, 2008; Perruchet & Vinter,

1998; Saffran et al., 1996). BTP for a given pair of words is defined as the occurrence probability of the previous word (w_{-1}) given the current word (w_0). It can be estimated for each word in a sentence in order to reveal peaks and dips in transitional likelihood, which reflect places where words are likely (peaks) or unlikely (dips) to co-occur.

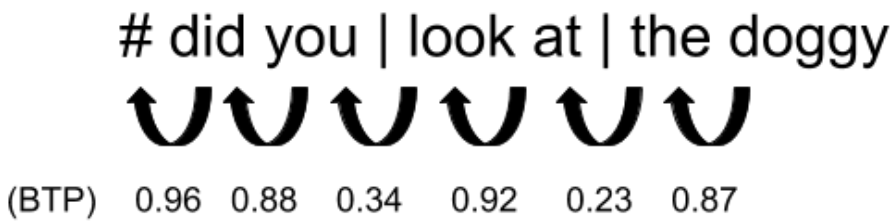

Figure 1: Example of a sentence with BTP between consecutive words. Chunks are split at points of low BTP (indicated by the vertical lines). "#" denotes a start-of-utterance marker.

The CBL model divides utterances into chunks, splitting the utterances whenever the BTP between two words drops below the running average BTP. In the example in Figure 1, the CBL might decide to split the sentence ("did you look at the doggy") into three chunks 'did you', 'look at', and 'the doggy', and store all three in its memory. As it sees more sentences, it would continue to add new chunks and track how often they co-occurred. **Once stored in memory, the chunks are not forgotten.**

The CBL was developed to model children's early speech production and comprehension without appealing to abstract grammatical categories. Specifically, it was designed as an implementation of the hypothesis that children detect and store multi-word chunks using BTP, and also use the stored chunks to parse speech and produce new utterances (see also Arnon & Snider, 2010, and Bannard & Matthews, 2008). The model's ability to simulate learning can be

measured by first training it on what children hear and then having the model reproduce what children say from the chunks that it learned.

We chose to build on the CBL model because it has successfully accounted for production data in multiple corpora, including **child language** datasets. For example: (a) it parsed text better than a shallow parser in three different languages (English, German, and French) when using individual words rather than word classes, (b) it was able to recreate up to 60% of child utterance productions in 13 different languages, and (c) it closely replicated data from an artificial grammar learning study (McCauley & Christiansen, 2011; Saffran, 2002). The model has also been able to replicate experimental data on children's multi-word utterance repetitions (Bannard & Matthews, 2008), over-regularization of irregular plural nouns (Arnon & Clark, 2011), and L2-learner speech (see also McCauley & Christiansen, 2014b, 2017). In sum, the CBL model appears to robustly predict the word-chunk patterns in children's speech when given information about what they hear in their input. **We extend this work by testing how the model performs with longitudinal data; it is not yet known how well it functions as a predictor of what children can say as they become more linguistically sophisticated.**

Testing for change with age

Following McCauley and colleagues (2011, 2014a, 2019) we tested the CBL model's ability to learn language by checking how well it can reconstruct children's utterances from the chunks discovered in their caregivers' speech. As we are interested in developmental change over the first three years of speech production, we analyzed the model's reconstruction ability with two measures:

- "Uncorrected": The binary (success/fail) reconstruction score originally used by McCauley and colleagues (2011, 2014a, 2019).
- "Corrected": A length-and-repetition-controlled reconstruction score that accounts for the fact that longer utterances have more opportunities for error reconstruction, and for the fact that some child utterances contain repetitions of **chunks**, making multiple reconstructions match the original utterance.

If BTP is an age-invariant mechanism, it should apply equally well across age. However, because children's utterances get longer as they get older, we would expect age invariance to only hold when we correct for utterance length. We therefore **test** for age invariance both with the original binary ("uncorrected") reconstruction score and a new ("corrected") score we **propose** to account for utterance length and word repetitions. If we find age-invariance, even while controlling for utterance length and word repetitions, it would strongly suggest that the mechanism is stable over the first three years of speech production and not simply influenced by other factors, e.g., utterance length. If not, it would suggest that use of the mechanism, in fact, changes with age (Bannard et al., 2009; Tomasello, 2005; Yang, 2016).

Predictions

With these previous findings as a starting point, we investigated whether the CBL could account for child speech production with equal precision over the first four years of life. **Taking for granted that children *eventually* develop abstract representations (e.g., as in Tomasello, 2008; Yang, 2016), w**e predicted that:

- The CBL would less accurately generate children's speech productions as they grew older; given the assumption that children gradually learn to abstract over the specific "chunks"

they encounter (Bannard et al., 2009; Tomasello, 2005; Yang, 2016) and, therefore, their speech should less often directly mirror their linguistic input at later ages. This finding would indicate that the immediate influence of children's language input statistics on their speech production decreases across development.

- Children will be more likely to use words that are not documented in the caregiver speech as they get older. These words could originate from other sources, such as peer speech or non-recorded caregiver speech (Hoff, 2010; Hoff-Ginsberg & Krueger, 1991; Mannle, Barton, & Tomasello, 1992; Roy, Frank, & Roy, 2009).
- Younger children's utterances would be reconstructed well on the basis of recently heard speech alone, whereas older children's utterances would be best constructed when considering a longer period of their historical input. Our reasoning was that older children's increased memory capacity (Bauer, 2005; Gathercole et al., 2004; Wojcik, 2013) allows them to draw on older input more easily in producing speech. If so, the findings would suggest that memory plays a critical role in the use of the same learning mechanism with age.

In sum, we expected to find that the CBL's ability to reconstruct children's speech decreases in-line with a concomitant increase in children's linguistic sophistication; an effect driven by children's use of more abstracted representations, words from other speech sources, and their increased ability to use historical input.

Methods

Model

The CBL model (McCauley & Christiansen, 2011) is an incremental computational model of language acquisition, that explores the possibility that infants and children parse their input into (multi-word) chunks during the process of acquiring language.

The model takes transcribed speech as input and divides the transcribed utterances into multi-word chunks. Each utterance begins with a start cue (denoted "#"). The exact placement of a chunk boundary within an utterance is determined by two factors: (1) the backward transitional probability (BTP) between consecutive words in the utterance, and (2) the inventory of already-discovered chunks. All newly discovered chunks are saved into the inventory, alongside the BTPs associated with each chunk. **The model tracks and stores:** the discovered chunks, the BTPs between words, and the BTPs between discovered chunks. For example, the model might parse the utterances "I see the **puppy**" and "did you look at the **puppy**?" into five different chunks, namely "I", "see", "the **puppy**", "did you", and "look at" based on the BTPs between these words compared to the average BTP found in the corpus so far.

