The specificity of statistical learning

Ansgar Endress

Abstract

Statistical Learning is ubiquitous across domains and species, and might be critical for the earliest stages of language acquisition, for example to identify and memorize words from fluent speech. However, other forms of associative learning are remarkably tuned to the ecological learning situations, and often dissociable from declarative memory. Here, we show that Statistical Learning selectively operates in certain learning situations, and is dissociable from (declarative) memory mechanisms that allow learners to place word-like items in memory. Statistical Learning predominantly operates in continuous speech sequences similar to those used in prior experiments, but not in discrete chunk sequences, even though the latter are likely encountered during language acquisition (due to the prosodic organization of language). Conversely, when exposed to continuous sequences in a memory recall experiment, participants are sensitive to probable syllable transitions, but, to the extent that they remember any items at all, they tend to initiate their productions with random syllables rather than with word onsets, leading to greater recall of *low*-probablility chunks. In contrast, familiarization with discrete sequences produces reliable memories of actual, high-probability forms. This dissociation between Statistical Learning and memory suggests that Statistical Learning might have a specialized role when distributional information can be accumulated (e.g., for predictive processing), and that it is separable from the (declarative) memory mechanisms needed to acquire words.

1 PNAS FORMAT

Research reports describe the results of original research of exceptional importance. The preferred length of these articles is 6 pages, but PNAS allows articles up to a maximum of 12 pages. A standard 6-page article is approximately 4,000 words, 50 references, and 4 medium-size graphical elements (i.e., figures and tables).

Templates are available at https://www.pnas.org/authors/submitting-your-manuscript#manuscript-formatting-guidelines

A manuscript file (in any format) including the following: * Title page (title, author list, classification, keywords) * Abstract (< 250 w) * Significance statement (< 120 w) * Main text - Introduction - Results - Discussion - Materials and methods (describe procedures in sufficient detail so that the work can be repeated) * Acknowledgments and funding sources * References

- Figures or tables with appropriate legends (may be uploaded separately)
- SI files (may be uploaded separately)
- Contact and competing interest information for all authors.
- Data sharing plans (for all data, documentation, and code used in analysis).
- Funding information and whether an open access license has been selected.
- A list of appropriate Editorial Board, NAS members, and qualified reviewers (minimum of three each) who are experts in the * paper's scientific area. A brief justification for suggested reviewers is welcome, particularly for interdisciplinary papers.

2 Significance statement (< 120w)

3 Introduction

Associative learning is remarkably widespread across species and domains [1–7], and might support a wide range of computations, especially during language acquisition [8, 9].

However, associative learning is also remarkably modular [10]. Humans have independent associative learning abilities in superficially similar domains, including associations of objects with landmarks vs. boundaries [11], associations among social vs. non-social objects [12] and associations among consonants vs. vowels [13, 14]. Likewise, preferential associations abound. For example, rats readily associate tastes with sickness and external stimuli with pain, but cannot associate taste with pain or external stimuli with sickness [15]. Such patterns of associations reflect the likely ecological sources of sickness vs. pain (i.e., food vs. external events), and can evolve in just 40 generations in fruit flies [16].

Critically, some associations can be detrimental, and are thus blocked. For example, taste-sickness associations (but not other associations) are blocked in a suckling context for rat pups with no exposure to solid food [17, 18], presumably because avoidance of the *only* food source is costly; in contrast, minimal exposure to solid food re-establishes taste-sickness associations [19].

While such results suggest that, over evolutionary times, the availability of associative learning can be modified for specific stimulus classes, it is less clear if associative learning is specialized for specific computational functions - or essentially a side effect of local neural processing [a "spandrel" in biological terms; 20] that is sometimes adaptive, sometimes neutral and sometimes detrimental. Here, we address this issue in a domain where the importance of associative learning has long been recognized: learning words from fluent speech. We suggest that associative learning is critical for predicting speech material and operates predominantly under conditions where prediction is possible. However, we also suggest that separate mechanisms are required to form (declarative) memories of the words learners need to acquire.

Speech is thought to be a continuous signal, and before learners can commit any words to memory, they need to learn where words start and where they end. They might rely on Transitional Probabilities (TPs) among items, that is, the conditional probability of a syllable σ_{i+1} given a preceding syllable σ_i, $P(\sigma_i \sigma_{i+1})/P(\sigma_i)$. Relatively predictable transitions are likely located inside words, while unpredictable ones straddle word boundaries. Early on, Shannon [21] showed that human adults are sensitive to such distributional information. Subsequent work demonstrated that infants and non-human animals share this ability [1–6], and that it might reflect simple associative mechanisms such as Hebbian learning [22].

Associative knowledge might be critical for predictive processing [23, 24], an ability that is critical for both language [25, 26] and other cognitive processes [27–29]. However, such knowledge does not imply that learners store words in (declarative) long-term memory. In fact, observers prefer high-TP items to low-TP items even if they have never encountered the items and thus could now have memorized them (because the items are played backwards; 7, 30, and sometimes even prefer high-TP items they have *never* encountered to low-TP items they have heard or seen [31]. Such results suggest that associative learning and memory for specific chunks may be dissociable (see also 32–34 and Discussion), a view that is reinforced by the suggestion that representations created by associative learning differ from those used for linguistic stimuli [31, 35].

Here, we explore the computational function of associative learning, focusing on the conditions under which it operates and its relation to memory processes. To explore its operating conditions, we note that speech does not come as a continuous signal but rather as a sequence of smaller units due to its prosodic organization [36–38]. This prosodic organization is perceived in unfamiliar languages [39–42] and even by newborns [43]. This prosodic information might affect the usefulness of statistical learning, because associative learning operates primarily *within* rather than across major prosodic boundaries [44]. As result, the learner's segmentation task is not so much to integrate distributional information over long stretches of continuous speech, but rather to decide whether the correct grouping in prosodic groups such as "*thebaby*" is "*theba + by*" or "*the + baby*" (though prosodic groups are often longer than just three syllables; Nespor and Vogel 37).

In Experiment 1, we thus ask whether associative learning operates in such smaller chunks, or only in longer stretches of continuous speech. In Experiment 2, we seek to elucidate the function of associative learning, asking (adult) participants to recall what they remember after being exposed to the speech stream from Saffran et al.'s [5] classic experiment, again with a continuous speech stream or a sequence of pre-segmented syllable sequences.

4 Results

In Experiment 1, participants listened to a speech sequence of tri-syllabic words. The words were either *pre-segmented* (i.e., with a silence after each word) or continuously concatenated. For half of the participants, both the TPs and the chunk frequency was higher between the the first two syllables of the word than between the last two syllables. An associative learner should thus split a triplet like ABC into an initial AB chunk followed by a singleton C syllable (hereafter $AB+C$ pattern). For the remaining participants, both the TPs and the chunk frequency favored an $A+BC$ pattern. Following this familiarization, they heard pairs of AB and BC items, and had to indicate which item was more like the familiarization items.

When the familiarization stream was pre-segmented, participants failed to split smaller utterances into their underlying components. As shown in Figure 1, the average performance did not differ significantly from the chance level of 50%, (M = 51.67, SD = 15.17), V = 216, p = 0.307. Likelihood ratio analysis favored the null hypothesis by a factor of 4.57 after correction with the Bayesian Information Criterion. As shown in Table S7, performance did not depend on the language condition. As shown in SI SOM4.1, this failure was replicated using a second voice (*en1*, British English male). The failure to use statistical learning to split pre-segmented units was conceptually replicated in a pilot experiment with Spanish/Catalan speakers using chunk frequency and backwards TPs as the primary cues (SI SOM5).

In contrast to the common finding that humans and other animals are sensitive to TPs, our participants failed to use TPs to split pre-segmented utterances into their underlying units. We thus asked if, in line with previous research, they can track TPs units are embedded into a *continuous* speech stream. That is, participants listened to the very same speech stream as in the pre-segmented condition, except that the stream was continuous.

As shown in Figure 1, the average performance differed significantly from the chance level of 50%, (M = 58.51, SD = 16.21), Cohen's d = 0.52, $CI_{.95}$ = 52.66, 64.35, V = 306.5, p = 0.0185. As shown in Table S7, performance did not depend on the language condition, and was marginally better than in the pre-segmented condition (p = .08).

We replicated the successful tracking of statistical information using a new sample of participants. As shown in Figure 1, the average performance differed significantly from the chance level of 50%, (M = 62.78, SD = 21.35), Cohen's d = 0.6, $CI_{.95}$ = 54.81, 70.75, V = 320, p = 0.00778. As shown in Table S7, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition (p = .013).

(As shown in SI SOM4.2, this result could not be replicated using a different voice (*en1*, male British English); participants seemed to prefer specific items, presumably because the synthesizer produced click-like sounds for some stops and fricatives that likely affected syllable grouping.)

Taken together, these results thus suggest that associative learning predominantly operates in continuous sequences, but less so in pre-segmented sequences. Such a result is compatible with the view that associative learning is important for predictive processing, given that continuous sequences are more conducive for prediction. In contrast, it raises doubts as to whether participants can use associative learning to memorize words, given that they do not seem to be able to do so in pre-segmented streams.

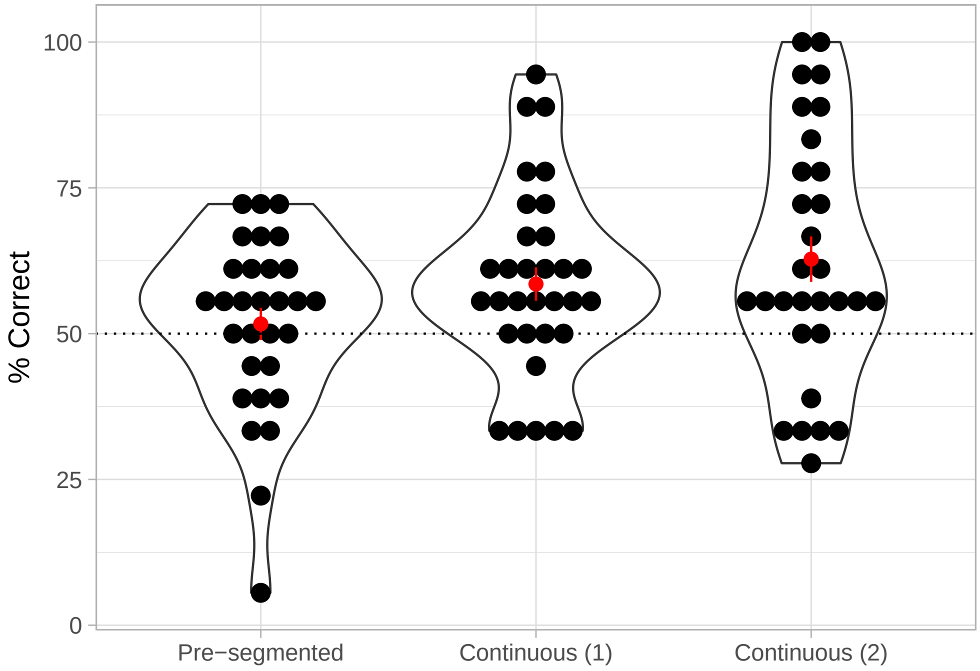

Figure 1: Results of Experiment 1. Each dot represents a participant. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a pre-segmented familiarization stream or (middle, right) a continuous familiarization stream. The two continuous conditions are replictions of one another.

