Supplementary Online Materials

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

SOM1 Measures and column names in the supplementary data file for Experiment 1

Table S1
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.

SOM2 Additional results for Experiment 1

Table S2

Supplementary analyses pertaining to the productions as well as test against their chances levels in the recall phase of Experiments 1a and 1b. The p value in the rightmost column reflects a Wilcoxon test comparing the continuous and the pre-segmented conditions.

	Continuous	Segmented	p(Continuous vs. Segmented).
Number of words			
lab-based (Exp. 1a)	M= 0.308, SE= 0.139, p= 0.0719	M= 1.85, SE= 0.308, p= 0.00224	0.005
online (Exp. 1b)	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 (Exp. 1a)	M= 0.308, SE= 0.139, p= 0.0719	M= 1.85, SE= 0.308, p= 0.00224	0.005
online (Exp. 1b)	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 (Exp. 1a)	M= 0.692, SE= 0.273, p= 0.031	M= 0, SE= 0, p= NaN	0.031
online (Exp. 1b)	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 (Exp. 1a)	M= 0.692, SE= 0.273, p= 0.031	M= 0, SE= 0, p= NaN	0.031
online (Exp. 1b)	M= 0.25, SE= 0.0657, p= 0.000717	M= 0, SE= 0, p= NaN	< 0.001
Actual vs. expected forward TPs			
lab-based (Exp. 1a)	M= -0.462, SE= 0.07, p= 0.000244	M= -0.315, SE= 0.0803, p= 0.00915	0.147
online (Exp. 1b)	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 (Exp. 1a)	M= 0.769, SE= 0.459, p= 0.181	M= 2.31, SE= 0.361, p= 0.00224	0.022
online (Exp. 1b)	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 (Exp. 1a)	M= 0.104, SE= 0.0601, p= 0.181	M= 0.615, SE= 0.0999, p= 0.00241	0.003
online (Exp. 1b)	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 (Exp. 1a)	M= 0.0769, SE= 0.0801, p= $>$ .999	M= 0, SE= 0, p= NaN	$>$ .999
online (Exp. 1b)	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 (Exp. 1a)	M= 0.011, SE= 0.0114, p= $>$ .999	M= 0, SE= 0, p= NaN	$>$ .999
online (Exp. 1b)	M= 0.0855, SE= 0.0198, p= 6.04e-05	M= 0.00846, SE= 0.00523, p= 0.181	< 0.001