Child utterance reconstruction task

Once the model has been trained on adult utterances, and thereby has discovered chunks in the adults' speech, we can test whether it closely matches the linguistic structures produced by the children in the same caregiver-child corpus. Following McCauley and Christiansen (2011), we use a child utterance reconstruction task to test whether the chunk statistics present in the adults' utterances are also present in the child's utterances. The model reconstructs the child utterances from the chunks and their related BTPs from the adult's utterances at the same age

point. This reconstruction process, which is slightly different from McCauley and Christiansen's (2011) process, is done in two steps (see Figure 2). First, a child utterance is converted into an unordered bag-of-chunks containing the set of largest possible chunks that had already been seen in the adults' speech, in line with the bag-of-words approach proposed in Chang, Lieven, and Tomasello (2008). Whenever the model encounters a word in the child utterance that is not present in the adult-based chunk inventory, it stops processing that utterance.[1] For instance, in the toy example in Figure 2, the child utterance "look at the puppy" would be broken down into a set of known chunks which were discovered in the adults' speech (e.g., "look at" and "the puppy", as in the step 2 speech bubble). If the utterance were "look at the poodle", and the model had not already added a chunk for the word "poodle" during training, then the word is unknown to the model and the utterance cannot be reconstructed; therefore the utterance would be rejected from further processing. However, in the case that the utterance *can* be broken down into known chunks, the model then tries to reconstruct the utterance by shuffling the chunks detected and reordering them based on their known transitional probabilities: the model begins with the utterance start cue and then finds the chunk that has the highest backwards transitional probability to the start cue, following that first chunk with the next one, which will be the remaining chunk with the highest backwards transitional probability to the first chunk, and again and again, until the set of chunks for that utterance is exhausted. So, the set of chunks "look at" and "the puppy"

---

[1] McCauley and Christiansen (2011) handle these cases differently. Our CBL implementation is identical to theirs up to this point. Therefore we also provide sentence reconstruction scores using their original method in the Supplementary Materials.

would be ordered depending on **the chunk that maximizes the BTP of the start cue (i.e., "look at"), followed by the chunk that maximizes the BTP of** "look at" (**i.e.,** "the **puppy**").

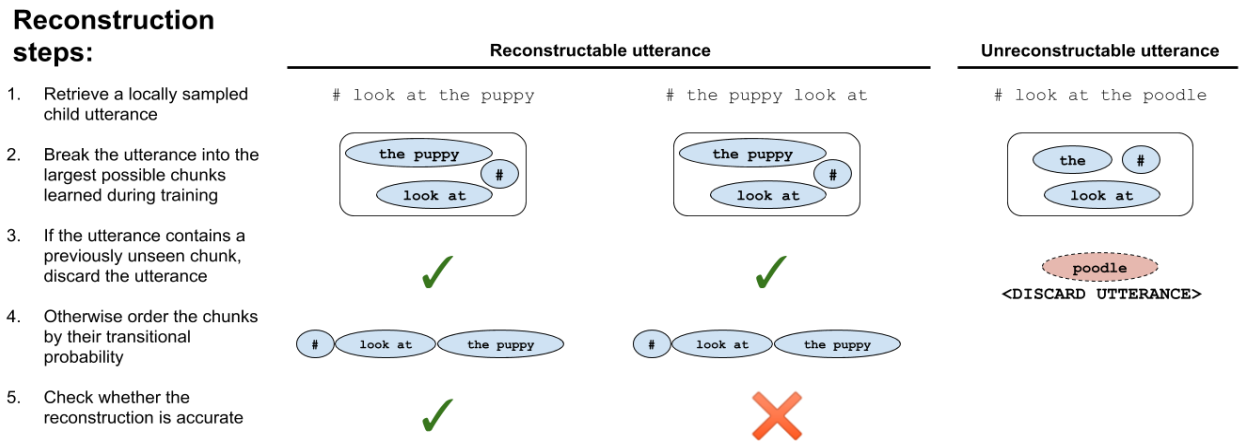

Figure 2: **Example of reconstruction attempts for three child utterances. The model tries to reconstruct the first two utterances using transitional probabilities of the chunks it finds, but it cannot do so with the third utterance, which contains a word ("poodle") that had not been previously seen during training.**

Materials and Procedure

As input to the model we used transcripts of 1–2-hour recordings of at-home interaction between six North American children and their caregivers who were recorded approximately every two weeks between ages 1;0 and 4;0 (the Providence corpus; Demuth, Culbertson, & Alter, 2006). We pre-processed the transcripts, which were formatted using CHAT conventions (MacWhinney, 2000), such that the input to the model only contained plain text orthographic

transcriptions of what was said.[2] We split the transcripts into two separate files, one with all the caregivers' utterances and one with all the child's utterances. Our pre-processing also added a "#" prefix to the start of each utterance.

The transcripts were sampled at approximate 6-month intervals between ages 1;0 and 4;0. We used two different sampling methods: a local data sampling method and a cumulative data sampling method. With the local data sampling method we selected data within a two-month interval around each age point. For example, for age point 1;6 we selected transcripts in which the child was between 1;5.0 and 1;6.31 years of age. This method led to ~800–4000 caregiver utterances at each age point. By design, the local sampling method focuses the model's training solely on *recent* linguistic input so that, when it tries to reconstruct children's utterances, the result is a test of how closely their current speech environment can account for what they say. **We sample *around* target age points and not *up-until* target age points because, while the Providence corpus is relatively densely sampled, recording sessions weren't frequent enough to guarantee a representative picture of each child's input in the month preceding each of the target age points. For this reason, we decided that training the model on input proximal to the tested age was a better method for getting a broad, but age-specific model of adult speech for each child at each age point.**

[2] All punctuation marks, grammatical notes, omitted word annotations, shortenings, and assimilations were removed from the utterances, such that only the text representing the spoken words of the utterance remained.

In contrast, the cumulative sampling method focuses the model's training on all previously heard linguistic input so that, when it tries to reconstruct children's utterances, the result is a test of how closely their current and previous speech environments can account for what they say. For the cumulative sample we selected data for each age point by taking all the available transcripts up to that age point. For example, for age 1;6 we selected all transcripts in which the child was 1;6 or younger. This method led to ~800–60,000 caregiver utterances across the different age points, with the number of caregiver utterances increasing (i.e., accumulating) with child age. As a consequence, the cumulative sample always contained more caregiver utterances than the local sample, except at age 1;0, the first sampled age point.

While we used two different sampling methods for training the model on adult data, all child utterances used for the reconstruction task were retrieved using the local sampling method for that particular age point. In other words, we only reconstructed the child utterances local to each tested age, regardless of the training strategy.

Analysis

We modeled two primary scores related to utterance reconstruction: the uncorrected (binary: success/fail) reconstruction score used by McCauley and colleagues (2011, 2014a, 2019) and the corrected reconstruction score we introduce in the current paper. The uncorrected reconstruction score (1: success, 0: fail) was computed for all child utterances that could be decomposed into previously seen chunks (see steps 4 and 5 in Figure 2). The corrected reconstruction score (defined below) was computed for the same set of utterances. We additionally included a third analysis: the likelihood that **a word encountered during the reconstruction task was not seen during training; utterances with unseen words**, by our version of the CBL, cannot be reconstructed (see step 3 in Figure Figure 2).

We used mixed-effects regression to analyze the effect of child age on both of the reconstruction scores and also whether **a word encountered during the reconstruction task was not encountered during training. All mixed-effects models included child age as a fixed effect and by-child random intercepts with random slopes of child age. By default, child age was modeled in years (1–4) so that the intercept reflects a developmental trajectory beginning at age 0. However, for the model of corrected reconstruction accuracy we had the additional advantage of being able to test whether the CBL performance significantly exceeded the baseline chance of correct reconstruction. We tested this difference at the average age in our longitudinal dataset (2;6) by centering age on zero in the statistical model (ages 1;0–4;0 are re-coded numerically as -1.5–1.5) such that the default model output would reflect the estimated difference from chance at the middle point of our age range.**

All analyses were conducted using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) and all figures were generated with the `ggplot2` package in `R` (R Core Team, 2014; Wickham, 2009). All code used to create the model and analyze its output is available at https://osf.io/ca8ts/. **Full tables of statistical model output are available in the Supplementary Materials.** Before turning to the main results we briefly describe the corrected reconstruction score and the analysis of previously unseen words in more detail.

