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 associative learning mechanisms are at least partially dissociable from those involved in 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 at random positions in the sequence rather than at the onsets of the words they are meant to remember, leading to greater recall of *low*-probablility items. 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.

Contents

```
# library(renv)
# renv::activate()
# renv::restore()
```

```
# Extract R file to accelarate segmentation

knitr::purl ('segmentation_recall_combined_for_revision4.Rmd',
      'segmentation_recall_combined_for_revision4.R')
```

1 House keeping

In the analyses below, we use the following parameters:

Parameter	Value
ALLOW.WORD.REPEATS.FOR.PART.WORDS	FALSE
ANALYZED.DATA.SETS	TRUE
ANALYZED.DATA.SETS	TRUE
EQUATE.N.SUBJ	TRUE
FILTER.SINGLE.SYLLABLES	TRUE
FILTER.UNATTESTED.ITEMS	TRUE
IGNORE.COL.PREFIXES	ITI_
IGNORE.COL.PREFIXES	presTime_
IGNORE.COL.PREFIXES	ISI_

L.BAD.SUBJ.CPUTIME	list(c(subj = "399612_200413_124119_2163c2ed5ac2dd37063193b689dafe82251f433e.csv" response = "dalonigtbdophophi dalobdakabdarobigopachu"), c(subj = "399612_210517_101428_d6832877cd6a16ecb1498f99ff25a2ee66096d93.csv", response = "be cu di tu dara pe gala du dopa,be cu di pe gala,be cu di pe gala bu dopa ,be cu di bu dopa"), c(subj = "399612_210517_100654_b3a2d858e648fab540f26d63cd60cdf40be0ad25.csv" response = "takahsakakakaratatataikokokokotatakatakatakatakatakatakataka"), c(subj = " 399612_210517_101201_0e335a2e6bcd07eaf32c06cd9a1a7c2e794601ee.csv", response = "dabroobitalooki,bkuti2,golab"), c(subj = "399612_210524_062929_10c3df303f8a5b47751465793bab45d638f079f4.csv" response = "matikulatatitulapapitularimatitulaatitula"), c(subj = "399612_210524_115828_a59856877975c71a1a7b9e9e3f776cb992aaebde.csv" response = "tu kalla ti palla tuti kulla papi pu tu kalla ti palla tuti kulla papi pu"), c(subj = "399612_210524_120014_a4d641ab312e71e300bd349948e4fcc7e936105e.csv" response = "tutopitulakatutopitoolaka"), c(subj = "399612_210524_120523_06795ef32f11db08be1b1336d62b8682d9f71cc5.csv" response = "pa-pikuchi,butalapapikuchi,kukala,pikala,budharapikuchi,chupapikachubudarap c(subj = "399612_210602_064236_2c0cec9dcb1be2a6bddff35a85c2e5d558fac9eb.csv", response = "da putty da raboo,da puppy da raboo,da raboo,da raboo,da puppy da rabooo"), c(subj = "399612_210602_064353_5daaf2d0bb79fd346179a286ad5f47980fb63672.csv" response = "rabi tiku ko kolada fabi ,rabi tiku ko la dafabi...."), c(subj = "399612_210602_072517_d3db595db75dedf7869e0e6a4a0e39c6541b361e.csv response = "dolapidolabu dolapidolatu doladiputipu doladipukipu dolakiputipu dolatipu"), c(subj = "399612_210524_115845_825bd2827d674aeb68ebc89a509b3cec63ab3a4c.csv response = "gola too,the rapi papi do,the rapi do,gola du the rapi papi do"
PRINT.INDIVIDUAL.FIGURES	FALSE
PRINT.INDIVIDUAL.PDFS	FALSE
PRINT.INDIVIDUAL.TABLES	FALSE
REMOVE.BAD.SUBJ	TRUE
REMOVE.INCOMPLETE.SUBJ	FALSE
RESEGMENT.RESPONSES	FALSE
start.time	2025-03-10 23:20:41.398923

2 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.

3 Significance statement ($<$ 120w)

4 Introduction

Associative learning is remarkably widespread across species and domains [**?**; **?**; Conway2005a; **?**; **?**; **?**; **?**; **?**; **?**], and might support a wide range of computations, especially during language acquisition [**??**].

However, associative learning is also remarkably modular [**?**]. Humans have independent associative learning abilities in superficially similar domains, including associations of objects with landmarks vs. boundaries [**?**], associations among social vs. non-social objects [**?**] and associations among consonants vs. vowels [**?**]. 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 [**?**]. 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 [**?**].

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 [**??**], presumably because avoidance of the *only* food source is costly; in contrast, minimal exposure to solid food re-establishes taste-sickness associations [**?**].

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; **?**] 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 [**?**] showed that human adults are sensitive to such distributional information.

Subsequent work demonstrated that infants and non-human animals share this ability [??????], and that it might reflect simple associative mechanisms such as Hebbian learning [??].

However, a sensitivity to distributional information 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 them and thus could now have memorized them [because the items are played backwards; ?; see also ?], and sometimes even prefer high-TP items they have *never* encountered to low-TP items they have heard or seen [?]. Such results suggest that associative learning and memory for specific chunks may be dissociable [see also ?; ?; ?; ? and Discussion). In fact, the types of representations created by associative learning might well be different from those used for linguistic stimuli [??]. Conversely, associative knowledge might be critical for predictive processing [??] that is critical for both language [??] and other cognitive processes [Bar2009; ?; ?; ?].

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 [???]. This prosodic organization is perceived in unfamiliar languages [????] ~~by infants [Hirsh-Pasek1987; Christophe1994; Gout2004]~~ and even by newborns [?]. This prosodic information might affect the usefulness of statistical learning, because associative learning operates primarily *within* rather than across major prosodic boundaries [?]. 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*".

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 [?] classic experiment, again with a continuous speech stream or a sequence of pre-segmented syllable sequences.

5 Methods summary (for main text)

Unless otherwise stated, stimuli were synthesized using mbrola [**?**] 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.

5.1 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.

5.2 Experiment 1 (Recognition experiment (London))

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.

5.3 Experiment 2 (Recall experiment)

Participants were instructed to listen to a monologue in "Martian", and to remember the Martian words. The languages were those from **?** 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.

5.4 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 XXX).

6 Methods (detailed, for SI)

6.1 Recognition experiment (London)

6.1.1 Participants

Table 2: 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 2. 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.

6.1.2 Design (London)

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 3.

Table 3: 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 3.

6.1.3 Stimuli

Stimuli were synthesized using the *us3* (American English male) voice from mbrola [**?**]. (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.)

Segment 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 all participants.

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).

6.1.4 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.

6.1.5 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.

6.2 Recall experiment

6.2.1 Materials

We re-synthesized the languages used in **?** Experiment 2. The four words in each language are given in Table 4. Stimuli were synthesized using the us3 (male American English) voice of the mbrola synthesizer [**?**], at a constant F_0 of 120 Hz and at a rate of 216 ms per syllable (108 ms per phoneme).

Table 4: Languages used Experiment 2. The words are the same as in **?** Experiment 2.

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 a 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.

6.2.2 Procedure

6.2.2.1 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.

6.2.2.2 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.

6.2.2.3 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.

7 Analysis

7.1 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 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 [**?**], 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 [**?**], 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.

7.2 Recall test

7.2.1 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 5.

Each recall response was analyzed in five steps. First, we applied pre-segmentation substitution rules to make the transcriptions more consistent (see Table 5, “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 5). 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.

7.2.1.1 Substitution rules compensating for potential misperceptions

All substitution rules are listed in Table 5. 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.

Table 5: 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			

7.2.2 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 10.

7.2.2.1 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);
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.

7.2.2.2 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 $1 \times 1 / 12$, 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 $(1/12 + 1/12 + 1/12)/3 = 1/12 \approx .083$.

7.2.2.3 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.

7.2.3 Demographics and missing subjects

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% ($N = 48$). Another 12 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 6.

Table 6: 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	7	7	0	19.7	18-22
segmented	both	12	12	0	19.4	18-22
Online						
continuous	L1	22	4	18	33.1	19-71
continuous	L2	22	8	14	29.7	19-71
segmented	L1	22	5	17	30.5	18-55
segmented	L2	22	2	20	27.2	18-62

8 Results

8.1 Recognition experiments (Results with the *us3* voice; the *en1* results are in the SI)

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.

Table 7: Descriptives for Experiment 1 (using the *us3* voice) and a pilot experiment (using the *en1* voice). !!!!TO BE MOVED TO THE SI!!!!

experimentID	N	M	SE	p
us2				
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				
Pre-segmented (en1)	30	0.543	0.047	0.268
Continuous (en1)	30	0.489	0.036	0.739

8.1.1 Can people recover words from pre-segmented prosodic units?

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), $t(29) = 0.6$, p~= 0.552, Cohen's d~= 0.11, $CI_{.95}$~= 46, 57.33, ns, 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 8, performance did not depend on the language condition. As shown in SI XXX, the failure to use statistical learning was also replicated using a second voice (*en1*, British English male).

The failure to use statistical learning to split pre-segmented units was replicated in a pilot experiment with Spanish/Catalan speakers using chunk frequency and backwards TPs as the primary cues (see SI XXX).

8.1.2 Can people recover words from a continuous stream? (1)

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), $t(31) = 2.97$, p~= 0.00573, Cohen's d~= 0.52, $CI_{.95}$~= 52.66, 64.35, V = 306.5, p = 0.0185. As shown in Table 8, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition

8.1.3 Can people recover words from a continuous stream? (2) (Replication)

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), $t(29) = 3.28$, p~= 0.00272, Cohen's d~= 0.6, $CI_{.95}$~= 54.81, 70.75, $V = 320$, $p = 0.00778$. As shown in Table 8, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition.

(As shown in SI XXX, 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 able to do so in pre-segmented streams.

