The specificity of statistical learning

Ansgar Endress

Abstract

Statistical Learning exists in many domains and species. It might be particularly crucial for the earliest stages of word learning, for example for recovering word-like chunks from fluent speech. However, other forms of associative learning are remarkably tuned to the ecological constraints of the learning situation. Here, we show that Statistical Learning is similarly constrained. It predominantly operates in continuous speech sequences similar to those used in prior experiments, but not in discrete chunk sequences similar to those likely encountered during language acquisition (due to the prosodic organization of language). Conversely, when exposed to continuous sequence in a memory recall experiment, participants tend to produce low-probability sequences because, to the extent that they remember any items at all, they initiate their productions at random positions in the sequence rather than at the onsets of statistically defined chunks. In contrast, familiarization with discrete sequences produces reliable memories of actual, high-probability forms. This dissociation between Statistical Learning and memory suggests that Statistical Learning might have a specialized role when distributional information can be accumulated (e.g., for predictive processing), and that it is separable from the (declarative) memory mechanisms needed to acquire words.

Contents

```
# Extract R file to accelarate segmentation

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

1 House keeping

In the analyses below, we use the following parameters:

Name	Value
ANALYZED.DATA.SETS	[illegible]
ANALYZED.DATA.SETS	[illegible]
FILTER.SINGLE.SYLLABLES	[illegible]
FILTER.UNATTESTED.ITEMS	[illegible]
IGNORE.COL.PREFIXES	[illegible]
IGNORE.COL.PREFIXES	[illegible]
IGNORE.COL.PREFIXES	[illegible]
L.BALLIST	list(subj = "399612_200413_124119_M056106.csv", response = "dalonigtbdophophi dalobdakabdarobigopachu"), c(subj = "399612_210517_101428_M070415.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_M038010.csv", response = "takahsakakakaratatataikokokokotatakatakatakatakatakatakataka"), c(subj = " 399612_210517_101201_M048600.csv", response ="dabroobitalooki,bkuti2,golab"), c(subj ="399612_210524_062929_M059506.csv",

Name	Value
response = “matiku- latat- itu- la- pa- pit- u- lari- mat- itu- laat- it- ula”), c(subj = “399612_210524_115845_M067482.csv”, re- sponse = “tu kalla ti palla tuti kulla papi pu tu kalla ti palla tuti kulla papi pu”), c(subj = “399612_210524_120014_M099076.csv”, re- sponse = “tu- to- pit- u- lakatu- to- pi- toolaka”), c(subj = “399612_210524_120523_M003515.csv”, re- sponse = “papikuchi,butalapapikuchi,kukala,pikala,budharapikuchi,chupapikachubudarapi”))	

Name	Value
PRINT.INDIVIDUAL.FIGURES	FALSE
PRINT.INDIVIDUAL.PDFS	TRUE
PRINT.INDIVIDUAL.TABLES	FALSE
REMOVE.BAD.SUBJ	TRUE
REMOVE.INCOMPLETE.SUBJ	YES
RESPONSE.LIST	ALL
start.time	2021-06-01 19:18:06

2 Introduction

Associative learning is widespread and exists in many species and domains [Aslin et al., 1998; Chen and Ten Cate, 2015; Conway2005a; Fiser and Aslin, 2002; Hauser et al., 2001; Saffran et al., 1996; Toro et al., 2005; Turk-Browne et al., 2005; Turk-Browne and Scholl, 2009]. This led to the conclusion that the computations for which it is critical might be similarly widespread, a view that has been particularly prominent in language acquisition [Aslin and Newport, 2012, Seidenberg et al., 2002, Thiessen, 2017].

However, associative learning is also remarkably modular [Endress, 2019]. For example, humans have independent associative learning abilities in superficially similar domains, including the learning of associations of objects with landmarks vs. boundaries [Doeller and Burgess, 2008, Doeller et al., 2008], associations among social vs. non-social objects [Tompson et al., 2019] and associations among consonants vs. vowels [Bonatti et al., 2005]. Further, associative learning abilities are generally not correlated across domains [Siegelman and Frost, 2015].