Corrected reconstruction accuracy

The corrected, length-and-repetition-controlled reconstruction score is a function of three factors: (a) whether the model successfully reconstructed the child utterance or not, (b) the number of chunks used to reconstruct the utterance, and (c) the number of duplicate chunks involved in the reconstruction. By taking the number of chunks into account, this reconstruction

score compensates for the fact that successful reconstruction is less likely for longer utterances. When an utterance contains duplicate chunks, the exact ordering of those duplicate chunks does not influence the correctness of the reconstruction. For example, if the utterance "I wanna, I wanna" is decomposed into the two chunks "I wanna" and "I wanna", it does not matter which of the two "I wanna" chunks is placed first when calculating the reconstruction accuracy of the utterance. Thus, utterances containing duplicate chunks are more likely to be reconstructed by chance alone than utterances with the same number of chunks **but** no duplicates. **Note that here we are detecting duplicate *chunks* in the utterance rather than duplicate *words*. At this post-training stage, the model is only able to parse the utterance into chunks; that is the relevant unit over which duplication may affect reconstruction accuracy.**

An utterance that is decomposed into N unique chunks can be reconstructed in $N!$ different orders. Hence, the **baseline probability of obtaining the correct order of N unique chunks equals** $1/N!$. When we take into account that chunks can be repeated within an utterance, chance level equals $(n_1!\, n_2!\, \ldots n_k!)/N!$, where N is the total number of chunks in the utterance, and $n_1, \ldots, n_k$ are the number of times a chunk is repeated for each of the k unique chunks found in the utterance (Figure 3).

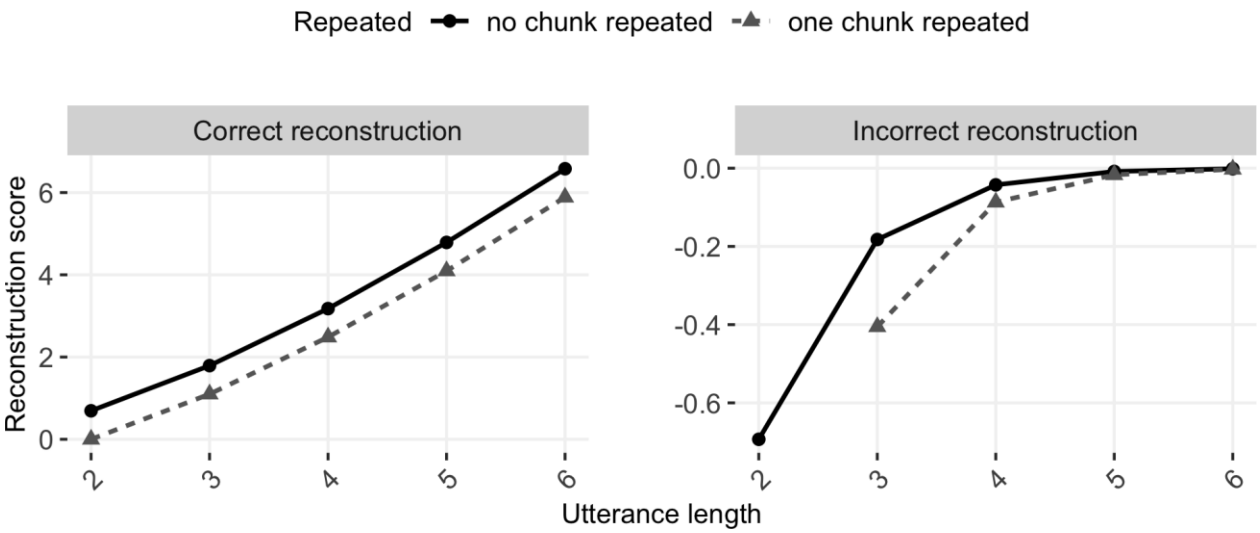

Figure 3: **Corrected reconstruction score for correct (left; positive values) and incorrect (right; negative values) reconstructions, as a function of utterance length (2–6 chunks). In this example, either no chunks are repeated (black/solid lines) or one chunk occurs twice in the utterance (gray/dashed lines).**

When probability of reconstruction was lower, we scored a correctly reconstructed utterance higher. We assigned a score of $-\log(chance)$ for each correct reconstruction and $\log(1 - chance)$ for each incorrectly reconstructed utterance. In layman's terms, this means that successfully reconstructed utterances were scored positively, but were weighed relative to the number of chunks and the number of repetitions they had, such that reconstructions of long utterances were given higher scores than reconstructions of short utterances. Along the same lines, incorrectly reconstructed utterances were scored negatively and were also weighed relative to the number of chunks they had, such that incorrect reconstructions of long utterances were given higher (i.e., less negative) scores than incorrect reconstructions of short utterances

To illustrate the corrected scoring method, let's compare two three-chunk utterances, one of which contains a duplicate chunk: "I wanna I wanna see" (chunks: "I wanna", "I wanna",

"see") and "I wanna see that" (chunks: "I wanna", "see", "that"). For the first utterance, chance level equals $(2! \times 1!)/(3!)$: The numerator is determined by the number of times each unique chunk is used, so because "I wanna" occurs two times and "see" occurs once, that is $2! \times 1!$. The denominator is determined by the factorial of total number of chunks (here: $3! = 3 \times 2 \times 1$). The resulting chance level is then $2/6$. For the second utterance, chance level equals $(1! \times 1! \times 1!)/(3!)$: The numerator is equal to $1! \times 1! \times 1!$ here because all chunks occur only once in the utterance. The denominator is the same as for the first utterance as the total number of chunks in the utterance is the same. Here, the resulting chance level is $1/6$. If the utterances are reconstructed correctly, the score is computed by $-\log(chance)$. So, the first utterance would get a positive score of $-\log(chance) = -\log(2/6) \approx 1.098$ and the second utterance would get a higher positive score of $-\log(chance) = -\log(1/6) \approx 1.791$ for increased reconstruction difficulty. If the utterances are reconstructed incorrectly, the score is computed by $\log(1 - chance)$. Thus, the first utterance would get a negative score of $\log(1 - chance) = \log(1 - (2/6)) \approx -0.405$ and the second utterance would get a less negative score of $\log(1 - chance) = \log(1 - (1/6)) \approx -0.182$.

Previously unseen words

Our third analysis focused on **the likelihood that words used in the child utterances were seen during training, given child age and sampling type. To prepare for this analysis we marked each word used by each child at each age point as having been seen during training (1) or not (0), given local and cumulative sampling.**

Results

Uncorrected reconstruction accuracy

The uncorrected score of accurate utterance reconstruction (McCauley & Christiansen, 2011, 2014a) showed that **the** model's average percentage of correctly reconstructed utterances across children and age points was similar for the locally and cumulatively sampled speech (local: mean = 65.4%, range across children = 59.9%–70.3%; cumulative: mean = 59.9%, range across children = 53.1%–68.2%). This is similar to, or slightly higher than, results reported by McCauley and Christiansen (2011) who found an average percentage of correctly reconstructed utterances of 59.8% over 13 typologically different languages with a mean age range of 1;8–3;6 years. Additionally, McCauley and Christiansen (2019) reported an average reconstruction percentage of 55.3% for 160 single-child corpora of 29 typologically different languages, including a performance of 58.5% for 43 English single-child corpora with a mean age range of 1;11–3;10.

