PNAS
www.pnas.org

Supplementary Information for

The specificity of Statistical Learning

Ansgar D. Endress & Maureen de Seyssel

Corresponding Author name.
E-mail: ansgar.endressm4x.org

This PDF file includes:

Supplementary text
Figs. S1 to S4 (not allowed for Brief Reports)
Tables S1 to S13 (not allowed for Brief Reports)
Legend for Dataset S1
SI References

Other supplementary materials for this manuscript include the following:

Dataset S1

Supporting Information Text

Methods

Recognition experiment.

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

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

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

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

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

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

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

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

Stimuli. Stimuli were synthesized using the *us3* (American English male) voice from mbrola (1). (We also used the *en1* (British English male) voice; however, as discussed below, this voice turned out to be of relatively low quality and introduced confounds in the data.)

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

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

For segmented streams, words were individually synthesized using mbrola. We then used a custom-made Perl script to randomize the words for each participant and concatenate them into a familiarization file using sox. The order of words was then randomized for each participant and concatenated into a single aiff file using sox. The silence among words was 540 ms

*http://sox.sourceforge.net/

†https://ffmpeg.org/

(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).

Apparatus. The experiment was run using Psyscope X[‡]. Stimuli were presented over headphones in a quiet room. Responses were collected from pre-marked keys on the keyboard.

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 item comprised the last two syllables of a word, and was a correct choice for Language 2. There were three items of each kind. They were combined into 9 test pairs. The test pairs were presented twice, with different item orders, for a total of 18 test trials.

Recall experiment.

Materials. We re-synthesized the languages used in (2) Experiment 2. The four words in each language are given in Table S3. Each word was composed for three syllables, which were composed of two segments in turn. Stimuli were synthesized using the us3 (male American English) voice of the mbrola synthesizer (1), at a constant F_0 of 120 Hz and at a rate of 216 ms per syllable (108 ms per phoneme).

Table S3. Languages used Experiment 2. The words are the same as in Experiment 2 in (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 words could not occur in immediate repetition. Each randomization was then (i) synthesized into a continuous speech stream using mbrola and then converted to mp3 using ffmpeg or (ii) used to concatenate words that had been synthesized in isolation, separated by silences of 222 ms into a segmented speech stream, which was then converted to mp3. Streams were faded in and out for 5 s using sox. For continuous streams, this yielded a stream duration of 1 min 57 s; for segmented streams, the duration was 2 min 37.

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

Procedure.

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

For the online experiments, participants watched a video with no clear objects during the familiarization.[§] 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.

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.

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.

[‡] http://psy.ck.sissa.it

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

1. Analysis

Recognition tests. Accuracy was averaged for each participant, and the scores were tested against the chance level of 50% using Wilcoxon tests. Performance differences across the languages (Language 1 vs. 2) and, when applicable, familiarization conditions (pre-segmented vs. continuous) were assessed using a generalized linear mixed model for the trial-by-trial data with the fixed factors language and, where applicable, familiarization condition, as well as random slopes for participants, correct items and foils. Following (3), 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 (4), 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.

Recall test.

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

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

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

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

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

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

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

The criteria were applied in this order.

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

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

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

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

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

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

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

Table S5. Analyses performed for the vocalizations

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

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

Measures For each underlying form, we calculate:

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

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

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.

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

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

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

Additional results

Experiment 1.

Experiment 2.

Additional tables and figures.

Fit of the number of participants producing words or part-words to a binomial distribution. We fit the data to two models, one where the learner successfully detected word-boundaries, and one where the learner successfully track TPs but initiates productions at a random position. We then calculate the likelihood of the data given these models.

