Dear Dr. Snyder,

We are writing to resubmit our manuscript, “Learning and long-term retention of large-scale artificial languages.” Thank you very much for your helpful 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

> *However, while the two reviewers have positive things to say about the study, one of the reviewers raises some important issues that should be addressed with care. In particular, it should be clarified how the stimuli used here relate to past statistical learning studies and the extent to which the cues are relevant to real-world language learning.*

Thank you for this feedback. In response to this comment, we have expanded our discussion of these issues.

> *Pg. 1. "statistical structure of the utterance" and "distributional structure of speech" are vague terms and could refer to any number of things in speech. This should be clarified for the typical reader, given that I think you mean a rather specific thing, i.e., the likelihood that one sound is followed by another. Similarly, please clarify "matched for overall frequency but vary in their statistical coherence." Despite my familiarity with this literature, I'm not sure what you mean by "coherence."*

We have clarified these references.

> *Pg. 1. Define "Zipfian frequency distributions"*

Done.

> *Pg. 2. "failure should place an upper bound on children's abilities as well" This assertion seems counter to the observation that children seem to be able to learn new languages better than adults. Also, depending on the age range tested in the studies cited here, is it not quite an extrapolation to assume that adults would have a better memory for new lexical items compared to young children?*

We have elaborated this argument in the manuscript. In particular, we now discuss the dissociation between what children are better at (morphology, syntax) and what adults are better at (remembering words) and give additional citations. Based on the literature on short-term and working memory—as well as on second-language learning—we believe it is a well-supported assertion that adults have better memory for new lexical items compared with children. (It is actually somewhat difficult to teach a toddler a new word so that they will remember it a day later; for example, Woodward, Markman, & Fitzsimmons, 1994 needed nine carefully-timed naming instances to promote retention.)

> *Pg. 2. "One additional participant" Does this reduce the number of participants to 3? Or is 4 the number of participants after excluding this one?*

There were 5 initially. We have clarified this detail.

> *Pg. 2. "no word repeated immediately" This is a very small point but isn't this a*

> *bit artificial since some words (not to mention homonyms and homophones) are sometimes repeated? Was there a reason not to let this happen a small proportion of the time?*

It has been customary in this literature to avoid immediate repetitions as they are extremely salient to listeners and are generally assumed to create an extra cue to word identities. We have now provided citations for this decision.

> *Pg. 2. "no word boundaries" Clarify that presumably this means no temporal gap between words. Also, were there any gaps between sentences? How many sentences were generated per subject? Any other details about the sentences (e.g., number of words per sentence) and how they were generated, stored, and presented to subjects should be included. As it stands, there are very few such details. Or perhaps, there is no such thing as a sentence per se? Do the 60,000 word tokens in a sense constitute a single very long sentence? Either way, please clarify.*

We have clarified that there are no temporal boundaries between words but that there were silences between sentences. Sentence lengths (Poisson distributed, with an overall mean of 4) were given in the previous submission but have been highlighted. Details of stimulus delivery have also been highlighted.

> *Pg. 2. "but could not read, talk" Did you instruct them to not think verbal thoughts? Did they report doing this while exercising etc.?*

We did not instruct them not to think verbal thoughts—we expect that this instruction would have been quite difficult to follow. They reported listening while exercising or walking from place to place, a detail which we now note in the manuscript.

> *Pg. 3. Define "Orthographically glossed"*

Done.

> *Pg. 3. In the text you use the terms "initial" and "interim," but in the figure legend you use "immediate" and "1-2 months." If these are the same terms, please clarify and use only one term per concept. Also, in Figure 1, there is a reference to "Experiment 2," but the text has not explained that (or whether) there is more than one experiment.*

Corrected.

> *Pg. 3. "Participants were not notified about the second interim test or the final test." Was this procedure approved by the IRB? If so, what was the rationale for not allowing participants to know about the tests ahead of time? If so, did you debrief subjects on why they were not told about this during the initial consent?*

This experiment was covered under a blanket protocol for research in the Gibson lab at MIT, which included language about the possibility of being invited back into the lab for future studies. All participants gave informed consent and were aware that they were not obligated to continue participating in the research.

> *Pg. 3. "or 2 months (1 participant, LB)."*

We assume that this was a request for clarification. We have noted that LB is the label given to the participant in Figures 1 and 2.

> *Pg. 3. Please state exactly what judgment participants were asked to make during the final 2AFC task. Also, explain the rationale for using 2AFC rather than the segmentation task, which makes it hard to compare across the two types of follow-up tests.*

Done. We note in the revision that we used the 2AFC task to provide comparability with previous studies of segmentation).