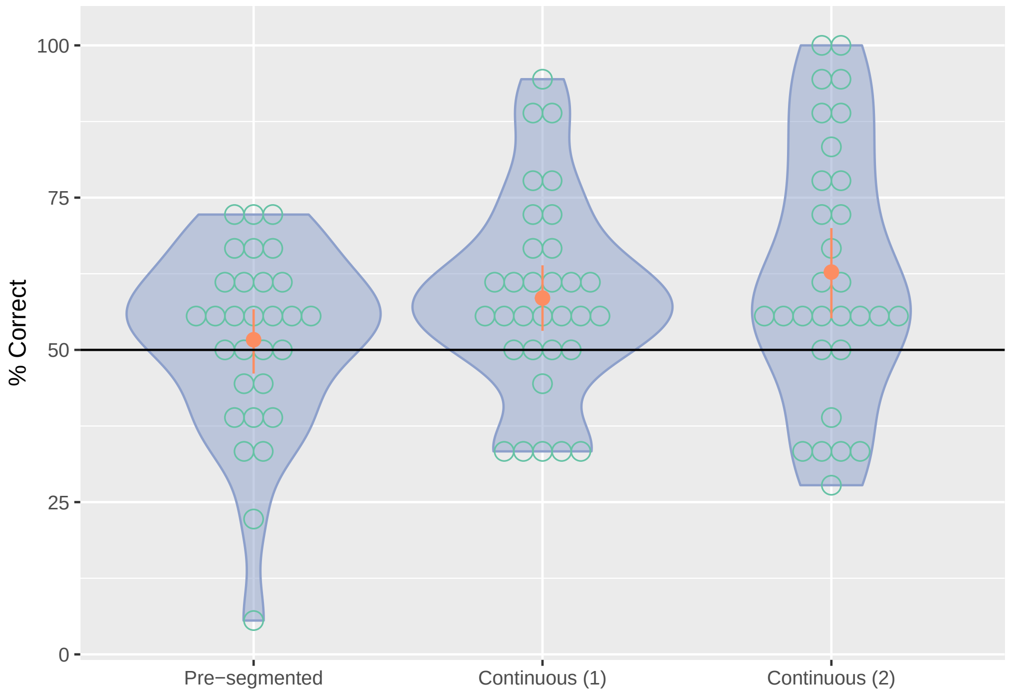

Figure 1: Results of Experiment 1. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a pre-segmented familiarization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

Table 8: Performance differences across familiarization conditions. 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					

langL2	0.114	0.673	-1.2, 1.43	0.170	0.865
Continuous familiarization (1)					
langL2	-0.184	0.480	-1.12, 0.757	-0.383	0.702
Continuous familiarization (2)					
langL2	0.317	0.786	-1.22, 1.86	0.403	0.687
Pre-segmented vs. continuous familiarization (1)					
langL2	-0.019	0.557	-1.11, 1.07	-0.033	0.973
segmsegmented	-0.328	0.188	-0.696, 0.0391	-1.752	0.080
Pre-segmented vs. continuous familiarization (2)					
langL2	0.215	0.657	-1.07, 1.5	0.327	0.743
segmsegmented	-0.608	0.244	-1.09, -0.13	-2.493	0.013

Table 9: Performance differences across familiarization conditions. 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.

	Log-odds					Odd ratios				
term	Estimate	SE	CI	t	p	Estimate	SE	CI	t	
Pre-segmented familiarization										
langL2	0.114	0.673	[-1.2, 1.43]	0.170	0.865	1.121	0.754	[0.3, 4.19]	0.170	0.86
Continuous familiarization (1)										
langL2	-0.184	0.480	[-1.12, 0.757]	-0.383	0.702	0.832	0.400	[0.325, 2.13]	-0.383	0.70
Continuous familiarization (2)										
langL2	0.317	0.786	[-1.22, 1.86]	0.403	0.687	1.372	1.079	[0.294, 6.41]	0.403	0.68
Pre-segmented vs. continuous familiarization (1)										
langL2	-0.019	0.557	[-1.11, 1.07]	-0.033	0.973	0.982	0.547	[0.329, 2.93]	-0.033	0.97
segmsegmented	-0.328	0.188	[-0.696, 0.0391]	-1.752	0.080	0.720	0.135	[0.499, 1.04]	-1.752	0.08
Pre-segmented vs. continuous familiarization (2)										
langL2	0.215	0.657	[-1.07, 1.5]	0.327	0.743	1.240	0.815	[0.342, 4.49]	0.327	0.74
segmsegmented	-0.608	0.244	[-1.09, -0.13]	-2.493	0.013	0.544	0.133	[0.337, 0.878]	-2.493	0.01

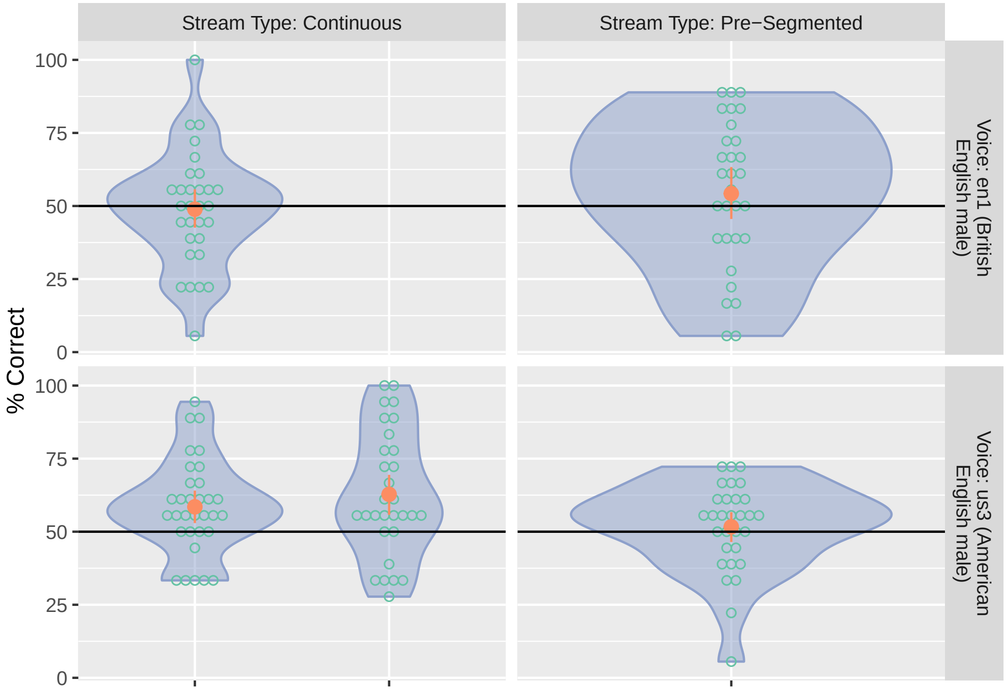

Figure 2: Results of Experiment 1. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) continuous familiarization stream or (right) a pre-segmented familiarization stream, synthesized with an American English voice (top) or a British English voice (bottom). The two continuous conditions are replictions of one another.

8.2 Recall experiment

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 **?** 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. 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. We ran both a lab-based and an online version of this experiment. 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.

Table 10: Analyses performed for the vocalizations

colName	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.

8.2.1 General measures

As shown in Table 11 and Figures 4a and b, participants produced about 4 items. Neither the number of items produced nor their lengths differed across the segmentation conditions.

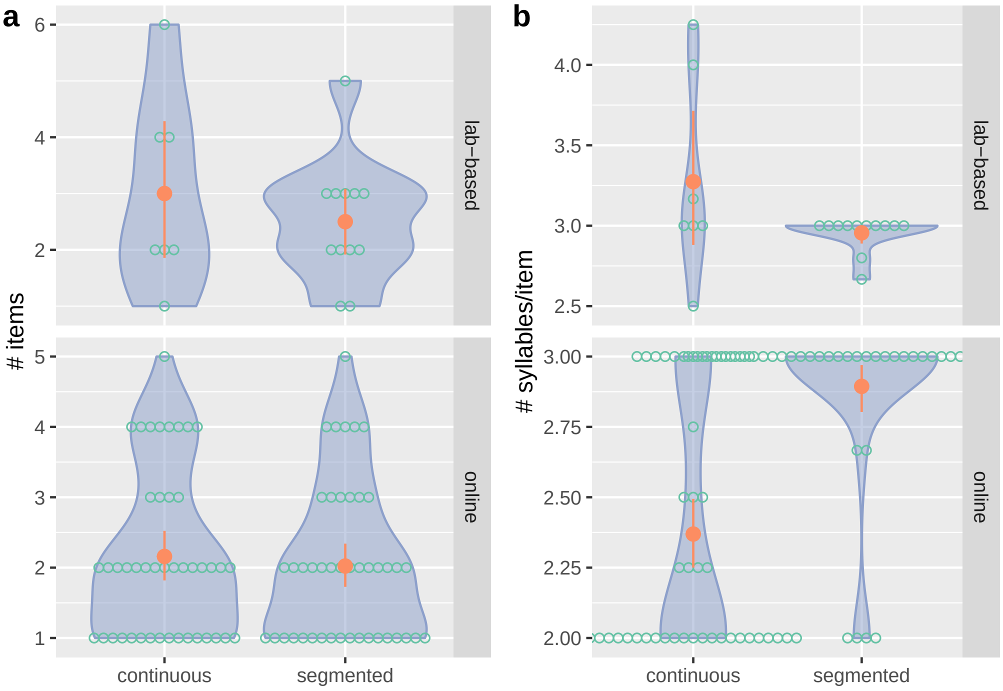

Figure 3: Number of items produced as well as their numbers of syllables.

8.2.2 TP-based analyses

Critically, and as shown in Table 11 and Figures 4c and d, forward and backward TPs in the participants' were significantly greater than the chance level of .083 in both segmentation conditions. Further, these TPs likely underestimate the participants' performance, as we included responses with unattested syllables that might reflect misperceptions; after removing such responses from analyses, 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.)

8.2.3 TP-based chunks analysis

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 7b, the proportion of high-TP among 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 1.283 (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.

8.2.4 Word vs. part-word analysis

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 a trisyllabic sequence from the stream, they can start that sequence on the first, second or third syllable of a word, and only the first possibility yields a word. As a result, if participants initiate their productions at a random position, a third of their productions should be words.

As shown in Table 11 and in Figure 7a, 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 0.807 more likely to perform at the chance level of 33% (2.636 for the lab-based experiments) than to perform at a level different from chance. This results thus suggest that participants in the continuous condition initiate their productions at random positions in the stream.

However, inspection of Figure 7a shows that the distribution after continuous sequences is clearly 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 a random position (and should produce words with a probability of 1/3). As shown in SI XXX, 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 13, these results are 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.) Given that participants produce few tri-syllabic items, we thus focus on shorter chunks.

8.2.5 Positional analyses

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, 4/12 should have correct final syllables.

As shown in Table 11 and Figure 7c 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 3.06 for initial syllables (2.559 for the lab-based experiment) and 3.444 for final syllables (2.026 for the lab-based experiment).

Table 11: Various analyses pertaining to the productions as well as
test against their chances levels.