Preferential associations between specific classes of stimuli abound [Seligman, 1970], and can evolve in just 40 generations in fruit flies [Dunlap and Stephens, 2014]. For example, rats readily associate tastes with sickness and external events such as sounds or light flashes with pain, but cannot (or only with great difficulty) associate taste with pain or external events with sickness [Garcia et al., 1974, 1976]. This pattern of associations reflects the likely ecological sources of sickness vs. pain (i.e., food vs. external events). Critically, the formation of taste-sickness associations (but not of other forms of associations) is blocked in a suckling context for rats who have not been exposed to solid food [Martin and Alberts, 1979, Alberts and Gubernick, 1984], presumably because avoidance of the only food source would be detrimental; in contrast, minimal exposure to solid food re-establish taste-sickness associations [Gubernick and Alberts, 1984].

While such results suggest that, over evolutionary times, the adaptiveness of associative learning is regulated by increasing or decreasing its availability for specific classes of stimuli, it is less clear if associative learning is specialized for specific computational functions, or whether it is essentially a side effect of local neural processing [a "spandrel" in biological terms; Gould et al., 1979], that is sometimes adaptive, somtimes neutral and sometimes detrimentl. Here, we address this issue in a domain where the importance of associative learning has long been recognized: word learning. We suggest that associative learning is critical for predicting speech material under conditions where it can be predicted, but that it does not lead to memory.

One of the most prominent cases for associative learning in language acquisition is word segmentation. 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, respectively. One strategy to find word boundaries relies on Transitional Probabilities (TPs) among items, that is, the conditional probability of a syllable σ_{i+1} given a preceeding syllable σ_{i+1}, $P(\sigma_i\sigma_{i+1})/P(\sigma_i)$. Early on, Shannon [Shannon, 1951] showed that human adults are sensitive to such distributional information. Subsequent work conclusively demonstrated that infants and non-human animals share this ability [XXX], and suggested that TPs can be extract using simple associative mechanisms such as Hebbian learning [Endress and Johnson, 2021].

However, a sensitivity to distributional information does not imply that learners extract chunks that can be stored in memory. In fact, such as sensitive is typically tested by comparing participants' preference for high-TP items over low-TP items. It turns out that participants sometimes prefer high-TP items thet have never seen or heard (and thus could not have memorize) over low-TP items they have heard [Endress and Langus, 2017], suggesting that associative learning and memory for specific chunks may be dissociable. In fact, the types of representations created by associative learning might well be different from those used for linguistic stimuli [Endress and Langus, 2017, Fischer-Baum et al., 2011], while associative knowledge might be critical for predictive processing that is critical for both language [Levy, 2008, Trueswell et al., 1999] and other cognitive processes (XXX).

Here, we test this idea, focusing on the conditions under which associative learning operates and on the function it might have. To elucidate the conditions under which associative learning operates, we note that speech does not come as an continuous signal but rather as a sequence of smaller units due to its prosodic organization [Beckman and Pierrehumbert, 1986, Cutler et al., 1997, Nespor and Vogel, 1986, Selkirk, 1986, Shattuck-Hufnagel and Turk, 1996]. This prosodic organization is perceived in unfamiliar languages [Brentari

et al., 2011, Endress and Hauser, 2010, Fenlon et al., 2008, Pilon, 1981], by infants [Hirsh-Pasek et al., 1987; Christophe1994; Gout et al., 2004] and even by newborns [Christophe et al., 2001]. Further, associative learning operates primarily *within* rather than across major prosodic boundaries [Shukla et al., 2007, 2011]. As result, the a learner's segmentation task is not so much to integrate distributional information over long stretches of continuous speech, but rather to decide whether the the correct grouping in prosodic groups such as "*thebaby*" is "*theba + by*" or "*the + baby*". We thus ask to whether associative learning operates in such smaller groups, or only in longer stretches of continuous sound.

To elucidate the function of associative learning, we ask adult participants what they recall after being exposed to the speech stream from [Saffran et al., 1996], again with a continuous speech stream or a sequence of pre-segmented syllable sequences.