In Experiment 2, we explored the computational function of associative learning, and asked if participants would remember the items that occurred in a speech stream. Adult participants listened to the artificial languages Saffran et al. [5] used with 8-months-olds, except that we doubled the exposure. The languages comprised four words, with a TP of 1.0 within words and 0.33 across word boundaries. The words were presented in a continuous stream or as a pre-segmented word sequence. We ran both a lab-based and an online version of this experiment. Lab-based participants just listened to the speech stream, while online participants watched an astronomical video at the same time.

Following a retention interval, participants had to repeat back the words they remembered from the speech stream. Lab-based participants responded vocally, while online participants typed their answer into a comment field. Finally, participant completed a recognition test during which we pitted words against part-words. Part-words are tri-syllabic items that straddle a word-boundary. For example, if *ABC* and *DEF* are two consecutive words, *BCD* and *CDE* are the corresponding part-words. If participants reliably choose words over part-words, they track TPs.

In the analyses below, we removed single syllable responses (and participants who did not produce any other other items). We also removed participants who did not perform at least 50% during the final recognition test.

As shown in Table 1 and Figures 2a and b, participants produced about 4 items. Neither the number of items produced nor their lengths differed across the segmentation conditions. Critically, and as shown in Table 1 and Figures 2c and d, forward and backward TPs in the participants' responses were significantly greater than the chance level of .083 in both segmentation conditions. These TPs likely underestimate the participants' performance, as we included responses with unattested syllables that might reflect misperceptions; after removing such responses, TPs in the participants' responses were about twice as large. Participants were thus clearly sensitive to the TPs in the speech stream. (TPs were somewhat higher in the pre-segmented condition. This finding does not contradict the results from the Experiment 1 above; after all, if participants faithfully recall familiarization items, the resulting TPs will be high as well.)

The traditional analysis of word segmentation experiments relies on the contrast between words and part-words. As mentioned above, part-words are tri-syllabic items that straddle a word-boundary. We thus calculated the proportion of words among words and part-words or concatenations of words and part-words. If participants faithfully produce trisyllabic sequences from the stream, they can start the sequences on the first, second or third syllable of a word, but only the first possibility yields words rather than part-words. As a result, if participants initiate their productions with a random syllable, a third of their productions should be words.

As shown in Table 1 and in Figure 3a, the proportion of words among words and part-words was close to 100% in the pre-segmented condition, but did not differ from the chancel level of 1/3 in the continuous condition. Likelihood ratio analysis suggests that, in the continuous condition, participants were 3.5 more likely to perform at the chance level of 33% (2.6 for the lab-based experiments) than to perform at a level different from chance. These results thus suggest that participants in the continuous condition initiate their productions at random positions in the stream, and that they did not remember any word forms.

However, inspection of Figure 3a shows that the distribution after continuous sequences is bimodal, with some participants producing only words, and others producing only part-words. Assuming that the number of participants producing words vs. part-words follows a binomial distribution, we can thus calculate the likelihood ratio of a model where learners identify word boundaries (and should thus produce words with probability 1), and a model where they track TPs and initiate productions at random positions (and should produce words with a probability of 1/3). As shown in SI SOM3.3, the likelihood ratio in favor of the first model is 3^{N_W} if participants produce no part-words (i.e., after a pre-segmented familiarization), where N_W is the number of participants producing words; otherwise, the likelihood ratio in favor of the second model is infinity. These results thus suggest that, despite their ability to track TPs, participants initiate productions at random positions in the sequence, and thus do not remember statistically defined words.

However, as shown in Figure S1, these results might be misleading because, in the continuous condition, many participants produce neither words *nor* part-words. In fact, on average, they produce only .4 words and part-words combined, respectively. (In the pre-segmented condition, most participants produce at least

Table 1: Various analyses pertaining to the productions as well as test against their chances levels.

	Continuous	Segmented	p(Continuous vs. Segmented)
Recognition accuracy			
lab-based	M= 0.615, SE= 0.048, p= 0.048	M= 0.923, SE= 0.046, p= 0.0012	0.012
online	M= 0.628, SE= 0.0318, p= 7.84e-05	M= 0.911, SE= 0.0193, p= 7.08e-15	< 0.001
Number of items			
lab-based	M= 4.23, SE= 0.756, p= 0.0016	M= 4.23, SE= 0.818, p= 0.00152	0.812
online	M= 4.03, SE= 0.292, p= 3.17e-14	M= 3.25, SE= 0.202, p= 2.74e-14	0.099
Number of syllables/item			
lab-based	M= 3.79, SE= 0.421, p= 0.0016	M= 2.97, SE= 0.0246, p= 0.0007	0.026
online	M= 2.65, SE= 0.0869, p= 2.29e-14	M= 2.93, SE= 0.0364, p= 1.04e-15	< 0.001
Proportion of words among words and part-words (or concatenations thereof)			
lab-based	M= 0.321, SE= 0.153, 0.798 (vs. 1/3)	M= 1, SE= 0, p= 0.0006 (vs. 1/3)	0.034
online	M= 0.417, SE= 0.105, p= 0.189 (vs. 1/3)	M= 1, SE= 0, p= 2.08e-13 (vs. 1/3)	< 0.001
Forward TPs			
lab-based	M= 0.301, SE= 0.0702, p= 0.0107	M= 0.634, SE= 0.092, p= 0.00159	0.006
online	M= 0.397, SE= 0.0316, p= 6.26e-12	M= 0.583, SE= 0.04, p= 3.82e-13	0.001
Backward TPs			
lab-based	M= 0.301, SE= 0.0702, p= 0.0107	M= 0.634, SE= 0.092, p= 0.00159	0.006
online	M= 0.397, SE= 0.0316, p= 6.26e-12	M= 0.583, SE= 0.04, p= 3.82e-13	0.001
Proportion of High-TP chunks among High- and Low-TP chunks			
lab-based	M= 0.75, SE= 0.289, p= 0.85 (vs. 2/3)	M= 1, SE= 0, p= 0.0006 (vs. 2/3)	1.000
online	M= 0.767, SE= 0.0459, p= 0.00154 (vs. 2/3)	M= 0.97, SE= 0.0187, p= 6.75e-13 (vs. 2/3)	< 0.001
Proportion of items with correct initial syllables			
lab-based	M= 0.333, SE= 0.105, p= 0.856	M= 0.809, SE= 0.0694, p= 0.00186	0.016
online	M= 0.419, SE= 0.0392, p= 0.0864	M= 0.738, SE= 0.0387, p= 1.58e-11	0.000
Proportion of items with correct final syllables			
lab-based	M= 0.456, SE= 0.125, p= 0.5	M= 0.818, SE= 0.0829, p= 0.00222	0.025
online	M= 0.386, SE= 0.043, p= 0.456	M= 0.7, SE= 0.0437, p= 4.14e-10	0.000

one word, with an average of 1.26.) Given that participants produce few tri-syllabic items, we thus focus on shorter chunks.

We first focus on bisyllabic chunks. They are either high-TP chunks that are part of a word, or low-TP chunks that straddle a word boundary. For example, with two consecutive words ABC and DEF, the high-TP chunks are AB, BC, ..., while the low-TP chunk is CD. As a result, two-syllable items have a 66% probability of being a high-TP chunk. As shown in Figure 3b, the proportion of high-TP among chunks high- and low-TP chunks exceeded chance in the pre-segmented condition, but not in the continuous condition. In the continuous condition, the likelihood ratio in favor of the null hypothesis is 0.652 (1.892 for the lab-based experiments). These results are thus consistent with the possibility that, in the continuous condition, participants do track TPs, but initiate their productions at random positions.

Finally, we analyze the productions in terms of correct initial final syllables. As there are four words with one correct initial and final syllable each, and 12 syllables in total, 4/12 of the productions should have "correct" initial syllables, and 4/12 should have correct final syllables.

As shown in Table 1 and Figure 3c and d, participants produced items with correct initial or final syllables at greater than chance level only in the segmented condition, but not the continuous condition. In the continuous condition, the likelihood ratio in favor of the null hypothesis was 0.785 for initial syllables (3.606 for the lab-based experiment) and 4.061 for final syllables (2.139 for the lab-based experiment).