	Continuous	Segmented
Recognition accuracy		
lab-based	N = 7, \M = 0.607, \SE = 0.0803, \p = 0.371	N = 12, \M = 0.917, \SE = 0.0491,
online	N = 44, \M = 0.71, \SE = 0.0328, \p = 1.06e-05	N = 44, \M = 0.943, \SE = 0.0181,
Number of items		
lab-based	N = 7, \M = 3, \SE = 0.707, \p = 0.0213	N = 12, \M = 2.5, \SE = 0.328, \p
online	N = 44, \M = 2.16, \SE = 0.18, \p = 5.3e-09	N = 44, \M = 2.02, \SE = 0.169, \p
Number of syllables/item		
lab-based	N = 7, \M = 3.27, \SE = 0.254, \p = 0.0215	N = 12, \M = 2.96, \SE = 0.0324, \
online	N = 44, \M = 2.37, \SE = 0.068, \p = 3.8e-09	N = 44, \M = 2.89, \SE = 0.0449, \
Proportion of words among words and part-words (or concatenations thereof)		
lab-based	N = 7, \M = 0.321, \SE = 0.153, \p = 0.322 (vs. 0.5); 0.798 (vs. 0.333333333333333)	N = 12, \M = 1, \SE = 0, \p = 0.00 0.333333333333333)
online	N = 16, \M = 0.562, \SE = 0.132, \p = 0.638 (vs. 0.5); 0.0353 (vs. 0.333333333333333)	N = 38, \M = 1, \SE = 0, \p = 7.46 0.333333333333333)
Forward TPs		
lab-based	N = 7, \M = 0.692, \SE = 0.0627, \p = 0.0156	N = 12, \M = 0.938, \SE = 0.0357,
online	N = 44, \M = 0.784, \SE = 0.0355, \p = 5.34e-09	N = 44, \M = 0.948, \SE = 0.0202,
Backward TPs		
lab-based	N = 7, \M = 0.692, \SE = 0.0627, \p = 0.0156	N = 12, \M = 0.938, \SE = 0.0357,
online	N = 44, \M = 0.784, \SE = 0.0355, \p = 5.34e-09	N = 44, \M = 0.948, \SE = 0.0202,
Proportion of High-TP chunks among High- and Low-TP chunks		
lab-based	N = 4, \M = 0.75, \SE = 0.289, \p = 0.424 (vs. 0.5); 0.85 (vs. 0.666666666666667)	N = 12, \M = 1, \SE = 0, \p = 0.00 0.666666666666667)
online	N = 40, \M = 0.764, \SE = 0.0553, \p = 9.26e-05 (vs. 0.5); 0.00762 (vs. 0.666666666666667)	N = 44, \M = 0.989, \SE = 0.0115, \ (vs. 0.666666666666667)
Proportion of items with correct initial syllables		
lab-based	N = 7, \M = 0.298, \SE = 0.149, \p = 0.672 (vs. 0.333333333333333); 0.495 (vs. 0.375)	N = 12, \M = 0.931, \SE = 0.0501, 0.333333333333333); 0.00123 (vs. 0.3
online	N = 44, \M = 0.412, \SE = 0.0638, \p = 0.354 (vs. 0.333333333333333); 0.645 (vs. 0.375)	N = 44, \M = 0.878, \SE = 0.0426, 0.333333333333333); 3.72e-09 (vs. 0.3
Proportion of items with correct final syllables		
lab-based	N = 7, \M = 0.429, \SE = 0.141, \p = 0.666 (vs. 0.333333333333333); 0.666 (vs. 0.375)	N = 12, \M = 0.9, \SE = 0.0627, \p 0.333333333333333); 0.00202 (vs. 0.3
online	N = 44, \M = 0.406, \SE = 0.0642, \p = 0.388 (vs. 0.333333333333333); 0.594 (vs. 0.375)	N = 44, \M = 0.886, \SE = 0.0434, 0.333333333333333); 2.01e-09 (vs. 0.3

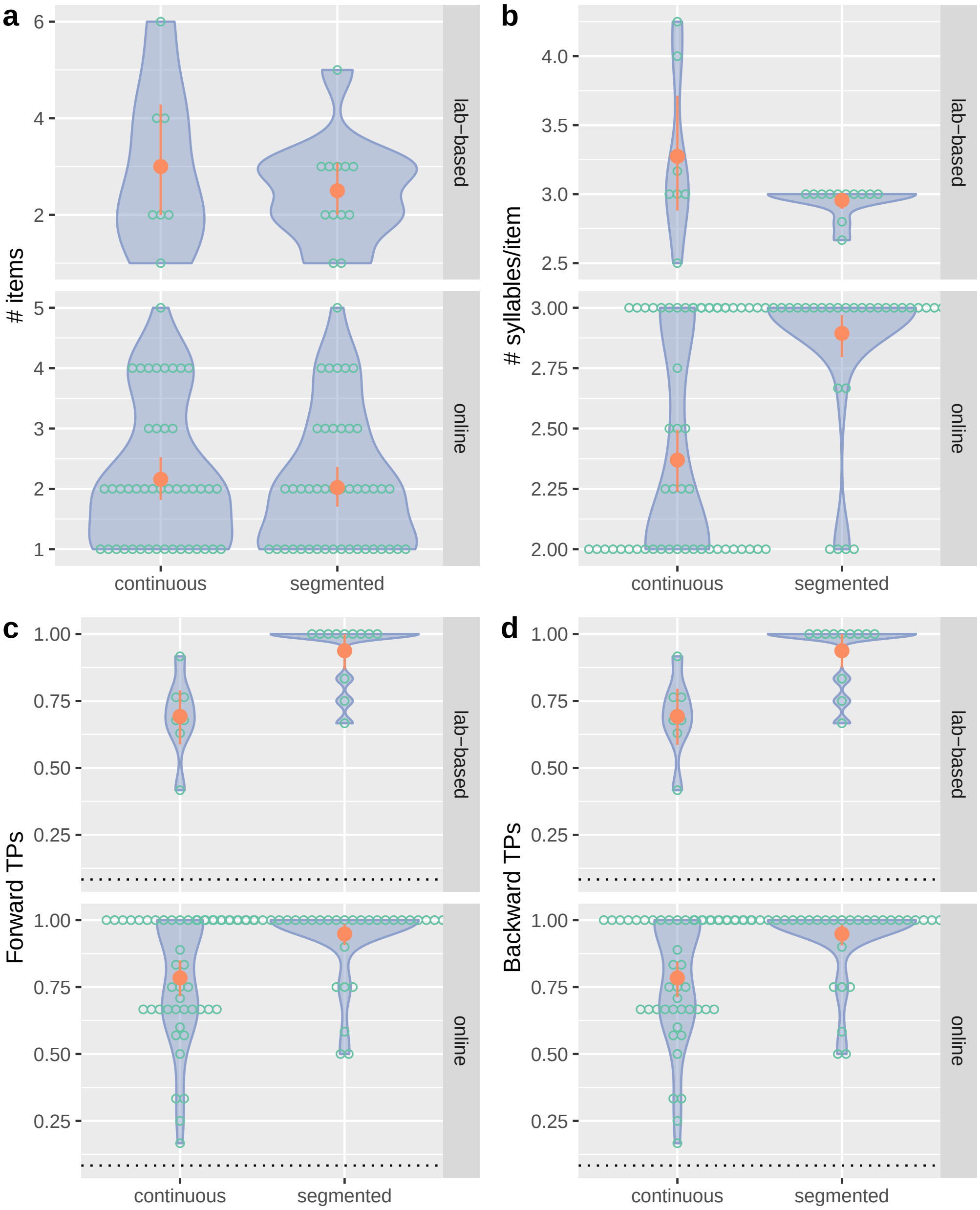

Figure 4: Number of items produced, number of syllables per item and forward and backward TPs. The dotted line represents the chance level for a randomly ordered syllable sequence.

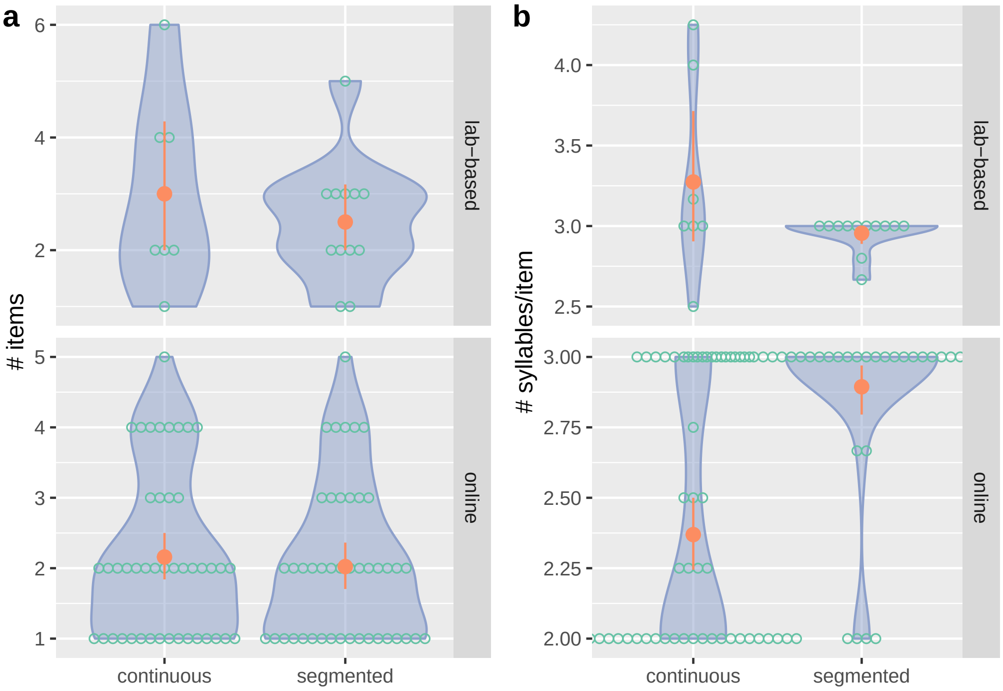

Figure 5: Number of items produced, number of syllables per item and forward and backward TPs. The dotted line represents the chance level for a randomly ordered syllable sequence.

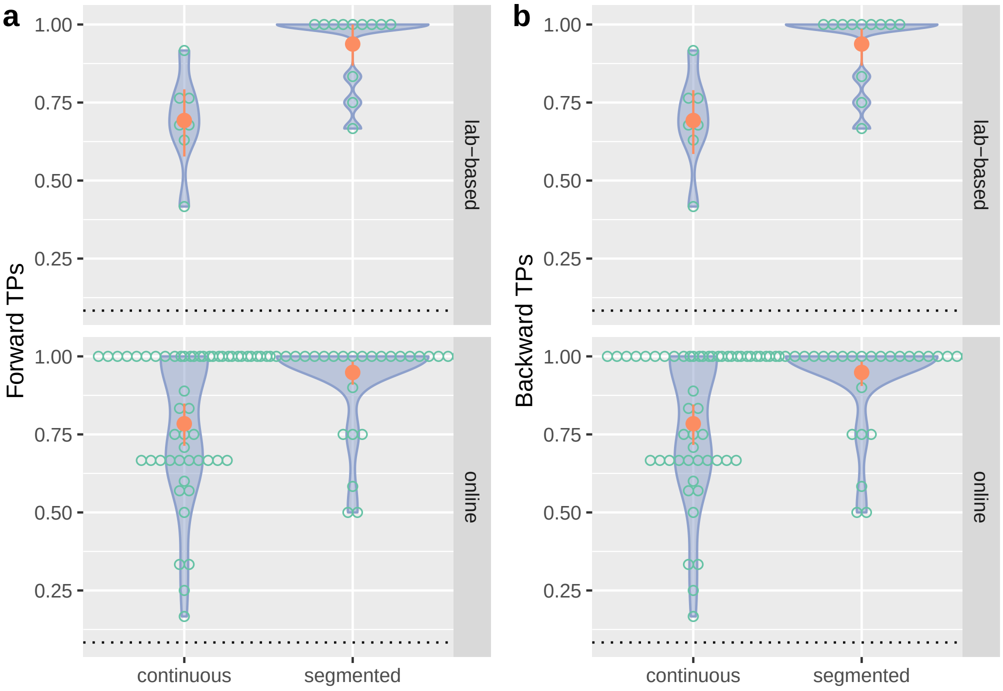

Figure 6: Number of items produced, number of syllables per item and forward and backward TPs. The dotted line represents the chance level for a randomly ordered syllable sequence.

