Learning and long-term retention of large-scale artificial languages

Michael C. Frank[1,*], Joshua B. Tenenbaum[2], Edward Gibson[2]
1 Department of Psychology, Stanford University, Palo Alto, CA, USA
2 Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
∗ **E-mail: mcfrank@stanford.edu**

Abstract

Recovering discrete words from continuous speech is one of the first challenges facing language learners. Infants and adults can make use of the statistical structure of utterances to learn the forms of words from unsegmented input, suggesting that this ability may be useful for bootstrapping language-specific cues to segmentation. It is unknown, however, whether performance shown in small-scale laboratory demonstrations of "statistical learning" can scale to allow learning of the lexicons of natural languages, which are orders of magnitude larger. Artificial language experiments with adults can be used to test whether the mechanisms of statistical learning are in principle scalable to larger lexicons. We report data from a large-scale learning experiment that demonstrates that adults can learn words from unsegmented input in much larger languages than previously documented and that they retain the words they learn for years. These results suggest that statistical word segmentation is scalable to the challenges of lexical acquisition in natural language learning.

Introduction

Spoken speech is a continuous acoustic waveform without consistent breaks at the boundaries between words in phrases or sentences. Although acoustic, phonetic, and prosodic features give partial evidence for where words begin and end, these cues vary widely between languages [1]. One source of information that is consistent across languages, however, is the statistical structure of the utterance itself [2]. In the absence of other information, infants and adults are able to use this structure to extract words from continuous speech [3, 4]. These findings of "statistical learning" suggest that learners are able to use the distributional structure of speech to identify coherent sequences.

In a typical statistical learning experiment, infants or adults listen to a stream of unsegmented speech, generated by randomly concatenating words from a language containing $4-6$ different word forms. After a very short exposure—sometimes as little as 2 minutes—listeners are then able to distinguish frequent sequences from less frequent or less statistically-coherent distractors, giving evidence that they can use the distributional pattern of syllables to identify words [3, 4]. Infants in this type of experiment can even distinguish between strings that are matched for overall frequency but vary in their statistical coherence [5].

What is the role that statistical learning plays in children's language acquisition? Some authors have suggested that it is an important part of the broader process of language acquisition [6–8], but others have questioned whether performance shown in short lab studies can scale to the challenges of lexical acquisition [9–11]. In particular, it is unknown whether a mechanism that has only been demonstrated to operate over highly restricted artificial languages with homogeneous lexicons can nevertheless be applied successfully to the complex and heterogeneous lexicons of natural languages.

Recent work has found that learners can map meanings to the outputs of statistical segmentation tasks [12, 13] and that statistical learning effects can be found using natural language stimuli [14]. In addition, statistical learning effects are robust to variation in word and sentence lengths [15] and to the Zipfian frequency distributions that are ubiquitous in natural languages [16]. But although the results of

these tests have been positive, they do not fully address concerns regarding whether statistical learning can scale to larger languages and longer retention intervals, because they still use small-scale experimental tasks.

We used adult learners to address this question, for two reasons. First, statistical learning abilities generally appear to be conserved across development [3, 4, 17], making adults a viable population for studying these abilities using large-scale and psychophysical paradigms not suited for infants and children. Second, unlike the ability to acquire complex syntactic and morphological regularities [18], memory for new lexical items increases considerably across development [19, 20]. Thus, if adults are unable to learn words from a particular language via statistical learning, this failure should place an upper bound on children's abilities as well.

We invited four individuals to listen to large corpora of synthesized speech, each over the course of a continuous ten-day period. Each participant listened for an hour a day on their iPod while they exercised, commuted to work, or relaxed, with the constraint that they did not read, speak, or otherwise use language during listening. The unique language that each participant heard was comprised of 1000 different words, which had the characteristic Zipfian frequency distribution of natural language, such that a few words were highly frequent while most others appeared only occasionally, and the lengths of words and sentences were Poisson distributed, also as in natural language. Words were concatenated randomly without immediate repetitions so there was no syntactic structure available, but all sentences had a minimum of two words and a mean of four. Each of these factors has been studied in isolation [15, 16]; our intention here was to combine them on a much larger scale than previously attempted.

Materials and Methods

All participants gave written consent to participate in this research, and the details of this consent procedure were approved by the MIT Committee on the Use of Humans as Experimental Subjects. Four naïve members of the MIT community (1 MIT undergraduate, 1 student at another local institution, and 2 employees) participated in the initial study and were matched with four yoked control participants. After three years, three participants in the experimental condition were located for followup testing via Facebook searches (the fourth could not be located). One additional participant was excluded for using an explicit strategy during the initial test phase (placing a segment boundary every two syllables without variation throughout the entire test).