In our statistical models of the uncorrected reconstruction accuracy[3], we first analyzed the CBL model's performance when it was trained on locally sampled caregiver speech. The number of correctly reconstructed utterances decreased with age ($b = -0.805, SE = 0.180, p < 0.001$); over time: The BTP statistics present in the caregivers' speech were less reflected in the child's own speech (Figure 4, left panel)**; as we shall see, this decrease is related to the uncorrected reconstruction score**.

---

[3] **accuracy ~ age + (age|child), family = binomial(link = 'logit').**

We then tested the model's performance when it was trained with a cumulative sample of caregiver speech, rather than just a local sample. As before, the number of correctly reconstructed utterances decreased with child age ($b = -0.821, SE = 0.146, p < 0.001$; Figure 4, right panel). These results indicate age-variance for the SL mechanism; its utility for modeling children's utterances changes with age.

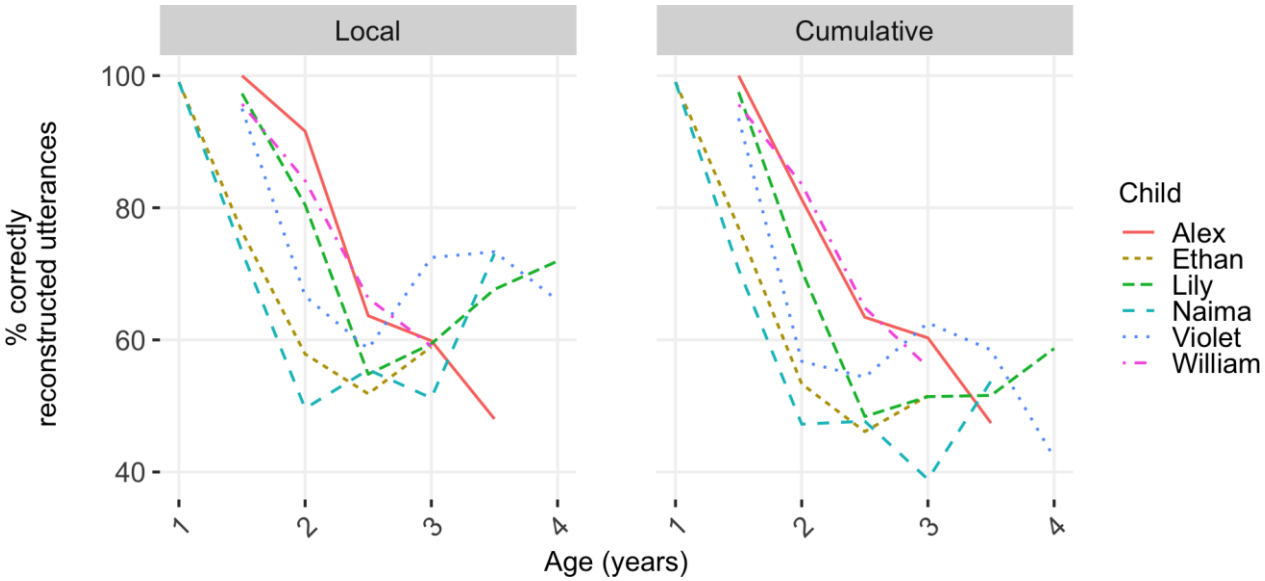

Figure 4: Percentage of correctly reconstructed utterances across the age range, using local (left) and cumulative (right) sampling.

Importantly, however, the length of the child utterances varied quite a lot (range = 1–44 words long; mean = 2.8, median = 2), and some of them contained repetitions of chunks (e.g., "I wanna, I wanna"), both of which influence the baseline probability of accurate reconstruction. Utterances from older children tended to contain more words than utterances from younger children (Figure 5, left panel). As a consequence, on average, utterances from older children are systematically less likely to be correctly reconstructed by chance, contributing to the decrease in the CBL's overall performance with age. Additionally, the percentage of child utterances that contained duplicate chunks decreased over time (Figure 5, right panel). Utterances with duplicate

chunks have a higher baseline probability of being accurately reconstructed by the model. So again, on average, utterances from older children were systematically more difficult, contributing to the age-related decrease in uncorrected reconstruction scores.

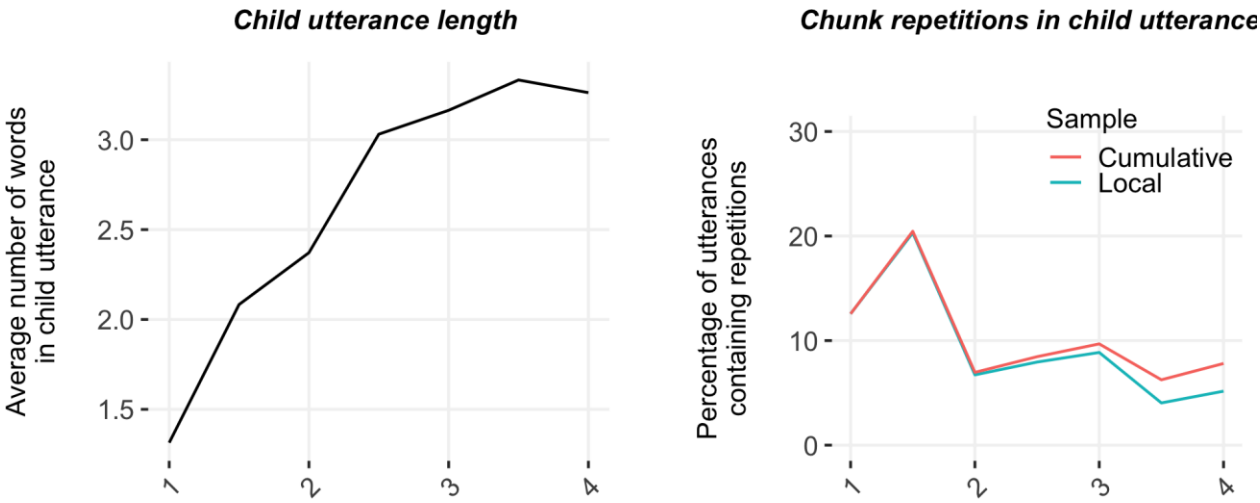

Figure 5: Children's utterances increased in length (number of words) with age (left) while simultaneously decreasing in the number of duplicate chunks used (right).