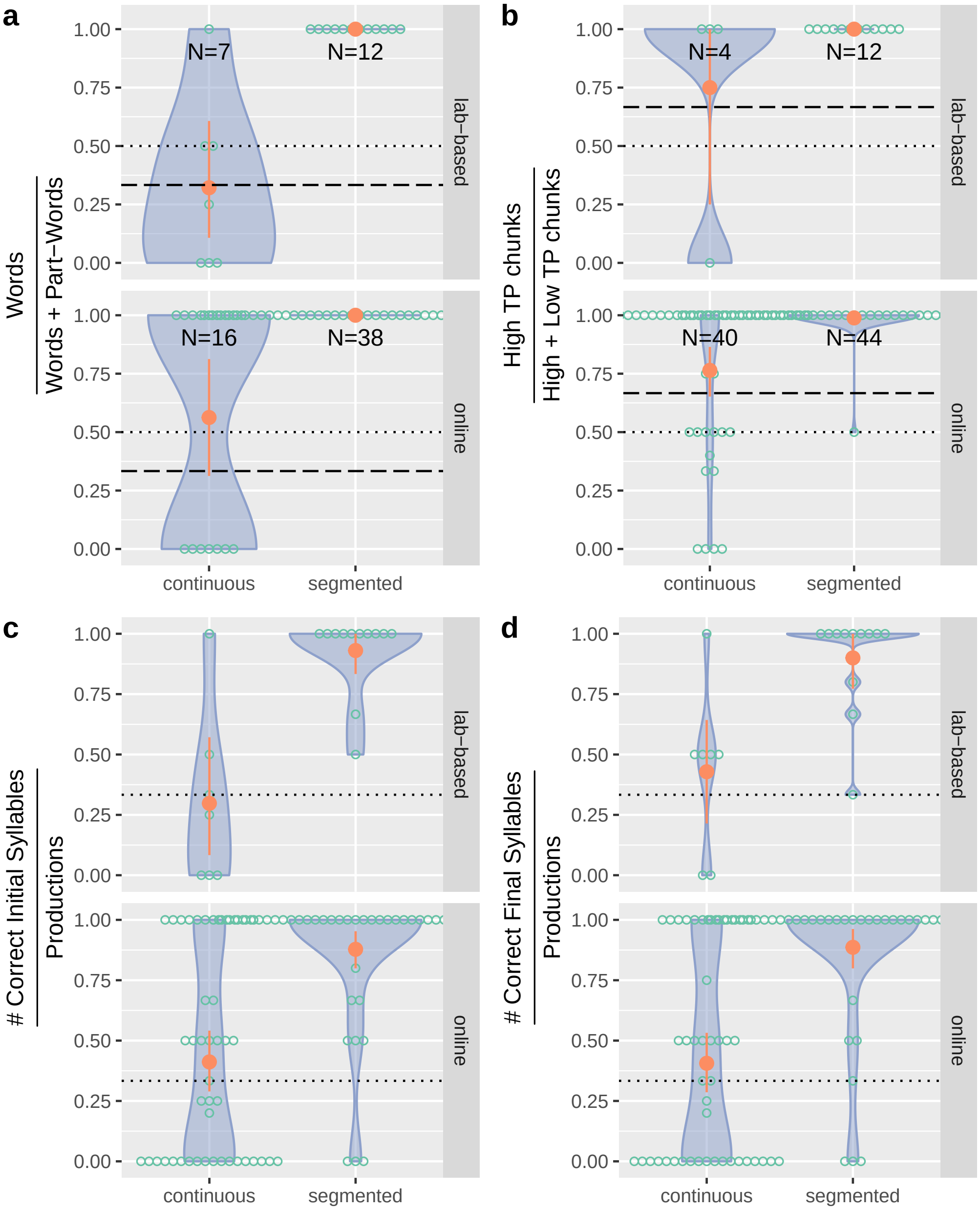

Figure 7: Analyses of the participants' productions. (a) Proportion of words among words and part-words. The dotted line represents the chance level of 50 percent in a two-alternative forced-choice task, while the dashed line represents the chance level of 33 percent that an attested 3 syllable-chunk 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 percent 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 percent.

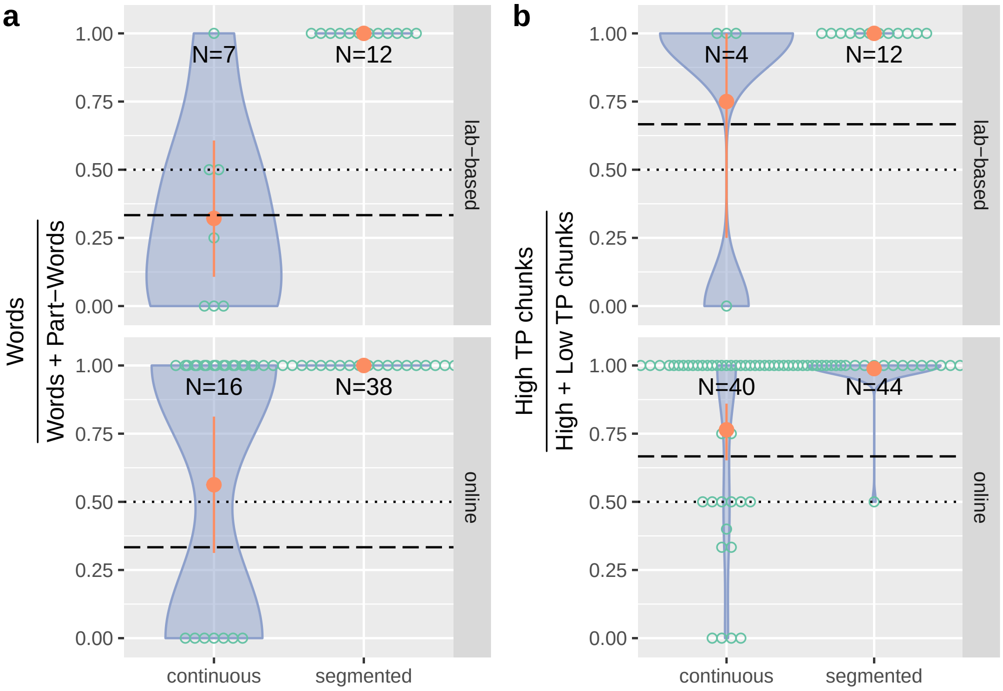

Figure 8: Analyses of the participants' productions. (a) Proportion of words among words and part-words. The dotted line represents the chance level of 50 percent in a two-alternative forced-choice task, while the dashed line represents the chance level of 33 percent that an attested 3 syllable-chunk 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 percent 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 percent.

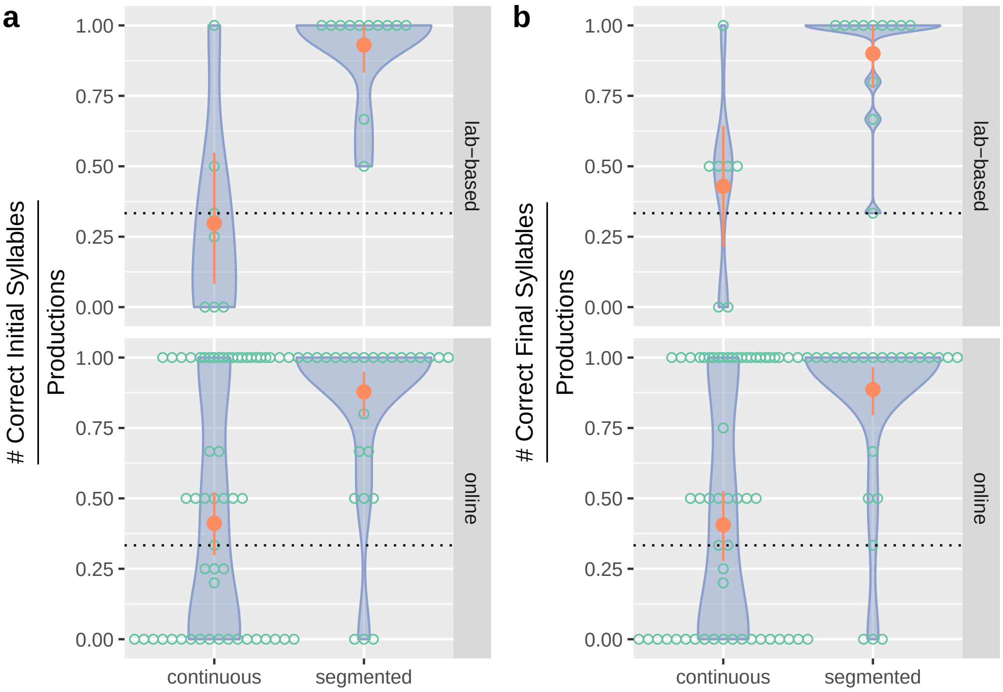

Figure 9: Analyses of the participants' productions. (a) Proportion of words among words and part-words. The dotted line represents the chance level of 50 percent in a two-alternative forced-choice task, while the dashed line represents the chance level of 33 percent that an attested 3 syllable-chunk 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 percent 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 percent.

8.3 Can a chunking model account for these results? Simulations with PARSER (added for revision for Cognitive Psychology)

Taken together, these results suggest that participants can learn statistical information from fluent speech, but that the information they retain does not only allow them to learn (statistically defined) chunks that might then be encoded as word candidates in declarative long-term memory. In fact, among those participants who produce words or part-words at all, about two thirds produce part-words. Such results suggest that statistical learning does not support the function it with which it was motivated originally – to learn words from fluent speech.

To illustrate the conclusion that a chunking model will not produce part-words rather than words, we attempted to bias PARSER [**?**], a prominent chunking model of word segmentation, to prefer part-words over words. PARSER segments continuous streams by recursively chunking units in the stream. These units are syllables or syllable combinations the model has encountered and retained in the speech stream. Units are built up recursively. For example, if a unit A is followed by a unit B, the model can create a new and larger unit AB that it can recognize later on. As a result, if this new unit AB is later followed by C, a new and still larger unit ABC might be created. The weight of recurring units is strengthened, while spurious units are eliminated through decay and interference.

We first familiarized the model with one of the speech streams used in Experiment 1 (i.e., one of the speech streams from [**?**] Experiment 2). Following this, we recorded the memory strength of words and of part-words. Specifically, we created 4 test trial pitting the two words against the two part-words, and, in each test trial, compared the weight of the word and that of the part-word. We assigned a value of 1 to the trial if the weight of the word in the lexicon was higher, of 0 with the weight of the part-word was higher, and of .5 if the two weights were the same. We then averaged these scores for all trials, and used this average as the performance of a simulated participant (see below).