[Bortfeld et al., 2005, Shi and Lepage, 2008] Frequent words [Ngon et al., 2013]: statistics within

Table 2: Demographics 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

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:

3 Methods

3.1 Recognition experiment (London)

3.1.1 Participants

Participants were recruited from the City, University London participant pool and received course credit or monetary compensation for their time. We targeted 30 participants per experiment (15 per language). The final demographic information is given in Table 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 ($N = 2$).

3.1.2 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 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

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 .333. 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.

3.1.3 Stimuli

Stimuli were synthesized using the mbrola system [Dutoit et al., 1996], using the *us3* (American English male) diphone base. (We also used the *en1* (British English male) diphone base; however, as discussed below, this diphone based turned out be of relatively low quality and introduced confounds in the data.)

We chose a constant segment duration of 60 ms (syllable duration 120 ms) with a constant F_0 of 120 Hz. These values were choosen to match recordings of natural speech that were intended to be used in an investigation 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).

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 (http://sox.sourceforge.net/). 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 duration 18 min 14 s).

3.1.4 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 keybaord.

3.1.5 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 listend 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 interstimulus interval of 500 ms, and had to choose which items was more like what they heard during familiarization. One item was comprised the first two syllables of a word, and was thus a correct choice for Language 1. The other items comprised the last two syllables of a word, and was thus 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.

3.2 Recall experiment

3.2.1 Materials

We resynthesized the languages used in Saffran et al. [1996] 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 [Dutoit et al., 1996], at with a constant F0 of 120 at a rate of 216 ms per syllable (108 ms per phoneme).

Table 4: Languages used in the recall experiment.

L1	L2
pabiku	bikuti
tibudo	pigola
daropi	tudaro
golatu	budopa

During familiarization, words were presented 45 times each. For each participant, we generated a random concatenation of 45 repetitions of the 4 words, with the constraint that a words could not occur in immediate reptition. 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 streams with different random orders of words.

3.2.2 Procedure

3.2.2.1 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 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 the muxmovie utility.

Following the familiarization, the was a 30 s retention interval. Participants were instructed to count backwards from 99 in time with a metronome beat at 3s / beat. Performance was not monitored.

(Note to self: This was the case for both psyscope and testable.)

3.2.2.2 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.

3.2.2.3 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

4 Analysis

4.1 Recognition experiment

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 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 [Baayen et al., 2008], random factors were removed from the model when they did not contribute to the model likelihood.

We use likelihood ratios to provide evidence for the null hypothesis that performance did not differ from the chance level of 50%. Following [Glover and Dixon, 2004], we fit the participant average to a linear model comprising only an intercept and the null model fixing the intercept to 50%, and evaluate the likelihood of these models after correcting for the difference in the number of parameters using the Bayesian Information Criterion.

Table 5: Descriptives for the recognition test

data.set	mySegmentationCond	lang	N	M	SE	p.wilcox
all						
city	continuous	L1	11	0.386	0.074	0.168
city	continuous	L2	11	0.500	0.094	1.000
city	segmented	L1	11	0.932	0.051	0.003
city	segmented	L2	11	0.909	0.040	0.003
tstbl	continuous	L1	46	0.538	0.043	0.336
tstbl	continuous	L2	30	0.558	0.058	0.275
tstbl	segmented	L1	28	0.893	0.033	0.000
tstbl	segmented	L2	29	0.845	0.053	0.000
>= 50%						
city	continuous	L1	7	0.536	0.039	1.000
city	continuous	L2	7	0.679	0.077	0.089
city	segmented	L1	8	0.938	0.067	0.011
city	segmented	L2	7	0.893	0.055	0.019
tstbl	continuous	L1	43	0.576	0.040	0.050
tstbl	continuous	L2	28	0.598	0.055	0.063
tstbl	segmented	L1	28	0.893	0.033	0.000
tstbl	segmented	L2	28	0.875	0.044	0.000

4.2 Recall experiment

4.2.1 Recognition test

4.2.2 Recall test (for the lab-based experiment or in general?)

The substitution rules employed in the current experiment are shown in Table 6.