Corrected reconstruction accuracy

Next, we used our corrected reconstruction score to assess the model's reconstruction accuracy while controlling for utterance length and the use of duplicate chunks. As explained above, the corrected score weighs whether each utterance was accurately reconstructed against its chance level of reconstruction, depending on the total number of chunks and number of duplicate chunks it contains. The model's average reconstruction score across children and age points was similar for the locally and cumulatively sampled speech (local: mean = 0.10, SE = 0.01; cumulative: mean = 0.06, SE = 0.01). Note again that one aim of this analysis was to test whether the corrected reconstruction score was above chance—here represented by a score of zero—so in

the statistical models we centered child age on zero so that the estimation would reflect the difference from zero for the average age in our sample (2;6).[4]

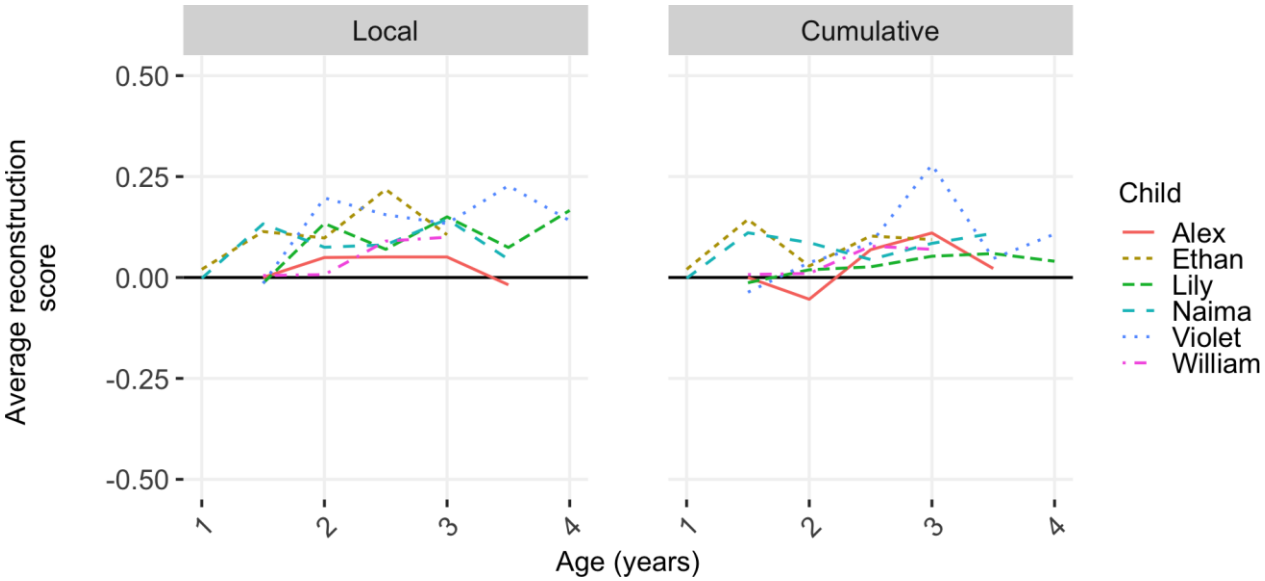

Figure 6: Corrected reconstruction scores across the age range, using local (left) and cumulative (right) sampling.

Again, we first analyzed the model's performance when it was trained on locally sampled caregiver speech. We found a significant positive intercept ($b = 0.11, SE = 0.02, t = 5.064$) and no significant change across age ($b = 0.030, SE = 0.018, t = 1.681$); the BTP statistics from the caregivers' speech were consistently reflected in the child's own speech (Figure 6, left panel).

As before, we created a parallel set of analyses to test the model's performance when it was trained with a cumulative sample of caregiver speech. We again found a significant positive intercept ($b = 0.06, SE = 0.010, t = 6.238$) and that accuracy did not change significantly across age ($b = 0.02, SE = 0.013, t = 1.590$; Figure 6, right panel).

---

[4] **accuracy ~ centered.age + (centered.age|child).**

In sum, contrary to the uncorrected reconstruction accuracy analysis, these corrected reconstruction score results indicate age-invariance for the SL mechanism. In addition, the model performed significantly above chance level in both the local and cumulative sampling contexts.

Children's use of unseen words

Utterances with words that were not encountered and stored as chunks during training were not included in the reconstruction task. We therefore also modeled whether child age and sampling type influenced the likelihood that a word in the child's speech had already been seen. For this analysis we compared the words used by each child at each age point to the words that *that* child had heard during training (local or cumulative), marking each word as having been seen during training (1) or not (0). For each sampling type, we then modeled the likelihood that a word was previously seen given a fixed effect of child age and random effect child with random slopes of child age.[5] With local sampling, words in the children's utterances were significantly less likely to have been previously seen as children got older ($b = -0.549, SE = 0.11, p < 0.001$; Figure 7, left panel). With cumulative sampling, this effect was neutralized; increasing age was associated with a small and non-significant decrease in the likelihood of previously seen words ($b = -0.022, SE = 0.121, p = 0.857$; Figure 7, right panel). By taking a longer history of linguistic input into account (i.e., by using cumulative sampling), **words that were not seen in the local sampling were indeed seen during cumulative sampling.**

[5] **prev_seen ~ age + (age|child) family = binomial(link = 'logit')**

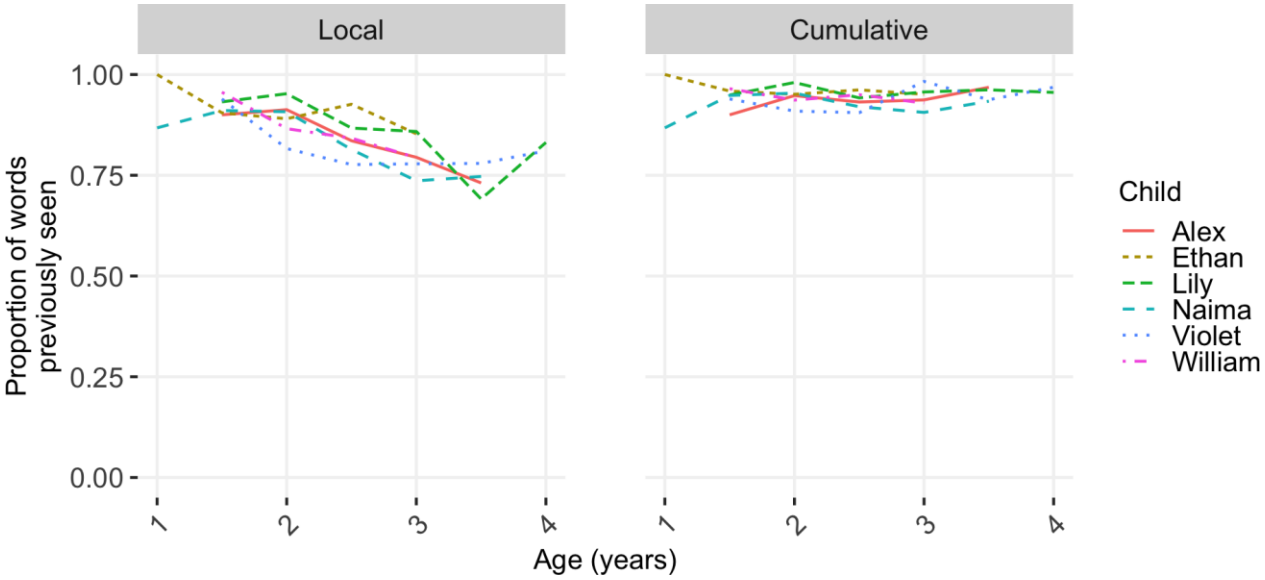

Figure 7: **Proportion of words in the local child utterances seen in the training data across age using local (left) and cumulative (right) sampling.**

Discussion

Our primary research question (as raised by, e.g., Arciuli & Simpson, 2011; Raviv & Arnon, 2018; Saffran et al., 1997; Shufaniya & Arnon, 2018) was whether the CBL **would change in its ability** to predict children's speech productions throughout development. We tested the model using both the original measure of accuracy as well as a new measure that takes into account utterance length and duplicate chunks in the utterance, which can make accurate reconstruction less likely (length) or more likely (duplicates). Using this corrected measure, we found that there was no significant change in the use of BTP with age. Notably, the CBL was able to construct utterances at above-chance levels despite these changes with age. Overall, **and against our predictions, the current** findings support the view that BTP is an age-invariant learning mechanism for speech production. In fact, the positive, but non-significant coefficients for the effect of age on corrected reconstruction accuracy indicate that, the CBL is, at least, not

getting worse at reconstructing children's utterances with age. **Also, the divergence in findings between the corrected and uncorrected accuracy scores illustrates how effects of length and chunk duplication can critically shift baseline performance during reconstruction; these features of natural speech should be controlled for in future work.**

