The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory

Abstract

Learning statistical regularities from the environment is ubiquitous across domains and species. It might support the earliest stages of language acquisition, including identifying and memorizing words from fluent speech (word-segmentation). We ask if the Statistical Learning mechanisms involved in word-segmentation are tuned to specific learning situations, and how they interact with the memory mechanisms needed to remember words. We show that Statistical Learning predominantly operates in continuous speech sequences like those used in earlier experiments, but not in discrete chunk sequences likely encountered during language acquisition. Conversely, when exposed to continuous sequences in a reproduction task, participants hardly remember any items at all and initiate their productions with random syllables (rather than word-onsets) despite being sensitive to probable syllable transitions. Only discrete familiarization sequences produce memories of actual items. Statistical Learning might thus be specialized to accumulate distributional information, but dissociable from the (declarative) memory mechanisms needed to acquire words.

Keywords: Statistical Learning; Declarative Memory; Predictive Processing; Language Acquisition

The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory

1 Introduction

The ability to learn statistical regularities from the environment is remarkably widespread across species and domains (Saffran, Aslin, & Newport, 1996; Hauser, Newport, & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Toro, Trobalon, & Sebastián-Gallés, 2005; Turk-Browne & Scholl, 2009; Chen & Ten Cate, 2015), and might support a wide range of computations, especially during language acquisition (Aslin & Newport, 2012). However, such Statistical Learning abilities are also remarkably modular (Endress, 2019). Humans have independent statistical learning abilities in superficially similar domains, including associations of objects with landmarks vs. boundaries (Doeller & Burgess, 2008), associations among social vs. non-social objects (Tompson, Kahn, Falk, Vettel, & Bassett, 2019) and associations among consonants vs. vowels (Bonatti, Peña, Nespor, & Mehler, 2005; Toro, Bonatti, Nespor, & Mehler, 2008).

Such preferential associations abound, can evolve in just 40 generations in fruit flies (Dunlap & Stephens, 2014), and likely reflect ecological constraints. For example, rats readily associate tastes with sickness and external stimuli with pain, but cannot associate taste with pain or external stimuli with sickness (Garcia, Hankins, & Rusiniak, 1974). However, taste-sickness associations (but not other associations) are blocked in a suckling context if rat pups had no exposure to solid food (Martin & Alberts, 1979; Alberts & Gubernick, 1984), presumably because avoidance of the *only* food source is costly. Over evolutionary times, Statistical Learning mechanisms can thus become tuned to specific learning situations, though they still might be a “spandrel” (Gould,

Lewontin, Maynard Smith, & Holliday, 1979) that evolved as a side effect of local neural processing and might undergo positive, negative or no selection in different brain pathways.

Here, we ask if the Statistical Learning mechanisms thought to be involved in learning words from fluent speech (i.e., in word-segmentation) show similar specializations, and how they relate to their presumed computational function, namely to store words in (declarative) memory. In brief, we suggest that these mechanisms are critical for predicting speech material and operate 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 often perceived as such in unknown languages), 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 syllables, 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 (1951) showed that human adults are sensitive to such distributional information. Subsequent work demonstrated that infants and non-human animals share this ability (Saffran et al., 1996; Hauser et al., 2001; Kirkham et al., 2002; Toro et al., 2005; Turk-Browne & Scholl, 2009; Chen & Ten Cate, 2015), and that it might reflect simple associative mechanisms such as Hebbian learning (Endress & Johnson, 2021).

Statistical Learning therefore supports predictive processing (Sherman & Turk-Browne, 2020; Turk-Browne, Scholl, Johnson, & Chun, 2010), a critical

ability for language (Levy, 2008; Trueswell, Sekerina, Hill, & Logrip, 1999) and other cognitive processes (Clark, 2013; Friston, 2010; Keller & Mrsic-Flogel, 2018). However, while words are clearly stored in declarative Long-Term Memory (after all, the point of knowing words is to "declare" them), statistical knowledge does not imply the formation of such memory representations. In fact, after exposure to sequences where some transitions are more likely than others, observers report greater familiarity with high-TP items than with low-TP items even when they have never encountered either of them and thus could not have memorized them (because the items are played backwards with respective to the familiarization sequence; Turk-Browne & Scholl, 2009; Jones & Pashler, 2007). Sometimes, observers even report greater familiarity with high-TP items they have *never* encountered than with low-TP items they have heard or seen (Endress & Langus, 2017).

Dissociations between Statistical Learning and declarative memory have long been documented behaviorally (Graf & Mandler, 1984), developmentally (Finn et al., 2016) and neuropsychologically (Knowlton, Mangels, & Squire, 1996; Poldrack et al., 2001; Squire, 1992), to the extent that statistical predictions can *impair* declarative memory encoding in healthy adults (Sherman & Turk-Browne, 2020). If Statistical Learning operates similarly in a word-segmentation context as in other learning situations, one would expect it to be dissociable from declarative Long-Term Memory, a view that is reinforced by the suggestion that the format of the representations created by Statistical Learning differs from that used for linguistic stimuli (Endress & Langus, 2017; Fischer-Baum, Charny, & McCloskey, 2011; Miozzo, Petrova, Fischer-Baum, & Peressotti, 2016).

Here, we explore the computational function of Statistical Learning in

word-segmentation, 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 (Cutler, Oahan, & van Donselaar, 1997; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996). This prosodic organization is perceived in unfamiliar languages (Brentari, González, Seidl, & Wilbur, 2011; Endress & Hauser, 2010; Pilon, 1981) and even by newborns (Christophe, Mehler, & Sebastian-Galles, 2001). It might affect the usefulness of Statistical Learning, because such speech cues tend to override statistical cues (Johnson & Jusczyk, 2001; Johnson & Seidl, 2009), and because Statistical Learning primarily operates *within* rather than across major prosodic boundaries (Shukla, Nespor, & Mehler, 2007; Shukla, White, & Aslin, 2011). As a 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 & Vogel, 1986).

