14 August 2020

Dear Dr. Cooper and Dr. Bonawitz,

We submit to you a second revision of our manuscript entitled “Modeling the influence of language input statistics on children's speech production”.

We have reviewed and responded to the comments, as detailed below. The data and manuscript + analysis code (.Rmd) are posted on the same OSF repository as before: <https://osf.io/ca8ts/?view_only=daaa1bcc71654842b0d70efe785a26b9> (simply click on the “Resubmitted-August2020” folder and download as zip). We have taken special care to explain in much more detail why the CBL is an appropriate starting point for investigating developmental change in statistical learning. We have also added more information about the limitations of the model, particularly concerning the psychological plausibility of chunk and transitional probability storage. As requested, we have also added some speculation on how those limitations might be addressed in future work. Finally, we made another pass of the manuscript to ensure that readers understand that BTP and the CBL are far from the only statistical learning phenomena worthy of studying for developmental change.

The main text of the manuscript is now 7939 words long (excluding abstract, references, and supplementary materials) and includes the same 7 figures (Figure 6 is now modified in response to a recommendation by Rev1). As before, we have placed figures near where they are referred to in the text for ease of review. We also provide our (unchanged) Supplementary Materials, which contain full statistical model output and model results using an alternative formulation of the model.

We thank you and the reviewers for another chance to improve the manuscript. Please do not hesitate to contact us if there are any additional requests or questions.

Sincerely, The Authors

########## REVIEWS ##########

Dear Authors,

Thank you for your revised manuscript “Modeling the influence of language input statistics on children's speech production” to Cognitive Science. I have sent the manuscript out for review from the original three reviewers, who were all generous enough to provide a response to your revisions.

As you’ll see, both Reviewer 1 and 2 are very positive about the revised manuscript, however Reviewer 3 raises concerns that the paper remains too narrowly focused and without enough discussion of alternative accounts. Having read through the paper as well as the original reviews again myself, I have come to the conclusion that these issues can be overcome in a revision. The substantive issue as I see it is in making sure that the framing of the paper is broad enough to appeal to the larger Cognitive Science audience. That means a little more effort will be required in justifying why the authors have chosen to focus more exclusively on BTPs. This is not to say the whole paper must be rewritten to include a new motivation – but simply that more justification must be provided, with broader framing perhaps, to help pull outside readers into the importance of the work and the reasons for the current scope of the work.

> We appreciate the opportunity to clarify the more general appeal of our study for Cognitive Science readers. We have added a paragraph toward the end of the introduction that, in brief and more general terms, presents the rationale for our focus on BTP, the CBL, and sentence production. It reads:

Our aim in the present study was to investigate the possibility of developmental change in SL by focusing on a single mechanism that is proposed to be at work over the longer arc of early language development (i.e., in speech segmentation *and* in utterance production and comprehension). Concomitantly, we focused on a developmental language phenomenon that shows gradual change over several years: children’s spontaneous utterances. Suiting our needs perfectly, BTP can be applied to the discovery and combination of linguistic chunks to predict patterns in sentence production (McCauley & Christiansen, 2011; Onnis & Thiessen, 2013; Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008). Further, BTP has been proposed as a continuous mechanism throughout development, influencing language processing from infancy to adulthood (Christiansen & Chater, 2016; McCauley & Christiansen, 2019a; Misyak et al., 2012). However, this hypothesis has to our knowledge not yet been tested with longitudinal data. While developmental change in SL could theoretically be tested with many other SL mechanisms and/or developmental language phenomena, the use of BTP and chunking to predict increasing utterance complexity presented a compelling starting place for the present work.

