

Dear Dr. Snyder,

We are writing to resubmit our manuscript, “Learning and long-term retention of large-scale artificial languages.” Thank you again for your feedback and guidance in this revision. You will find below point-by-point responses to your own comments and those of the reviewers.

Please feel free to contact us with any questions or concerns. Thank you very much for your consideration,

Sincerely,

Michael C. Frank, Joshua B.  
Tenenbaum, and Edward Gibson

## Editorial Comments

*In several places, there should be a comma after "e.g."*

Done.

*Details about how the participants were recruited should be provided. I assume this was not a random sample of the community.*

Done.

*Pg. 3. "sequences. Thus, there were no temporal or" or "sequences; thus, there were no temporal or"*

Corrected.

*Pg. 3. "In the first interim test session"*

Corrected.

*Pg. 3. "one month (3 participants) or 2 months"*

Corrected.

## Reviewer #1

*1. p3, para 6: "In the first interm test session ("immediate test")" the word interim is misspelled.*

Corrected (see above).

*2. p3, para 7: "was administered after one month (3 participants) and 2 months". Shouldn't "and" be changed to "or"?*

Corrected (see above).

*3. p5, para 1: "due to their presentation in the visual modality OF THE TEST."*

Done.

*4. ref 46 is published and the proceedings paper should be dropped in favor of Mintz, T.H., Newport, E.L., & Bever, T.G. (2002). The distributional structure of grammatical categories in speech to young children. Cognitive Science, 26, 393-425.*

Done. Thank you for the reference.

## Reviewer #2

*I do not feel that the authors have addressed my criticisms adequately. In many cases, they agreed with my major points (that research with artificial languages of the type they use has little relevance) but can only point out the existence of other similar work (which is equally irrelevant) to justify the current approach.*

We regret that the reviewer does not feel that we have addressed his/her points. While we agree with the reviewer that artificial language research has flaws (as does every experimental strategy), we strongly disagree with the premise that such research is irrelevant. On the contrary, work using artificial languages has been extremely productive in creating new insights about adults' and children's learning mechanisms and has spawned lines of research that connect directly with natural language learning abilities. For a more detailed review of artificial language research and the productive inferences that can be made from such research, we suggest Aslin & Newport (2012, *Current Directions in Psychological Science*) and Romberg & Saffran (2010, *WIREs*).

*A critical empirical flaw is that "Since the language is constructed to favor short sentences which further favor the use of more frequent words, how are we sure that the higher performance on high frequency words is not the result of boundary effects-as they will more likely to appear at boundaries? Of course, with short sentences (e.g., 2-3 words), boundaries alone give cues for word segmentation, without needing to use any statistical information."*

*The authors said "First, short sentences do not favor frequent words. Second, words were placed in sentences at random; hence the probability of appearing at a boundary was the same for a high frequency word and a low frequency word. Nevertheless, individual high frequency words did appear more often at boundaries than did individual low frequency words (since by definition they appeared more often overall); hence the current experiment cannot measure boundary effects."*

*But shorter words ARE favored. Since both words and sentences are created from a Poisson distribution with a mean of 2, there are more shorter word and sentence TYPES in the language. Furthermore, word TOKEN frequencies are drawn from a Zipfian distribution, which favors a small subset of word TYPES. Assuming that this small subset of types is chosen randomly from the set of all types, it follows that word TOKEN frequencies favor shorter words. Furthermore, since the generation of words in a sentence is independent—see my criticism above, as no natural language works this way—the drawing of words in the construction of sentences favors shorter words which are more frequent.*

*This has the effect of having shorter and more frequent words more likely to appear in shorter sentences; thus, these words are more likely to appear at utterance boundaries.*

As we stated in our writeup, there is a Poisson distribution over the lengths of words in our language. This distribution was independent of all other distributions in the construction of the language, as is now made more explicit in the manuscript. While this distribution directly implies that there were more shorter words in our language (both in token frequencies and type frequencies), the causal inference above is still incorrect. Nothing about the generative process for the language made short words or frequent words more likely to occur at boundaries. They occurred at boundaries at the same frequency that they did in any other location.

*... it is well known that words at boundaries are easy to segment. For instance, 6 month old infants can segment words at boundaries WITHOUT any training (Seidl & Johnson 2006, Dev. Sci.) That “the current study cannot measure boundary effects” is a fatal flaw of this paper.*

The Seidl & Johnson study shows that utterance boundaries do provide information for segmentation (though this paradigm still followed the standard familiarization/test structure—albeit with a slightly shorter familiarization than other studies—so it is confusing to suggest that there is no training involved). More generally, the inference about the importance of boundaries is supported by a variety of other studies, including our own adult artificial language work. Nevertheless, that work (and other work, including our own) suggests that boundaries *and* statistical chunking both provide information relevant for segmentation. The contribution of the current study is that we show that these mechanisms lead to learning even in much larger-scale languages than have been investigated previously.

*(Also, note that Zipf’s law favors shorter words, as Zipf famously observed. If the authors followed the strict interpretation of Zipf’s law, then they would directly building in a bias that favors short words. The text is not clear on what they did.)*

We have clarified the detail that frequencies and lengths were chosen independently.