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.