Soon after this paragraph we expand on three specific reasons for focusing on BTP and the CBL, all in the “Backward transitional probability and the Chunk-Based Learner” sub-section. We have designed these paragraphs to not only make our reasoning more clear, but to also enable other Cognitive Science readers to make connections with the work. The additions related to this change are copied below:

First, as mentioned, we were interested in pursuing a model based on backward transitional probability. BTP is one of multiple approaches for dividing streams of continuous speech into coherent and/or re-combinable units; other approaches include, for example, forward transitional probability (FTP) 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). While both BTP and FTP have been shown to effectively enable infants, adults, and simulated learners to segment chunks from continuous speech, direct comparisons between the two for planning and parsing whole spoken utterances suggests an asymmetry in their performance. For example, BTPs outperform FTPs in predicting phonetic word durations in spoken English for both function and content words (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), in shallow parsing of English, French, and German child-directed speech (McCauley & Christiansen, 2019a), and in reconstructing child-produced sentences in 29 languages (McCauley & Christiansen, 2019a).

Second, among models using BTP, the CBL was of particular interest in the current study because, at the computational level (Marr, 1982), it is designed to be psycholinguistically plausible for utterance processing (see McCauley & Christiansen, 2019a, for a review). It uses BTP to incrementally build up an inventory of speech chunks (e.g., “doggy”, “look at”), and stores the chunks and their co-occurrence frequencies such that the accumulated chunk inventory can be used to both parse and produce utterances on the basis of what the model has encountered so far. By only storing shallow information about how chunks combine, its performance in processing multi-chunk utterances also depends exclusively on local relations in the speech signal, mirroring the transitory and sequential nature of spontaneous speech (Christiansen & Chater, 2016). The model can also utilize its BTP-based chunks to engage in *predictive* processing during parsing tasks (McCauley & Christiansen, 2019a). This “recognition-based prediction” method, together with the central use of multi-word chunks and the parallelism between comprehension and production, renders the CBL impressively consistent with findings from both spontaneous and elicited language processing by adults and children (e.g., Arnon & Snider, 2010; Arnon & Clark, 2011; Diessel & Tomasello, 2000; Ferreira & Patson, 2007; Pickering & Garrod, 2013). Of course, this psycholinguistic plausibility only extends to the computational level of analysis—translations of this model to the algorithmic level will be essential to its long-term utility (Griffiths, Lieder, & Goodman, 2015)—but the CBL’s attention to the incremental, local, and structurally parallel nature of natural language use increased its appeal for the present study.

Third, the CBL has previously succeeded at modeling naturalistic speech production, the task we target in the current paper. For example: (a) as mentioned above, it parsed text better than a shallow parser in three different languages (English, German, and French), (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, 2019a; 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 (McCauley & Christiansen, 2017; McCauley & Christiansen, 2014b; McCauley and Christiansen, 2019a). The model’s performance on utterance production tasks over developmental time is of prime interest as a next theoretical step. Instability in performance over developmental time would hint at significant influences of children’s growing language knowledge, cognitive resources (e.g., working memory, speed of processing), or a combination of the two, on the overt utility of the mechanism.

In addition, in the General Discussion, the authors should take at least a paragraph or two to discuss in more details the concerns raised by Reviewer 3 regarding the seeming paradox between what we know is maturing in development and why this is not reflected in the CBL results, as well as some speculation on how TP storage etc might be psychologically plausibly implemented. These revisions can be seen as opportunities to take a little more space to discuss these deeper cognitive issues. This revision will help also bring the manuscript’s contributions to bear to a broader Cognitive Science readership.

> We have summarized Reviewer 3’s concerns in two new paragraphs in the “Limitations and Future Work” subsection of the Discussion. To these same paragraphs, we also added speculation about how the CBL’s limitations might be approached in future work. The relevant changes are copied here:

Although the CBL was perfectly suited for this initial investigation (see Introduction), it is unclear how this model could be implemented at the neural level. In particular, the CBL does not specify how BTP (between chunks, and the running average) is stored in the brain, nor how the comparison mechanism that inserts chunk boundaries is implemented. The model’s requirement for access to precise estimates of BTP between any two chunks may, with accumulated natural input, hugely increase its memory requirements. That said, these probabilities could potentially be approximated more efficiently in a neural net, which would also yield more graded, abstract chunks.