In Experiment 1, we thus ask whether Statistical Learning operates in such smaller chunks, or only in longer stretches of continuous speech. Participants listened to a speech sequence of tri-syllabic non-sense words synthesized using mbrola (Dutoit, Pagel, Pierret, Bataille, & van der Vreken, 1996). The words were either *pre-segmented* (i.e., with a silence after each word) or continuously concatenated. These continuous vs. pre-segmented presentation modes trigger different sets of memory processes (Peña, Bonatti, Nespor, & Mehler, 2002; Endress & Bonatti, 2016), but it is unknown if either of these processes involves declarative memory.

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 (TPs of 1.0 vs. .33). A Statistical Learner should thus split triplets 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. To make the learning task as simple as possible, the statistical pattern of the words was thus consistent for each participant. Following this familiarization, participants heard pairs of *AB* and *BC* items, and had to indicate which item was more like the familiarization items.

In Experiment 2, we sought to elucidate the function of Statistical Learning, asking (adult) participants to recall what they remember after being exposed to the speech stream from Saffran et al.'s (1996) classic experiment, again with a sequence of pre-segmented "words" or with a continuous speech stream.

2 Methods summary

Unless otherwise stated, stimuli were synthesized using mbrola (Dutoit et al., 1996) 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`.

2.1 Participants

In Experiment 1, 30, 30 and 31 participants were retained for analysis for the pre-segmented condition, the continuous condition and its replication. The sample size was chosen to be feasible within a third year psychology student project. In Experiment 2, 26 participants were retained for the lab-based

version, and 157 for the online version. As we had no prior expectation about the effect size, we targeted a sample at least 30 participants for each of the conditions (i.e., continuous vs. pre-segmented × Language 1 vs. Language 2). Participants reported to be native speakers of English.

2.2 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 above). 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.

2.3 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.'s (1996) 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.

2.4 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 SOM2).

3 Results

In Experiment 1, participants listened to a speech sequence of tri-syllabic non-sense 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. We thus expected learners to split a triplet like ABC into an $AB+C$ pattern. For the remaining participants, both the TPs and the chunk frequency favored an $A+BC$ pattern. Following this

familiarization, participants 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 1, performance did not depend on the language condition. As shown in SOM4, 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 (SOM5).

Table 1
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

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 artificial speech stream as in the pre-segmented condition, except that the stream was continuous and had no silences between words.

As shown in Figure 1, the average performance differed significantly from the chance level of 50%, ($M = 58.51$, $SD = 16.21$), Cohen's $d = 0.52$, $CI_{.95} = 52.66, 64.35$, $V = 306.5$, $p = 0.02$. As shown in Table 1, 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 $d = 0.6$, $CI_{.95} = 54.81, 70.75$, $V = 320$, $p = 0.008$. As shown in Table 1, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition ($p = .013$).

(This result could not be replicated using a different voice (*en1*, male British English; see SOM4); participants seemed to prefer specific items, irrespective of the language they had been familiarized with, 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 Statistical Learning mechanisms predominantly operate in continuous sequences, but less so in pre-segmented sequences (see also Shukla et al., 2007, 2011). Such a result is

compatible with the view that Statistical 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 Statistical Learning mechanisms to memorize words, given that they do not seem to be able to do so in pre-segmented streams.

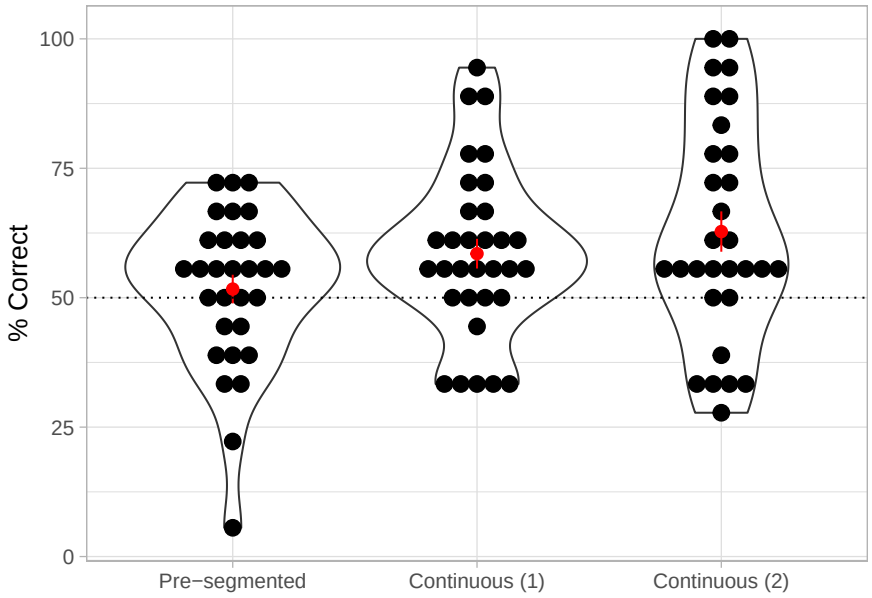

Figure 1. Results of Experiment 1. Each dot represents a participant. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a pre-segmented familiarization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

In Experiment 2, we explored the computational function of Statistical Learning, and asked if participants would remember the items that occurred in a speech stream. Adult participants listened to the artificial languages from Saffran et al.'s (1996) Experiment 2 with 8-months-old infants, except that we doubled the exposure to increase the opportunity for learning the statistical structure of the speech stream. The languages comprised four tri-syllabic 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 the experiment.

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). To focus the analyses on participants who learned the statistical structure of the speech stream, we also removed participants who did not perform at least 50% during the final recognition test; as a result, performance during the recognition phase cannot be evaluated against a chance level.