Different words at different ages

We also analyzed the number of utterances with previously unseen words in them, arguing that older children's increased memory capacity (Bauer, 2005; Gathercole et al., 2004; Wojcik, 2013) would possibly allow them to draw upon older input more easily in producing speech. Indeed, we found an increase in the number of utterances containing previously unseen words with age in the local sample but a decrease when taking their longer linguistic history into account. The change in word usage we find here could be partly due to a change in linguistic input not captured in the transcripts. The corpus we used is relatively dense: multi-hour at-home recordings made approximately every two weeks for 2–3 years. However, this corpus still only contained a small fraction of what each child heard during the represented periods of time (i.e., 2 hours of ~200 waking hours in a fortnight). Non-recorded caregiver speech may contribute an increasing amount of lexical diversity. Consider, for example, that input from peers containing different lexical items could have increased as children became old enough to independently socialize with other children or attend daycare or preschool (Hoff, 2010; Hoff-Ginsberg & Krueger, 1991; Mannle et al., 1992), which may help to account for the increased presence of words not found in the caregiver's speech. This problem is difficult to address directly since, even with cutting-edge tools and significant supporting resources, it is still nearly impossible to collect and transcribe a child's complete language environment (Casillas & Cristia, 2019; Roy et al., 2009). This effect could instead be simulated in future work by feeding speech from other

children or adults into the model to mimic speech from peers and other caregivers. That said, our results showed that the likelihood of previously unseen words actually decreased with age for the cumulative sample, suggesting that the "missing" words *are* present in caregiver speech, just not always in the recently recorded input.

Additionally, an improvement in memory capacity with age provides a potential explanation for the current findings. Throughout childhood, including the first few years, SL-relevant cortical regions continue maturing (Casey et al., 2000; Diamond, 2002; Rodríguez-Fornells et al., 2009; Uylings, 2006) with concurrent increases in long-term memory (Bauer, 2005; Wojcik, 2013), working memory, and speed of processing (Gathercole et al., 2004; Kail, 1991). By ages three and four, the children in the current study may have been able to much more reliably draw upon information they were exposed to in the more distant past. If so, we would expect no significant increase in the use of previously unheard words as children get older with the cumulative sampling method—consistent with what we found here (Figure 7, right panel). This pattern of results indicates that children's developing memory could play an important role in the way they use environmental input statistics over age.

Abstraction and complex utterances

Our findings are not consistent with a representational shift toward abstraction during the early language learning process. For instance, if children schematized their constructions or switching to rule-based representations (Bannard et al., 2009; Tomasello, 2005; Yang, 2016), we would expect a decrease in reconstruction accuracy over time, given that the CBL's reconstructions are limited to the immediate statistics of the child's language environment. In contrast, we saw that the model's ability to reconstruct child utterances from caregivers' speech was age-invariant when taking into account utterance length and chunk duplicates. These results

do fall in line with SL theories proposing that the mechanisms for processing, storing, and deploying information **stay** constant over age, even though SL behavior on the surface might **seem** to change over time (e.g., Misyak et al., 2012).

As the CBL model only employs a single, simple mechanism for creating and tracking linguistic units, it is impressive that it performs at above-chance levels when accounting for children's speech productions in the first few years. If the mechanism is truly age-invariant, it should be able to handle both young children's speech and adults' speech; here we see that it handles the developing linguistic inventory of children ages 1;0 to 4;0, during which time children's utterances **become** much more sophisticated and much closer to adult-like form.

Going beyond the scope of this paper, a next step would be to explore how the CBL could be modified to augment its performance, particularly on more complex utterances. For example, the CBL model does not include the use of semantics when dividing the caregivers' speech into chunks or when reconstructing the child utterances. However, the meaning of what both caregiver and child are trying to convey plays a fundamental role in selecting words from the lexicon and in constructing utterances—they are interacting, and not just producing speech. The same set of words, ordered in different ways, can have entirely different meanings (e.g., "the dog bites the man" vs. "the man bites the dog"). Additionally, the CBL currently works on text-only transcriptions of conversations, but speech prosody could potentially critically change how children detect chunks. Prosodic structures within an utterance highlight syntactic structures and help to distinguish between pragmatic intentions, for example, distinguishing between questions, imperatives, and statements (e.g., Bernard & Gervain, 2012; Speer & Ito, 2009). Ideally, the CBL model would also be tested on a (more) complete corpus of what children hear in the first few years to further investigate the origins of the "previously unseen" words in children's utterances;

though we appreciate that densely sampled and transcribed collections of audio recordings are extremely costly to create (Casillas & Cristia, 2019; Roy et al., 2009).

In principle, the "next steps" proposed above—indeed the whole idea of analyzing chunking performance across developmental time—are not limited to the CBL, or even BTP. Rather, we make here a general call for dealing with richer data, regardless of the core underlying mechanism (Aslin et al., 1998; Cleeremans & Elman, 1993; French et al., 2011; Mareschal & French, 2017; Onnis & Thiessen, 2013; Pelucchi et al., 2009; Perruchet & Desaulty, 2008; Perruchet & Vinter, 1998; Saffran et al., 1996). In the current study, we decided to use the CBL because it had previously been successful in reconstructing children's utterances within our target age range (McCauley & Christiansen, 2011, 2014a, 2019) and had not yet been tested for age-invariance. However, given the required memory and comparison (executive function) components of this model, as well as its requirement of discrete (not gradable) chunks, other approaches—particularly those inspired by maturational neurocognitive development (Cleeremans & Elman, 1993; e.g., Mareschal & French, 2017; Perruchet & Vinter, 1998)—would be welcome comparisons to the present findings. Notably, while the CBL here performed above chance on average, there is still room to improve in modeling what the children said based on what they heard in the recordings.

Conclusion

In this study, we investigated whether the CBL model—a computational learner using one SL mechanism (BTP)—could account for children's speech production with equal accuracy across ages 1;0 to 4;0 given information about their speech input. **This work extended previous**

CBL studies by testing the robustness of utterance reconstruction across an age range featuring substantial grammatical development and while also introducing a new controlled accuracy measure for reconstruction. The model's ability to reconstruct children's utterances remained stable with age when controlling for utterance length and duplicate chunks, both when taking into account recent and cumulative linguistic experience. These findings suggest that this particular mechanism for segmenting and tracking chunks of speech may be age-invariant (Raviv & Arnon, 2018; Shufaniya & Arnon, 2018). A rich and growing literature on SL in development has demonstrated that similar mechanisms can account for much of children's early language behaviors; knowing whether the use of these mechanisms changes as children get older is a crucial piece of this puzzle. To explore this topic further, future work will have to address additional cues to linguistic structure and meaning, the density of data needed to get reliable input estimates, and the interaction of SL with other developing skills that also impact language learning.

Acknowledgements

<Retracted for review>

References

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science, 14*(3), 464–473.

Arnon, I., & Clark, E. V. (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.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

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.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

Bannard, C., & Matthews, D. (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.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284–17289.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bauer, P. J. (2005). Developments in declarative memory: Decreasing susceptibility to storage failure over the second year of life. *Psychological Science*, *16*(1), 41–47.