4.2.2.1 Substitution rules compensating for potential misperceptions

- O might be perceived as A (but probably not vice versa)
- Voiced and unvoiced consonants can be confused:
  - g and k
  - d and t
  - b and p
- b might be perceived as v

In some cases, these rules give you several possible matches. For example, in line 64, rapidala might be rOpidAlA or rOpidOlA

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

1. If one option provides more or longer existing chunks, choose it. For example, rOpidAlA has the chunk rOpi (pidA isn't possible in the stream), while rOpidOlA contains rOpi, so in this case the rule doesn't discrminate between the two :)
2. If one option requires fewer changes with respect to the original transcription, choose that.

I would apply the rules in this order, but I can see why you might want to use the opposite order as well.

4.2.2.2 Identify closest matches (for testable)

Table 6: Substitution rules applied to the participants vocalizations before and after the input was segmented into chunks. The patterns are given as regular expressions

Before segmentation		After segmentation	
Pattern	Replacement	Pattern	Replacement
-		u	o
2	tu	v	b
two	tu	p	b
([aeou])ck	\1k	b	p
ar([,\s+])	a\1	t	d
ar$	a	d	t
tyu	tu	k	g
ph	f	g	k
th	t	a	o
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			

Each recall response was analyzed in five steps. First, we applied pre-segmentation substitution rules to make the transcriptions more consistent (see Table 6). For example, *ea* (presumably as in *tea*) was replaced with *i*.

Second, responses were segmented into their underlying units. If a response contained a semicolon (;) or comma character (,), 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 subdividing resulted in any single-syllable response (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 like to reflect the words *bikuti*, *tudaro* and *budopa*.

Finally, if the response did not contain any commata or semicolons, it was segmented based on its spaces (if any).

Third, we removed geminate consonants and applied another set of substitution rules that to take into account possible misperceptions (see 6)). 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 substrings were created by clipping the first or the last syllable from the underlying form, and the search was repeated recursively for each of these substrings until a match was found. We then selected the longest match for all substrings.

Fifth, for each surface form, we selected the underlying form using the criteria (in this order) that, among the candidate underlying form of each surface form, the selected underlying form had (i) had the maximal number of attested syllables, (ii) the maximal length, and (iii) the shortest derivation.

4.2.2.3 Change columns to categorize transcriptions

We then computed various properties for each underlying form, given the "target" language the participant had been exposed to. Specifically, we calculated: (1) the number of syllables, (2) whether it was a word from the target language, (3) whether it was a concatentation 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 a "class-word" with the two initial syllables coming from one word and the final syllables from another word or vice versa (i.e., class-words had the form *A_iB_iC_j* or *A_iB_jC_j*, (9) whether it was high-TP chunk (i.e., a word or a word with the first or the last syllable missing, after removing any leading or trailing unattested syllables), (10) whether it was a low-TP chunk (i.e., a chunk of the form *C_iA_j*, after removing lead or trailing unattested syllables), (11) whether it had a "correct" initial syllable, (12) whether it had a "correct" final syllable, (13) 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), (14) whether it was a backward word from the target language (i.e., a word with the syllable order reversed), (15) whether it was a concatenation of backward words, (16) whether it was a backward word or a concatenation thereof (i.e., the disjunction of (14) and (15)), (17) whether it was a backward part-word, (18) whether it was a concatenation of backward part-words, (19) whether it was a backward word or a concatenation thereof (i.e., the disjunction of (17) and (18)), (19) whether it was high-*backward*-TP chunk (i.e., a backward word or a backward word with the first or the last syllable missing, after removing any leading or trailing unattested syllables), (20) whether it was a low-*backward*-TP chunk (i.e., a chunk of the form *A_iC_j*, after removing

lead or trailing unattested syllables, (21) the average forward TP of the transitions in the form, (22) the *expected* forward TP of the form if form is attested in the speech stream (see below for the calculation), and (23) the average backward TP of the transitions in the form.

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

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

Table 7: Number of unattested items

data.set	streamType	N.total.M	N.total.min	N.total.max	N.unattested.M	N.unattested.min	N.unattested.max
city	continuous	4.21	1	9	2.50	0	5
city	segmented	4.21	2	13	1.86	0	10
testable	continuous	4.73	1	10	2.11	0	8
testable	segmented	3.69	1	15	1.75	0	13

As shown in Table 7, there was a considerable number of recall responses containing unattested syllables. The complete list of unattested items is in `segmentation_recall_unattested.xlsx`. Unattested items are items that are not word, 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.