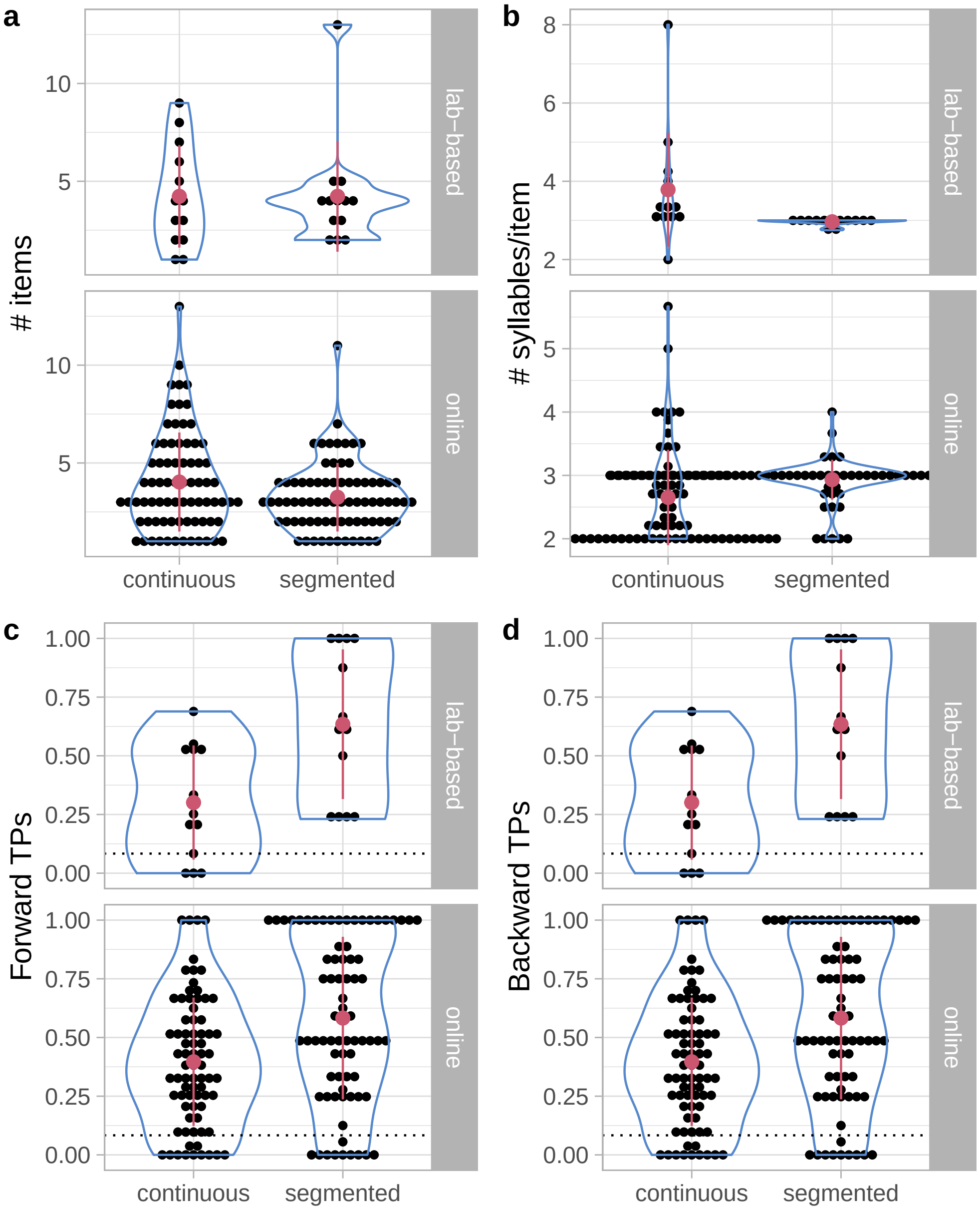

Figure 2: Number of items produced, number of syllables per item and forward and backward TPs. The dotted line represents the chance level for a randomly ordered syllable sequence.

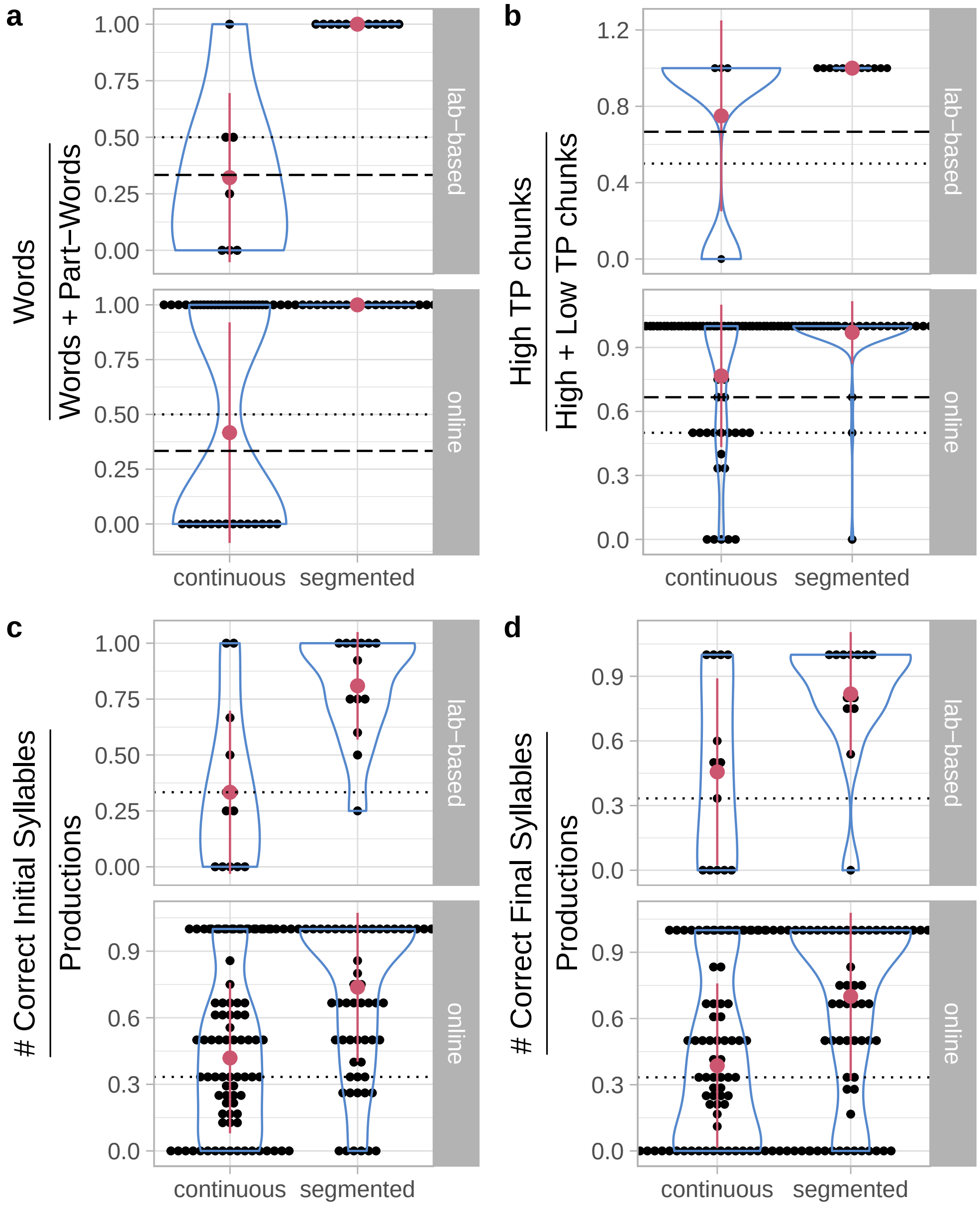

Figure 3: Analyses of the participants' productions. (a) Proportion of words among words and part-words. The dotted line represents the chance level of 50% in a two-alternative forced-choice task, while the dashed line represents the chance level of 33% that an attested 3 syllable-chunk is a word rather than a part-word. (b) Proportion of high-TP chunks among high- and low-TP chunks. The dashed line represents the chance level of 66% that an attested 2 syllable-chunk is a high-TP rather than a low-TP chunk. (c) proportion of productions with correct initial syllables and (d) with correct final syllables. The dotted line represents the chance level of 33%.

5 Discussion

Taken together, Experiments 1 and 2 suggest that associative learning and (declarative) memory might fulfill different computational functions. In Experiment 1, participants tracked statistical dependencies predominantly when they were embedded in a continuous speech stream, but not across pre-segmented chunk sequences. These results are consistent with Shukla et al's [44] finding that associative learning predominantly occurs within major prosodic groups, and, within these groups, predominantly at the edges of those groups; we show that, with shorter and better separated groups, associative learning can be abolished altogether. In line with results from conditioning experiments [15, 17–19], associative learning can thus be enhanced or suppressed depending on the learning situation. The enhanced associative learning in continuous sequences is consistent with the view that associative learning is important for predictive processing [23, 24], given that prediction is arguably more useful in lengthy chunks. It is also consistent with the view that associative learning may be less important for memorizing utterances, especially given that, due to its prosodic organization, speech tends to be pre-segmented into smaller groups such as those used by Shukla et al [44] or in Experiment 1 [36, 38–43].

Experiment 2 showed that, even when participants successfully track associative information, they remember familiarization items only when familiarized with a pre-segmented sequence; in contrast, when familiarized with a continuous sequence, their productions started with random syllables rather than actual word onsets. Given that the memory representations of linguistic items are based on their initial and final syllables [31, 35], this result thus suggests that associative learning did not lead to the creation of declarative memory representations

The combined results of Experiments 1 and 2 echo dissociations between associative learning and declarative memory. Such dissociations have long been documented behaviorally [45], developmentally [46] and neuropsychologically [32–34, 47], to the extent that statistical predictions can *impair* declarative memory encoding in healthy adults [23]. The standard conclusion is that the (cortical) declarative memory system might be independent of a (neostriatal) system for associative learning [32–34]. In line with earlier proposals [23, 24], we thus suggest that the computational function of associative learning might be distinct from that of (declarative) memory encoding, and that associative learning might be more important for predictive processing. The relative salience of these mechanisms might depend on how adaptive they are for the learning problem at hand.

These results also have implications for the more specific problem of word segmentation. If learners cannot use associative learning to encode word candidates in (declarative) memory, they need to use other cues. Possible cues include using known words as delimiters for other words [48–50], attentional allocation to beginnings and ends of utterances [44, 51, 52], legal sound sequences [53, 54] and universal aspects of prosody [39–43, 55]. Such cues might plausible support declarative memories of words because they (but not transition-based associative information) are consistent with how linguistic sequences are encoded in declarative long-term memory, where linguistic sequences are encoded with reference to their first and their last element [31, 35].

Taken together, associative learning and declarative memory might thus have separable functions, the former for predictive processing and the latter for remembering objects and episodes.

6 Methods summary

Unless otherwise stated, stimuli were synthesized using mbrola [56] and the *us3* (American English male) voice. Lab-based experiments were run using Psyscope X (http://psy.ck.sissa.it) in a quiet room. Online experiments were run on https://testable.org.

6.1 Participants

In Experiment 1, 30, 30 and 31 participants were retained for analysis for the pre-segmented condition, the continuous condition and its replication. In Experiment 2, 26 participants were retained for the lab-based version, and 157 for the online version. Participants reported to be native speakers of English.

6.2 Experiment 1

Participants were instructed to listen to a monologue in "Martian", and to remember the Martian words. Following this, they listened to a sequence of tri-syllabic words (Language 1: *w3:legu:*, *w3:levOI*, *w3:lenA:*, *faIzO:gu:*, *faIzO:vOI*, *faIzO:nA:*, *rVb{gu:*, *rVb{vOI*, *rVb{nA:*; Language 2: *w3:legu:*, *faIlegu:*, *rVlegu:*, *w3:zO:vOI*, *faIzO:vOI*, *rVzO:vOI*, *w3:b{nA:*, *faIb{nA:*, *rVb{nA:*). In Language 1 and 2, both TPs and the chunk frequency favored *AB+C* and *A+BC* patterns, respectively (TPs of 1.0 vs. 1/3; see main text). Segments lasted 60 ms and had an F_0 of 120 Hz. Sequences (45 repetitions/word) were either continuous or had 540 ms silences between words. Sequences were then played thrice (total familiarization: 7 min 17s (continuous); 18 min 14 s (pre-segmented)).

Following this familiarization, participants listened to pairs of items and had to choose the more "Martian" one. One item comprised the *first two* syllables of a word, one the *last two* syllables. The three items of each kind were combined into 9 test pairs. The test pairs were presented twice.

6.3 Experiment 2

Participants were instructed to listen to a monologue in "Martian", and to remember the Martian words. The languages were those from Saffran et al. [5] Experiment 2 (Language 1: *pAbiku*, *tibudO*, *dArOpi*, *gOLAtu*; Language 2: *bikuti*, *pigOLA*, *tudArO*, *budOpA*). Segments lasted 108 ms at an F_0 of 120 Hz. The words were combined into 20 sequences (45 repetitions/word) with different random orders, either continuously or with 222 ms silences between words. Sequences were played twice (total familiarization: 3 min 53 (continuous) and 5 min 13 (pre-segmented)). Online participants watched a nebula during familiarization.