> *Pg. 3. Please clarify in intuitive (and if you like mathematical) terms what is the difference between precision and recall, and how the harmonic mean of these two values is meaningful.*

We have added a footnote describing these measures, which are standard in the literature on word segmentation.

> *Figure 2 caption: "e.g.,"*

Corrected.

> *The Figures might be better placed later in the text in general, so they are closer to where they are discussed.*

The Figures are now later in the text.

> *Pg. 5. In what sense are the final test results consistent with "previous work on Zipfian frequency distributions"?*

In the revised manuscript, we highlight that both this work and our previous work on Zipfian distributions show strong frequency effects.

> *Pg. 6. Define "transition statistics" and "frequency or transitional probability."*

Done.

Reviewer #1

> *This is a really innovative and important study. ...*

Thank you very much for your comments.

> *1. The word "up" should be added after "scale" in line 5 of the Abstract.*

Done.

> *2. In addition to refs 12 & 13, for completeness, Shukla, White & Aslin (2011) should be added.*

Done.

> *3. In paragraph 2 of Materials and Methods, you need to clarify that "Word lengths" refer to the number of syllables, and "sentence lengths" refer to the number of words.*

Done.

> *4. Caption for Fig. 1 should change Experiment 1 to 2 (there IS no Exp. 2).*

Done. We apologize for the error.

> *5. In the second to last paragraph of Materials and Methods, you say "the day after they finished listening." I think you need to clarify that this was after the last (i.e., 10th) day of listening. You also need to make it clear that after the 10th day of listening, there was no further exposure to the corpus prior to the second test.*

Done.

> *6. I think you need to raise another possible reason for the "bump" in performance after 2-3 months post-exposure – namely, seeing the orthographic transcription during the first test could have triggered a re-coding of the input, which then facilitated performance upon retesting.*

Thank you for this suggestion; it is now included.

> *7. In the second to last paragraph of the Results, you say "we saw a strong relationship between word frequency and performance for word." This makes no sense. Is there something missing here?*

Apologies for this typo. It is now corrected; we saw a relationship between performance and frequency.

> *8. In the last paragraph of the Discussion, you use the term "passive exposure," which some readers might mistakenly interpret as not requiring attention to the speech materials. There is evidence from several studies that attention is required,*

> *so you should clarify this point.*

Corrected, thank you.

Reviewer #2:

> *The scale of the project–one hour training for a year–is unprecedented and the authors should be commended for their efforts. Unfortunately, as I discuss below, the experimental design, methodology and evaluation are very far removed from the research on how word segmentation actually works in human children, rendering this work of little empirical relevance. I can only recommend rejection.*

We regret that the reviewer feels that we did not make sufficient efforts to describe the connections between our work and the broader literature on word segmentation. We have attempted to rectify this issue in the revision.

We note that there is an error in the reviewer's summary of the experiments: participants were not trained one hour per day for a year, but instead were trained one hour per day for ten days, and then tested 1 – 2 months later and three years later.

> *1. The authors aim to test "the role that statistical learning plays in child language acquisition" (p1). However, this experiment, even if the results were entirely valid, has nothing to do with statistical learning, which can be interpreted either in a narrow or broad sense.*

We have clarified that, although the broader question that frames our work is indeed about the role of statistical learning in language acquisition, the current study primarily addresses the question of whether statistical learning mechanisms are scalable in adults.

> *In the narrow sense, all research on statistical learning in child word segmentation focuses on the effectiveness of transitional probabilities over syllables, following the well known paper by Saffran et al. (1996, Science). This requires the stimulus to be constructed such as word boundaries are predictable from transitional probability changes. This is plainly not the way the artificial language is constructed here.*

This contention is not correct. Although transition probabilities (TPs) are not a perfect strategy for segmentation in our language, a local minimum-finding strategy using TPs can achieve F-scores of .85 or more at test. In fact, in Saffran et al.'s original adult study, TPs were a strong but imperfect cue to word boundaries.

> *In a broad sense, statistical learning can be interpreted as *some* kind of statistical information (other than transitional probability). But that's not what's done here either. The authors show that at least some adults can segment words in large scale language, but there is NO discussion of what type of statistical learning they may be following, or whether the results are NOT due to some non-statistical*

> *strategy (e.g. perceptual chunking heuristics; Endress et al. 2005. JEP: General) in the first place.*

We did include some discussion of what strategies adults may be following in our previous submission, citing work by Perruchet, Goldwater, and French among others. But we did not feel that the current results provided a strong constraint on such proposals (as mentioned above, a TP strategy does succeed in the current experiment). Our other work (e.g. Frank et al., 2010; Kurumada, Meylan, & Frank, 2011 and under review, cited in the manuscript) provides extensive commentary on what class of models best fits adult performance in statistical word segmentation tasks.