We thus decided to restrict ourselves to responses that can be clearly interpreted and removed all items containing unattested syllables. Here, `FILTER.UNATTESTED.ITEMS` was set to FALSE, while `FILTER.SINGLE.SYLLABLES` was set to FALSE.

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.

4.2.3 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 = $, 14). Another 8 participants were excluded because parsing their productions took an excessive amount of computing time, though their productions did not seem to ressemble the familiarization items in the first place. The responses were dalonigtbdophophi dalobdakabdarobigopachu // 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 // takahsakakakaratatataikokokokotatakatakatakatakatakatakataka // dabroobitalooki,bkuti2,golab // matikulatatitulapapitularimatitulaatitula // tu kalla ti palla tuti kulla papi pu tu kalla ti palla tuti kulla papi pu // tutopitulakatutopitoolaka // papikuchi,butalapapikuchi,kukala,pikala,budharapikuchi,chupapikachubudarapi.

The final demographic information is given in Table 8.

Table 8: Demographics of the final sample. Note that the City participants completed both segmentation conditions.

data.set	streamType	lang	N	Females	Males	Age.m	Age.range
city							
city	continuous	both	14	14	0	17.9	0-22
city	segmented	both	14	14	0	17.9	0-22
tstbl							
tstbl	continuous	L1	37	8	29	31.0	18-71
tstbl	continuous	L2	26	10	16	30.5	18-71
tstbl	segmented	L1	25	5	20	29.7	18-48
tstbl	segmented	L2	26	5	21	32.2	18-62

4.2.3.1 Save categorized data

4.2.3.2 Measures for productions in the recall phase

We will use the following measures to analyze the participants' productions in the recall phase. Some analyses below rely on within-participant averages [A], within-participant sums [S] or other measures [O].

- General measures
  - [S] Number of items produced. To be compared across segmentation conditions, and against zero.
  - [A] Average length of items produced. To be compared across segmentation conditions.
  - [S,A] Number and proportion (among productions) of words (and concatenations thereof)
  - [S,A] Number and proportion (among productions) of part-words (and concatenations thereof)
  - [O] Performance in the two alternative forced-choice test. To be compared across segmentation conditions.
- TP-based analyses (raw TPs)
  - [A] Average forward TP in items
    - Compare across segmentation conditions
    - Compare to expected TPs for a random string. The expected TPs for a random concatenation are the TPs in a random bigram. For an A or a B syllable, the random TP is 1 × 1 / 12, as there is only 1 (out of 12) non-zero TP continuations. For a C syllable, the random TP is 3 × 1/3 / 12, as there are 3 possible concatenations. On average, the random TP is thus $(1/12 + 1/12 + 1/12)/3 = 1/12 \approx .083$.
    - Calculate difference *expected* TPs for correctly reproduced items, given the item's initial position. The expected TPs for items of at least 2 syllables starting on an initial syllable are c(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.
  - [A] Average backward TP in items
- TP-based analyses (chunks). In addition to th raw TPs above, we also counted high- and low-TP *chunks*. As mentioned above, high-TP chunks are words or words with the first or the last syllable missing, after removing any leading or trailing unattested syllables, while low-TP chunks are chunks of the form C_iA_j, after removing lead or trailing unattested syllables.
  - [S,A] Number and proportion of high TP-chunks.
  - [S,A] Number and proportion of low TP-chunks.
  - [0] Proportion of high-TP chunks among high and low-TP chunks.