As shown in Table 2 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 2 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' actual performance, as we included responses with unattested syllables that might reflect misperceptions (and thus lower TPs); 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.)

We next examined the production of two-syllable chunks. Such chunks can be either high-TP chunks (if they are part of a word) or low-TP chunks (if they 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 both the pre-segmented condition and the continuous condition (at least in the online version), with a significantly higher proportion in the pre-segmented versions. These results thus confirm that participants are sensitive to TPs or high frequency chunks (which are confounded in the current design).

We now turn to the question of whether a sensitivity to TPs implies memory for words. We address this issue in two ways, by using the traditional contrast between words and part-words and by turning to the question at the heart of word segmentation — do participants know where words start and where they end?

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 recalled by the participants. 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 a word rather than a part-word. As a result, if participants

initiate their productions with a random syllable, a third of their productions should be words.

As shown in Table 2 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 chance level of 33% in the continuous condition. Likelihood ratio analysis suggests that, in the continuous condition, participants were 3.5 times 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 do not remember any word forms.

However, inspection of Figure 3a shows that the distribution in the continuous condition 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 is binomially distributed, we calculated the likelihood ratio of a model where learners identify word boundaries (and should 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 SOM3.1, 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. Given that the overwhelming majority of participants produce words only after a pre-segmented familiarizations, 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 one word, with an average of 1.26.)

We thus turn to question of whether participants know where words start and end, asking if participants produce correct initial and final syllables. If participants use Statistical Learning to remember words, they should know where words start and where they end. In contrast, if they just track TPs, they should initiate the responses with random syllables. As there are four words with one correct initial and final syllable each, and 12 syllables in total, $4/12 = 1/3$ of the productions should have "correct" initial syllables, and 1/3 should have correct final syllables. Given that participants tend to produce high-TP two-syllable chunks (i.e., AB and BC rather than CD chunks), the actual baseline level is somewhat higher. (For example, participants in the continuous condition produce about 75% high-TP chunks; if they initiate their productions with high-TP chunks, one would expect them to produce about $75\%/2 = 3/8$ rather than 1/3 items with correct initial syllables.) However, to evaluate the group performance, we keep the baseline of 1/3.

As shown in Table 2 and Figure 3c and d, participants produced items with correct initial or final syllables at greater than chance level only in the pre-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 (3.61 for the lab-based experiment) and 4.06 for final syllables (2.14 for the lab-based experiment). While it is possible that performance in the continuous condition might exceed the chance-level of 1/3 with more than the 78 participants currently included, the actual chance-level is somewhat higher

(about 38.4%). Critically, only 42% of the productions have a correct initial syllable, which is unexpected if participants knew where words start and where they end. Together with the finding that the overwhelming majority of participants produce no word at all, these results thus suggest that TPs do not allow learners to reliably detect onsets and offsets of words.

Table 2
Various analyses pertaining to the productions as well as test against their chances levels. The p value in the rightmost column reflects a Wilcoxon test comparing the continuous and the pre-segmented conditions.

	Continuous	Pre-segmented	p(continuous vs. pre-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
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 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
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

4 Discussion

Taken together, Experiments 1 and 2 suggest that Statistical Learning and (declarative) memory might fulfill different computational functions in the process of word-segmentation. In Experiment 1, participants tracked statistical dependencies predominantly when they were embedded in a continuous speech stream, but not across pre-segmented chunk sequences. This is consistent with

earlier findings that Statistical Learning predominantly occurs within major prosodic groups, and, within these groups, predominantly at the edges of those groups (Shukla et al., 2007; Seidl & Johnson, 2008). We show that, with shorter and better separated groups, Statistical Learning can be abolished altogether. In line with results from conditioning experiments (Alberts & Gubernick, 1984; Garcia et al., 1974; Gubernick & Alberts, 1984; Martin & Alberts, 1979), Statistical Learning, and maybe associative learning in general, can thus be enhanced or suppressed depending on the learning situation. The enhanced Statistical Learning in continuous sequences is consistent with the view that Statistical Learning is important for predictive processing (Turk-Browne et al., 2010; Sherman & Turk-Browne, 2020), given that prediction is arguably more useful in lengthy chunks. It is also consistent with the view that Statistical 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 (Cutler et al., 1997; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996; Brentari et al., 2011; Endress & Hauser, 2010; Pilon, 1981; Christophe et al., 2001).

Experiment 2 provided the first direct test of the contents of the participants' (episodic or semantic) declarative memory after exposure to an Statistical Learning task. The results suggest that, even when participants successfully track statistical information, they remember familiarization items only when familiarized with a pre-segmented sequence. In contrast, when familiarized with a continuous sequence, their productions start with random syllables rather than actual word onsets. Given that the memory representations of linguistic items are based on their initial and final syllables (Endress & Langus, 2017; Fischer-Baum et al., 2011; Miozzo et al., 2016), this result thus

suggests that Statistical Learning did not lead to the creation of declarative memory representations.

Contrary to this conclusion, some authors suggest that Statistical Learning might lead to declarative memories for chunks (Graf-Estes, Evans, Alibali, & Saffran, 2007; Isbilen, McCauley, Kidd, & Christiansen, 2020). Such experiments generally proceed in two phases. During a Statistical Learning phase, participants are exposed to some statistically structured sequence. Then, they are exposed to items presented in isolation, and show some processing advantage for isolated high-probability items compared to isolated low-probability items. However, we proposed that such experiments have a two-step explanation that does not involve declarative memory (Endress & Langus, 2017). First, during the Statistical Learning phase, participants acquire statistical knowledge without remembering any specific items. When experimenters subsequently provide participants with *isolated* chunks, the accumulated statistical knowledge facilitates processing of the experimenter-provided chunks (e.g., due to predictive processing), without these chunks having been acquired before being supplied by the experimenter. In contrast to such indirect designs, we provide a direct measure of declarative knowledge of sequence items, and show that participants do not form declarative memories of sequence items unless the sequence is pre-segmented.