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

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

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

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

Pilot Experiment 1: Using the *en1* voice

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

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

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

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

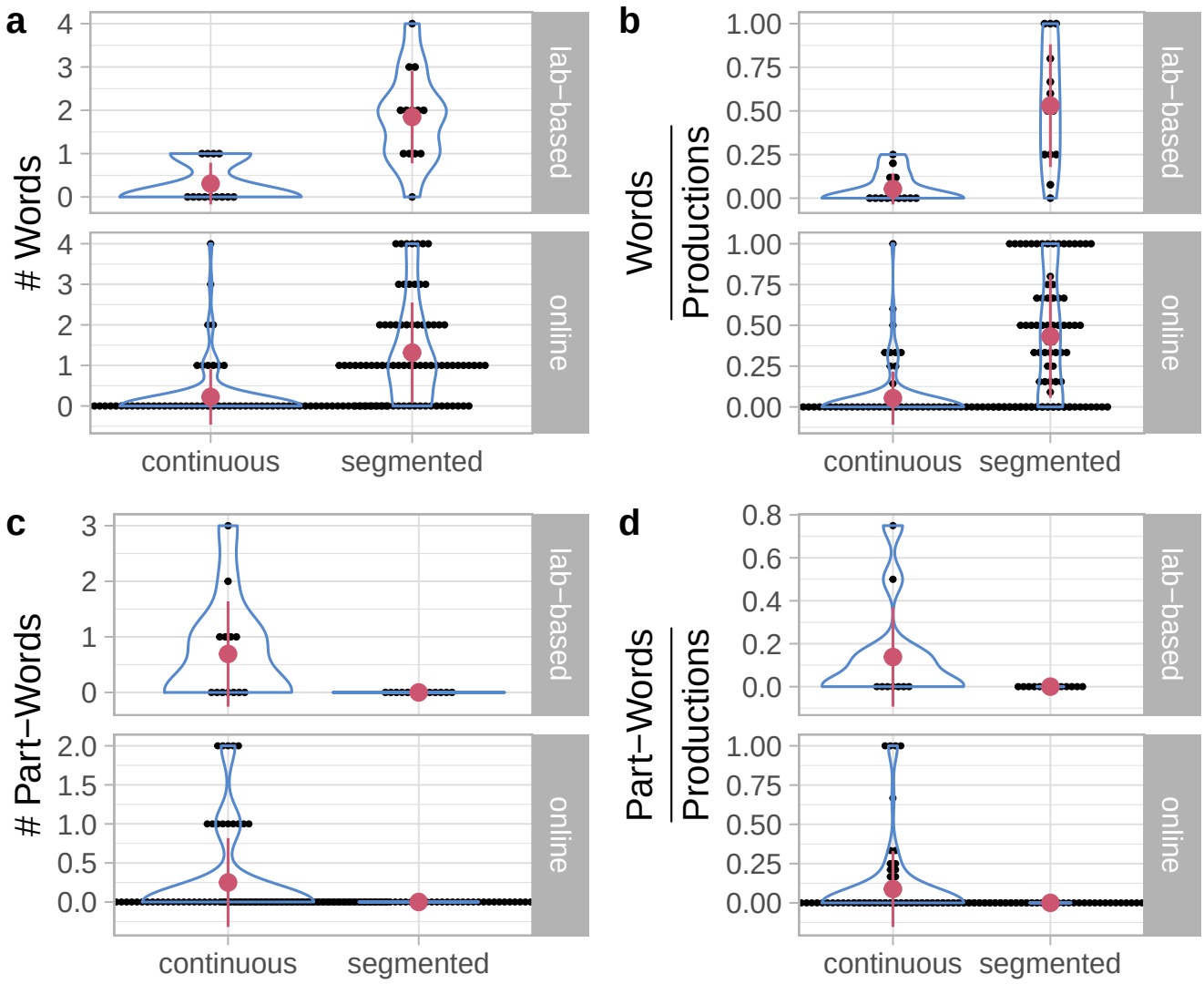

Fig. S1. Number and proportion (among vocalizations) of words and part-words.