Following the familiarization and a 30 s filled retention interval, participants completed the recall test. Lab-based participants had 45 s to repeat back the words they remembered; their vocalizations were recorded for offline analysis. Online participants had 60 s to type their answer into a comment field. Finally, participants completed a recognition test during which we pitted words against part-words.

6.4 Analysis of productions

The responses were transformed using a set of substitutions rules to allow for misperceptions (e.g., confusion between /b/ and /p/) or orthographic variability (e.g., *ea* and *ee* both reflect the sound /i/). Finally, we selected the best matches to the familiarization stimuli (see SI SOM2.2.1).

References

[1] Richard N Aslin, Jenny R Saffran, and Elissa L Newport. Computation of conditional probability statistics by 8-month-old infants. *Psychol Sci*, 9:321–324, 1998.

[2] Jiani Chen and Carel Ten Cate. Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behav Processes*, 117:29–34, Aug 2015. doi: 10.1016/j.beproc.2014.09.004.

[3] József Fiser and Richard N Aslin. Statistical learning of new visual feature combinations by infants. *Proc Natl Acad Sci U S A*, 99(24):15822–6, 2002. doi: 10.1073/pnas.232472899.

[4] Marc D Hauser, Elissa L Newport, and Richard N Aslin. Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78(3):B53–64, 2001.

[5] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–8, 1996.

[6] Juan M Toro, Josep B Trobalon, and Núria Sebastián-Gallés. Effects of backward speech and speaker variability in language discrimination by rats. *J Exp Psychol Anim Behav Process*, 31(1):95–100, Jan 2005. doi: 10.1037/0097-7403.31.1.95.

[7] Nicholas B Turk-Browne and Brian J Scholl. Flexible visual statistical learning: Transfer across space and time. *J Exp Psychol: Hum Perc Perf*, 35(1):195–202, 2009.

[8] Richard N. Aslin and Elissa L. Newport. Statistical learning. *Current Directions in Psychological Science*, 21(3):170–176, 2012. doi: 10.1177/0963721412436806.

[9] Mark S. Seidenberg, Maryellen C. MacDonald, and Jenny R. Saffran. Does grammar start where statistics stop? *Science*, 298(5593):553–554, 2002.

[10] Ansgar D. Endress. Duplications and domain-generality. *Psychological Bulletin*, 145(2), 2019. doi: 10.1037/bul0000213.

[11] Christian F Doeller and Neil Burgess. Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proc Natl Acad Sci U S A*, 105(15):5909–14, Apr 2008. doi: 10.1073/pnas.0711433105.

[12] Steven H Tompson, Ari E Kahn, Emily B Falk, Jean M Vettel, and Danielle S Bassett. Individual differences in learning social and nonsocial network structures. *Journal of experimental psychology. Learning, memory, and cognition*, 45:253–271, February 2019. ISSN 1939-1285. doi: 10.1037/xlm0000580.

[13] Luca L Bonatti, Marcela Peña, Marina Nespor, and Jacques Mehler. Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychol Sci*, 16(8): 451–459, 2005.

[14] Juan M. Toro, L.L. Bonatti, M. Nespor, and J. Mehler. Finding words and rules in a speech stream: functional differences between vowels and consonants. *Psychol Sci*, 19:137–144, 2008.

[15] J. Garcia, W. G. Hankins, and K. W. Rusiniak. Behavioral regulation of the milieu interne in man and rat. *Science*, 185(4154):824–31, Sep 1974.

[16] Aimee S. Dunlap and David W. Stephens. Experimental evolution of prepared learning. *Proceedings of the National Academy of Sciences*, 111(32):11750–11755, 2014. doi: 10.1073/pnas.1404176111. URL http://www.pnas.org/content/111/32/11750.abstract.

[17] L T Martin and J R Alberts. Taste aversions to mother's milk: the age-related role of nursing in acquisition and expression of a learned association. *Journal of comparative and physiological psychology*, 93:430–445, June 1979. ISSN 0021-9940.

[18] Jeffrey R. Alberts and David J. Gubernick. Early learning as ontogenetic adaptation for ingestion by rats. *Learn Motiv*, 15(4):334 – 359, 1984. ISSN 0023-9690. doi: 10.1016/0023-9690(84)90002-X.

[19] D J Gubernick and J R Alberts. A specialization of taste aversion learning during suckling and its weaning-associated transformation. *Dev Psychobiol*, 17:613–628, November 1984. ISSN 0012-1630. doi: 10.1002/dev.420170605.

[20] S. J. Gould, R. C. Lewontin, J. Maynard Smith, and Robin Holliday. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):581–598, 1979. doi: 10.1098/rspb.1979.0086.

[21] C. E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64, jan 1951. doi: 10.1002/j.1538-7305.1951.tb01366.x.

[22] Ansgar D Endress and S P Johnson. When forgetting fosters learning: A neural network model for statistical learning. *Cognition*, 104621, 2021. doi: 10.1016/j.cognition.2021.104621.

[23] Brynn E. Sherman and Nicholas B. Turk-Browne. Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences of the United States of America*, 117:22760–22770, September 2020. ISSN 1091-6490. doi: 10.1073/pnas.2013291117.

[24] Nicholas B Turk-Browne, Brian J Scholl, Marcia K Johnson, and Marvin M Chun. Implicit perceptual anticipation triggered by statistical learning. *Journal of neuroscience*, 30:11177–11187, 2010. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.0858-10.2010.

[25] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, mar 2008. doi: 10.1016/j.cognition.2007.05.006.

[26] J. C. Trueswell, I. Sekerina, N. M. Hill, and M. L. Logrip. The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition*, 73(2):89–134, Dec 1999.

[27] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, may 2013. doi: 10.1017/s0140525x12000477.

[28] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, jan 2010. doi: 10.1038/nrn2787.

[29] Georg B. Keller and Thomas D. Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, oct 2018. doi: 10.1016/j.neuron.2018.10.003.

[30] Jason Jones and Harold Pashler. Is the mind inherently forward looking? comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, 14:295–300, April 2007. ISSN 1069-9384. doi: 10.3758/bf03194067.

[31] Ansgar D. Endress and A Langus. Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, 92:37–64, 2017. doi: 10.1016/j.cogpsych.2016.11.004.

[32] B J Knowlton, J A Mangels, and L R Squire. A neostriatal habit learning system in humans. *Science*, 273:1399–1402, September 1996. ISSN 0036-8075.

[33] R A Poldrack, J Clark, E J Paré-Blagoev, D Shohamy, J Creso Moyano, C Myers, and M A Gluck. Interactive memory systems in the human brain. *Nature*, 414:546–550, November 2001. ISSN 0028-0836. doi: 10.1038/35107080.

[34] Larry R. Squire. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99(2):195–231, 1992. doi: 10.1037/0033-295x.99.2.195.

[35] Simon Fischer-Baum, Jonathan Charny, and Michael McCloskey. Both-edges representation of letter position in reading. *Psychon Bull Rev*, 18(6):1083–1089, Dec 2011. doi: 10.3758/s13423-011-0160-3.

[36] A Cutler, D Oahan, and Wilma van Donselaar. Prosody in the comprehension of spoken language: A literature review. *Lang Speech*, 40(2):141–201, 1997.

[37] M. Nespor and I. Vogel. *Prosodic Phonology.* Dordrecht, Foris, 1986.

[38] S. Shattuck-Hufnagel and A. E. Turk. A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res*, 25(2):193–247, Mar 1996.

[39] Diane Brentari, Carolina González, Amanda Seidl, and Ronnie Wilbur. Sensitivity to visual prosodic cues in signers and nonsigners. *Lang Speech*, 54(1):49–72, 2011.

[40] Ansgar D. Endress and Marc D. Hauser. Word segmentation with universal prosodic cues. *Cognit Psychol*, 61(2):177–199, 2010.

[41] Jordan Fenlon, Tanya Denmark, Ruth Campbell, and Bencie Woll. Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2):177–200, 2008.

[42] Robert Pilon. Segmentation of speech in a foreign language. *J. Psycholinguist. Res.*, 10(2):113 – 122, 1981. ISSN 0090-6905.

[43] Anne Christophe, Jacques Mehler, and Nuria Sebastian-Galles. Perception of prosodic boundary correlates by newborn infants. *Infancy*, 2(3):385–394, 2001.

[44] Mohinish Shukla, Marina Nespor, and Jacques Mehler. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognit Psychol*, 54(1):1–32, Feb 2007. doi: 10.1016/j.cogpsych.2006.04.002.

[45] Peter Graf and George Mandler. Activation makes words more accessible, but not necessarily more retrievable. *Journal of Verbal Learning and Verbal Behavior*, 23(5):553–568, 1984. doi: 10.1016/s0022-5371(84)90346-3.

[46] Amy S. Finn, Priya B. Kalra, Calvin Goetz, Julia A. Leonard, Margaret A. Sheridan, and John D.E. Gabrieli. Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology*, 142:212–220, feb 2016. doi: 10.1016/j.jecp.2015.09.027.

[47] Nuttida Rungratsameetaweemana, Larry R Squire, and John T Serences. Preserved capacity for learning statistical regularities and directing selective attention after hippocampal lesions. *Proc. Natl. Acad. Sci. U.S.A.*, 116:19705–19710, September 2019. ISSN 1091-6490. doi: 10.1073/pnas.1904502116.

[48] Heather Bortfeld, James L Morgan, Roberta Michnick Golinkoff, and Karen Rathbun. Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychol Sci*, 16(4):298–304, 2005. doi: 10.1111/j.0956-7976.2005.01531.x.

[49] MR Brent and JM Siskind. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–44, 2001.

[50] Karima Mersad and Thierry Nazzi. When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, 8(3):303–315, 2012. doi: 10.1080/15475441.2011.609106.

[51] Padraic Monaghan and Morten H. Christiansen. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *J Child Lang*, 37(3):545–564, Jun 2010. doi: 10.1017/S0305000909990511.

[52] Amanda Seidl and Elizabeth K Johnson. Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *J Child Lang*, 35(1):1–24, Feb 2008.

[53] James M. McQueen. Segmentation of continuous speech using phonotactics. *J Mem Lang*, 39(1):21–46, 1998.

[54] Anne Pier Salverda, Delphine Dahan, Michael K Tanenhaus, Katherine Crosswhite, Mikhail Masharov, and Joyce McDonough. Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, 105(2):466–76, Nov 2007. doi: 10.1016/j.cognition.2006.10.008.

[55] Elizabeth K Johnson and Amanda H Seidl. At 11 months, prosody still outranks statistics. *Dev Sci*, 12 (1):131–41, Jan 2009. doi: 10.1111/j.1467-7687.2008.00740.x.

[56] T Dutoit, V Pagel, N Pierret, F Bataille, and O van der Vreken. The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 3, pages 1393–1396, Philadelphia, 1996.