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

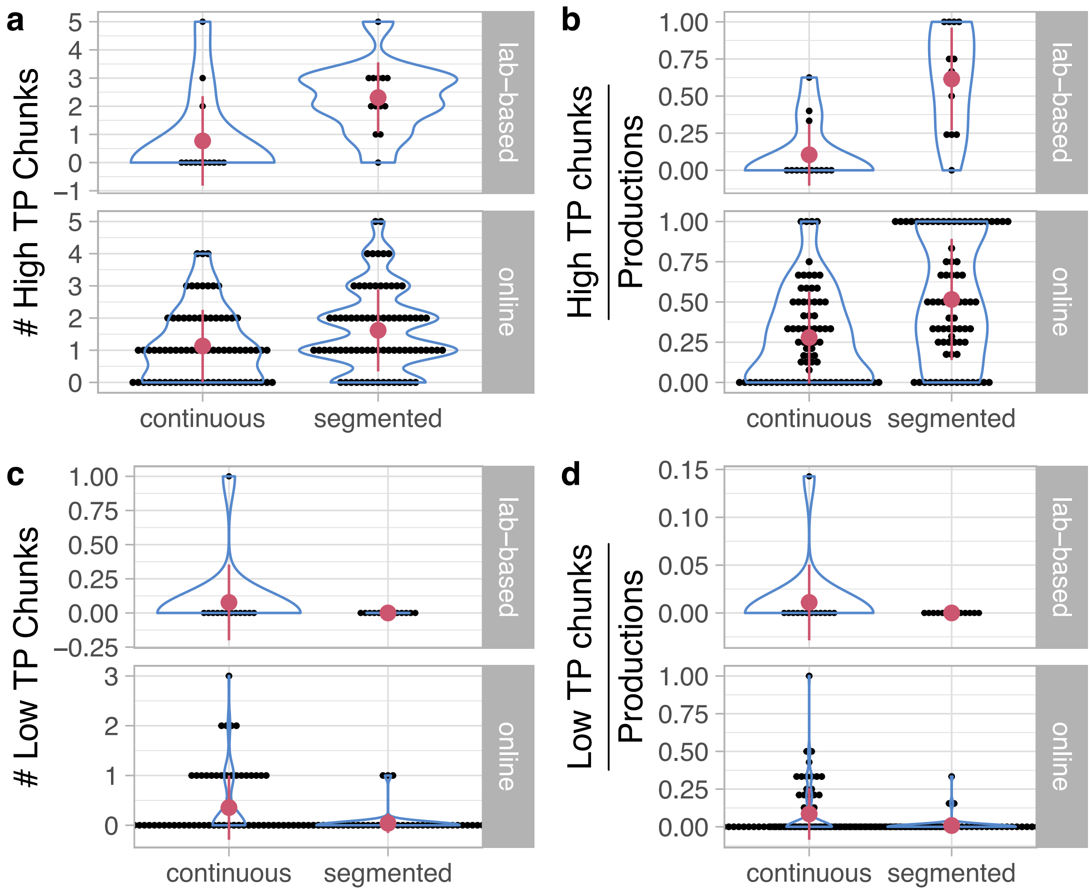

Figure S1. Plot of High and Low TP chunks.

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

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

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

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

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

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

SOM4 Analyses of Experiment 2 after removing outliers

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 in the main analyses above, we first present the results for the British English (en1) voice and then those for the American English (us3) voice.

SOM4.1 Experiment 2a (British English voice)

Figure S2 shows the results for the pre-segmented familiarization. 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 S3, performance did not depend on the language condition.

We next asked if, in line with previous research, they can track TPs units that 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.

Figure S2 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 S3, 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 sounds might have prevented participants from using statistical learning. We thus decided to replicate the results with a different, American English voice.

SOM4.1.1 Experiment 2b (American English voice). Figure S2 shows the results for the pre-segmented condition with the American English (us3) voice. 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 S3, performance did not depend on the language condition.

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 S2, 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 S3, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition.

Given the unexpected results with the *en1* voice above, we replicated the successful tracking of statistical information using a new sample of participants. As shown in Figure S2, 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 S3, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition.

The results obtained after removing outliers are thus similar to those reported in the main text.

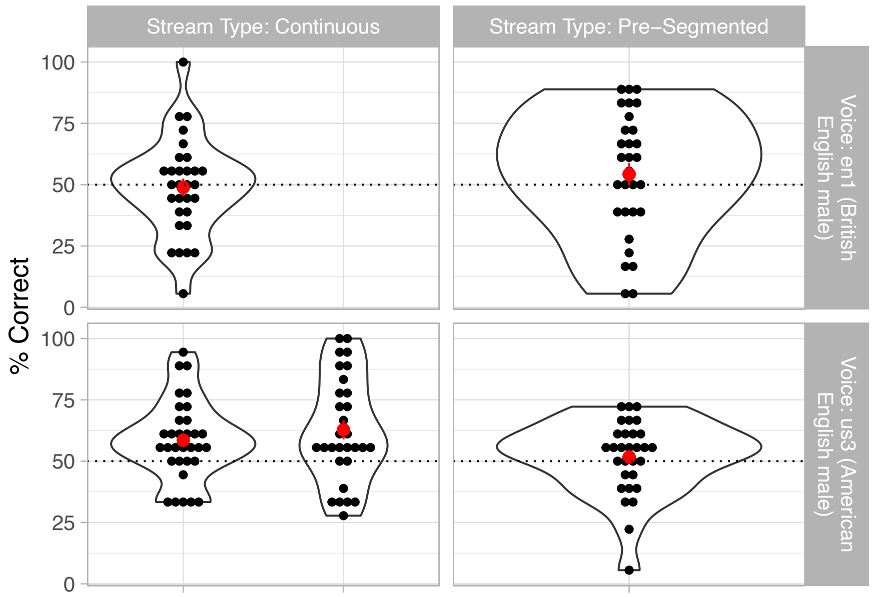

Figure S2. Results of Experiment 1 after outliers of more than 2.5 standard deviations from each condition mean were excluded. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) continuous familiarization stream or (right) a pre-segmented familiarization stream, synthesized with a British English voice (top) or an American English voice (bottom). The two continuous conditions are replictions of one another.