Perhaps more troubling is the BTP comparison mechanism, which presumably relies on functions of executive control, working memory, and/or long-term memory, and which is likely influenced by the child’s speed of processing, all of which are known to change dramatically during the developmental period tested here (Bauer, 2005; Casey et al., 2000; Diamond, 2002; Gathercole et al., 2004; Kail, 1991; Rodríguez-Fornells et al., 2009; Uylings, 2006; Wojcik, 2013). Why, then, do we find no age effect here? We propose two possibilities that could be explored further: (a) while these memory, processing, and executive control functions *do* improve with age, they are already sufficient early on for the foundational computations of the model, and their increased functioning only comes into play once children begin to produce highly complex utterances; (b) caregiver linguistic input itself, perhaps via the child’s signs of comprehension, closely parallels these maturational gains (e.g., via “fine-tuning”; Roy et al., 2009; Snow, 2017). Again, neural networks may be a natural option for exploring how changes in these maturational factors interact with changing input in the creation and storage of chunks. If further research did find that developmental change alters the CBL’s ability to reproduce children’s utterances, it would raise questions about the age-invariant influence of BTP over development. A similar approach could be taken to comparably investigate age-related change in the use of other mechanisms, including FTP.

We underscore the limitations and future ways forward in the immediately following paragraph, also pointing to other models that might be fruitfully investigated along similar lines to what we present here, as copied here:

In principle, these “next steps”—calling for the use of richer data—and potential neural-net implementations—to better simulate storage and processing limitations—could be explored using a number of different SL mechanisms for speech segmentation, comprehension, and production (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 fact, a number of existing models already take closer inspiration from neurocognitive maturational findings (e.g., Mareschal & French, 2017; Cleeremans & Elman, 1993; Perruchet & Vinter, 1998), and a side-by-side comparison of their longitudinal performance on natural language data with the CBL would be a worthwhile follow-up to the present research. Notably, while the CBL here performed above chance on average, there was still room to improve; another model may show even better performance, or the CBL might improve upon the addition of some of these maturational features.

To this end, I am giving the paper another Revise and Resubmit. However, I do NOT anticipate turning it back out to the reviewers. If you choose to revise (which I hope you will!), I would need a response letter that clarifies point-by-point where you have made edits to the paper to address Reviewer 3’s more substantial concerns. Please also read over and correct the more minor requests as noted by Reviewer 1 and 2. I will then go through your revision and letter myself and render a (hopefully final) decision.

> Thank you for the clear explanation and for reviewing this second revision. We have added point-by-point summaries of our edits below, and have indicated all edited text in **orange** in the manuscript. We have kept the manuscript anonymous just in case you do choose to send it back out to reviewers.

I am quite optimistic that a revision that addresses these points will be successful and I hope you find this feedback helpful.

Sincerely,

Elizabeth Bonawitz

**Reviewer #1:** Elise Hopman

The authors did an excellent job addressing everything I raised in my review. I also appreciate the changes made to the manuscript based on the other suggestions, and think the manuscript is much improved and ready for publication.

Some suggestions (in order of the manuscript):

- under 'analysis' it says (see step 3 in Figure Figure 2), duplicate word Figure

- I was eager to view the osf archive but didn't have access (yet). I assume everything the authors mention is on there.

- below Figure 3, I personally prefer 'layperson' to 'layman' (though I realize there's an argument for 'layman' being gender neutral).

- below Figure 4, 'utterances from older children tended to contain more words ….', suggestion to insert (and thus on average more chunks) since, while more words is what is being plotted in Fig 5, more chunks is what affects the model.

- in Figure 6, it could be helpful also plotting the average model prediction for each age +- SE or CI in addition to the 6 lines for individual children, to make it more visually clear that it is overall different from 0 for both local and cumulative.

- there's a comma missing in footnote 5 after the random effects of the model and before its family specification.