Bernard, C., & Gervain, J. (2012). Prosodic cues to word order: What level of representation? *Frontiers in Psychology*, *3*, 451.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132.

Casey, B. J., Giedd, J. N., & Thomas, K. M. (2000). Structural and functional brain development and its relation to cognitive development. *Biological Psychology*, *54*(1-3), 241–257.

Casillas, M., & Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra*, *5*(1), 24. doi:doi:10.1525/collabra.209

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77.

Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 154–159.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, *9*(3), 198–213.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551.

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

Cleeremans, A., & Elman, J. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT press.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24–39.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, *49*(2), 137–173. doi:doi:10.1177/00238309060490020201

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. Stuss & R. Knights (Eds.), *Principles of frontal lobe function* (pp. 466–503). New York: Oxford University Press.

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *The Quarterly Journal of Experimental Psychology*, *64*(5), 1021–1040.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614.

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: High frequency words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, Advance online publication.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177–190.

Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. *First Language*, *30*(3-4), 461–472.

Hoff-Ginsberg, E., & Krueger, W. M. (1991). Older siblings as conversational partners. *Merrill-Palmer Quarterly*, *37*(3), 465–481.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8-and 11-month-olds. *Infancy*, *14*(1), 2–18.

Jost, E., & Christiansen, M. H. (2016). Statistical learning as a domain-general mechanism of entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 227–244). Washington D.C.: Mouton de Gruyter.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, *109*(3), 490–501.

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual differences in snfant speech segmentation: Achieving the lexical shift. *Infancy*, *23*(6), 770–794.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247–1252.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Psychology Press.

Mannle, S., Barton, M., & Tomasello, M. (1992). Two-year-olds' conversations with their mothers and preschool-aged siblings. *First Language*, *12*(34), 57–71.

Mareschal, D., & French, R. M. (2017). TRACX2: A connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160057.

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1619–1624.

McCauley, S. M., & Christiansen, M. H. (2014a). Acquiring formulaic language: A computational model. *The Mental Lexicon*, *9*(3), 419–436.

McCauley, S. M., & Christiansen, M. H. (2014b). Reappraising lexical specificity in children's early syntactic combinations. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 1000–1005.

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*(3), 637–652.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Pyschological Review*, *126*(1), 1–51.

Misyak, J. B., Goldstein, M. H., & Christiansen, M. H. (2012). Statistical-sequential learning in development. *Statistical Learning and Language Acquisition*, 13–54.

Monroy, C. D., Gerson, S. A., & Hunnius, S. (2017). Toddlers' action prediction: Statistical learning of continuous action sequences. *Journal of Experimental Child Psychology*, *157*, 14–28.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284.

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

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*(4), e12593.

Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & Diego-Balaguer, R. de. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3711–3735.

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 2106–2111.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172–196.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181–203.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science*, *42*(8), 3100–3115.

Slone, L. K., & Johnson, S. P. (2015). Infants' statistical learning: 2-and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, *133*, 47–56.

Speer, S. R., & Ito, K. (2009). Prosody in first language acquisition–Acquiring intonation as a tool to organize information in conversation. *Language and Linguistics Compass*, *3*(1), 90–110.

StClair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*(3), 341–360.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, 21.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition* (1st ed.). Harvard University Press.

Tomasello, M. (2008). Acquiring linguistic constructions. In Damon W, R. Lerner, D. Kuhn, & R. Siegler (Eds.), *Child and adolescent development* (pp. 263–297). New York: Wiley.

Uylings, H. B. M. (2006). Development of the human cortex and the concept of "critical" or "sensitive" periods. *Language Learning*, *56*(1), 59–90.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer Publishing Company, Incorporated.

Wojcik, E. H. (2013). Remembering new words: Integrating early memory development into word learning. *Frontiers in Psychology*, *4*, 151.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language* (1st ed.). MIT Press.

Supplementary materials: Modeling the influence of language input statistics on children's speech production

Full model output of main text analyses

Below readers will find the full statistical model output for each of the six models reported in the main text (that is, one for local input and one for cumulative input for each of the three analyses: uncorrected likelihood of accurate reconstruction, corrected reconstruction accuracy, and likelihood of unseen words).

Full model output for the analysis of uncorrected reconstruction accuracy using local speech input.

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	2.59	0.50	5.19	0.00
fixed	NA	age	−0.81	0.18	−4.44	0.00
ran_pars	child	sd__(Intercept)	1.36	NA	NA	NA
ran_pars	child	sd__age	0.49	NA	NA	NA
ran_pars	child	cor__(Intercept).age	−0.97	NA	NA	NA

Full model output for the analysis of uncorrected reconstruction accuracy using cumulative speech input.

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	2.40	0.39	6.09	0.00
fixed	NA	age	−0.82	0.14	−5.83	0.00
ran_pars	child	sd__(Intercept)	1.18	NA	NA	NA
ran_pars	child	sd__age	0.41	NA	NA	NA
ran_pars	child	cor__(Intercept).age	−0.96	NA	NA	NA

Full model output for the analysis of corrected reconstruction accuracy using local speech input.

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	0.11	0.02	5.06	NA
fixed	NA	recentered_age	0.03	0.02	1.68	NA
ran_pars	child	sd__(Intercept)	0.05	NA	NA	NA
ran_pars	child	sd__recentered_age	0.04	NA	NA	NA
ran_pars	child	cor__(Intercept).recentered_age	0.63	NA	NA	NA
ran_pars	Residual	sd__Observation	0.63	NA	NA	NA

Full model output for the analysis of corrected reconstruction accuracy using cumulative speech input.

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	0.06	0.01	6.24	NA
fixed	NA	recentered_age	0.02	0.01	1.59	NA
ran_pars	child	sd__(Intercept)	0.02	NA	NA	NA
ran_pars	child	sd__recentered_age	0.03	NA	NA	NA
ran_pars	child	cor__(Intercept).recentered_age	−0.71	NA	NA	NA
ran_pars	Residual	sd__Observation	0.59	NA	NA	NA

Full model output for the analysis of likelihood that a word in the child's speech was seen during training.

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	3.13	0.33	9.64	0.00
fixed	NA	age	−0.55	0.11	−5.00	0.00
ran_pars	child	sd__(Intercept)	0.74	NA	NA	NA
ran_pars	child	sd__age	0.25	NA	NA	NA
ran_pars	child	cor__(Intercept).age	−0.94	NA	NA	NA

Full model output for the analysis of number of utterances with previously unheard words using cumulative speech input.

effect	group	term	estimate	std.error	statistic	p.value
fixed	NA	(Intercept)	2.85	0.36	7.82	0.00
fixed	NA	age	−0.02	0.12	−0.18	0.86
ran_pars	child	sd__(Intercept)	0.78	NA	NA	NA
ran_pars	child	sd__age	0.25	NA	NA	NA
ran_pars	child	cor__(Intercept).age	−0.96	NA	NA	NA

Results using the original McCauley and Christiansen (2011) reconstruction method

While our implementation of the CBL learner is identical to McCauley and Christiansen's, our implementation of the ***reconstruction task*** diverges slightly from **theirs: we discard utterances with unknown words and instead provide a second analysis focused on the number of these 'un-reconstructable' utterances across age. Our reasoning for discarding utterances with unknown words was that there is no obvious way to give them a valid default transition matrix with other existing chunks. In contrast, McCauley and Christiansen (2011) built a new chunk for each unknown word when reconstructing utterances. This chunk with the unknown word was then assigned a BTP equal to zero with respect to any other chunk in the utterance it originated from. In what follows,** we present results using McCauley and Christiansen's (2011) original reconstruction task method. **Because**

their reconstruction task attempts to reconstruct all utterances, we do not also provide analyses of the number of utterances containing unknown words.