A unique artificial language was generated for each participant. Each language had 1000 word types and 60,000 word tokens (for ∼10 hours of speech). Frequencies of words were distributed as a power-law with exponent value 1 [21], such that there were a few highly frequent words and many more with lower frequencies (max = ∼8000, min = 10 tokens). We used a classic Zipfian frequency distribution: $f(x) \propto 1/r(x)$, where $f(x)$ is the frequency of word x and $r(x)$ is its rank. Word lengths were generated by drawing from a Poisson distribution with mean 2 and adding 1 to avoid lengths of zero (mean=3); sentence lengths were generated by drawing from a Poisson with mean 2 and adding 2 to avoid sentences of length 1 (mean=4).

Words were created by combining 24 consonants and 14 vowels into 336 CV syllables and concatenating randomly. Sentences were then created by randomly concatenating words according to the frequency distribution of word types, with no word repeated immediately. Each training sentence was synthesized with no word boundaries using the MBROLA speech synthesis package with the `us3` diphone database, with a duration of 250 ms per syllable and a constant F0 of 100 Hz [22]; test materials were synthesized with the same settings (flat prosody).

Participants listened on their personal iPods for approximately one hour each day over 10 days. They were instructed that they did not need to pay attention while listening but could not read, talk, or otherwise use language; instead they were encouraged to listen while exercising or walking from place to place. To improve compliance, participants kept journals of listening activity.

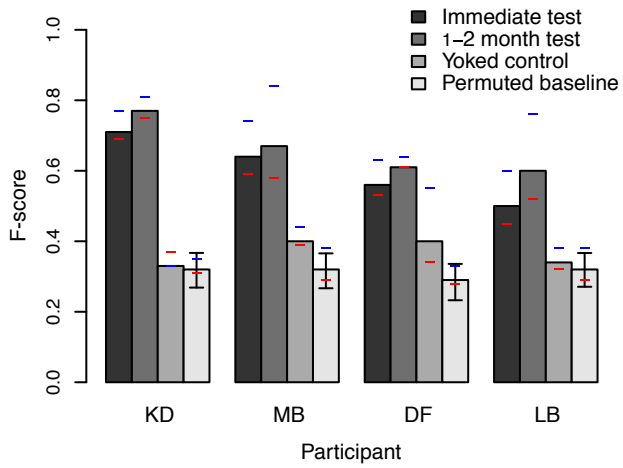

Figure 1. Results of the interim tests for Experiment 2. Bars show F-scores (the harmonic mean of precision and recall) for the initial and 1–2 month test sessions, along with permuted baseline and yoked control scores. Blue and red lines give precision and recall scores respectively for each participant and condition (means for permuted baseline). Error bars show 95% confidence intervals.

Because two-alternative forced choice (2AFC) trials impart information about what correct answers may be, it is not possible to conduct multiple testing sessions using a 2AFC paradigm. We therefore used an orthographic segmentation paradigm that tested participants' performance in making explicit word segmentation decisions to probe performance immediately after training [16, 23]. In the first interim test session, which occurred the day after they finished listening, participants were tested on their ability to segment 400 tokens (∼100 sentences). Orthographically glossed sentences (e.g. "go lah bu pa doh ti") were presented on a computer screen; participants were instructed to listen to the sentence as many times as they wanted and to click between syllables where they thought there was a break between words. Each of the four yoked control participants completed the same initial test as one participant in the study, but without completing the training session. The second interim test was identical to the first and was administered after one month (3 participants) and 2 months (1 participant, LB).

The final test was a 2AFC, administered approximately 3 years after the initial testing session (36 – 37 months). Participants listened to 64 MP3 files of pairs of words, synthesized as above. One word was from the lexicon of the language they had initially heard during training, while the other was a frequency-matched word from the lexicon of another participants' language (but contained syllables that were present in both languages). The words were sampled uniformly across the log frequency range spanned by the training sample, but all words above frequency 1000 were tested. Participants were not notified about the second interim test or the final test until several days beforehand, when they were contacted for scheduling.