Table S3

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 ratios				
term	Voice	Estimate	SE	CI	Estimate	SE	CI	t	p
Pre-segmented familiarization, British English voice (Exp. 2a)									
language = L2	en1	-0.097	0.441	[-0.96, 0.767]	0.908	0.400	[0.383, 2.15]	-0.220	0.826
Continuous familiarization, British English voice (Exp. 2a)									
language = L2	en1	-0.842	0.221	[-1.28, -0.409]	0.431	0.095	[0.279, 0.665]	-3.807	0.000
Pre-segmented vs. continuous familiarization, British English voice (Exp. 2a)									
language = L2	en1	-0.903	0.369	[-1.63, -0.179]	0.406	0.150	[0.197, 0.836]	-2.446	0.014
stream type = segmented	en1	-0.090	0.347	[-0.77, 0.591]	0.914	0.317	[0.463, 1.81]	-0.258	0.796
language = L2 × stream type = segmented	en1	0.810	0.487	[-0.144, 1.76]	2.248	1.094	[0.866, 5.84]	1.664	0.096
Pre-segmented familiarization, American English voice (Exp. 2b)									
language = L2	us3	-0.048	0.654	[-1.33, 1.23]	0.953	0.624	[0.264, 3.44]	-0.074	0.941
Continuous familiarization (1), American English voice (Exp. 2b)									
language = L2	us3	-0.184	0.480	[-1.12, 0.757]	0.832	0.400	[0.325, 2.13]	-0.383	0.702
Continuous familiarization (2), American English voice (Exp. 2b)									
language = L2	us3	0.317	0.786	[-1.22, 1.86]	1.372	1.079	[0.294, 6.4]	0.403	0.687
Pre-segmented vs. continuous familiarization (1), American English voice (Exp. 2b)									
language = L2	us3	-0.102	0.551	[-1.18, 0.978]	0.903	0.497	[0.307, 2.66]	-0.185	0.853
stream type = segmented	us3	-0.243	0.167	[-0.571, 0.0843]	0.784	0.131	[0.565, 1.09]	-1.456	0.145
Pre-segmented vs. continuous familiarization (2), American English voice (Exp. 2b)									
language = L2	us3	0.115	0.652	[-1.16, 1.39]	1.122	0.732	[0.313, 4.03]	0.177	0.859
stream type = segmented	us3	-0.509	0.224	[-0.949, -0.0693]	0.601	0.135	[0.387, 0.933]	-2.269	0.023

SOM5 Pilot Experiment: Testing the use of chunk frequency

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

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

SOM5.1 Methods

Table S4
Demographics of the final sample in the pilot experiment.

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

SOM5.1.1 Participants. Demographic information of the pilot experiment is given in Table S4. Participants were native speakers of Spanish and Catalan and were recruited from the Universitat Pompeu Fabra community.

SOM5.1.2 Stimuli. Stimuli transcriptions are given in Table S5. They were synthesized using the *es2* (Spanish male) voice of the mbrola (Dutoit et al., 1996) speech synthesized, using a segment duration of 225 ms and an fundamental frequency of 120 Hz.

SOM5.1.3 Apparatus. Participants were test individually in a quiet room. Stimuli were presented over headphones. Responses were collected from pre-marked keys on the keyboard. The experiment with 3 repetitions per word (see below) were run using PsyScope X; the other experiments were run using Experyment (`https://www.expyriment.org/`).

SOM5.1.4 Familiarization. The design of the pilot experiment is shown in Table S5. 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 5 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 S5
Design of the pilot experiment. (Left) Language structure. (Middle) Structure of test items. Correct items for Language 1 are foils for Language 2 and vice versa. (Right) Actual items in SAMPA format; dashes indicate syllable boundaries

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

SOM5.1.5 Test. Following this familiarization, participants were informed that they would hear new items, and had to decide which of them was in Martian. Following this, they heard pairs of two syllabic items with an ISI of 1000 ms. One was a word-initial chunk and one a word-final chunk.