The combined results of Experiments 1 and 2 echo dissociations between associative learning and declarative memory (Graf & Mandler, 1984; Finn et al., 2016; Knowlton et al., 1996; Poldrack et al., 2001; Squire, 1992), suggesting that the (cortical) declarative memory system might be independent of a (neostriatal) system for associative learning (Knowlton et al., 1996; Poldrack et al., 2001; Squire, 1992), though other authors propose that both types of memory involve

the hippocampus (Sherman & Turk-Browne, 2020; Ellis et al., 2021). In line with earlier proposals (Turk-Browne et al., 2010; Sherman & Turk-Browne, 2020), 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 useful and 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 Statistical 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 (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Brent & Siskind, 2001; Mersad & Nazzi, 2012), attentional allocation to beginnings and ends of utterances (Monaghan & Christiansen, 2010; Seidl & Johnson, 2008; Shukla et al., 2007), legal sound sequences (McQueen, 1998) and universal aspects of prosody (Brentari et al., 2011; Christophe et al., 2001; Endress & Hauser, 2010; Pilon, 1981). 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 (Endress & Langus, 2017; Fischer-Baum et al., 2011; Miozzo et al., 2016).

To the extent that Statistical Learning reflects implicit memory systems (Christiansen, 2018), this suggestion mirrors earlier proposals that implicit and declarative memory systems might have different roles during language acquisition, with declarative memory systems supporting the acquisition of words and implicit memory system supporting the grammar-like regularities

(Ullman, 2001; Pinker & Ullman, 2002). While we are agnostic about the extent to which Statistical Learning can support grammar acquisition, such results, together with the current ones, suggest that Statistical Learning and declarative memory might have separable functions, the former for predictive processing and the latter for remembering objects and episodes.

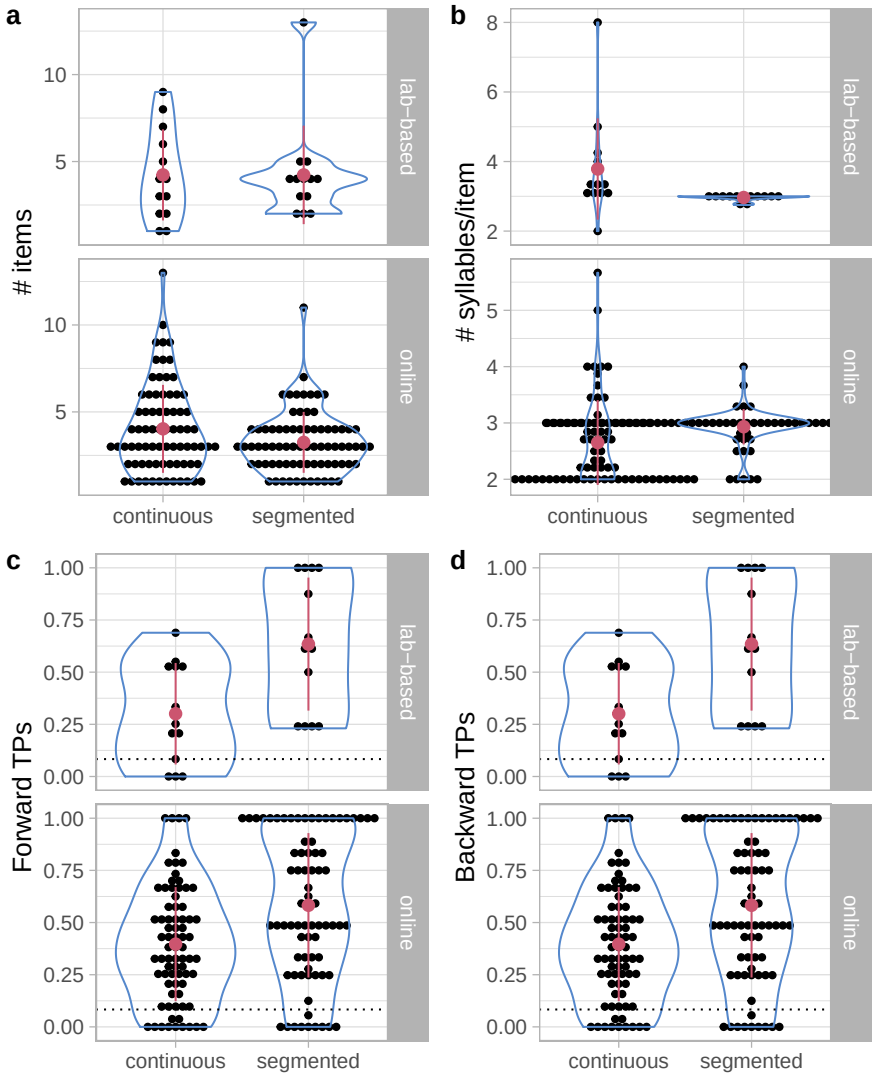

Figure 2. Number of items produced, number of syllables per item and forward and backward TPs. The dotted line represents the chance level for a randomly ordered syllable sequence.