- I strongly prefer wording along the lines of 'the model could reconstruct children's speech productions' (with reconstruct instead of account for, as the authors do in the abstract), e.g. in the first sentence of the conclusion, and in the first sentence of the discussion paragraph starting with 'As the CBL model only employs … ', and several other places. Reconstruct is a more precise word for what the CBL model does, since it only reorders some chunks and doesn't actually come up with a produced utterance 'from scratch' (e.g. based on semantics).

> We thank the reviewer for their comments and have corrected each of the above issues where indicated by the reviewer. We have also made the suggested addition of model estimates to Figure 6. We are sorry the OSF archive was not accessible. We have tested the link in this new submission and can verify that it gives access to visitors without further requirement. We note that the full repository of code and materials are already also on GitHub and will be made public upon paper acceptance.

=========================================

**Reviewer #2:** The authors have done a good job and they have satisfactorily answered to my main comments.

I think that the manuscript is now ready for publication in Cognitive Science.

- p. 14, first line: "overt" should be "over"

- p. 14, BTP and the Chunk-Based Learner: "Backward transitional probability" should be replaced by "BTP"

- p. 20: "their known transitional probabilities should be replaced by "their known BTPs"

- p. 20: "the highest backwards transitional probability" should be replaced by "the highest BTP" (2 times)

- p. 21, in the caption of Figure 2: "using transitional probabilities" should be replaced by "using BTPs"

> We thank the reviewer for their comments and have corrected each of the above issues where indicated.

=========================================

**Reviewer #3:** I must confess to being disappointed in the revised version of this paper. The authors have responded to the criticism that in their original ms. they basically didn't cite any other computational model by CBL and very little empirical work related to statistical learning in infants and children. However, by adding the references that were suggested, which they did to my satisfaction, I, nonetheless, feel that they have not responded adequately, or in some cases, at all, to the substantive issues raised in my review.

I wrote: "I would like the notion of BTPs as a driving mechanism 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?" Their whole paper is based only on BTPs, but there is no explanation of this. Predictive coding has been shown to be an extremely important and powerful cognitive mechanism. And predictive processing involves FORWARD transitional probabilities and, yet, this is only referred to in a single added comment: "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..." In other words, the authors are only concerned with BTPs. WHY is it that these should be studied to the exclusion of the others? This needs more justification than, "Because it seems to work". Pelucchi et al. (2009), for example, showed that infants are sensitive to BTPs, but made no claims about this being the only mechanism involved. Is CBL's claim that FTPs play no role? If not, then how could CBL use these latter types of TPs?

> We have made further edits clarifying why we here selected BTP (see the text copied in response to the AE’s comments); note that these additions include references to FTP vs. BTP in tasks relating to sentence-processing (BTP performs better) and a brief explanation of how BTP can, in fact, be used to do a sort of forward prediction (using chunks). With our expanded justification, and with our reiteration at multiple points that our choices were particular to a preliminary investigation, we hope that it is now more obvious that we make no claim about BTP being the only mechanism involved in sentence processing, nor in signal segmentation more generally.

CBL is not an engineering model, it is supposed to be a \*cognitive\* model, and, as such, it cannot just sweep under the rug issues like how is TP information stored, what comparison mechanisms are there for it to know when TPs are higher or lower than previously seen TPs? This issue is not discussed at all, even though I specifically asked it to be. And it leads to one of the thorniest -- but centrally important to their paper -- issues that the authors have left basically untouched in their reply. As I said in my review: "We know that 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 that 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." But the authors simply claim that "CBL does not at all model maturational changes" and hope that this is sufficient. Granted, there are no explicit "maturational mechanisms" included in the model, but we are given no clue as to why this age-invariant property might occur in their model or, for that matter, in young children. These claims need to be teased out. Is the claim that BTPs are not dependent on executive control mechanisms, whereas, say, FTPs are? Other authors have modeled these changes -- and the changes in chunking (see Mareschal & French, 2017) -- with learning rates, for example, but is the claim here that BTPs are unaffected by the maturational changes that we know occur? If so, this seems very strange, but at the very least, it needs to be clearly justified.