We analyzed the effect of child age on the model's reconstruction abilities for the child utterances with a mixed-effects model, including age as a fixed effect and a by-child random intercept with random slopes of age.

First, we used the binary (1: reconstructed correctly, 0: not reconstructed correctly) measure from McCauley and Christiansen (2011, 2014). The model's average percentage of correctly reconstructed utterances across children and age points was similar for the locally and cumulatively sampled speech (local: mean = 61.3 %, range across children = 51.6%–69.6%; cumulative: mean = 59.4%, range across children = 50.8%–68.4%). The number of correctly reconstructed utterances decreased with age, regardless of the sampling methods (local: $b = -0.939, SE = 0.174, p < 0.001$; cumulative: $b = -0.848, SE = 0.138, p < 0.001$; see Figure 1).

Second, we used our corrected, length-and-repetition-controlled reconstruction score. The model's average reconstruction score across children and age points was similar for the locally and cumulatively sampled speech (local: mean = 0.12, SE = 0.01; cumulative: mean = 0.08, SE = 0.01). As in the main text, we centered age in the model so that we could investigate whether reconstruction was greater than chance level at the average age in our sample (2;6 years). Using both sampling methods, we found a significantly positive intercept (local sampling: $b = 0.130, SE = 0.016, t = 7.911$; cumulative sampling: $b = 0.0789, SE = 0.012, t = 6.426$), and the model's reconstruction score increased over age, significantly in the case of the cumulative sampling method (local sampling: $b = 0.029, SE = 0.016, t = 1.854$; cumulative sampling: $b =$

$0.031, SE = 0.013, t = 2.333$); see Figure 2). These results show that the model performed at above-chance levels, and indicates age-invariance with the corrected reconstruction score.

Importantly, these results are highly similar to those from our implementation of the CBL model in the main text, which does not attempt to reconstruct utterances with previously unseen words. These findings suggest that the CBL is not significantly impacted in its ability to reconstruct children's utterances in the first four years, regardless of the minor algorithmic differences in how new words are treated between the original and current CBL models.

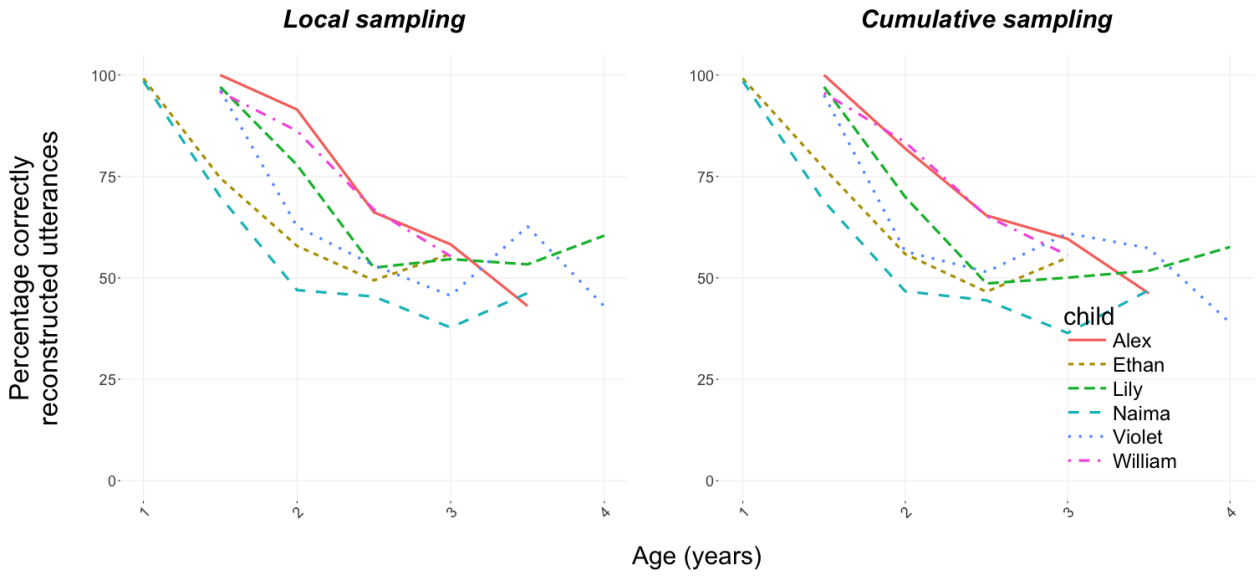

Figure 1: Uncorrected reconstruction scores across the analyzed age range for local (left) and cumulative (right) sampling while using McCauley and Christiansen's method for handling new words.

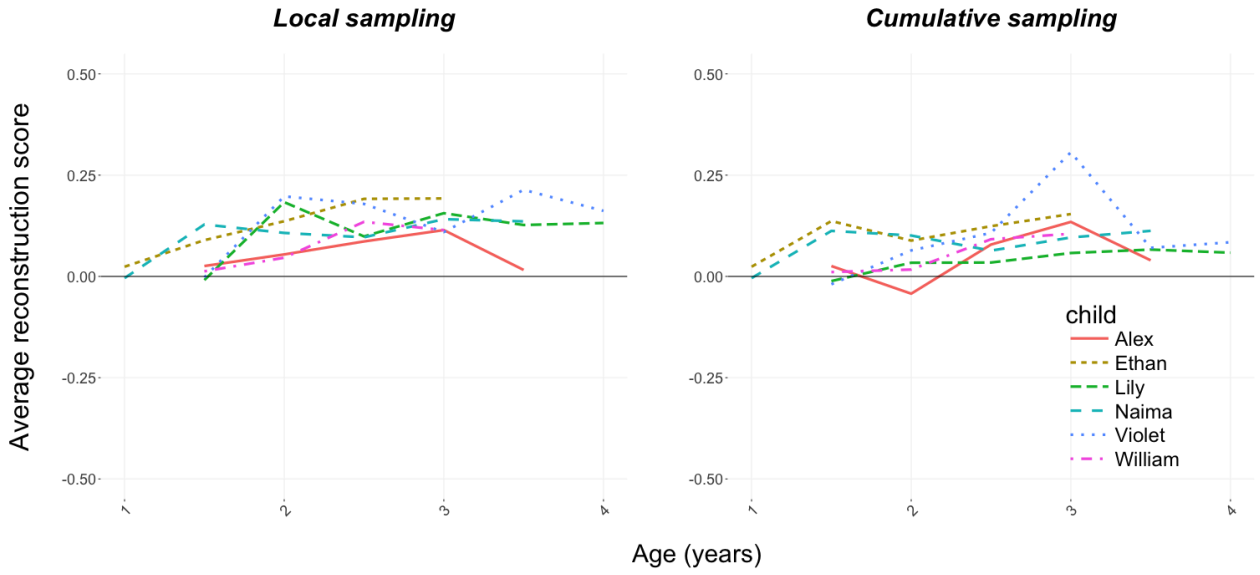

Figure 2: Corrected reconstruction scores across the analyzed age range for local (left) and cumulative (right) sampling while using McCauley and Christiansen's method for handling new words.

References

McCauley, S. M., & Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The CAPPUCCINO model. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1619–1624.

McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *The Mental Lexicon*, *9*(3), 419–436.