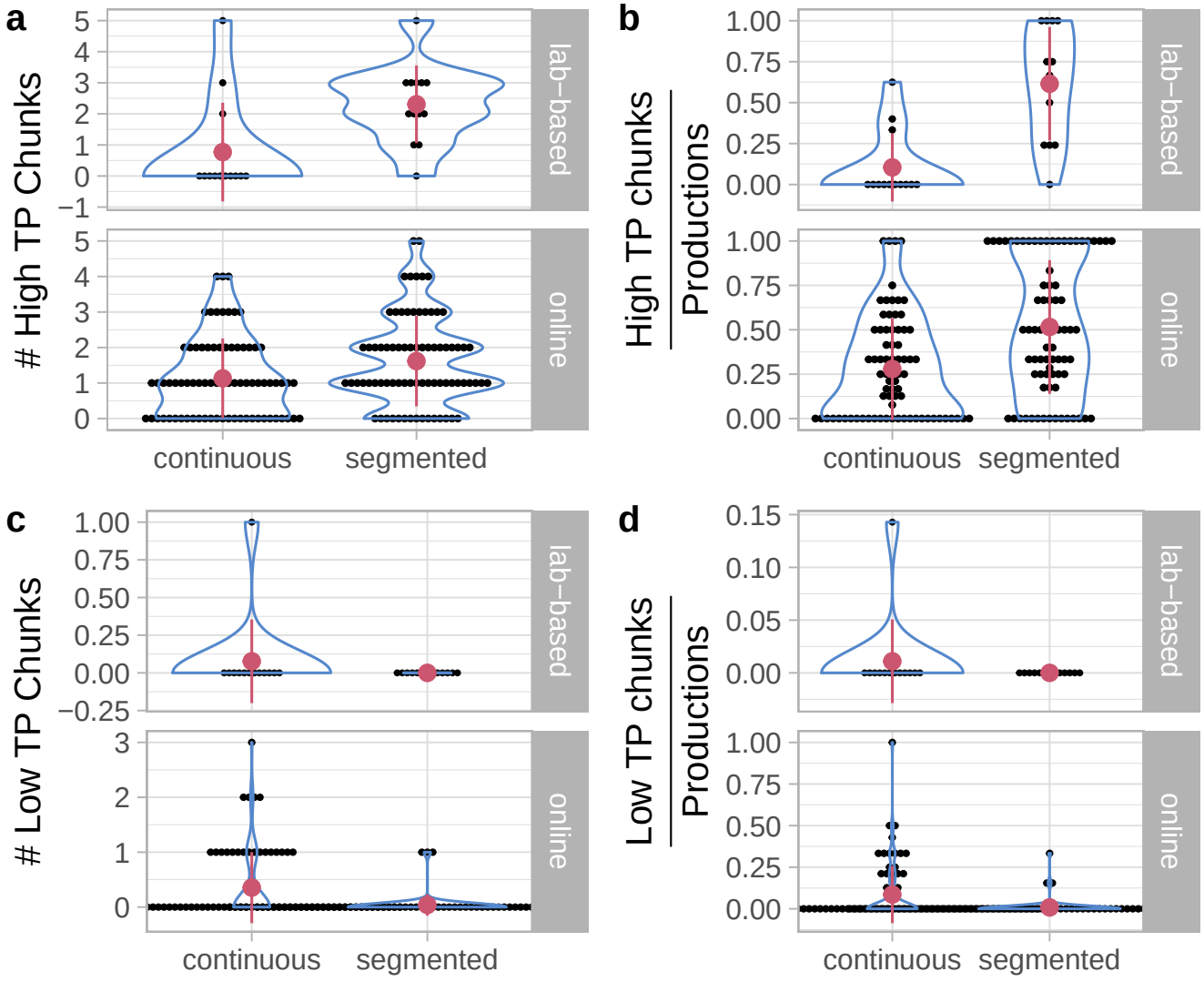

Fig. S2. Plot of High and Low TP chunks.