[57] R. Harald Baayen, D.J. Davidson, and D.M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390 – 412, 2008. ISSN 0749-596X. doi: 10.1016/j.jml.2007.12.005.

[58] Scott Glover and Peter Dixon. Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychon Bull Rev*, 11(5):791–806, Oct 2004.

Supplementary Online Materials

SOM1 Methods

SOM1.1 Recognition experiment

SOM1.1.1 Participants

Table S1: Demographics of the final sample for Experiment 1.

Familiarization Condition	N	Females	Males	Age (M)	Age (range)
Pre-segmented	30	18	12	26.3	18-43
Continuous (1)	32	26	6	20.1	18-44
Continuous (2)	30	20	10	23.2	18-36

Participants were recruited from the City, University London participant pool and received course credit or monetary compensation for their time. We targeted 30 participants per experiment (15 per language). The final demographic information is given in Table S1. An additional six participants took part in the experiment but were not retained for analysis because they had taken part in a prior version of this experiment ($N = 4$), were much older than the rest of our sample ($N = 2$), or used their phone during the experiment or were visibly inattentive ($N = 2$). Participants reported to be native speakers of English.

SOM1.1.2 Design

Participants were familiarized with a sequence of tri-syllabic words. In Language 1, both the TPs and the chunk frequency was higher in the bigram formed by the first two syllables than in the bigram formed by the last two syllables; as a result, an associative learner should split a triplet like *ABC* into an initial *AB* chunk followed by a singleton *C* syllable (hereafter *AB+C* pattern). In Language 2, both the TPs and the chunk frequency favored an *A+BC* pattern. The basic structure of the words is shown in Table S2.

Table S2: Design of Experiment 1. (Left) Language structure. (Middle) Structure of test items. Correct items for Language 1 are foils for Language 2 and vice versa. (Right) Actual items in SAMPA format; dashes indicate syllable boundaries.

Word structure for		Test item structure for		Actual words for	
Language 1	Language 2	Language 1	Language 2	Language 1	Language 2
ABC	ABC	AB	BC	w3:-le-gu:	w3:-le-gu:
ABD	FBC	FG	GD	w3:-le-vOI	faI-le-gu:
ABE	HBC	HJ	JE	w3:-le-nA:	rV-le-gu:
FGC	AGD			faI-zO:-gu:	w3:-zO:-vOI
FGD	FGD			faI-zO:-vOI	faI-zO:-vOI
FGE	HGD			faI-zO:-nA:	rV-zO:-vOI
HJC	AJE			rV-b{-gu:	w3:-b{-nA:
HJD	FJE			rV-b{-vOI	faI-b{-nA:
HJE	HJE			rV-b{-nA:	rV-b{-nA:

As result, in Language 1, the first bigram has a (forward and backward) TP of 1.0, while the second bigram has a (forward and backward) TP of .33. In contrast, in Language 2, the first bigram has a forward TP of .33, while the second bigram has a forward TP of 1.0. Likewise, the initial bigrams were three times as frequent as the final ones for Language 1, while the opposite holds for Language 2.

We asked whether participants would extract initial bigrams or final bigrams. The test items are given in Table S2.

SOM1.1.3 Stimuli

Stimuli were synthesized using the *us3* (American English male) voice from mbrola [56]. (We also used the *en1* (British English male) voice; however, as discussed below, this voice turned out to be of relatively low quality and introduced confounds in the data.)

Segments had a constant duration of 60 ms (syllable duration 120 ms) with a constant F_0 of 120 Hz. These values were chosen to match recordings of natural speech that were intended to be used in investigations of prosodic cues to word segmentation.

For continuous streams, a single file with 45 repetitions of each word was synthesized for each language (2 min 26 s duration). It was faded in and out for 5 s using sox (http://sox.sourceforge.net/) and then compressed to an mp3 file using ffmpeg (https://ffmpeg.org/). The stream was then presented 3 times to a participant (total familiarization duration 7 min 17 s). The random order of the words was different for every participant.

For segmented streams, words were individually synthesized using mbrola. We then used a custom-made Perl script to randomize the words for each participant and concatenate them into a familiarization file using sox. The order of words was then randomized for each participant and concatenated into a single aiff file using sox. The silence among words was 540 ms (1.5 word durations). The total stream duration was 6 min 12s. The stream was then presented 3 times to a participant (total familiarization: 18 min 14 s).

SOM1.1.4 Apparatus

The experiment was run using Psyscope X (http://psy.ck.sissa.it). Stimuli were presented over headphones in a quiet room. Responses were collected from pre-marked keys on the keyboard.

SOM1.1.5 Procedure

Participants were informed that they would listen to a monologue by a talkative Martian, and instructed to try to remember the Martian words. Following this, they listened to three repetitions of the familiarization stream described above, for a total familiarization duration of 7 min 17 s (continuous stream) or 18 min 14 s (segmented stream).

Following this familiarization, participants were presented with pairs of items with an inter-stimulus interval of 500 ms, and had to choose which items was more like what they heard during familiarization. One item comprised the first two syllables of a word, and was a correct choice for Language 1. The other items comprised the last two syllables of a word, and was a correct choice for Language 2. There were three items of each kind. They were combined into 9 test pairs. The test pairs were presented twice, with different item orders, for a total of 18 test trials.

SOM1.2 Recall experiment

SOM1.2.1 Materials

We re-synthesized the languages used in Saffran et al. [5] Experiment 2. The four words in each language are given in Table S3. Stimuli were synthesized using the us3 (male American English) voice of the mbrola synthesizer [56], at a constant F_0 of 120 Hz and at a rate of 216 ms per syllable (108 ms per phoneme).

During familiarization, words were presented 45 times each. We generated random concatenations of 45 repetitions of the 4 words, with the constraint that words could not occur in immediate repetition. Each randomization was then (i) synthesized into a continuous speech stream using mbrola and then converted to mp3 using ffmpeg (https://ffmpeg.org/) (ii) used to concatenate words that had been synthesized in isolation, separated by silences of 222 ms into a segmented speech stream, which was then converted to mp3. Streams were faded in and out for 5 s using sox (http://sox.sourceforge.net/). For continuous streams, this yielded a stream duration of 1 min 57 s; for segmented streams, the duration was 2 min 37.

We created 20 versions of each stream with different random orders of words.

Table S3: Languages used Experiment 2. The words are the same as in [5] Experiment 2.

L1	L2
pabiku	bikuti
tibudo	pigola
daropi	tudaro
golatu	budopa

SOM1.2.2 Procedure

SOM1.2.2.1 Familiarization

Participants were informed that they would be listening to an unknown language and that they should try to learn the words from that language. Following, the familiarization stream was presented twice, leading to a total familiarization duration of 3 min 53 for the continuous streams and 5 min 13 for the segmented streams. They could proceed to the next presentation of the stream by pressing a button.

For the online experiments, participants watched a video with no clear objects during the familiarization (panning of the Carina nebula, obtained from https://esahubble.org/videos/heic0707g/). The video was combined with the speech stream using the muxmovie utility.

Following the familiarization, there was a 30 s retention interval. In both the lab-based and the online experiments, participants were instructed to count backwards from 99 in time with a metronome beat at 3s / beat. Performance was not monitored.

SOM1.2.2.2 Recall test

Following the retention interval, participants completed the recall test. During the lab-based experiments, participants had 45 s to repeat back the words they remembered; their vocalizations were recorded using ffmpeg and saved in mp3 format. During the web-based experiments, participants had 60 s to type their answer into a comment field, during which they viewed a progress bar.

SOM1.2.2.3 Recognition test

Following the recall test, participant completed a recognition test during which we pitted words against part-words. The (correct) test words for Language 1 (and part-words for Language 2) were /pAbiku/ and /tibudO/; the (correct) test words for Language 2 (and part-words for Language 1) were /tudArO/ and /pigOlA/. These items were combined into 4 test pairs.

SOM2 Analysis

SOM2.1 Recognition tests

Accuracy was averaged for each participant, and the scores were tested against the chance level of 50% using Wilcoxon tests. Performance differences across the languages (Language 1 vs. 2) and, when applicable, familiarization conditions (pre-segmented vs. continuous) were assessed using a generalized linear mixed model for the trial-by-trial data with the fixed factors language and, where applicable, familiarization condition, as well as random slopes for participants, correct items and foils. Following [57], random factors were removed from the model when they did not contribute to the model likelihood.

We use likelihood ratios to provide evidence for the null hypothesis that performance did not differ from the chance level of 50%. Following [58], we fit the participant averages to (i) a linear model comprising only an intercept and (ii) the null model fixing the intercept to the appropriate baseline level, and evaluated the likelihood of these models after correcting for the difference in the number of parameters using the Bayesian Information Criterion.

SOM2.2 Recall test

SOM2.2.1 Analysis procedure

Participants in Experiment 2 had to recall what they remembered from the familiarization streams. Lab-based participants were recorded and their productions were transcribed by two independent observers. Disagreements were resolved by discussion. Online participants typed their responses directly into a comment box. We then applied a number of substitution rules to allow for misperceptions (e.g., a confusion between /p/ and /b/) and orthographic variability (e.g., *tea* and *tee* are both pronounced as /ti/). The complete list of substitution rules is shown in Table S4.

Each recall response was analyzed in five steps. First, we applied pre-segmentation substitution rules to make the transcriptions more consistent (see Table S4, "before segmentation"). For example, *ea* (presumably as in *tea*) was replaced with *i*. These substitutions were not considered when calculating the derivation length (see below).

Second, responses were segmented into their underlying units. If the response did not contain any commata (,) or semicolons (;), any spaces in the response were used to delineate units. If a response contained a semicolon or comma, these were used to delineate units. For each of the resulting units, we verified if they contained additional spaces. If they did, these spaces were removed if further segmenting the units based on the spaces resulted in one or more single-syllable units (operationalized as a string with a single vowel); otherwise, the units were further sub-divided based on the spaces. The rationale for this algorithm is that responses such as *bee coo tee,two da ra,bout too pa* were likely to reflect the words *bikuti*, *tudaro* and *budopa*.

Third, we removed geminate consonants and applied another set of substitution rules to take into account possible misperceptions (see Table S4). For example, we treated the voiced and unvoiced variety of stop consonants as interchangeable. Specifically, for each "*surface*" form produced by the participants, we generated candidate "*underlying*" forms by recursively applying all substitutions rules and keeping track of the number of substitution rules that were applied to derive an underlying form from a surface form. For each unique candidate underlying form, we kept the shortest derivation.

Fourth, for each candidate underlying form, we identified the longest matching string in the familiarization stream. The algorithm first verified if a form was contained in a speech stream starting with an A, B or C syllable; if the underlying form contained unattested syllable, one syllable change was allowed with respect to the speech streams. If no matches were found, two sub-strings were created by clipping the first or the last syllable from the underlying form, and the search was repeated recursively for each of these sub strings until a match was found. We then selected the longest match for all sub strings.