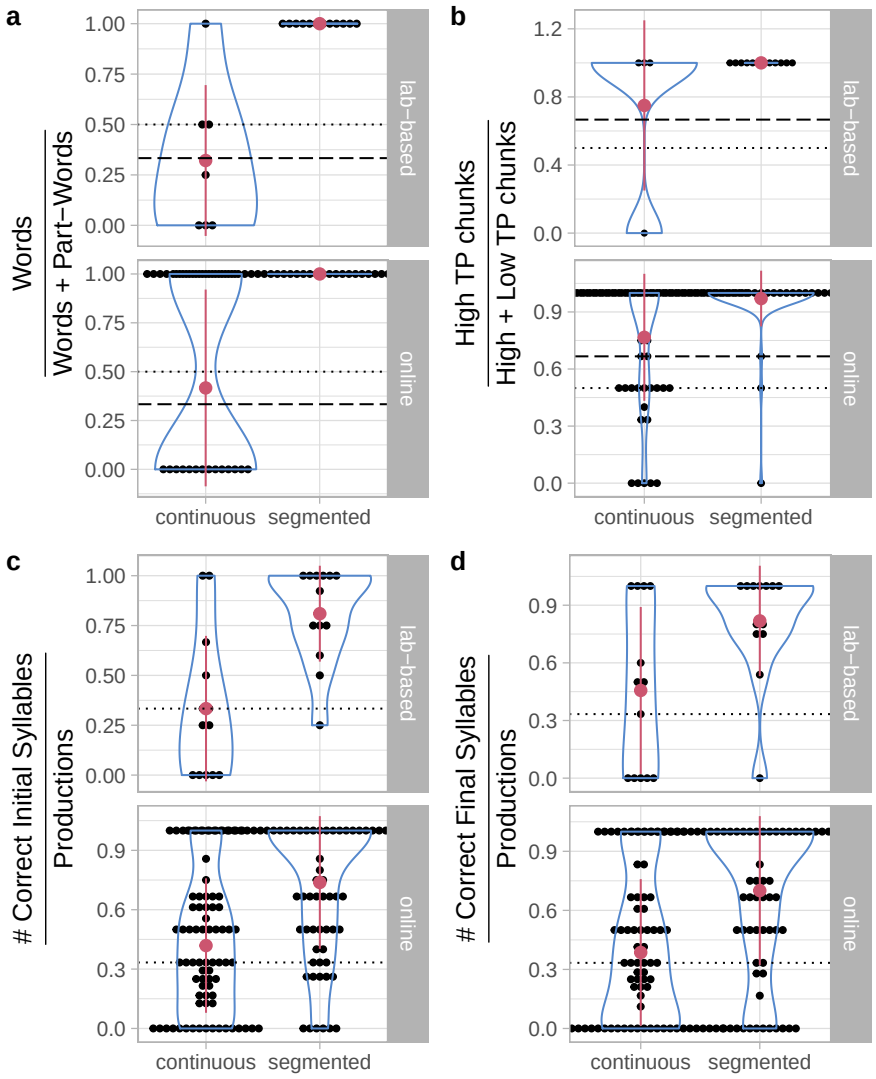

Figure 3. Analyses of the participants' productions. (a) Proportion of words among words and part-words. The dotted line represents the chance level of 50% in a two-alternative forced-choice task, while the dashed line represents the chance level of 33% that an attested 3 syllable-chunk is a word rather 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%.

References

Alberts, J. R., & Gubernick, D. J. (1984). Early learning as ontogenetic adaptation for ingestion by rats. *Learn Motiv*, *15*(4), 334 - 359. doi: 10.1016/0023-9690(84)90002-X

Aslin, R. N., & Newport, E. L. (2012). Statistical learning. *Current Directions in Psychological Science*, *21*(3), 170-176. doi: 10.1177/0963721412436806

Baayen, R. H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390 - 412. doi: 10.1016/j.jml.2007.12.005

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychol Sci*, *16*(8), 451-459.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychol Sci*, *16*(4), 298-304. doi: 10.1111/j.0956-7976.2005.01531.x

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-44.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Lang Speech*, *54*(1), 49–72.

Chen, J., & Ten Cate, C. (2015, Aug). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behav Processes*, *117*, 29–34. doi: 10.1016/j.beproc.2014.09.004

Christiansen, M. H. (2018, apr). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468–481. doi: 10.1111/tops.12332

Christophe, A., Mehler, J., & Sebastian-Galles, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, *2*(3), 385-394.

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

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

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

Dunlap, A. S., & Stephens, D. W. (2014). Experimental evolution of prepared learning. *Proceedings of the National Academy of Sciences*, *111*(32), 11750-11755. Retrieved from `http://www.pnas.org/content/111/32/11750.abstract` doi: 10.1073/pnas.1404176111

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vreken, O. (1996). 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* (Vol. 3, pp. 1393–1396). Philadelphia.

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2021). Evidence of hippocampal learning in human infants. *Curr Biol*, *31*, 3358–3364.e4. doi: 10.1016/j.cub.2021.04.072

Endress, A. D. (2019). Duplications and domain-generality. *Psychological Bulletin*, *145*(12), 1154-1175. doi: 10.1037/bul0000213

Endress, A. D., & Bonatti, L. L. (2016). Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(1), 19–35. doi: 10.1002/wcs.1376

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognit Psychol*, *61*(2), 177-199. doi: 10.1016/j.cogpsych.2010.05.001

Endress, A. D., & Johnson, S. P. (2021). When forgetting fosters learning: A neural network model for statistical learning. *Cognition*, *104621*. doi: 10.1016/j.cognition.2021.104621

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

Finn, A. S., Kalra, P. B., Goetz, C., Leonard, J. A., Sheridan, M. A., & Gabrieli, J. D. (2016, feb). Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology*, *142*, 212–220. doi: 10.1016/j.jecp.2015.09.027

Fischer-Baum, S., Charny, J., & McCloskey, M. (2011, Dec). Both-edges representation of letter position in reading. *Psychon Bull Rev*, *18*(6), 1083–1089. doi: 10.3758/s13423-011-0160-3

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

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

Glover, S., & Dixon, P. (2004, Oct). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychon Bull Rev*, *11*(5), 791–806.

Gould, S. J., Lewontin, R. C., Maynard Smith, J., & Holliday, R. (1979). 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. doi: 10.1098/rspb.1979.0086

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

Graf-Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007, Mar). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychol Sci*, *18*(3), 254-60. doi: 10.1111/j.1467-9280.2007.01885.x

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

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020, July). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive science*, *44*, e12848. doi: 10.1111/cogs.12848

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *J Mem Lang*, *44*(4), 548–567.