> *But the artificial language here is created WITHOUT statistical dependence between words. No natural language behaves like that. And it follows that the results of the study in principle cannot have any bearing on statistical learning.*

This criticism applies equally to nearly all current work on segmentation using artificial languages. In fact, in our recent work (e.g. Meylan et al., CogSci 2012), we do investigate word-to-word dependencies. Artificial language paradigms are used to reduce the complexity of human language and bring phenomena under experimental control—one goal of our research program has been to gradually introduce some of this complexity back into such languages.

> *2. There is now emerging consensus in the field that the artificial language learning paradigm, where all language-like properties are withheld, has run its course. Human children do not learn languages in isolation or without other sources of information: they exploit the sound-meaning associations (Hay et al. 2011, Cog. Psych), pay attention to utterance boundaries (Seidl & Johnson 2006, Dev. Science) and a variety of prosodic cues (Johnson & Jusczyk 2001, Cog. Psy, Thiessen & Saffran 2003, Dev. Science). Note that the role of statistical learning is still very much relevant, but fewer and fewer labs are using artificial languages, including the most prominent proponents of the paradigm (e.g., Hay et al. 2011, Cog. Psy., Pelucchi et al. 2009 Cognition). Natural but unfamiliar languages can dominate recent child word segmentation research, which brings the results closer to the actual mechanisms of language learning. Even artificial language studies have made efforts to make them sound as natural language like as possible e.g., by adding prosody (Hay & Saffran 2012 Infancy).*

We generally agree with the reviewer that artificial language research has flaws. In fact, the goal of the current study was to address one of those flaws: the small size of artificial languages. Other work, such as that cited above, has been influential in addressing others. Due to the laborious, hand-constructed nature of the stimuli for "unfamiliar language" experiments, this method is not practicable for investigations of large-scale language learning. Nevertheless, we do wish to note that artificial language research is still quite popular: a quick search on Google Scholar suggests that there have been dozens of studies using artificial language paradigms published in the best journals in the last two years.

> *3. 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.*

There are several errors in this comment. 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. Experiment 1 of Frank et al., 2010 suggests that boundary effects do contribute to segmentation, as has been posited by other researchers, but that they do not account for statistical learning effects (as was shown by the initial Saffran et al. work).

> *4. It is not clear to me whether the authors' baseline metric is a reasonable one. It does not mitigate against simple but likely effective baselines if the subject develops some sense of the average length of words, according to which he/she will be able to estimate the total number of words–thus the total number of word boundaries–in each test sentence. Recall that the artificial language is designed to favor short sentences as well as short words.*

This comment is also incorrect. As we wrote in the manuscript, "... because this baseline randomizes individual participants' decisions within each sentence, it ensures that participants' accuracy was not due to guessing based on assumptions about the distribution of word lengths (as opposed to actual knowledge of word forms)." Randomizing the arrangement of boundaries within a sentence preserves the average word length.

> *5. As noted by Adriaans & Kager (J Mem. Lg, 2009), the scoring of word segmentation by boundaries should use signal detection metrics such as d', which provides better indication of how well the results improve over baselines. The information retrieval metrics such as F-score, which the authors adopt, are appropriate for word-based, rather than boundary-based, evaluation.*

We use F-score as our metric because it has been the standard metric in the computational literature on word segmentation for both word-based and boundary-based evaluation (e.g. Goldwater, Griffiths, & Johnson, 2009). Adriaans & Kager (J Mem. Lang., 2009) write that "such metrics [e.g., F-score] do not necessarily assign low scores to random models. Specifically, random models (in which no learning takes place) will obtain high precision scores whenever there are many potential boundaries to be found" (p. 320). For this reason, we provided a random baseline (which does indeed receive low F-scores in our paradigm).

> *6. The authors MUST report the formulas of their mixed-effect model and how*

> *the significance is computed.*

Done.

> *7. Lastly, while I appreciate the tremendous amount of time involved in the study (and probably because of it), the paper reports results from 3-4 subjects. Experimental studies of language learning typically require a much large number of subjects.*

We agree that this sort of case study is somewhat unusual for work on artificial language learning (though see Bahrick et al., 1993, which also had four participants). It is the norm in other fields such as psychophysics, where large amounts of data can be collected from individual participants (e.g. Burr & Ross, 2008, which had 4 subjects, to take a recent influential example).