> In our additions (see text copied in response to the AE’s comments), we now highlight these limitations of the CBL and provide some speculation as to how they could be addressed in more algorithmic-level translations of the model (e.g., in a neural net). We disagree with the characterization of the CBL as an “engineering” model, but do draw out explicitly that it is a computational-level model and that it needs more algorithmic-level interpretations to be of long-term utility in cognitive science.

The question, and plausibility, of the "chunk inventory", as described by the authors needs to be discussed, beyond simply saying "The model depends on a highly simplified and idealized version of reality." Well, yes and no. It is taking input from real utterances of children. and so questions about graded chunks and chunk strength are important ones. And again, since CBL is a \*cognitive\* model, is it reasonable to explicitly store all of the discovered chunks, the BTPs between words, and the BTPs between discovered chunks"? I asked: So, does CBL also store how strong a chunk is, along with all of the other information? Also, keeping track of the "BTPs between words" 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. But again this issue was simply sidestepped by saying that they had done what they could to bring attention to some of the model's limitations.

> We agree that there are serious limitations to consider in children’s storage of chunks and their co-occurrences (note that the model keeps track of co-occurrences and computes BTP on the fly, which is somewhat different from what the reviewer assumes; though, in both cases, the storage required is significant). As mentioned in response to the previous comment, we bring more attention to this issue in the “Limitations” section, proposing that neural networks may be a promising approach to investigating these problems of storage; they also offer a way to more easily introduce and investigate the consequences of maturational change in, e.g., working memory, on sentence reconstruction.

In short, the authors did, to a significant and adequate extent, correct the original version's major flaw of not citing any of the considerable modeling and empirical literature surrounding segmentation and chunking of words by young children. However, all of the other key questions that I wanted to see discussed or justified were largely, if not completely, ignored. CBL is supposed to be a \*cognitive\* model, not an engineering model where issues like TP storage and comparison can be ignored, as long as the model's output corresponds to empirical data.

The point is that it would be almost trivial to create a language in which BTPs provided no segmentation cues. All it would take to produce such a language, call it No-BTP-Language, is a simple modification of Perruchet and Desaulty's 2008 language in which FTPs provided no segmentation information (and which Pelucchi et al. used for infants). So, if CBL is a model of how infants do segmentation and chunking using only BTPs, then presumably this model would not be able to do segmentation and chunking of No-BTP-Language. But, of course, real infants would be able to. And this is a fundamental problem with the model presented in this paper. So, do we need another model to do FTPs? Is the CBL approach only applicable to BTPs? Is there any cognitive plausibility to this? (I think not.) But the authors must discuss these issues, instead of sweeping them under the carpet as they have done in their revision.

> Surely the reviewer is right about creating a language in which BTP doesn’t work. However, we would prefer to evaluate the model in its performance on real human languages. This is a real strength of BTP: as a reminder, McCauley and Christiansen (2019; Appendix C) compare BTP and FTP on 29 languages, showing that BTP outperforms FTP in 25 of them, with small differences on the 4 other languages. Of course, other mechanisms, such as FTP, are likely also at work, this evidence suggests that BTPs are rather good at capturing segment boundaries in real human languages, which is the cognitive phenomenon of interest here.

Cognitive models must be in the business of explaining empirical data, not just reproducing it; engineering models are not saddled with this constraint. I want to know why BTPs might be age-invariant and why, presumably, FTP's aren't, when what we know about the differences in executive functions with age would argue against this. I want to know how TP storage and comparison happens in a psychologically plausible manner in order to allow word boundaries to be delineated. And so on.

Until these questions are dealt with in a reasonable manner, I do not think this paper should be accepted. I still maintain that the results presented and the accompanying simulations are extremely limited in scope and, by themselves, do not justify publication of the paper, especially since the authors have not dealt with the larger issues of cognitive plausibility of their model raised above.

> We hope that the changes introduced in this new revision will satisfy the reviewer’s concerns regarding the plausibility limitations of the model, the role of maturation in the present results, and the future potential of exploring *other* SL mechanisms in a similar vein to what we present here.