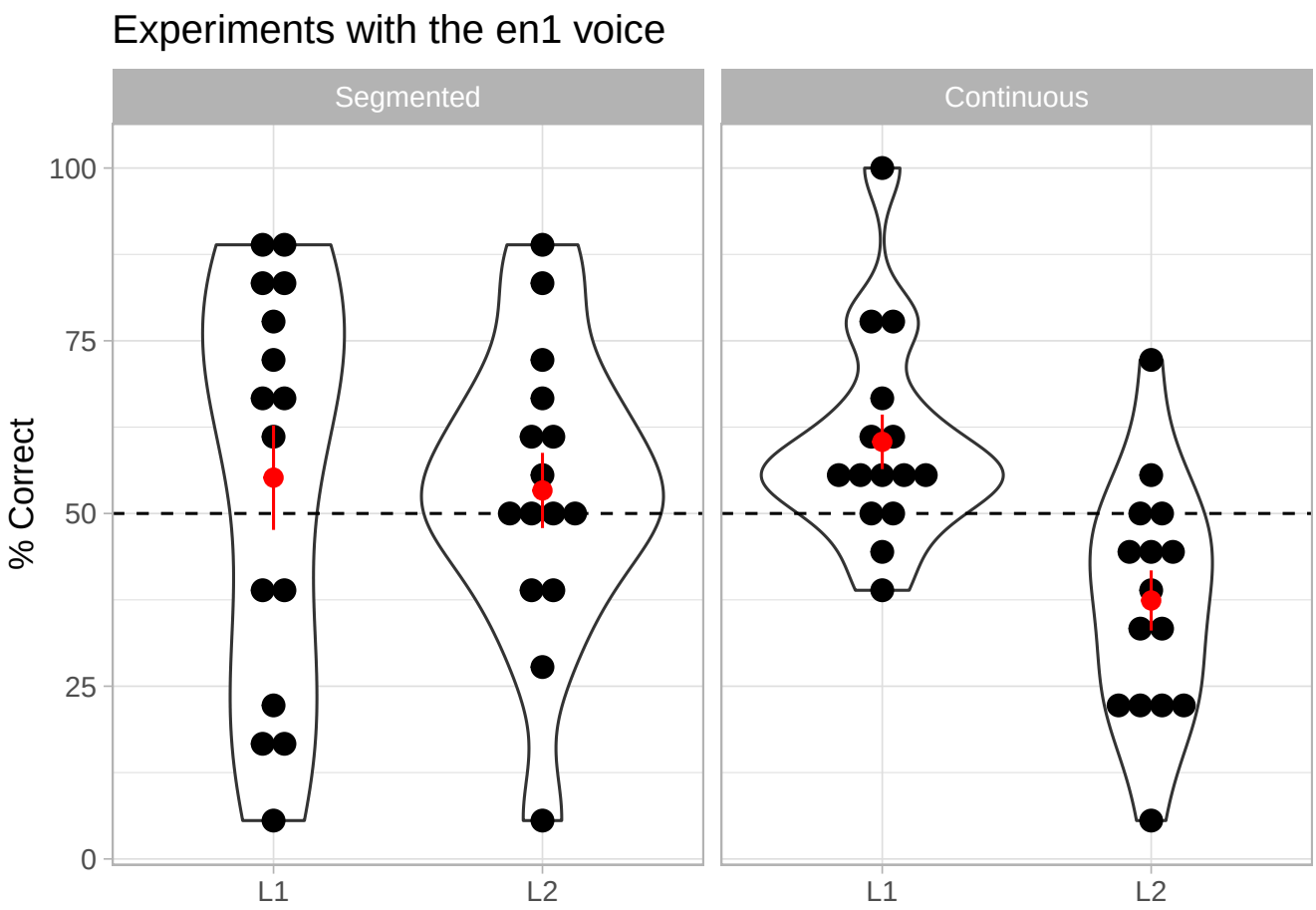

Fig. S3. Results for a pre-segmented presentation of the stream (540 ms silences, left) and continuous presentation of the stream (right). Each word was repeated 45 times. The voice was *en1*.

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

Pilot Experiment 2: Testing the use of chunk frequency

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

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

Methods.

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

Stimuli. Stimuli transcriptions are given in Table S12. They were synthesized using the *es2* (Spanish male) voice of the mbrola (1) speech synthesized, using a segment duration of 225 ms and an fundamental frequency of 120 Hz.

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/).

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

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

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

Results. As shown Table S13, a generalized linear model revealed that performance depended neither on the amount of familiarization nor on the familiarization language. As shown in Figure S4, a Wilcoxon test did not detect any deviation from the chance level of 50%, neither for all amounts of familiarization combined, M= 53.5, SE= 2.71, p= 0.182, nor for the individual familiarization conditions (3 repetitions per word: M= 54.1, SE= 4.81, p= 0.416; 15 repetitions per word: M= 54.6, SE= 4.52, p= 0.325; 30 repetitions per word: M= 51.9, SE= 4.98, p= 0.63). Following **(author?)** (4), 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.