Fifth, for each surface form, we selected the underlying form among the candidate underlying forms using three criteria:

1. The winning underlying form had had the maximal *number of attested syllables* among candidate underlying forms;
2. The winning underlying form had the *maximal length* among candidate underlying forms;
3. The winning underlying form had the *shortest derivation* among candidate underlying forms.

The criteria were applied in this order.

SOM2.2.1.1 Substitution rules compensating for potential misperceptions

All substitution rules are listed in Table S4. We now motivate the substitution rules compensating for potential misperceptions:

- /O/ might be perceived as /A/
- Voiced and unvoiced consonants can be confused; that is /g/ can be confused with /k/, /d/ with /t/ and /b/ and /p/.
- /b/ might be perceived as /v/.

In some cases, these rules result in multiple possible matches. For example, the transcription *rapidala* might correspond to /rOpidAlA/ or /rOpidOlA/.

In such cases, we apply the following criteria (in the following order) to decide which match to choose.

1. Choose the option leading to more or longer chunks that are attested in the speech stream.
2. If multiple options lead to chunks of equal length, choose the option requiring fewer changes with respect to the original transcription.

SOM2.2.2 Measures of interest

We computed various properties for each underlying form, given the "target" language the participant had been exposed to. All measures provided in the raw data are described in Table S5.

SOM2.2.2.1 Measures

For each underlying form, we calculate:

1. the number of syllables;
2. whether it was a word from the target language;
3. whether it was a concatenation of words from the target language;
4. whether it was a single word or a concatenation of words from the target language (i.e., the disjunction of (2) and (3));
5. whether it was a part-words from the target language,
6. whether it was a *complete* concatenation of part-words from the target language (i.e., the number of syllables of the item had to be a multiple of three, without any unattested syllables);
7. whether it was a single part-word or a concatenation of part-words from the target language;
8. whether it was high-TP chunk (i.e., a word with the first or the last syllable missing, after removing any leading or trailing unattested syllables);
9. whether it was a low-TP chunk (i.e., a chunk of the form $C_i A_j$, after removing lead or trailing unattested syllables;
10. whether it had a "correct" initial syllable
11. whether it had a "correct" final syllable;
12. whether it is part of the speech stream (i.e., the disjunction of being an attested syllable, being a word or a concatenation thereof, being a part-word or a concatenation thereof, being a high-TP chunk or a low-TP chunk);

Table S4: Substitution rules applied to the participants vocalizations before and after the input was segmented into chunks. The patterns are given as Perl regular expressions. Substitutions prior to segmentation were not counted when calculating the derivation length.

Before segmentation		After segmentation	
Pattern	Replacement	Pattern	Replacement
\.{3,}		u	o
-		v	b
2	tu	p	b
two	tu	b	p
([aeou])ck	\1k	t	d
ar([,\s+])	a\1	d	t
ar$	a	k	g
tyu	tu	g	k
ph	f	a	o
th	t		
qu	k		
ea	i		
ou	u		
aw	a		
ai	a		
ie	i		
ee	i		
oo	u		
e	i		
c	k		
w	v		
y	i		
h			

Table S5: Analyses performed for the vocalizations

Column name in data file	Meaning
n.items	Number of recalled items
n.syll	Mean number of syllables of the recalled items
n.words	Number of recalled words
p.words	Proportion (among recalled items) of words
n.words.or.multiple	Number of recalled words or concatenation of words
p.words.or.multiple	Proportion (among recalled items) of words or concatenation of words
n.part.words	Number of recalled part-words
p.part.words	Proportion (among recalled items) of part-words
n.part.words.or.multiple	Number of recalled part-words or concatenation of part-words
p.part.words.or.multiple	Proportion (among recalled items) of part-words or concatenation of part-words
p.words.part.words	Proportion of words among (recalled) words and part-words. This is used for comparison to the recognition test.
p.words.part.words.or.multiple	Proportion of words among (recalled) words and part-words or concatenation thereof. This is used for comparison to the recognition test.
n.high.tp.chunk	Number of high TP chunks. High TP chunks are defined as two-syllabic chunk from a word
p.high.tp.chunk	Proprtion (among recalled items) of high TP chunks. High TP chunks are defined as two-syllabic chunk from a word
n.low.tp.chunk	Number of low TP chunks. Low TP chunks are defined as two-syllabic word transitions
p.low.tp.chunk	Proportion (among recalled items) of low TP chunks. Low TP chunks are defined as two-syllabic word transitions
p.high.tp.chunk.low.tp.chunk	Proportion of high-TP chunks among high and low-TP chunks. High TP Chunks are defined as two-syllabic chunks from words; low TP chunks are two-syllabic word transitions
average_fw_tp	Average (across recalled items) of average forward TPs among transitions in a given item.
average_fw_tp_d_actual_expected	Average (across recalled items) of the difference between the average ACTUAL forward TPs among transitions in a given item and the EXPECTED forward TP in that item, based on the items first element. See calculate.expected.tps.for.chunks for the calculations
average_bw_tp	Average (across recalled items) of average backward TPs among transitions in a given item.
p.correct.initial.syll	Proportion (among recalled items) that have a correct initial syllable.
p.correct.final.syll	Proportion (among recalled items) that have a correct final syllable.
p.correct.initial.or.final.syll	Proportion (among recalled items) that have a correct initial or final syllable.

13. the average forward TP of the transitions in the form;
14. the *expected* forward TP of the form if form is attested in the speech stream (see below for the calculation);
15. the average backward TP of the transitions in the form.

SOM2.2.2.2 Expected TPs

For items that are *correctly* reproduced from the speech stream, the expected TPs depend on the starting position. For example, the expected TPs for items of at least 2 syllables starting on an initial syllable are (1,

Table S6: Demographics of the final sample. The lab-based participants completed both segmentation conditions.

Sequence Type	Language	N	Females	Male	Age (M)	Age (range)
Lab-based						
continuous	both	13	13	0	17.8	0-22
segmented	both	13	13	0	17.8	0-22
Online						
continuous	L1	38	8	30	31.7	18-71
continuous	L2	38	18	20	29.7	19-71
segmented	L1	38	11	27	28.8	18-55
segmented	L2	38	4	34	29.0	18-62

1, 1/3, 1, 1, 1/3, 1, 1, 1/3, . . .); if the item starts on a word-medial syllable, these TPs are (1, 1/3, 1, 1, 1/3, 1, 1, 1/3, 1, . . .).

In contrast, the expected TPs for a random concatenation of syllables are the TPs in a random bigram. For an A or a B syllable, the random TP is 1 × 1 / 12, as there is only 1 (out of 12) non-zero TP continuations. For a C syllable, the random TP is 3 × 1/3 / 12, as there are 3 possible concatenations. On average, the random TP is thus $(1/12 + 1/12 + 1/12)/3 = 1/12 \approx .083$.

SOM2.2.2.3 Exclusion of responses and participants

There was a considerable number of recall responses containing unattested syllables. The complete list of unattested items is in `segmentation_recall_unattested.xlsx` in the supplementary data. Unattested items are items that are not words, part-words (or concatenations thereof), high- or low-TP chunks, or a single syllable. However, it is unclear if these unattested syllables reflect misperceptions not caught by our substitution rules, typos, memory failures or creative responses. This makes it difficult to analyze these responses. For example, the TPs from and to an unattested syllable are zero. However, if the unattested syllable reflects a misperception or a typo, the true TP would be positive, and our estimates would underestimate the participant's statistical learning ability.

Here, we decided to include items with unattested syllables to avoid excluding an excessive number of participants. However, the results after removing such items are essentially identical, with the exception of the TPs in the participants' responses. Given that TPs to and from unattested syllables are zero by definition, TPs after removal of responses containing unattested syllables are much higher.

We also decided to remove single syllable responses, as it is not clear if participants volunteered such responses because they thought that individual syllables reflected the underlying units in the speech streams or because they misunderstood what they were ask to do.

SOM2.2.3 Demographics

To reduce performance differences between the pre-segmented and the continuous familiarization conditions, participants were excluded from analysis if their accuracy in the recognition test was below 50% (N = 19). Another 11 participants were excluded because parsing their productions took an excessive amount of computing time, though their productions did not seem to resemble the familiarization items in the first place. The final demographic information is given in Table S6.

SOM3 Additional results

SOM3.1 Experiment 1

Table S7: Performance differences across familiarization conditions in Experiment 1. The differences were assessed using a generalized linear model for the trial-by-trial data, using participants, correct items and foils as random factors. Random factors were removed from the model when they did not contribute to the model likelihood.

Effect	Estimate	Std. Error	CI	*t*	*p*
Pre-segmented familiarization					
Language = L2	0.114	0.673	-1.2, 1.43	0.170	0.865
Continuous familiarization (1)					
Language = L2	-0.184	0.480	-1.12, 0.757	-0.383	0.702
Continuous familiarization (2)					
Language = L2	0.317	0.786	-1.22, 1.86	0.403	0.687
Pre-segmented vs. continuous familiarization (1)					
Language = L2	-0.019	0.557	-1.11, 1.07	-0.033	0.973
Pre-segmentation: Yes	-0.328	0.188	-0.696, 0.0391	-1.752	0.080
Pre-segmented vs. continuous familiarization (2)					
Language = L2	0.215	0.657	-1.07, 1.5	0.327	0.743
Pre-segmentation: Yes	-0.608	0.244	-1.09, -0.13	-2.493	0.013

SOM3.2 Experiment 2

SOM3.2.1 Additional tables and figures

Table S8: Various supplementary analyses pertaining to the productions as well as test against their chances levels.