Results

All participants were able to segment novel sentences into their component words. Following the methods commonly used to evaluate computational studies of segmentation [24, 25], we compared participants' responses to the correct segmentation, computing precision (proportion of reported segmentation decisions that were correct) and recall (proportion of all correct segmentation decisions that were reported). We then took the harmonic mean of these numbers to produce an F-score. Figure 1 shows these measures,

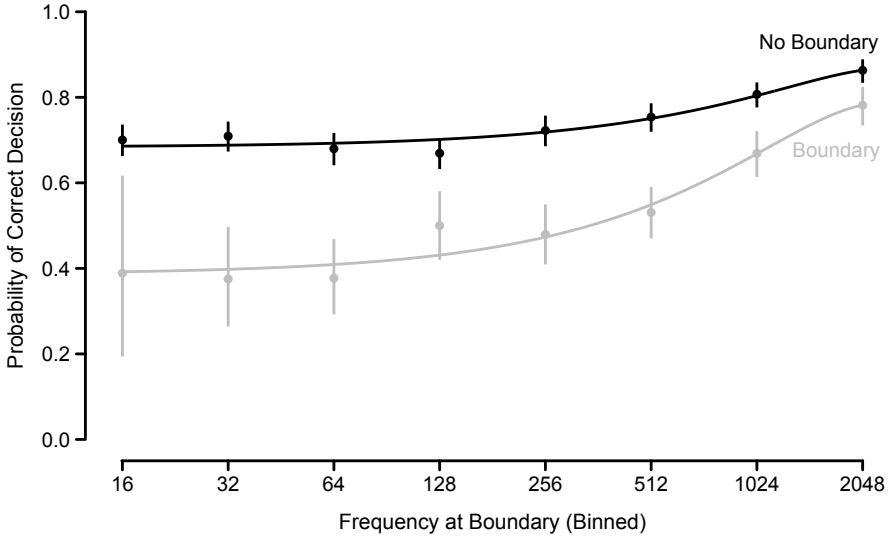

Figure 2. Probability of making a correct segmentation decision at a particular location in a sentence, plotted by whether there was a boundary at that location. Results are averaged across participants, and binned by the logarithm of the highest frequency word at the boundary (e.g. at the boundary between two words, the higher of the two word frequencies). Points show means, intervals show binomial 95% confidence intervals with a non-informative Beta prior, and lines show a loess smoother.

both immediately after exposure and in a surprise 1–2 month followup test session.

In general performance was relatively high, with F-scores generally above .5 and precision and recall relatively close to one another. Overall, precision was higher than recall in all cases, suggesting that participants placed fewer boundaries than was appropriate, but that the boundaries they did place were accurate (in some cases over 80% correct). In addition, performance increased slightly from the first test to the 1–2 month followup. Although our small sample precludes making any inferences on the basis of this numerical increase, it is suggestive of a potential memory consolidation effect [26].

To create a chance baseline for the F-score measure, we randomly permuted participants' own segmentation decisions. We created 10,000 simulated segmentations of each sentence for each participant: we took their initial segmentation of the sentence and shuffled the positions of the boundaries while keeping the number of boundaries constant. We then computed F-scores for each of these random segmentations and empirical 95% confidence intervals on these permuted F-scores. Both immediately and $1-2$ months later, participants performed considerably above baseline (empirical $p < .0001$). This result suggests that participants learned and retained the forms of the words and were able to apply this knowledge to make sensible decisions about how to segment speech in the language. In addition, 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).

Performance was also well above the performance of the yoked controls, who received testing but no training. Although some of the yoked controls' performance was higher than baseline, even the most successful was still well below the performance of the least successful trained participant. This result suggests that performance in the initial segmentation task was not due to learning only the most frequent words (those that could be learned during the test session alone).

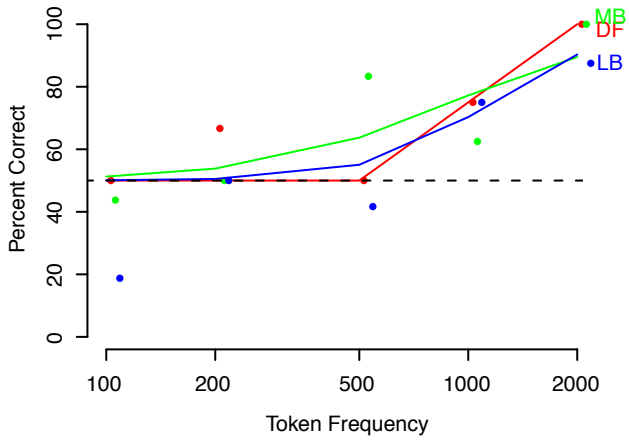

Figure 3. Percent correct on 2AFC test trials, three years after beginning Experiment 2. Dots show individual participants' performance in one frequency range and are jittered slightly on the horizontal to avoid overplotting. Lines show best fitting half-logit regression models for individual participants.