Johnson, E. K., & Seidl, A. H. (2009, Jan). At 11 months, prosody still outranks statistics. *Dev Sci*, *12*(1), 131-41. doi: 10.1111/j.1467-7687.2008.00740.x

Jones, J., & Pashler, H. (2007, April). Is the mind inherently forward looking? comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, *14*, 295–300. doi: 10.3758/bf03194067

Keller, G. B., & Mrsic-Flogel, T. D. (2018, oct). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435. doi: 10.1016/j.neuron.2018.10.003

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002, mar). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. doi: 10.1016/s0010-0277(02)00004-5

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996, September). A neostriatal habit learning system in humans. *Science*, *273*, 1399–1402.

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

Martin, L. T., & Alberts, J. R. (1979, June). 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.

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

Mersad, K., & Nazzi, T. (2012). 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. doi: 10.1080/15475441.2011.609106

Miozzo, M., Petrova, A., Fischer-Baum, S., & Peressotti, F. (2016, May). Serial

position encoding of signs. *Cognition*, *154*, 69–80. doi: 10.1016/j.cognition.2016.05.008

Monaghan, P., & Christiansen, M. H. (2010, Jun). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *J Child Lang*, *37*(3), 545–564. doi: 10.1017/S0305000909990511

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Foris: Dordrecht.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-7. doi: 10.1126/science.1072901

Pilon, R. (1981). Segmentation of speech in a foreign language. *J. Psycholinguist. Res.*, *10*(2), 113 - 122.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends Cogn Sci*, *6*(11), 456-463.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001, November). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. doi: 10.1038/35107080

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-8.

Seidl, A., & Johnson, E. K. (2008, Feb). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *J Child Lang*, *35*(1), 1-24.

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

Shattuck-Hufnagel, S., & Turk, A. E. (1996, Mar). A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res*, *25*(2),

193-247.

Sherman, B. E., & Turk-Browne, N. B. (2020, September). 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. doi: 10.1073/pnas.2013291117

Shukla, M., Nespor, M., & Mehler, J. (2007, Feb). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognit Psychol*, *54*(1), 1-32. doi: 10.1016/j.cogpsych.2006.04.002

Shukla, M., White, K. S., & Aslin, R. N. (2011, Apr). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proc Natl Acad Sci U S A*, *108*(15), 6038–6043. doi: 10.1073/pnas.1017617108

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

Tompson, S. H., Kahn, A. E., Falk, E. B., Vettel, J. M., & Bassett, D. S. (2019, February). Individual differences in learning social and nonsocial network structures. *Journal of experimental psychology. Learning, memory, and cognition*, *45*, 253–271. doi: 10.1037/xlm0000580

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

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

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

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

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of neuroscience*, *30*, 11177–11187. doi: 10.1523/JNEUROSCI.0858-10.2010

Ullman, M. T. (2001, Oct). A neurocognitive perspective on language: The declarative/procedural model. *Nat Rev Neurosci*, *2*(10), 717-26. doi: 10.1038/35094573

Supplementary Online Materials

SOM1 Detailed Methods

SOM1.1 Experiment 1

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

SOM1.1.1 Participants. 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, a Statistical 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.

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 (Dutoit et al., 1996). (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

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:

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[1] and then compressed to an mp3 file using ffmpeg[2]. 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[3]. 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

[1] `http://sox.sourceforge.net/`

[2] `https://ffmpeg.org/`

[3] `http://psy.ck.sissa.it`

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 item 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 Experiment 2

SOM1.2.1 Materials. We re-synthesized the languages used in (Saffran et al., 1996) Experiment 2. The four words in each language are given in Table S3. Each word was composed for three syllables, which were composed of two segments in turn. Stimuli were synthesized using the us3 (male American English) voice of the mbrola synthesizer (Dutoit et al., 1996), at a constant F_0 of 120 Hz and at a rate of 216 ms per syllable (108 ms per phoneme).

Table S3
Languages used Experiment 2. The words are the same as in Experiment 2 in Saffran et al. (1996).

L1	L2
pabiku	bikuti
tibudo	pigola
daropi	tudaro
golatu	budopa

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 or (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. 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.

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. 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.[4] 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.

---

[4] A panning of the Carina nebula, obtained from `https://esahubble.org/videos/heic0707g/`.

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 (Baayen, Davidson, & Bates, 2008), 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 (Glover & Dixon, 2004), 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_iA_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);
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, 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 $3 \times 1/3$ / 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. Once the final sample of participants in the continuous condition was established, we randomly removed participants from the pre-segmented condition to equate the number of participants across the conditions. The final demographic information is given in Table S6.

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.

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

SOM3 Additional results for Experiment 2

Table S7
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.

SOM3.1 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 $P(\mathrm{model}|\mathrm{data}) = P(\mathrm{data}|\mathrm{model})\frac{P(\mathrm{model})}{P(\mathrm{data})}$. 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:

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

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.