	Continuous	Segmented	*p*(Continuous vs. Segmented).
Number of words			
lab-based	*M*= 0.308, *SE*= 0.139, *p*= 0.0719	*M*= 1.85, *SE*= 0.308, *p*= 0.00224	0.005
online	*M*= 0.224, *SE*= 0.0791, *p*= 0.00482	*M*= 1.32, *SE*= 0.143, *p*= 7.32e-11	< 0.001
Proportion of words among productions			
lab-based	*M*= 0.308, *SE*= 0.139, *p*= 0.0719	*M*= 1.85, *SE*= 0.308, *p*= 0.00224	0.005
online	*M*= 0.224, *SE*= 0.0791, *p*= 0.00482	*M*= 1.32, *SE*= 0.143, *p*= 7.32e-11	< 0.001
Number of part-words			
lab-based	*M*= 0.692, *SE*= 0.273, *p*= 0.031	*M*= 0, *SE*= 0, *p*= NaN	0.031
online	*M*= 0.25, *SE*= 0.0657, *p*= 0.000717	*M*= 0, *SE*= 0, *p*= NaN	< 0.001
Proportion of part-words among productions			
lab-based	*M*= 0.692, *SE*= 0.273, *p*= 0.031	*M*= 0, *SE*= 0, *p*= NaN	0.031
online	*M*= 0.25, *SE*= 0.0657, *p*= 0.000717	*M*= 0, *SE*= 0, *p*= NaN	< 0.001
Actual vs. expected forward TPs			
lab-based	*M*= -0.462, *SE*= 0.07, *p*= 0.000244	*M*= -0.315, *SE*= 0.0803, *p*= 0.00915	0.147
online	*M*= -0.42, *SE*= 0.0329, *p*= 1.3e-12	*M*= -0.352, *SE*= 0.0365, *p*= 7.56e-11	0.120
Number of High-TP chunks			
lab-based	*M*= 0.769, *SE*= 0.459, *p*= 0.181	*M*= 2.31, *SE*= 0.361, *p*= 0.00224	0.022
online	*M*= 1.13, *SE*= 0.13, *p*= 5.35e-10	*M*= 1.62, *SE*= 0.147, *p*= 6.19e-12	0.014
Proportion of High-TP chunks among productions			
lab-based	*M*= 0.104, *SE*= 0.0601, *p*= 0.181	*M*= 0.615, *SE*= 0.0999, *p*= 0.00241	0.003
online	*M*= 0.279, *SE*= 0.0331, *p*= 1.08e-09	*M*= 0.516, *SE*= 0.0435, *p*= 8.27e-12	< 0.001
Number of Low-TP chunks			
lab-based	*M*= 0.0769, *SE*= 0.0801, *p*= > .999	*M*= 0, *SE*= 0, *p*= NaN	> .999
online	*M*= 0.355, *SE*= 0.0747, *p*= 2.41e-05	*M*= 0.0395, *SE*= 0.0226, *p*= 0.149	< 0.001
Number of Low-TP chunks among productions			
lab-based	*M*= 0.011, *SE*= 0.0114, *p*= > .999	*M*= 0, *SE*= 0, *p*= NaN	> .999
online	*M*= 0.0855, *SE*= 0.0198, *p*= 6.04e-05	*M*= 0.00846, *SE*= 0.00523, *p*= 0.181	< 0.001

[*] The expected TPs for items of at least 2 syllables starting on an initial syllable are 1, 1/3, 1, 1, 1/3, 1, 1, 1/3, ... The difference between the actual and the expected TP needs to be compared to zero, as the expected TP differs across items.

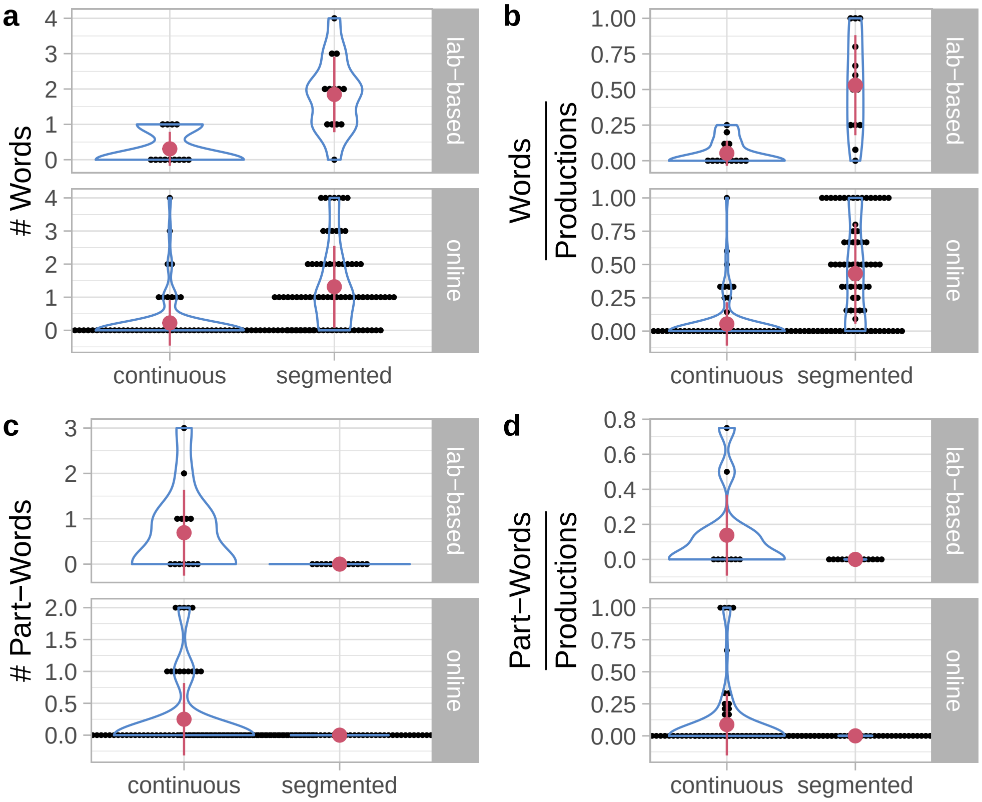

Figure S1: Number and proportion (among vocalizations) of words and part-words.

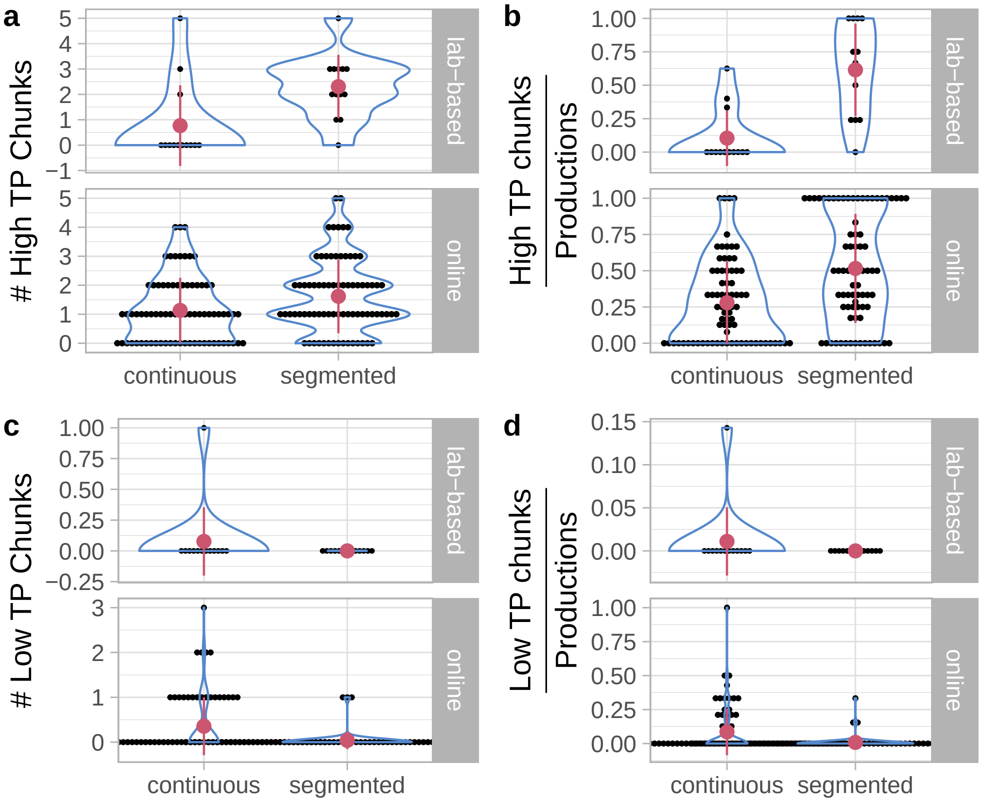

Figure S2: Plot of High and Low TP chunks.

SOM3.3 Fit of the number of participants producing words or part-words to a binomial distribution

We fit the data to two models, one where the learner successfully detected word-boundaries, and one where the learner successfully track TPs but initiates productions at a random position. We then calculate the likelihood of the data given these models.

According to the first model, the probability of producing words rather then part-words is $p^1_{\mathrm{W}} = 1$, and the probability of using part-words is $p^1_{\mathrm{PW}} = 1 - p^1_{\mathrm{W}} = 0$. According to the second model, the learner has one chance in three to initiate a production on a word-initial syllable. As a result, the probability of producing words is $p^2_{\mathrm{W}} = \frac{1}{3}$, and the probability of using part-words is $p^2_{\mathrm{PW}} = 1 - p^2_{\mathrm{W}} = \frac{2}{3}$.

Assuming that participants produce either words or part-words, the probability of N_{W} producing words and N_{PW} producing part-words is given by a binomial distribution. We can then use Bayes' theorem to calculate the model likelihood $P(\text{model}|\text{data}) = P(\text{data}|\text{model})\frac{P(\text{model})}{P(\text{data})}$. If both models are equally likely a priori, the likelihood ratio of the models given the data is the likelihood ratio of the data given the models:

$$
\begin{aligned}
\Lambda_{1,2} &= \frac{P(\text{model}_1|\text{data})}{P(\text{model}_2|\text{data})} = \frac{P(\text{data}|\text{model}_1)}{P(\text{data}|\text{model}_2)} \\
&= \frac{\begin{pmatrix} N_{\mathrm{W}} + N_{\mathrm{PW}} \\ N_{\mathrm{W}} \end{pmatrix}}{\begin{pmatrix} N_{\mathrm{W}} + N_{\mathrm{PW}} \\ N_{\mathrm{W}} \end{pmatrix}} \frac{1^{N_{\mathrm{W}}} 0^{N_{\mathrm{PW}}}}{\frac{1}{3}^{N_{\mathrm{W}}} \frac{2}{3}^{N_{\mathrm{PW}}}} \\
&= \begin{cases} 3^{N_{\mathrm{PW}}} & N_{\mathrm{PW}} = 0 \\ 0 & N_{\mathrm{PW}} > 0 \end{cases}
\end{aligned}
$$

For $N_{\mathrm{PW}} = 0$, the likelihood ratio in favor of the first model is $3^{N_{\mathrm{PW}}}$; $N_{\mathrm{PW}} > 0$ the likelihood ratio in favor of the second model is infinite.

SOM4 Pilot Experiment 1: Using the *en1* voice

We ran an experiment identical to the pre-segmented condition of Experiment 1, except that materials were synthesized using the *en1* (British English male) voice.