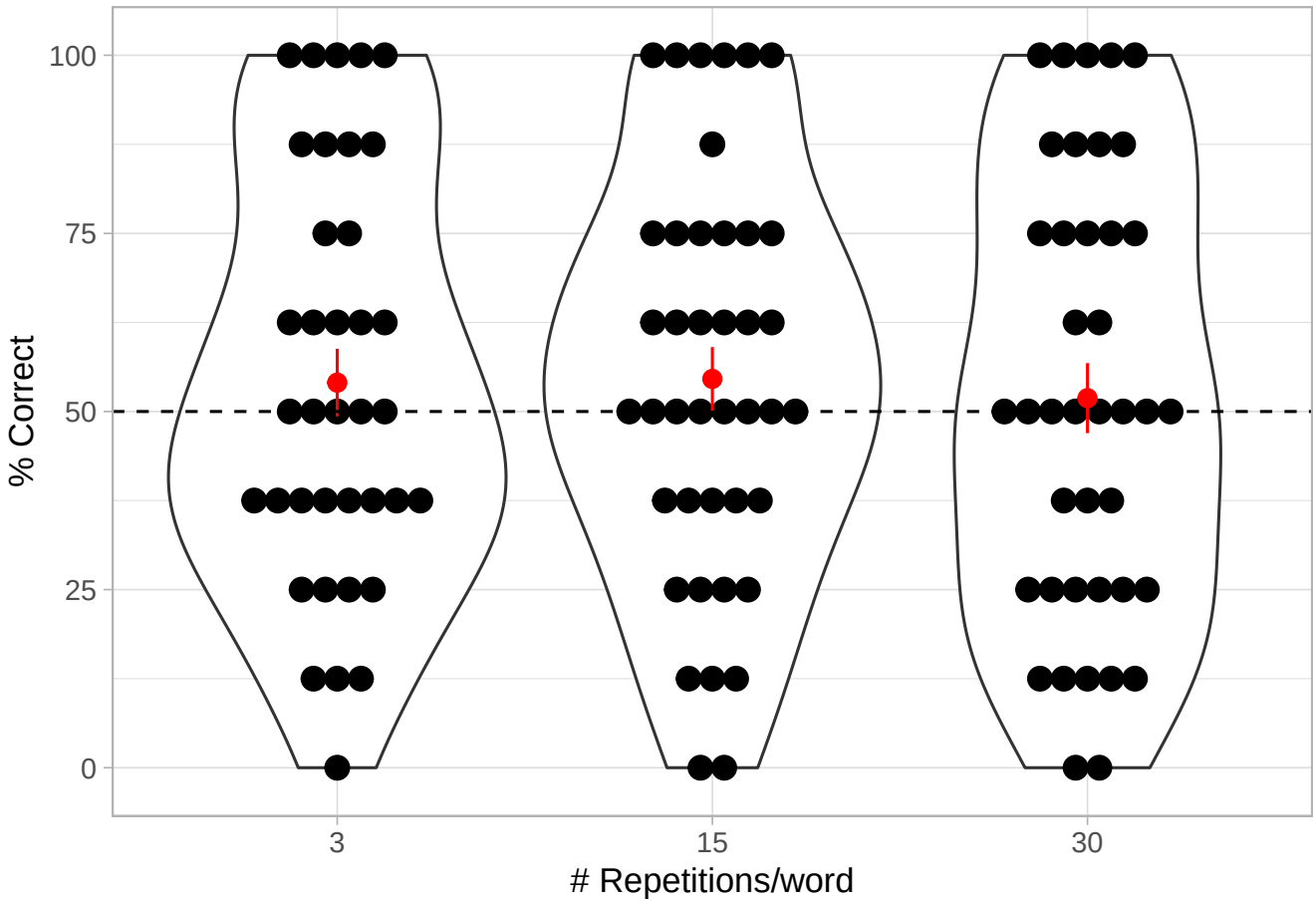

Fig. S4. Results of Pilot Experiment 2. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right)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

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

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

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

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

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

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

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

Table S11. Demographics of Pilot Experiment 2.

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

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

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

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

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

SI Dataset S1 (segmentation_recall_unattested.xlsx)
Type or paste legend here.

References

1. T Dutoit, V Pagel, N Pierret, F Bataille, O van der Vreken, The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes in *Proceedings of the Fourth International Conference on Spoken Language Processing*. (Philadelphia), Vol. 3, pp. 1393–1396 (1996).
2. JR Saffran, RN Aslin, EL Newport, Statistical learning by 8-month-old infants. *Science* **274**, 1926–8 (1996).
3. RH Baayen, D Davidson, D Bates, Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390 – 412 (2008).
4. S Glover, P Dixon, Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychon Bull Rev* **11**, 791–806 (2004).