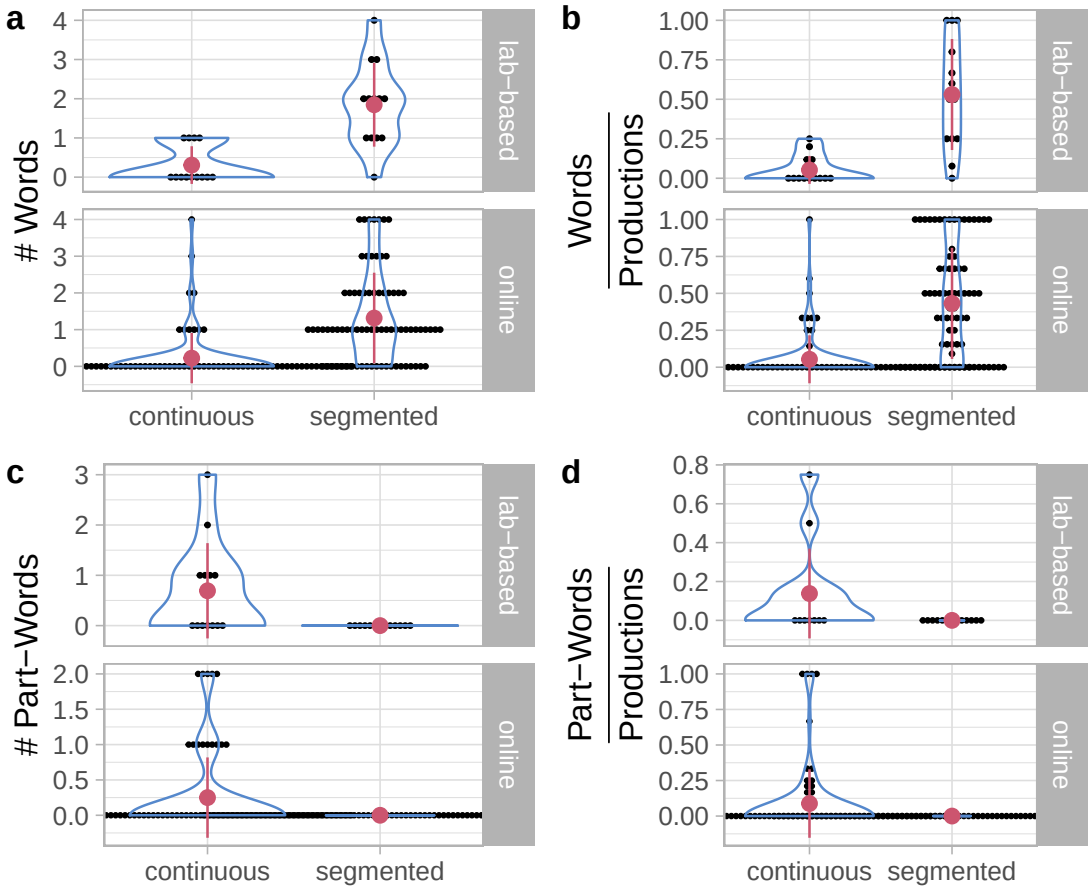

Figure S1. Number and proportion (among vocalizations) of words and part-words.

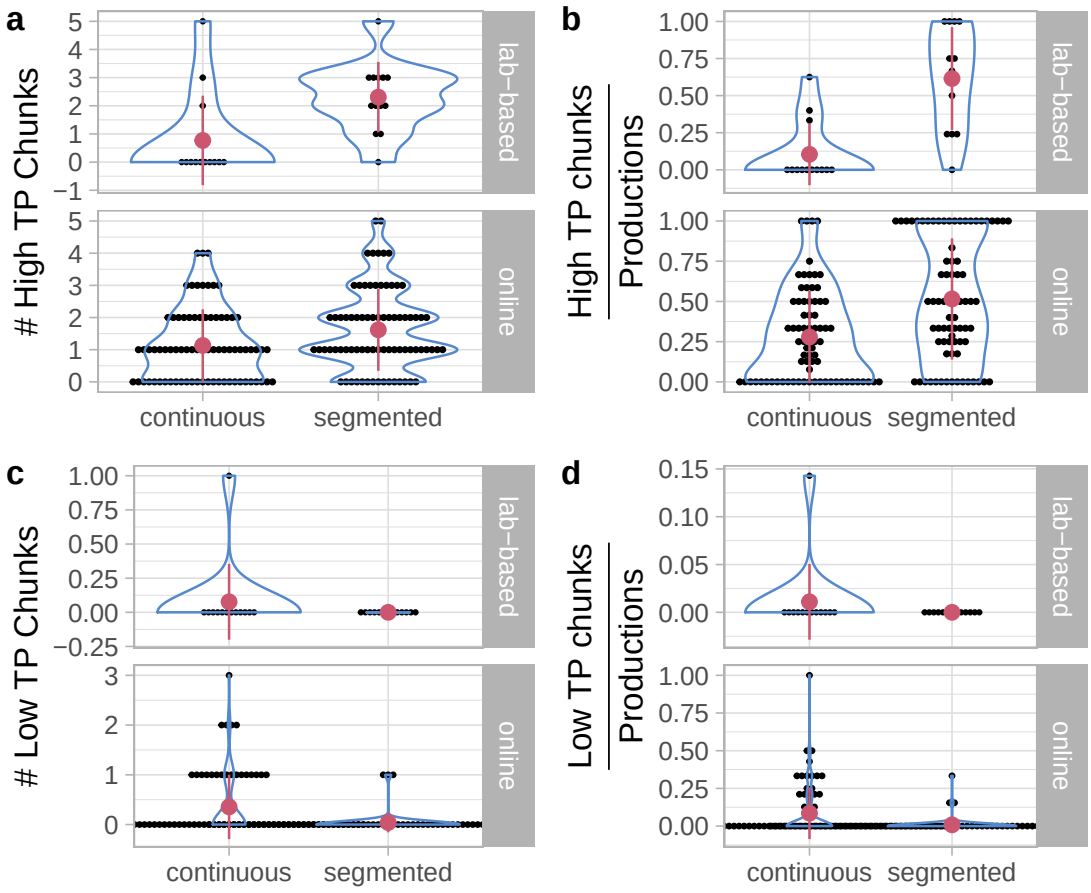

Figure S2. Plot of High and Low TP chunks.