We attempted to bias the model to prefer part-words in two ways. First, we deleted the first two syllable from each speech stream. Speech streams thus started with a part-word. Second, at each time step, PARSER reads in a randomly determined number of units. We forced it to read in three units on the first time step, and thus to create a part-word in its lexicon.

PARSER has five parameters: the maximal number of units considered, the increment in memory strength upon encountering a unit, the weight threshold for an item to be removed from the lexicon, the initial weights of the syllables, the forgetting rate and the interference rate. We varied the forgetting rate and the interference rate and kept the original values of the other variables. We used forgetting rates from 0 to .1 and interference rates from 0 to .01, both in 101 equidistant steps. (In the original model, the forgetting rate was .05 and the interference rate of .005.) These parameter combinations thus yielded $101 \times 101 = 10,201$ simulated "experiments.' ' Each experiment was run with 50 random initializations, representing 50 participants. The one of 40 randomizations of the words in the speech stream was randomly chosen for each participant.

The results revealed that all 507,965 simulated participants for which we obtained data (i.e., who had either words or part-words in the lexicon) had a preference for words over part-words.

Further, all 10201 simulated experiments showed a statistical significant preference for words. . Across experiments, the average effect size (Cohen's d) was 1.616 (range 0.344, 3.672), with the smaller effect sizes mainly occurring for high forgetting rates (see Figure 10). With [**?**]'s original parameters, the effect size was 1.833.

These results thus show that at least one prominent chunking model will never prefer part-words over words. Given that the majority of those participants who produced either words or part-words produced words, these suggest that chunk models either cannot account for the current results, or, to the extent that other chunking models might account for them, that these models learn information that does not allow them to recover words from fluent speech.

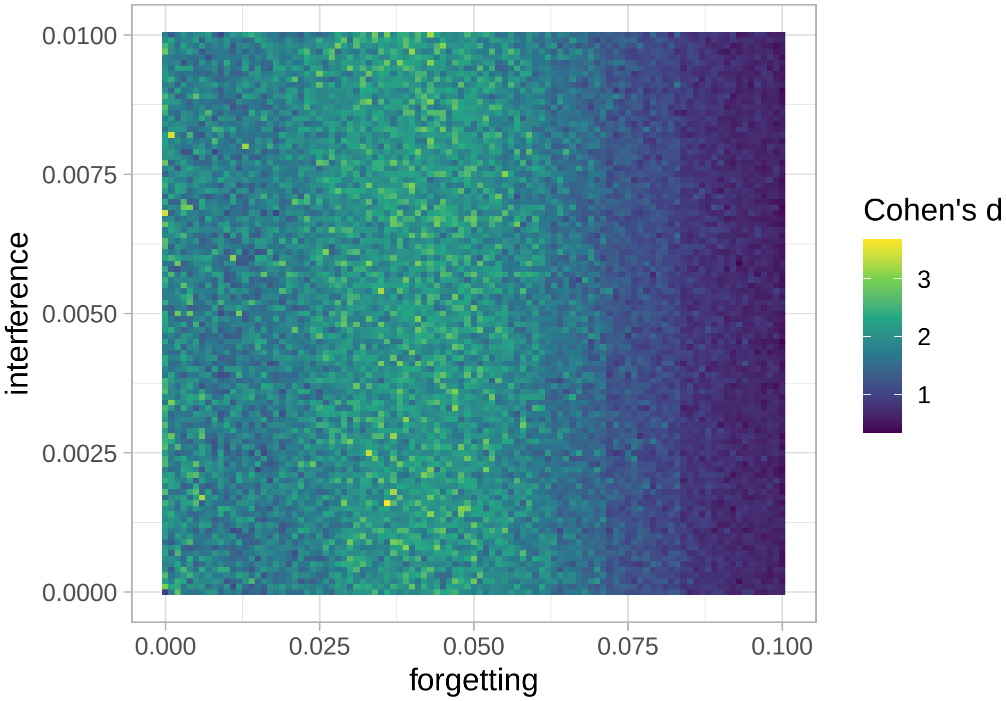

Figure 10: Effect sizes (Cohen's d) of the preference for words over part-words in PARSER as a function of the forgetting rate and the inteference rate. All simulated experiments yielded a significant preference for words.

8.4 Correlations (added for revision for Cognitive Psychology)

The results so far suggest that the information extracted in statistical learning tasks does not allow participants to identify word boundaries. Further, the pattern of performance cannot be explained by a prominent chunking model of word segmentation [**?**].

Statistical learning performance (as measured in the recognition test) might still be related to memory for words candidates (as measured by the productions)[1], albeit indirectly. For example, in analogy to electrophysiological findings suggesting that statistically structured sequences can elicit periodic brain activity [**?**, @Flo2022,@Kabdebon2015,@Moser2021], participants who produced part-words might have focused on those syllables at the beginning of part-words (if they reflect the period onset of an oscillation), and those who produced words might have focused on word-initial syllables, and the syllables participants happen to focus on might be chosen randomly. While such periodic activity can result from Hebbian learning mechanisms that do not place any items in long-term memory (Endress & Fló, under review), it might still direct the participants' attention. Given that attention affects statistical learning [**?**, @Turk-Browne2005], participants who happen to rhythmically entrain to part-words would focus on statistically less cohesive syllable sequences, while participants who happen to entrain to words would focus on statistically more cohesive syllables, which might affect recognition performance in turn. (We cannot exclude the possibility that recognition performance might be directly linked to production performance, without the mediation of other processes such as attentions. However, this view would imply that statistical learning does not allow the majority of the participants to learn words from fluent speech, given that two thirds of the participants produced part-words rather than words.)

Comparing recognition and recall performance is problematic, because our recall data is discrete rather than continuous. We will thus link recognition and recall performance through two analyses. First, and as mentioned above, two-thirds of the participants in the continuous condition produced part-words, while only one third produced words. We will compare performance in the recognition phase between those participants producing part-words and those producing words.

Second, it turned out that, during the recall phase, the proportion of productions with "correct" initial or final syllables was reasonably continuous (see Figure XXX). We will thus correlate these proportions as well as the TPs in the strings produced by the participants with their performance in the recognition phase.

8.4.1 Discrete measures

The overwhelming majority of participants who produced words or part-words produced either exclusively words or exclusively part-words (or concatenations thereof). For our analysis, we thus excluded a total of 3 participants who had intermediate proportions.

Further, we excluded participants who produced neither words nor part-words. The counts are shown in Table 12. Further, since no participants in the pre-segmented conditions produced part-words, some statistical comparisons are not available for the pre-segmented condition.

Table 12: Counts of participants producing exlusively words, exclusively part-words, neither words nor part-words, or a mixture of both. For the comparison of the recognition performance of participants who produced part-words vs. words, we excluded participants who produced neither of these item types or a mixture thereof.

		Participants producing			
data.set	streamType	Part-words	Words	Neither (excluded)	Mixture (excluded)
lab-based	continuous	3	1	0	3
lab-based	segmented	0	12	0	0
online	continuous	7	9	28	0

[1] We thank an anonymous reviewer for suggesting this possibility

online	segmented	0	38	6	0

Table 13: Recognition performance as a function of whether participants produced words or part-words. The p value reflects a Wilcoxon test comparing participants producing words and participants producing part-words, respectively.

			Recognition performance		
Segmentation Condition	Productions	N	M	SE	p
lab-based					
continuous	Words	1	50.0	NA	0.637
continuous	Part-Words	3	75.0	17.68	NA
segmented	Words	12	91.7	4.91	NA
online					
continuous	Words	9	88.9	6.42	0.003
continuous	Part-Words	7	53.6	3.86	NA
segmented	Words	38	94.1	2.01	NA

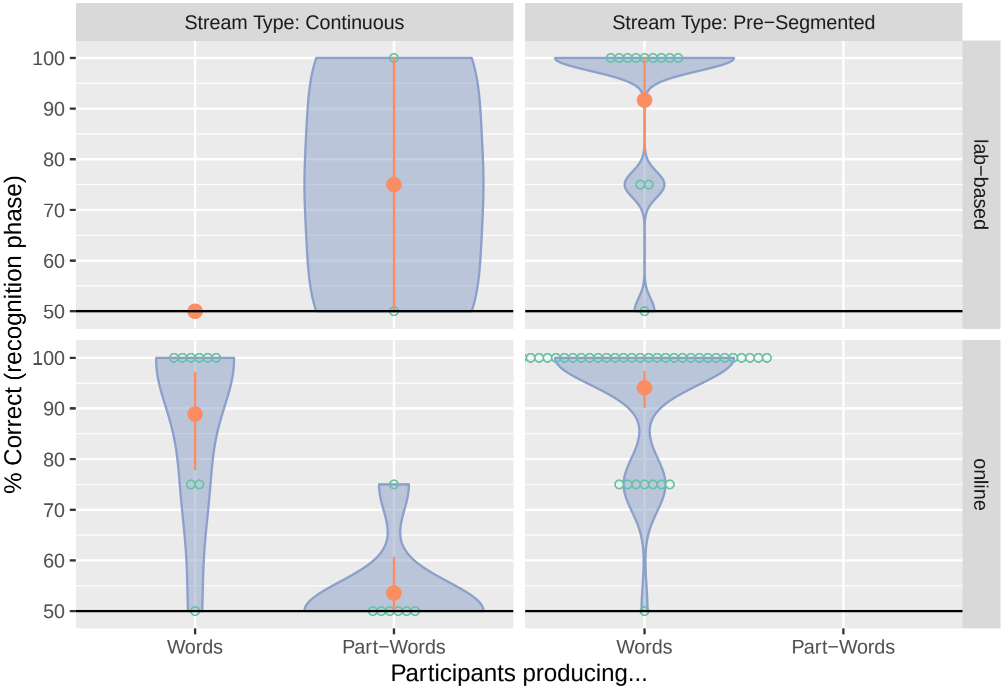

Figure 11: Recognition performance in Experiment 1 as a function of whether a participant produces words or part-words. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a continuous familiarization stream or (right) a pre-segmented familiarization stream, in the lab-based version of the experiment (top) or in the online version (bottom).

As shown in Table 13 and Figure 11, participants in the continuous condition of the online experiment who produced words performed statistically better in the recognition test as well. (In the lab-based experiments, there were only 4 participants in total who produced either words or part-words, making statistical comparisons unreliable.)

8.4.2 Continuous measures

We next correlate recognition performance with the continuous measures of the recall performance, that is, the average forward TPs of the production, the proportion of productions with correct initial syllables, and the proportion of productions with correct final syllables.

As shown in Figure 12, recognition performance in the continuous condition was correlated both with the proportion of correct initial and final syllables in the participants productions, though not with the average TPs in their productions. These correlation were not significant in the pre-segmentation conditions, presumably because of the very high level of performance.