Further evidence that participants gained partial knowledge of many words—rather than learning just a few high frequency words—comes from an analysis of participants' boundary decisions at individual locations in sentences (Figure 2). We examined each decision on the basis of whether there was actually a word boundary at that location. Since most words were longer than two syllables, there were more instances of correct rejections at word-internal locations than instances of hits at word boundaries, hence the overall higher performance on word-internal no-boundary locations. We classified decisions by word frequency; for boundary locations, we used the higher frequency of the two words adjacent to the boundary. Overall, we saw a strong relationship between word frequency and performance for word, and a linear mixed-effects model [27] with maximal random effect structure confirmed this conclusion, finding effects of log frequency ($\beta = .34$, $p < .001$), boundary presence ($\beta = 2.51$, $p < .0001$), and their interaction ($\beta = .25$, $p = .001$).

Three years after the initial experiment, we located three of four participants and administered a surprise test, asking them to distinguish words from novel length-matched distractors. A logistic mixed-effects model showed a highly significant effect of log frequency on performance ($\beta = 1.33$, $p < .0001$), congruent with Experiment 1 and previous work on Zipfian frequency distributions [16]. Overall, while there was no evidence for retention of low-frequency words, retention of the high-frequency words was close to perfect despite the long period between training and test (Figure 3).

Discussion

Our experiment was designed to test whether the abilities demonstrated in "statistical learning" tasks can be applied to large-scale lexicons. The evidence presented here suggests that they can. After ten days of passive exposure, learners acquired partial knowledge about many words in a massive artificial language, and retained the most frequent words across a three-year delay.

How does the scale of our experiment compare to natural language learning? Children hear $\sim 250,000$ – 1,000,000 word tokens per month, for a total of ~ 3 – 12 million words by their first birthday. These tokens are in a Zipfian distribution across 20,000 – 60,000 word types. The most frequent word will then be heard $\sim 250,000$ – 3,000,000 times, and the hundredth most frequent will still be heard more than

2000 times.[1] Thus our data provide an in-principle demonstration that ambiguous contexts can lead to learning within both a frequency range and retention interval comparable to natural language learning. Nevertheless, developmental experiments will be necessary to test whether statistical learning is a viable route to large-scale word learning for infants and children.

Exposure frequency was the primary determinant of retention in our data. Previous work on word segmentation has suggested that learners compute transition statistics [3, 4], but the experimental data across multiple modalities of statistical learning experiments are consistent with many possible psychological mechanisms, not just the transition probability computation [15,16,31]. One class of models relies on memory mechanisms to extract and retain an internally-consistent segmentation of the input into frequent chunks [25, 32, 33]. Chunking models that have interference effects or parsimony biases could provide a good explanation for the frequency dependence of learners' performance, while also capturing transitional probability effects. Thus, "frequency or transitional probability" may be the wrong question: Instead, future research should investigate proposed mechanisms that capture both smaller-scale transition probability effects and large-scale frequency dependence.

Although our experiments were not directly designed to test the connection between memory mechanisms and statistical learning, there are nevertheless similarities between our results and several studies of language learning and long-term memory. First, the dependence of performance on log word frequency parallels the relationship found by Anderson [34] and others. Second, the scale of learning is consistent with previous work on long-term lexical memory [35]. Third, many models of language learning assume that only the highest-frequency forms are retained and used for inferences [36, 37]. Finally, although comparable studies have not been performed, children's retention of novel word forms and meanings over intervals of weeks or months has been well-documented [38, 39].

Despite limited experimental evidence, the utility of passive exposure—via television, radio, podcast, or overheard speech—is widely debated in informal discussions of second language learning. Our results show that for adults, passive exposure can promote the long-term retention of high-frequency, statistically-coherent chunks of language, albeit without any links to meaning. This kind of exposure may create a baseline competence for future comprehension in meaningful settings, useful both for prelinguistic infants who hear large amounts of speech before they begin producing or comprehending language and for adults learning to parse an unfamiliar language.

Acknowledgments

Thanks to the participants in the study for generously volunteering their time.

References

1. Jusczyk PW (1997) The discovery of spoken language. Cambridge, MA: The MIT Press.

2. Harris ZS (1951) Methods in structural linguistics. Chicago, IL: University of Chicago Press.

3. Saffran JR, Aslin R, Newport E (1996) Statistical learning by 8-month-old infants. Science 274: 1926.

4. Saffran JR, Newport EL, Aslin RN (1996) Word segmentation: The role of distributional cues. Journal of Memory and Language 35: 606-621.

[1] Hart and Risley [28] give an input range of 10 – 35 million words by age 3. The Human Speechome Corpus [29] contains approximately 16 million words in 15 months, for $\sim$ 1 million words per month, again 36 million words by age 3. Average English vocabulary is around 60,000 words [30], though this may be significantly limited in child-directed speech.