SOM4.1 Familiarization with a pre-segmented stream

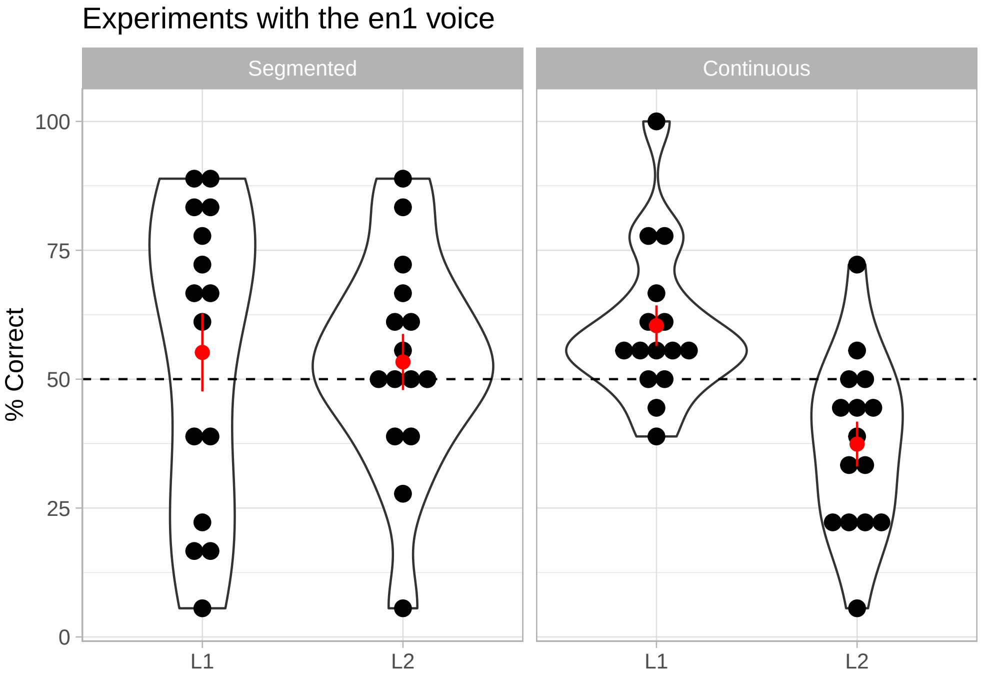

Figure S3: Results for a pre-segmented presentation of the stream (540 ms silences, left) and continuous presentation of the stream (right). Each word was repeated 45 times. The voice was *en1*.

Table S9: Descriptives for Experiment 1 (using the *us3* voice) and Pilot Experiment 1 (using the *en1* voice).

Condition	*N*	*M*	*SE*	*p*
us2 voice				
Pre-segmented	30	0.517	0.028	0.307
Continuous (1)	32	0.585	0.029	0.018
Continuous (2)	30	0.628	0.040	0.007
en1 voice				
Pre-segmented (en1)	30	0.543	0.047	0.268
Continuous (en1)	30	0.489	0.036	0.739

As shown in Figure S3, when the speech stream was pre-segmented, the average performance did not differ significantly from the chance level of 50%, (M = 54.26, SD = 25.09), Cohen's d = 0.17, $CI_{.95}$ = 44.89, 63.63, ns, . Likelihood ratio analysis favored the null hypothesis by a factor of 3.555 after correction with the Bayesian Information Criterion. Further, as shown in Table S10, performance did not depend on the language condition.

Table S10: Performance differences across language conditions in Pilot Experiment 1. The differences were assessed using a generalized linear model for the trial-by-trial data, using participants, correct items and foils as random factors. Random factors were removed from the model when they did not contribute to the model likelihood

Effect	Estimate	Std. Error	CI	*t*	*p*
Pre-segmented					
Language = L2	-0.097	0.441	-0.96, 0.767	-0.22	0.826
Continuous					
Language = L2	-1.024	0.410	-1.83, -0.22	-2.50	0.013

SOM4.2 Familiarization with a continuous stream

As shown in Figure S3, when the speech stream was continuous, the average performance did not differ significantly from the chance level of 50%, (M = 48.89, SD = 19.65), $t(29)$ = -0.31, p = 0.759, Cohen's d = 0.057, $CI_{.95}$ = 41.55, 56.23, ns, V = 166, p = 0.818. Likelihood analyses revealed that the null hypothesis was 5.221 than the alternative hypothesis after a correction with the Bayesian Information Criterion. However, as shown in Table S10, performance was much better for Language 1 than for Language 2, presumably due to some click-like sounds the synthesizer produced for some stops and fricatives (notably /f/ and /g/). These sound likely affected grouping, and prevented participants from using statistical learning.

SOM5 Pilot Experiment 2: Testing the use of chunk frequency

In Pilot Experiment 2, we asked if participants could break up tri-syllabic items by using the chunk frequency of sub-chunks. The artificial languages were designed such that, in a trisyllabic item such as *ABC*, chunk frequency (and backwards TPs) favor in the initial *AB* chunk for half of the participants, and the final *BC* chunk for the other participants.

Across participants, we also varied the exposure to the languages, with 3, 15 or 30 repetitions per word, respectively.

SOM5.1 Methods

SOM5.1.1 Participants

Table S11: Demographics of Pilot Experiment 2.

# Repetitions/word	*N*	Age (*M*)	Age (Range)
3	37	21.1	18-35
15	41	21.0	18-27
30	40	20.8	18-26

Demographic information of Pilot Experiment 2 is given in Table S11. Participants were native speakers of Spanish and Catalan and were recruited from the Universitat Pompeu Fabra community.

SOM5.1.2 Stimuli

Stimuli transcriptions are given in Table S12. They were synthesized using the *es2* (Spanish male) voice of the mbrola [56] speech synthesized, using a segment duration of 225 ms and an fundamental frequency of 120 Hz.

SOM5.1.3 Apparatus

Participants were test individually in a quiet room. Stimuli were presented over headphones. Responses were collected from pre-marked keys on the keyboard. The experiment with 3 repetitions per word (see below) were run using PsyScope X; the other experiments were run using Experyment (https://www.expyriment.org/).

SOM5.1.4 Familiarization

The design of Pilot Experiment 2 is shown in Table S12. The languages comprise trisyllabic items. All forward TPs were 0.5. However, in Language 1 the chunk composed of the first two syllables (e.g., *AB* in *ABC*) were twice as frequent as the chunk composed of the last two syllables (e.g., *BC* in *ABC*); the backward TPs were twice as high as well. Language 2 favored the word-final chunk. Participants were informed that they would listen to a sequence of Martian words, and then listened to a sequence of the eight words in S2 with an ISI of 1000 ms and 3, 15 or 30 repetitions per word. Due to programming error, the familiarization items for 15 and 30 repetitions per word were sampled with replacement.

Table S12: Design of the Pilot Experiment 2. (Left) Language structure. (Middle) Structure of test items. Correct items for Language 1 are foils for Language 2 and vice versa. (Right) Actual items in SAMPA format; dashes indicate syllable boundaries

Word structure for		Test item structure for		Actual words for	
Language 1	Language 2	Language 1	Language 2	Language 1	Language 2
ABC	ABC	AB	BC	ka-lu-mo	ka-lu-mo
DEF	DEF	DE	EF	ne-fi-To	ne-fi-To
ABF	DBC			ka-lu-To	ne-lu-mo
DEC	AEF			ne-fi-mo	ka-fi-To
AGJ	JBG			ka-do-ri	ri-lu-do
AGK	KBG			ka-do-tSo	tSo-lu-do
DHJ	JEH			ne-pu-ri	ri-fi-pu
DHK	KEH			ne-pu-tSo	tSo-fi-pu

Table S13: Performance in Pilot Experiment 2 for different amounts of exposure. The differences were assessed using a generalized linear model for the trial-by-trial data, using participants as a random factor.

Effect	Estimate	Std. Error	CI	t	p
Language = L2	0.337	0.493	-0.629, 1.3	0.684	0.494
# Word repetitions	0.017	0.018	-0.018, 0.0513	0.942	0.346
Language = L2 × # Word repetitions	-0.042	0.025	-0.0916, 0.00698	-1.682	0.093

SOM5.1.5 Test

Following this familiarization, participants were informed that they would hear new items, and had to decide which of them was in Martian. Following this, they heard pairs of two syllabic items with an ISI of 1000 ms. One was a word-initial chunk and one a word-final chunk.

The test items shown in Table S2 were combined into four test pairs, which were presented twice with different item orders. A new trial started 100 ms after a participant response.

SOM5.2 Results

As shown Table S13, a generalized linear model revealed that performance depended neither on the amount of familiarization nor on the familiarization language. As shown in Figure S4, a Wilcoxon test did not detect any deviation from the chance level of 50%, neither for all amounts of familiarization combined, *M*= 53.5, *SE*= 2.71, *p*= 0.182, nor for the individual familiarization conditions (3 repetitions per word: *M*= 54.1, *SE*= 4.81, *p*= 0.416; 15 repetitions per word: *M*= 54.6, *SE*= 4.52, *p*= 0.325; 30 repetitions per word: *M*= 51.9, *SE*= 4.98, *p*= 0.63). Following Glover and Dixon [58], the null hypothesis was 4.696 times more likely than the alternative hypothesis after corrections with the Bayesian Information Criterion, and 1.217 more likely after correction with the Akaike Information Criterion.

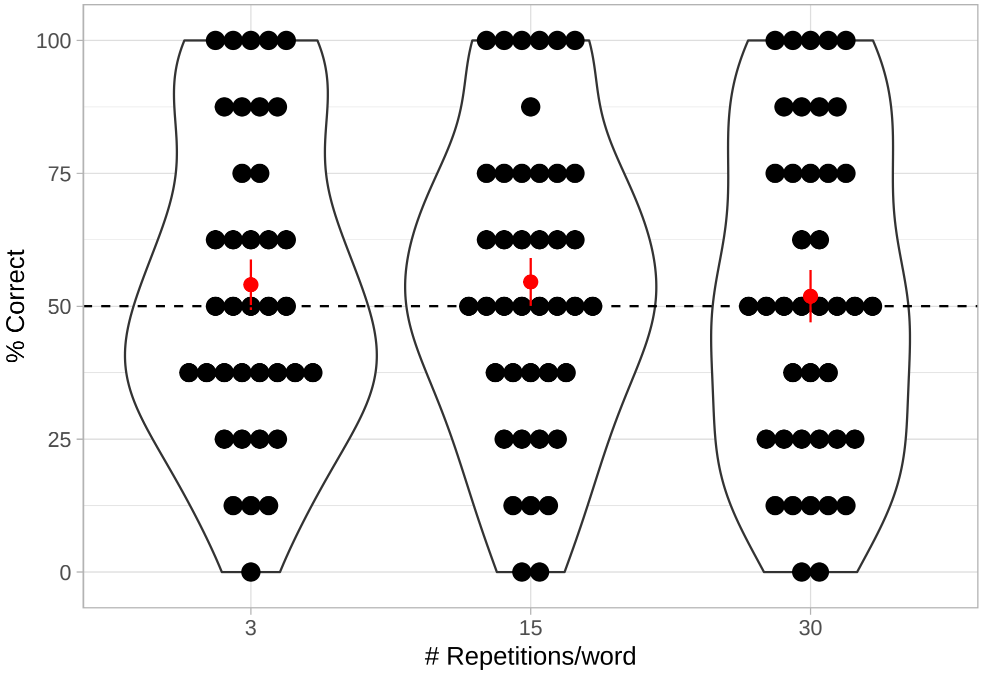

Figure S4: Results of Pilot Experiment 2. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right) 30 repetitions per word.