Add to figure legend: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

While these results suggest that recognition and recall performance are related, the underlying causal pathway is unclear. On the one hand, and as mentioned above, participants who happen to focus on those syllables corresponding to words rather than part-words would also focus on more statistically cohesive syllable sequences, which, in turn, would lead to better recognition performance as well. Alternatively, recall and production performance might also be linked directly. Critically, under either hypothesis, statistical learning would not allow help participants with the problem that motivated sequentail statistical learning approaches to word segmentation in the first instance, namely to identify word boundaries in fluent speech.

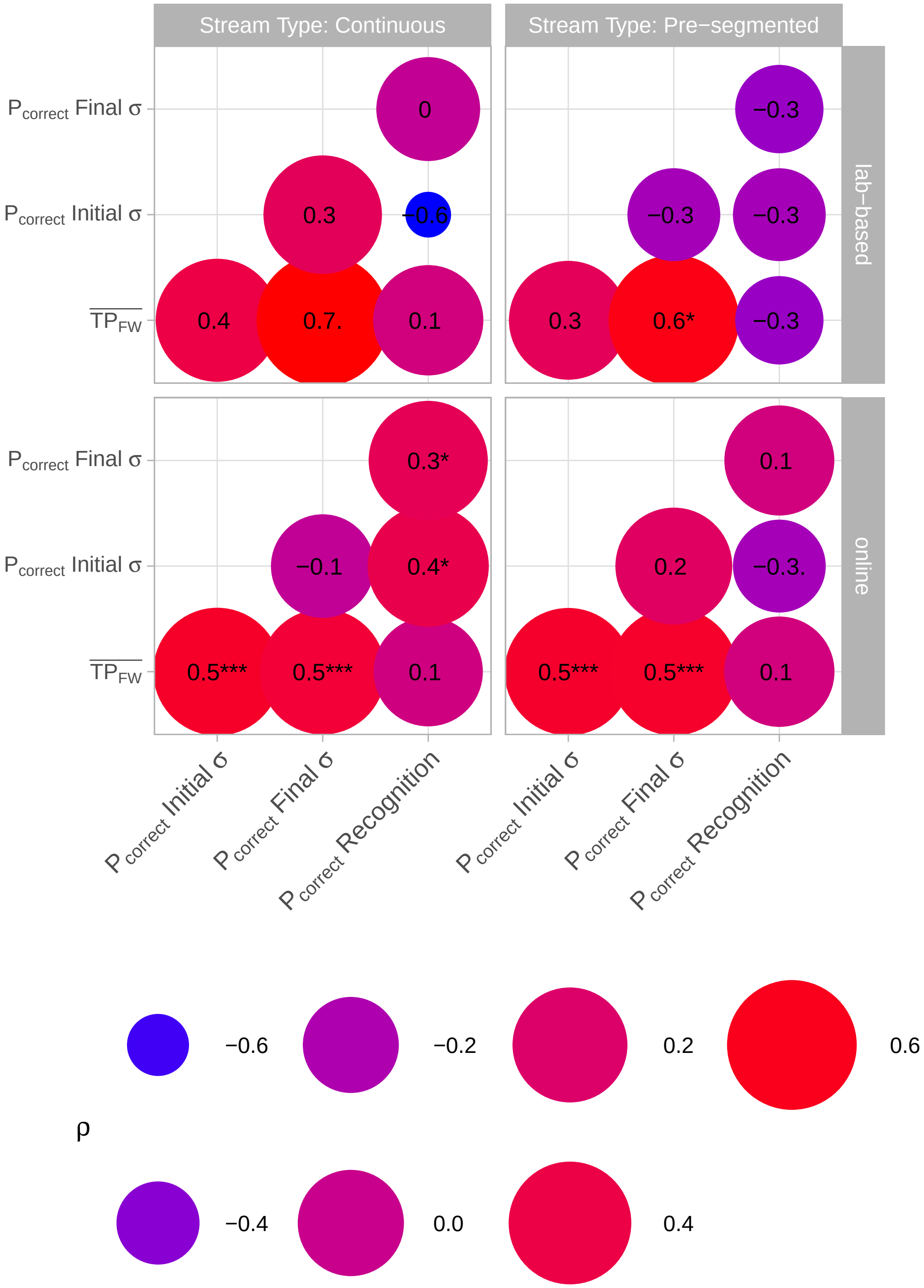

Figure 12: Spearman correlations between the performance in the recognition test (Pcorrect Recognition) three measures of the participants' productions: The proportion of correct initial syllables (P correct Initial sigma) and of final syllables (P correct Final sigma) as well as the average forward TPs in the participants' productions (bar TP FW).

9 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. This result is consistent with the possibility that associative learning is important for predictive processing, but maybe less so for memorizing utterances [**??**], especially given that speech streams tend to be pre-segmented due to their prosodic organization [**???????**]. These results echo those from conditioning experiments suggesting that, depending on the learning situation, some associations are enhanced, while others are prevented [**?????**]. Experiment 1 suggests that associative learning predominantly occurs in continuous sequences. While prediction is arguably more useful in lengthy chunks, it is unclear if there are specific triggers that suppress the integration of associative information across chunks in temporal sequences. If so, associative learning might well be implicit, but still under the control of some stimulus features.

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 at random positions with respect to actual word boundaries, suggesting 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 [**?**], developmentally [**?**] and neuropsychologically [**?????**], to the extent that statistical predictions can *impair* declarative memory encoding in healthy adults [**?**]. The standard conclusion is that the (cortical) declarative memory system might be independent of a (neostriatal) system for associative learning [**???**]. In line with earlier proposals [**???**], 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 [**???**] ~~[Shi2008; Weijer1999]~~, attentional allocation to beginnings and ends of utterances [**???**], legal sound sequences [**??**] and universal aspects of prosody [**??????**].

Such cues have a critical advantage over transition-based associative information: they are consistent with the (declarative) memory encoding of linguistic sequences. In fact, the order of linguistic sequences is encoded with respect to their first and their last element [**??**]. To the extent that learners use learning mechanisms that are adaptive for a learning problem, using declarative memory mechanisms thus seems more conducive to word segmentation than relying on associative information.

10 Appendix

10.1 Additional results for the recall experiments

10.1.1 Additional tables and figures

Table 14: Various supplementary analyses pertaining to the productions as well as test against their chances levels.

	Continuous	Segmented
Number of words		
lab-based	N = 7, \M = 0.571, \SE = 0.218, \p = 0.0719	N = 12, \M = 2, \SE = 0.287, \p =
online	N = 44, \M = 0.318, \SE = 0.118, \p = 0.00709	N = 44, \M = 1.66, \SE = 0.186, \p
Proportion of words among productions		
lab-based	N = 7, \M = 0.571, \SE = 0.218, \p = 0.0719	N = 12, \M = 2, \SE = 0.287, \p =
online	N = 44, \M = 0.318, \SE = 0.118, \p = 0.00709	N = 44, \M = 1.66, \SE = 0.186, \p
Number of part-words		
lab-based	N = 7, \M = 1.29, \SE = 0.388, \p = 0.031	N = 12, \M = 0, \SE = 0, \p = NaN
online	N = 44, \M = 0.205, \SE = 0.0777, \p = 0.0177	N = 44, \M = 0, \SE = 0, \p = NaN
Proportion of part-words among productions		
lab-based	N = 7, \M = 1.29, \SE = 0.388, \p = 0.031	N = 12, \M = 0, \SE = 0, \p = NaN
online	N = 44, \M = 0.205, \SE = 0.0777, \p = 0.0177	N = 44, \M = 0, \SE = 0, \p = NaN
Actual vs. expected forward TPs		
lab-based	N = 7, \M = -0.118, \SE = 0.0508, \p = 0.0579	N = 12, \M = -0.0486, \SE = 0.0312
online	N = 44, \M = -0.0477, \SE = 0.0218, \p = 0.00175	N = 44, \M = -0.0386, \SE = 0.0176
Number of High-TP chunks		
lab-based	N = 7, \M = 1.43, \SE = 0.812, \p = 0.181	N = 12, \M = 2.5, \SE = 0.328, \p
online	N = 44, \M = 1.5, \SE = 0.176, \p = 1.05e-07	N = 44, \M = 2, \SE = 0.171, \p =
Proportion of High-TP chunks among productions		
lab-based	N = 7, \M = 0.369, \SE = 0.19, \p = 0.181	N = 12, \M = 1, \SE = 0, \p = 0.00
online	N = 44, \M = 0.676, \SE = 0.0601, \p = 6.65e-08	N = 44, \M = 0.989, \SE = 0.0115,
Number of Low-TP chunks		
lab-based	N = 7, \M = 0.143, \SE = 0.154, \p = 1	N = 12, \M = 0, \SE = 0, \p = NaN
online	N = 44, \M = 0.455, \SE = 0.111, \p = 0.000395	N = 44, \M = 0.0227, \SE = 0.023,
Number of Low-TP chunks among productions		
lab-based	N = 7, \M = 0.0714, \SE = 0.0772, \p = 1	N = 12, \M = 0, \SE = 0, \p = NaN
online	N = 44, \M = 0.211, \SE = 0.0506, \p = 0.000613	N = 44, \M = 0.0114, \SE = 0.0115