- Positional analyses
  - [A] Proportion of items with syllables in correct postions
    a. Items with correct initial syllables. Chance level: 4/12
    b. Items with correct final syllables. Chance level: 4/12
    c. Disjunction of *a* and *b*. Chance level: $2 \times 4/12 - 4/12 \times 4/12 = 5/9$

Table 9: Descriptives. Check exclusion files

voice	experimentID	segm	lang	N	M	SE
en	stats.3x.en.cont	continuous	L1	15	0.604	0.041
en	stats.3x.en.cont	continuous	L2	15	0.374	0.045
en	stats.3x.en.segm	segmented	L1	15	0.552	0.078
en	stats.3x.en.segm	segmented	L2	15	0.533	0.056
us	stats.3x.us.cont	continuous	L1	16	0.608	0.042
us	stats.3x.us.cont	continuous	L2	16	0.562	0.042
us	stats.3x.us.cont2	continuous	L1	15	0.600	0.057
us	stats.3x.us.cont2	continuous	L2	15	0.656	0.058
us	stats.3x.us.segm	segmented	L1	15	0.507	0.053
us	stats.3x.us.segm	segmented	L2	15	0.526	0.024

5 Results

5.1 Recognition experiments (Results with the us3 diphone base; the en1 results are in the appendix)

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

We first asked if learners can split utterances that likely result from prosodic presegmentation into its 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. Further, as shown in Table 10, performance did not depend on the language condition. As shown in Appendix XXX, the failure to use statistical learning was replicated using a second diphone base.

Further, 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 Supplementary Information XXX).

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

While observers failed to split presegmented words into their underlying units, we now ask if they can do so when the words are embedded into a continuous speech stream.

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 10, performance did not depend on the language condition. As shown in Table 10, performance was significantly better compared to when the stream was pre-segmented.

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

We replicated the sucessful 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 10, performance did not depend on the language condition. As shown in Table 10, performance was significantly better compared to when the stream was pre-segmented.

However, as shown in Appendix XXX, when trying to replicate these results using a different diphone base, we obtained a preference for specific items due to the poor quality of the diphone base.

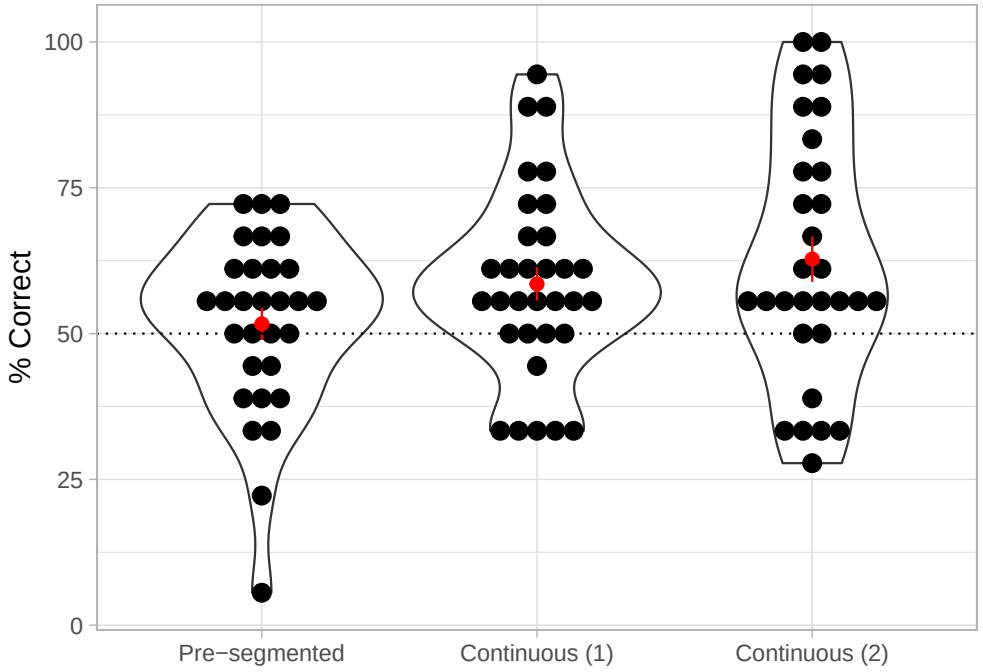


Figure 1: Results of Experiment 1. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a pre-segmented familiarization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

Table 10: 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

5.2 Recall experiment

In the analyses below, we removed all items that contained syllables not attested in the speech stream as it is unclear how these items should be analyzed. As a result, we also removed participants who did not produce any items that contained attested syllables only.

Table 11: 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.