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

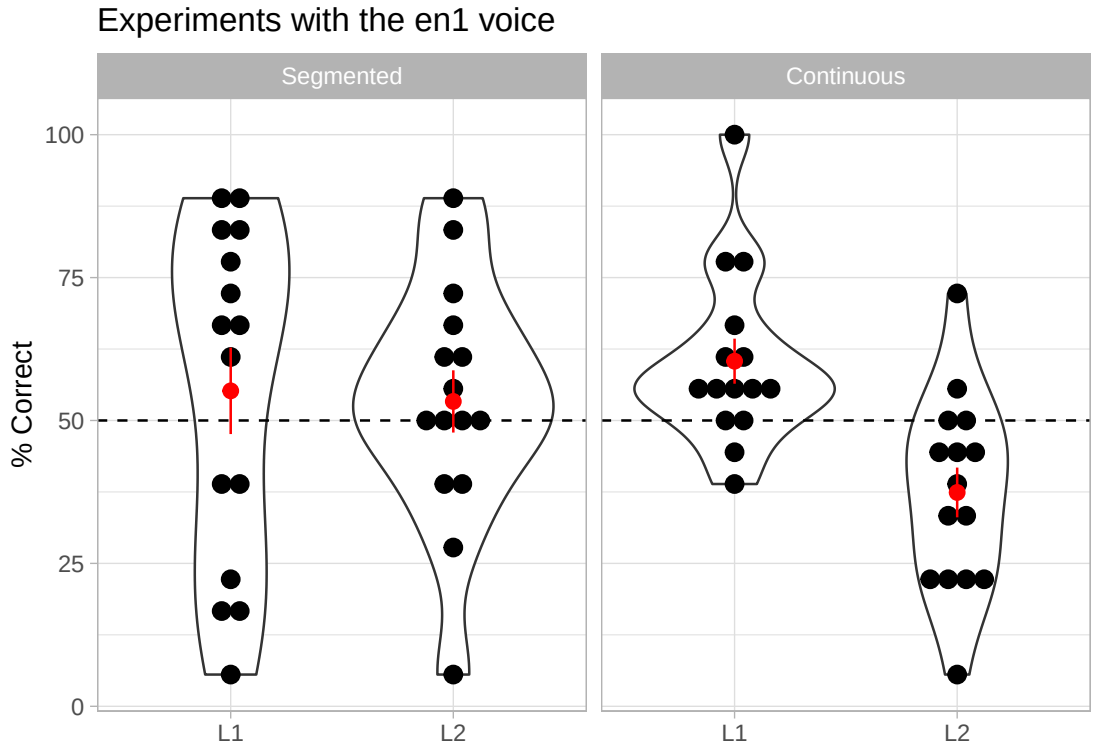

Figure S3. Results for a pre-segmented presentation of the stream (540 ms silences, left) and continuous presentation of the stream (right). Each word was repeated 45 times. The voice was *en1*.

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

S9, performance did not depend on the language condition.

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 S9, 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.

Table S8
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

Table S9

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

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

Table S10
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

SOM5.1.1 Participants. Demographic information of Pilot Experiment 2 is given in Table S10. 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 S11. They were synthesized using the *es2* (Spanish male) voice of the mbrola (Dutoit et al., 1996) 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 S11. 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 S11
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

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

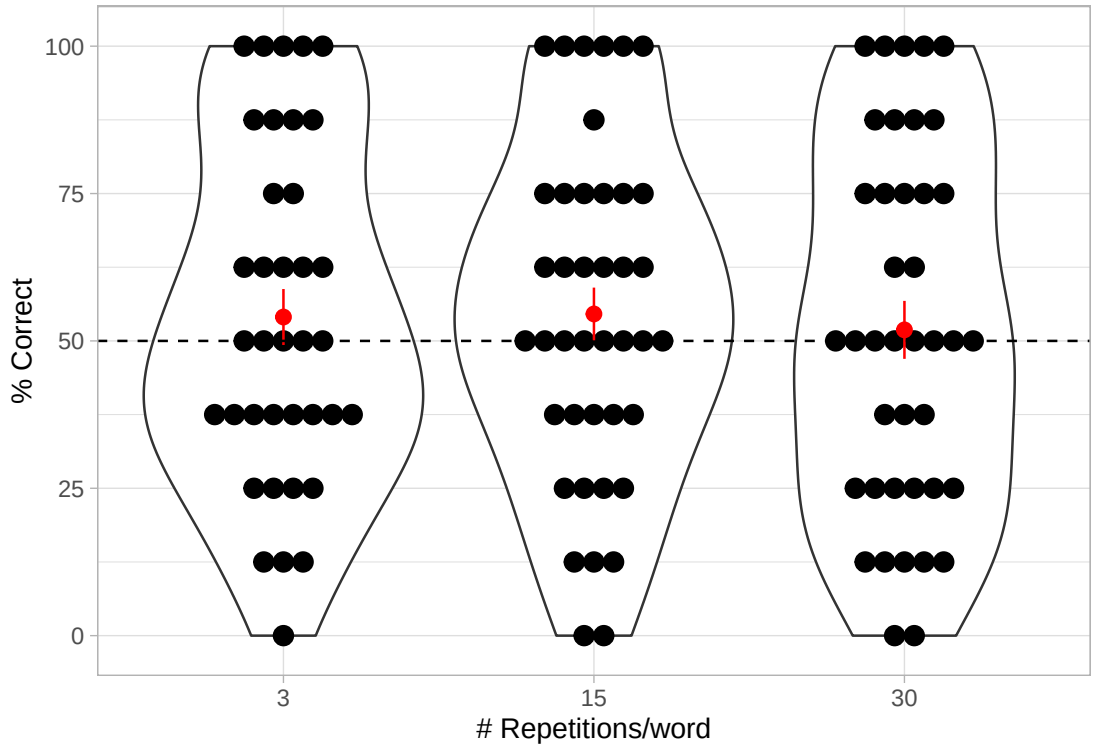

Figure S4. Results of Pilot Experiment 2. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right)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

As shown Table S12, 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

Table S12
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

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 (2004), 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.