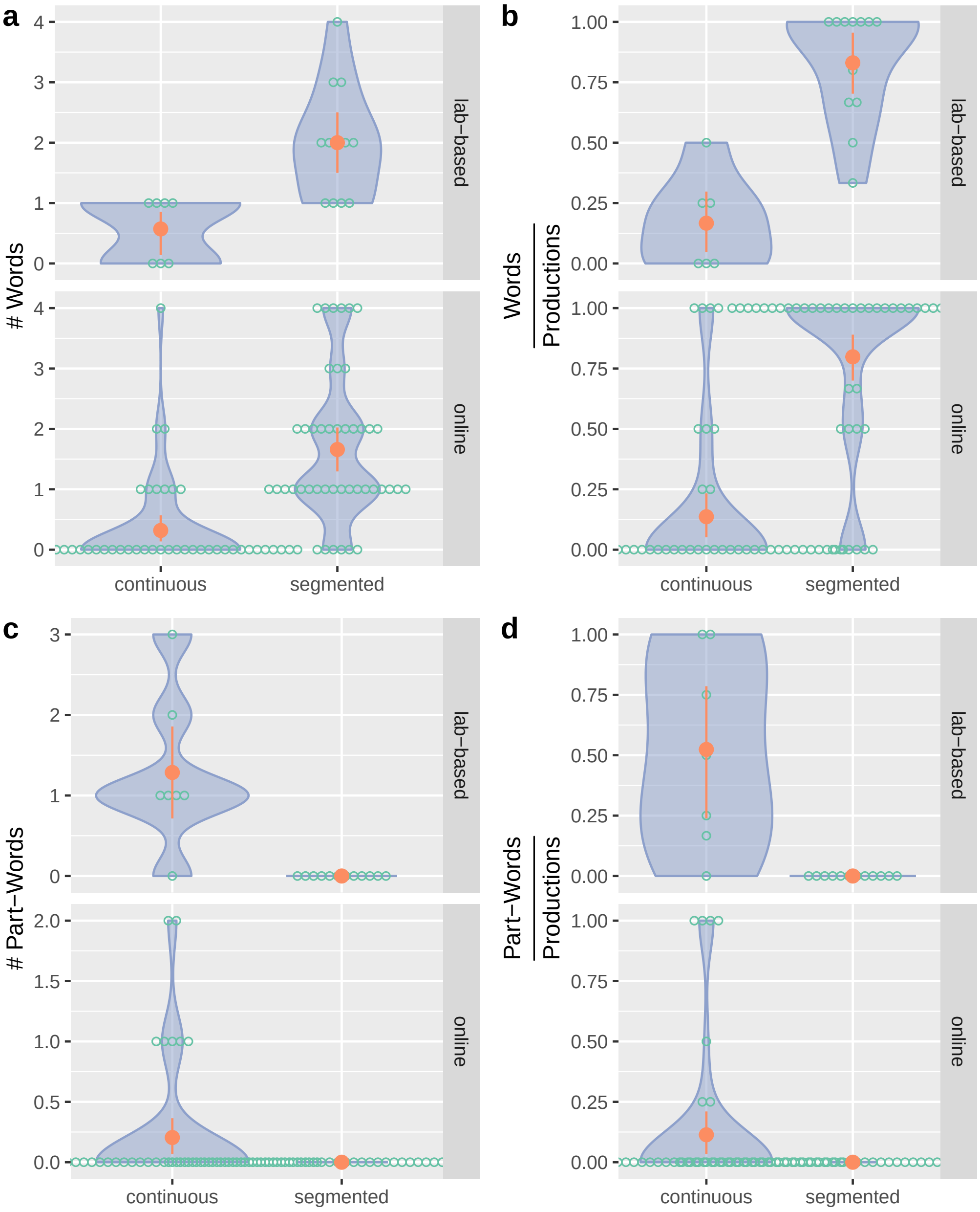

Figure 13: Number and proportion (among vocalizations) of words and part-words.

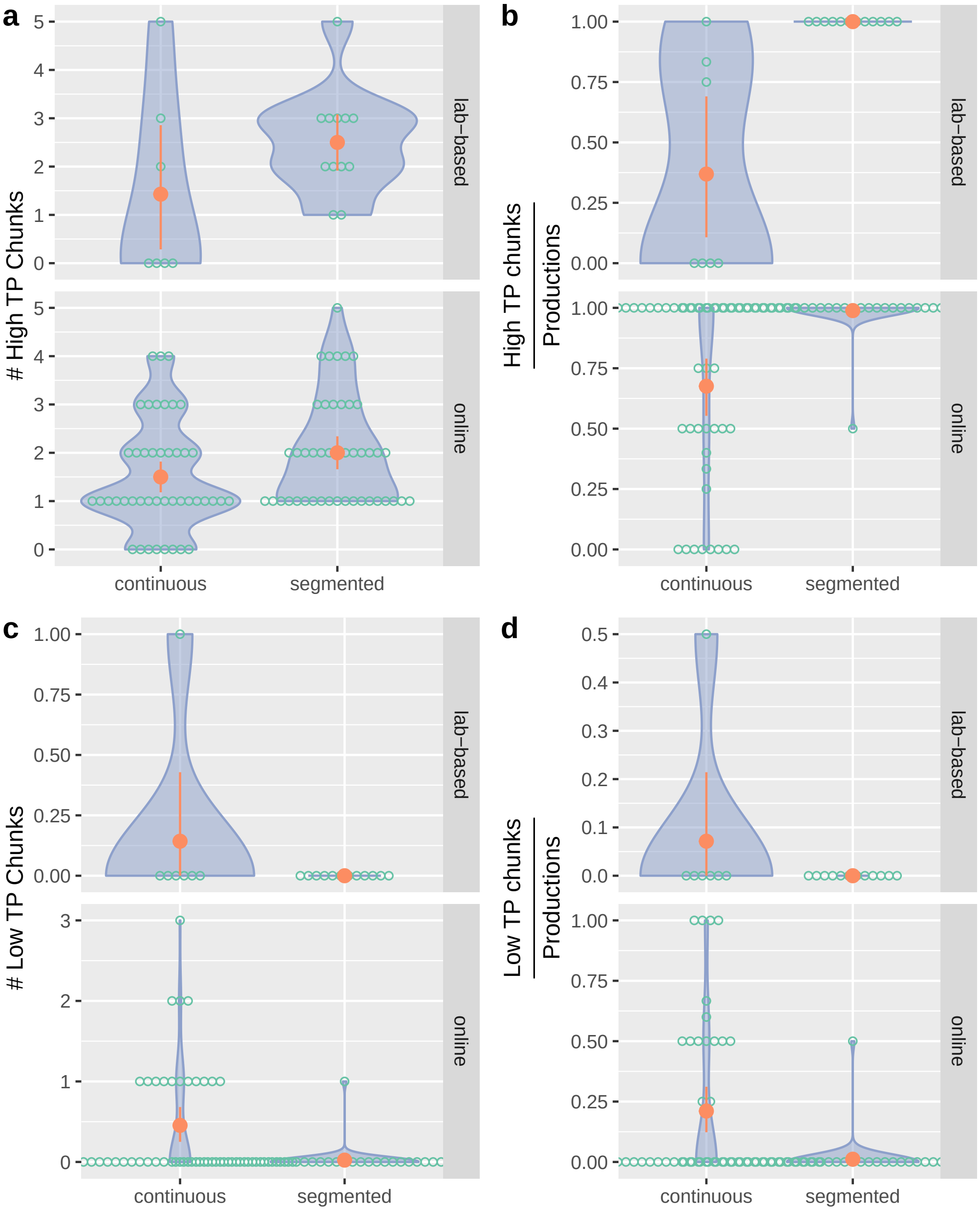

Figure 14: Plot of High and Low TP chunks.

10.1.2 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 $p^1_{\text{W}} = 1$, and the probability of using part-words is $p^1_{\text{PW}} = 1 - p^1_{\text{W}} = 0$. 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 $p^2_{\text{W}} = \frac{1}{3}$, and the probability of using part-words is $p^2_{\text{PW}} = 1 - p^2_{\text{W}} = \frac{2}{3}$.

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:

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

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

10.2 Experiments with the *en1* voice

10.2.1 Segmented stream, 3 repetitions of the stream, en1 voice

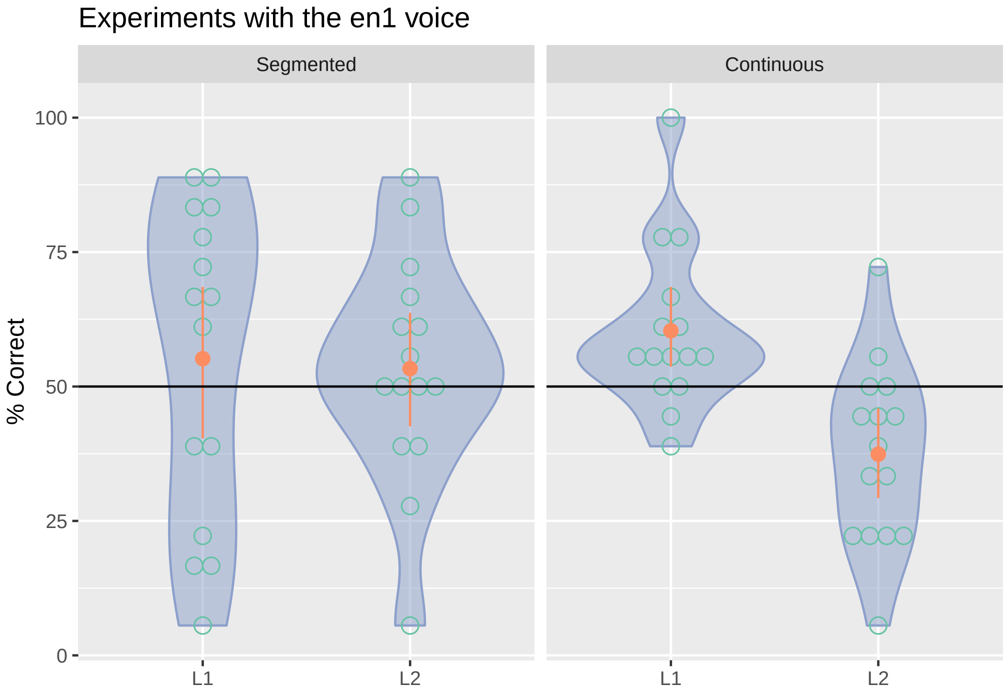

Figure 15: Results for a pre-segmented presentation of the stream (540 ms silences, left) and continuous presentation of the stream (right). Each word was repeated 45 times. The voice was *en1*.

As shown in Figure 15, the average performance did not differ significantly from the chance level of 50%, (M~= 54.26, SD~= 25.09), $t(29) = 0.93$, p~= 0.36, 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 15, performance did not depend on the language condition.

10.2.2 Continuous stream, 3 repetitions of the stream, en1 voice

xxx

As shown in Figure 15, 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 15, 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 might have prevent participants from using statistical learning.

Table 15: Performance differences across language conditions. 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

	Log-odds					Odd rat		
term	Estimate	SE	CI	t	p	Estimate	SE	CI

stats.3x.en.segm								
langL2	-0.097	0.441	[-0.96, 0.767]	-0.220	0.826	0.908	0.400	[0.383, 2
stats.3x.en.cont								
langL2	-1.024	0.410	[-1.83, -0.22]	-2.496	0.013	0.359	0.147	[0.161, 0
stats.3x.en.segm.cont								
langL2	-1.061	0.382	[-1.81, -0.313]	-2.779	0.005	0.346	0.132	[0.164, 0
experimentIDstats.3x.en.segm	-0.242	0.360	[-0.949, 0.464]	-0.673	0.501	0.785	0.283	[0.387, 1
langL2:experimentIDstats.3x.en.segm	0.967	0.508	[-0.0292, 1.96]	1.902	0.057	2.631	1.338	[0.971, 7

10.3 Pilot recognition experiment testing the use of chunk frequency

In Pilot Experiment 1, 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.

10.3.1 Methods

10.3.1.1 Participants

Table 16: Demographics of Pilot Experiment 1.

# 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 1 is given in Table 16. Participants were native speakers of Spanish and Catalan and were recruited from the Universitat Pompeu Fabra community.

10.3.2 Stimuli

Stimuli transcriptions are given in Table 17. They were synthesized using the *es2* (Spanish male) voice of the mbrola [**?**] speech synthesized, using a segment duration of 225 ms and an fundamental frequency of 120 Hz.

10.3.2.1 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/).

10.3.2.2 Familiarization The design of Pilot Experiment 1 is shown in Table 17. 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 3 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 17: Design of the Pilot 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	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

10.3.2.3 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 3 were combined into four test pairs, which were presented twice with different item orders. A new trial started 100 ms after a participant response.