5. Aslin RN, Saffran JR, Newport EL (1998) Computation of conditional probability statistics by 8-month-old infants. Psychological Science 9: 321-324.

6. Bates E, Elman J (1996) Learning rediscovered. Science 274: 1849.

7. Saffran JR (2003) Statistical language learning. Current Directions in Psychological Science 12: 110.

8. Kuhl P (2004) Early language acquisition: Cracking the speech code. Nature reviews neuroscience 5: 831–843.

9. Johnson E, Tyler M (2010) Testing the limits of statistical learning for word segmentation. Developmental Science 13: 339–345.

10. Yang C (2004) Universal Grammar, statistics or both? Trends in Cognitive Sciences 8: 451–456.

11. Yang C (2008) The great number crunch. Journal of Linguistics 44: 205–228.

12. Graf Estes K, Evans J, Alibali M, Saffran J (2007) Can infants map meaning to newly segmented words? Psychological Science 18: 254.

13. Mirman D, Magnuson J, Graf Estes K, Dixon J (2008) The link between statistical segmentation and word learning in adults. Cognition 108: 271–280.

14. Pelucchi B, Hay J, Saffran J (2009) Statistical learning in a natural language by 8-month-old infants. Child development 80: 674–685.

15. Frank MC, Goldwater S, Griffiths T, Tenenbaum J (2010) Modeling human performance in statistical word segmentation. Cognition 117: 107–125.

16. Kurumada C, Meylan S, Frank M (under review) Zipfian frequency distributions facilitate word segmentation in context.

17. Saffran JR, Newport E, Aslin R, Tunick R, Barrueco S (1997) Incidental language learning: Listening (and learning) out of the corner of your ear. Psychological Science 8: 101.

18. Johnson J, Newport E (1989) Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. Cognitive psychology 21: 60–99.

19. Snow C, Hoefnagel-Höhle M (1978) The critical period for language acquisition: Evidence from second language learning. Child Development : 1114–1128.

20. Gathercole S, Pickering S, Ambridge B, Wearing H (2004) The structure of working memory from 4 to 15 years of age. Developmental Psychology 40: 177.

21. Zipf G (1965) Human behavior and the principle of least effort: An introduction to human ecology. New York, NY: Hafner.

22. Dutoit T, Pagel V, Pierret N, Bataille F, Van Der Vrecken O (1996) The MBROLA project: towards a set of high quality speechsynthesizers free of use for non commercial purposes. In: Proceedings of the Fourth International Conference on Spoken Language. Philadelphia, PA, volume 3, pp. 1393–1396.

23. Frank MC, Tily H, Arnon I, Goldwater S (2010) Beyond transitional probability: Human learners impose a parsimony bias in statistical word segmentation. In: Proceedings of the Annual Meeting of the Cognitive Science Society.

24. Brent MR (1999) An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning 34: 71-105.

25. Goldwater S, Griffiths T, Johnson M (2009) A Bayesian framework for word segmentation: Exploring the effects of context. Cognition 112: 21-54.

26. McGaugh J (2000) Memory–a century of consolidation. Science 287: 248–251.

27. Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press.

28. Hart B, Risley T (1995) Meaningful differences in the everyday experience of young American children. Baltimore, MD: Brookes Publishing Company.

29. Roy BC, Frank MC, Roy D (2009) Exploring word learning in a high-density longitudinal corpus. In: Proceedings of the 30th Annual Conference of the Cognitive Science Society.

30. Pinker S (1994) The Language Instinct. New York: Morrow.

31. Orbán G, Fiser J, Aslin RN, Lengyel M (2008) Bayesian learning of visual chunks by human observers Bayesian learning of visual chunks by human observers. Proceedings of the National Academy of Sciences 105: 2745–2750.

32. Perruchet P, Vinter A (1998) PARSER: A model for word segmentation. Journal of Memory and Language 39.

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

34. Anderson JR, Schooler LJ (1990) Reflections of the environment in memory. Psychological Science 2: 396–408.

35. Bahrick HP, Bahrick LE, Bahrick AS, Bahrick PE (1993) Maintenance of foreign language vocabulary and the spacing effect. Psychological Science 4: 316–323.

36. Swingley D (2005) Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology 50: 86–132.

37. Mintz T, Newport E, Bever T (1995) Distributional regularities of form class in speech to young children. In: Proceedings of the Northeastern Linguistics Society. Citeseer, volume 25, pp. 43–54.

38. Markson L, Bloom P (1997) Evidence against a dedicated system for word learning in children. Nature 385: 813–815.

39. Jusczyk PW, Hohne EA (1997) Infants' memory for spoken words. Science 277: 1984–1986.