The test items shown in Table 5 were combined into four test pairs, which were presented twice with different item orders. A new trial started 100 ms after a participant response.

SOM5.2 Results

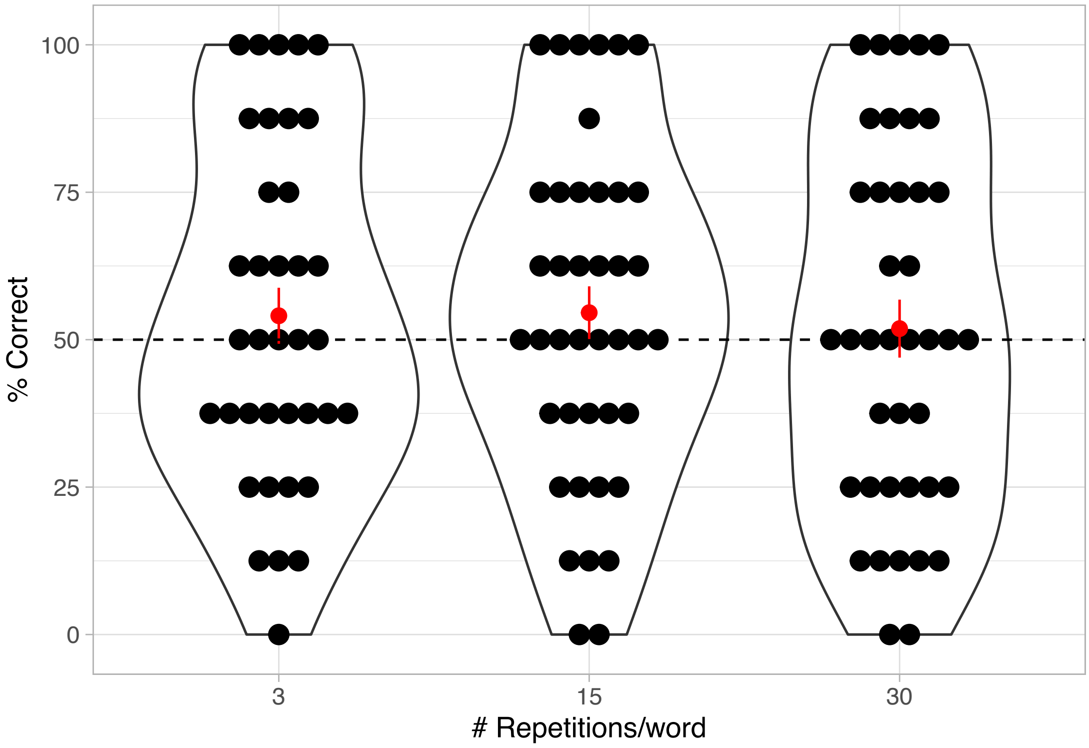

Figure S3. Results of the pilot experiment. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right) 30 repetitions per word.

As shown Table S6, a generalized linear model revealed that performance depended neither on the amount of familiarization nor on the familiarization language. As shown in Figure S3, 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

Table S6

Performance in the pilot experiment 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					Odds ratios				
term	Estimate	SE	CI	t	p	Estimate	SE	CI	t	p
language = L2	0.337	0.493	[-0.629, 1.3]	0.684	0.494	1.401	0.691	[0.533, 3.68]	0.684	0.494
number of repetitions/word	0.017	0.018	[-0.018, 0.0513]	0.942	0.346	1.017	0.018	[0.982, 1.05]	0.942	0.346
language = L2 × number of repetitions/word	-0.042	0.025	[-0.0916, 0.00698]	-1.682	0.093	0.959	0.024	[0.912, 1.01]	-1.682	0.093

conditions (3 repetitions per word: $M = 54.1$, $SE = 4.81$, $p = 0.416$; 15 repetitions per word: $M = 54.6$, $SE = 4.52$, $p = 0.325$; 30 repetitions per word: $M = 51.9$, $SE = 4.98$, $p = 0.63$). Following Glover and Dixon (2004), the null hypothesis was 4.696 times more likely than the alternative hypothesis after corrections with the Bayesian Information Criterion, and 1.217 more likely after correction with the Akaike Information Criterion.