10.3.3 Results

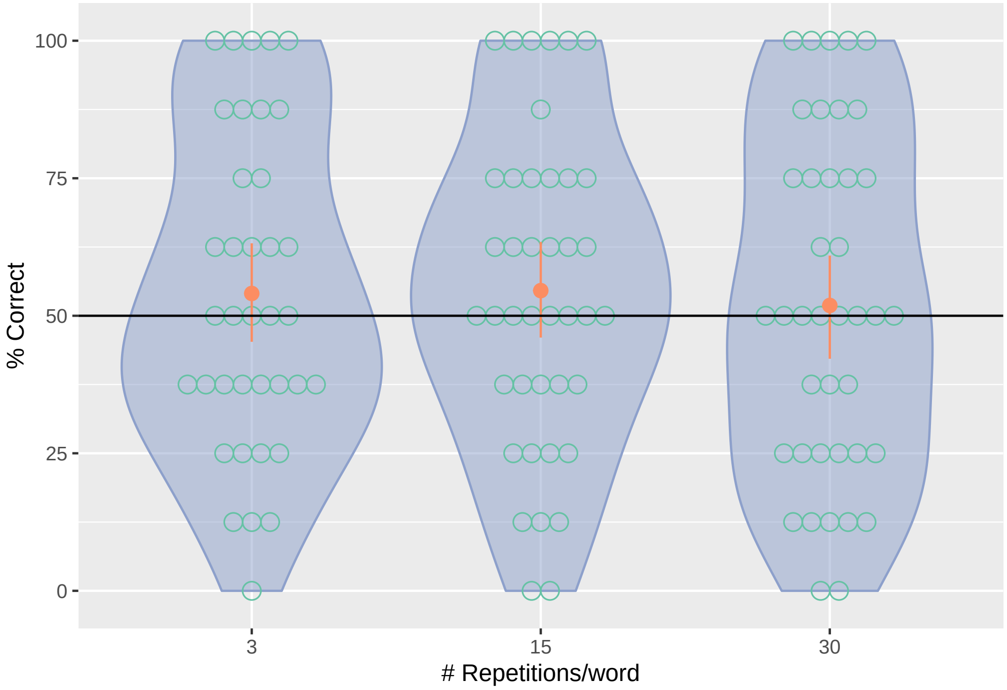

Figure 16: Results of Pilot Experiment 1. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right) 30 repetitions per word.

Table 18: Performance in Pilot Experiment 1 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
langL2	0.337	0.493	-0.629, 1.3	0.684	0.494
n.rep.word	0.017	0.018	-0.018, 0.0513	0.942	0.346
langL2:n.rep.word	-0.042	0.025	-0.0916, 0.00698	-1.682	0.093

Table 19: Performance in Pilot Experiment 1 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.

	Log-odds					Odd ratios				
term	Estimate	SE	CI	t	p	Estimate	SE	CI	t	
langL2	0.337	0.493	[-0.629, 1.3]	0.684	0.494	1.401	0.691	[0.533, 3.68]	0.684	0.4
n.rep.word	0.017	0.018	[-0.018, 0.0513]	0.942	0.346	1.017	0.018	[0.982, 1.05]	0.942	0.3
langL2:n.rep.word	-0.042	0.025	[-0.0916, 0.00698]	-1.682	0.093	0.959	0.024	[0.912, 1.01]	-1.682	0.0

As shown Table 18, a generalized linear model revealed that performance depended neither on the amount of familiarization nor on the familiarization language. As shown in Figure 16, 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 **?**, 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.

10.4 Analyses of Experiments 2 (Recognition) after removing outliers

Table 20: Descriptives for Experiment 1 (using the *us3* voice) and a pilot experiment (using the *en1* voice) after removing outliers. !!!!TO BE MOVED TO THE SI!!!!

experimentID	N	M	SE	p
us3				
Pre-segmented	29	0.533	0.024	0.151
Continuous (1)	32	0.585	0.029	0.018
Continuous (2)	30	0.628	0.040	0.007
en1				
Pre-segmented (en1)	30	0.543	0.047	0.268
Continuous (en1)	29	0.471	0.033	0.480

10.4.1 Can people recover words from pre-segmented prosodic units?

We repeat the analyses of Experiment 2 after removing outliers differing by more than 2.5 standard deviations from the mean in each condition ($N = 2$). As before, when the familiarization stream was pre-segmented and synthesized with the us3 voice, participants failed to split smaller utterances into their underlying components.

As shown in Figure 17, the average performance did not differ significantly from the chance level of 50%, (M~= 53.26, SD~= 12.64), $t(28) = 1.39$, p~= 0.176, Cohen's d~= 0.26, $CI_{.95}$~= 48.45, 58.07, ns, $V = 216$, $p = 0.151$. Likelihood ratio analysis favored the null hypothesis by a factor of 2.058 after correction with the Bayesian Information Criterion. As shown in Table 21, performance did not depend on the language condition.

The failure to use statistical learning was also replicated using a second voice (*en1*, British English male). As shown in Figure 17, the average performance did not differ significantly from the chance level of 50%, (M~= 54.26, SD~= 25.09), $t(29) = 0.93$, p~= 0.36, Cohen's d~= 0.17, $CI_{.95}$~= 44.89, 63.63, ns, $V = 222$, $p = 0.242$. 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 21, performance did not depend on the language condition.

10.4.2 Can people recover words from a continuous stream? (1)

We next 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 17, when the *us3* voice was used, the average performance differed significantly from the chance level of 50%, (M~= 58.51, SD~= 16.21), $t(31) = 2.97$, p~= 0.00573, Cohen's d~= 0.52, $CI_{.95}$~= 52.66, 64.35, $V = 306.5$, $p = 0.0185$. As shown in Table 21, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition

10.4.3 Can people recover words from a continuous stream? (2) (Replication)

Given the unexpected results with the *en1* voice below, we replicated the successful tracking of statistical information using a new sample of participants.

As shown in Figure 17, the average performance differed significantly from the chance level of 50%, (M~= 62.78, SD~= 21.35), $t(29) = 3.28$, p~= 0.00272, Cohen's d~= 0.6, $CI_{.95}$~= 54.81, 70.75, $V = 320$, $p = 0.00778$. As shown in Table 21, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition.

10.4.4 Replication with a different voice (en1)

However, when the *en1* voice was used, Figure 17 shows that the average performance did not differ significantly from the chance level of 50%, (M~= 47.13, SD~= 17.42), $t(28)$ = -0.89, p~= 0.382, Cohen's d~= 0.16, $CI_{.95}$~= 40.5, 53.75, ns, V = 140, p = 0.551. Likelihood analyses revealed that the null hypothesis was 3.629 than the alternative hypothesis after a correction with the Bayesian Information Criterion. However, as shown in Table 21, 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 might have prevent participants from using statistical learning.

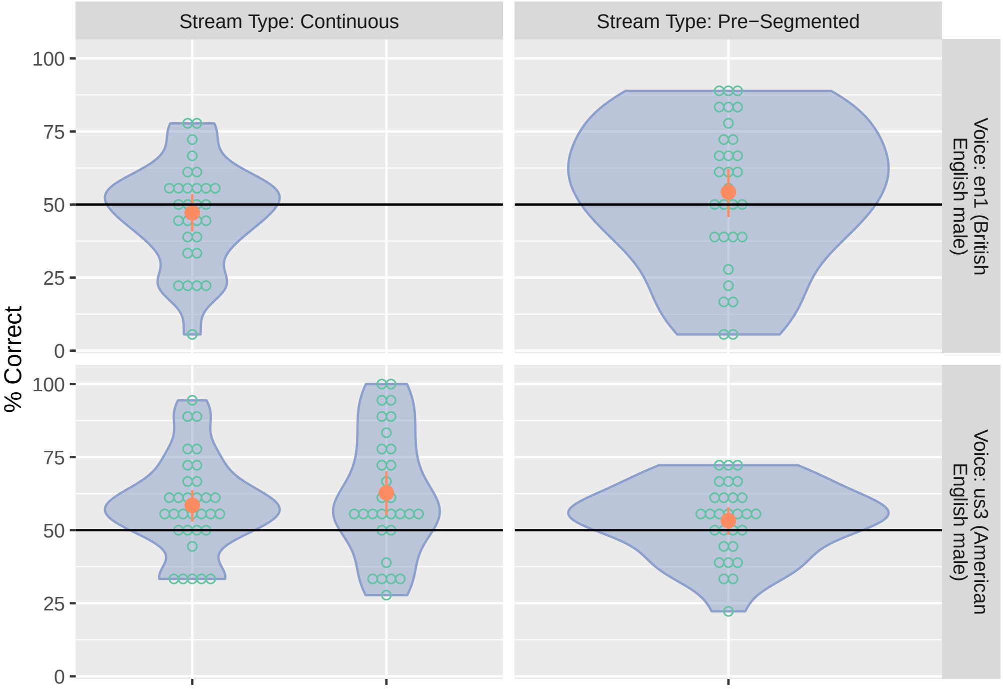

Figure 17: Results of Experiment 1 after outliers of more than 2.5 standard deviations from each condition mean were excluded. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) continuous familiarization stream or (right) a pre-segmented familiarization stream, synthesized with an American English voice (top) or a British English voice (bottom). The two continuous conditions are replictions of one another.

Table 21: Performance differences across familiarization conditions in Experiment 2 after removal of outliers differing more thang 2.5 standard deviations from the mean. 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.

		Log-odds			Odd	
term	Voice	Estimate	SE	CI	Estimate	S
Pre-segmented familiarization (us3)						
langL2	American English (us3)	-0.048	0.654	[-1.33, 1.23]	0.953	0.62
Continuous familiarization (us3) (1)						

langL2	American English (us3)	-0.184	0.480	[-1.12, 0.757]	0.832	0.40
Continuous familiarization (us3) (2)						
langL2	American English (us3)	0.317	0.786	[-1.22, 1.86]	1.372	1.07
Pre-segmented vs. continuous familiarization (us3) (1)						
langL2	American English (us3)	-0.102	0.551	[-1.18, 0.978]	0.903	0.49
segmsegmented	American English (us3)	-0.243	0.167	[-0.571, 0.0843]	0.784	0.13
Pre-segmented vs. continuous familiarization (us3) (2)						
langL2	American English (us3)	0.115	0.652	[-1.16, 1.39]	1.122	0.73
segmsegmented	American English (us3)	-0.509	0.224	[-0.949, -0.0693]	0.601	0.13
Pre-segmented familiarization (en1)						
langL2	British English (en1)	-0.097	0.441	[-0.96, 0.767]	0.908	0.40
Continuous familiarization (en1)						
langL2	British English (en1)	-0.842	0.221	[-1.28, -0.409]	0.431	0.09
Pre-segmented vs. continuous familiarization (en1)						
langL2	British English (en1)	-0.903	0.369	[-1.63, -0.179]	0.406	0.15
experimentIDstats.3x.en.segm	British English (en1)	-0.090	0.347	[-0.77, 0.591]	0.914	0.31
langL2:experimentIDstats.3x.en.segm	British English (en1)	0.810	0.487	[-0.144, 1.76]	2.248	1.09

```
# library(renv)
# renv::snapshot()
```