After computing these counts and averages, we asked which counts were significantly different from zero in a one-tailed Wilcoxon test, either for the continuous or the segmented condition. These counts were n.items,

n.syll, n.words, n.words.or.multiple, n.part.words, n.part.words.or.multiple, n.high.tp.chunk, n.low.tp.chunk. (Note: These counts are currently restricted to the testable data set.)

As shown in Table 5.2.4, participants produced on average 1.338 words in the segmented condition, and 0.156 in the continuous condition.

5.2.1 General measures

We first calculate the number of items produced as well their average, and compare them against the zero as well as across segmentation conditions. As shown in Table 5.2.4 and Figures 2a and b, participants produced positive number of items. Neither the number of items produced nor their number of syllables differed across the segmentation conditions.

5.2.2 TP-based analyses

We first computed the average forward TPs in the produced items, and separately for each segmentation condition, compared it to both the expected TPs for random strings and the expected TPs given the starting syllable.

- The expected TPs for items of at least 2 syllables starting on an initial syllable are c(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.
- The expected TPs for a random concatenation are the TPs in a random bigram. For an A or a B syllable, the random TP is 1×1 / 12, as there is only 1 (out of 12) non-zero TP continuations. For a C syllable, the random TP is $3 \times 1/3$ / 12, as there are 3 possible concatenations. On average, the random TP is thus $(1/12 + 1/12 + 1/12)/3 = 1/12 \approx .083$.

We compared these measures across segmentation conditions.

As shown in Table 5.2.4 and Figures 2c and d, forward and backward TPs were significantly greater than expected for a random string in both the continuous and the segmented condition, with greater TPs in the segmented conditions. However, they were significantly *lower* than the TPs expected if items recalled faithfully, given their starting position.

As shown in Figure 5.2.4b, participants produced a positive number of high-TP chunks in both the segmented and the continuous condition, with a significantly greater number in the segmented condition. In contrast, they produced a positive number of low-TP chunks only in the continuous condition. Accordingly, the proportion of high-TP chunks among high- and low-TP chunks exceeded 50% only in the segmented condition.

5.2.3 Word vs. part-word analysis

We next calculate the number and proportion of among (productions) of words and part-words respectively; we also accept concatenations of words and part-words. The proportions will be compared across stream types as well as to zero.

Finally, we calculate the proportion of words among the word and part-word productions. This proportion will be compared across segmentation types, as well as to the chance level of 50%.

As shown in Table 5.2.4 and in Figure 5.2.4a, participants produced a positive number of words only in the segmented condition, but not in the continuous condition. In contrast, they produced a positive number of part-words only in the continuous condition, but not in the segmented condition. Accordingly, the proportion of words among words and part-words was significantly greater than 50% in the segmented condition, but numerically (though not significantly) smaller than 50% in the continuous condition. The latter result is consistent with participants randomly picking a syllable to start their vocalizations; if so, part-words should be 2 times as likely as words.

5.2.4 Positional analyses

Finally, we analyze the productions in terms of correct initial final positions. As there are four initial and final positions, respectively, 4/12 of the productions should have "correct" initial positions, 4/12 should have correct final positions, while $2 \times 4/12 - (4/12)^2 = 5/9$ should have either correct initial or final positions.

As shown in Table 5.2.4 and Figure 5.2.4c and d, participants produced items with correct initial or final positions at great than chance level only in the segmented condition, but not the continuous condition.

\begin{table}

\caption{Various analyses pertaining to the productions as well as test against their chances levels. Number of items produced, their numbers of syllables, number of words, number of part-words (chance level: 0), proportion of words among productions, proportion of part-words among productions, proportion of words among words and part-words (chance level 50%), average forward TPs (chance level: 1/12), difference between positionally expected and actual TPs, average backward TPS. CHUNKS }

	Continuous	Segmented	*p* (Continuous vs. Segmented)
Recognition accuracy			
lab-based	*M*= 0.607, *SE*= 0.0448, *p*= 0.0477	*M*= 0.911, *SE*= 0.0439, *p*= 0.000898	0.008
online	*M*= 0.595, *SE*= 0.0335, *p*= 0.00411	*M*= 0.887, *SE*= 0.0268, *p*= 5.38e-10	0.000
Number of items			
lab-based	*M*= 4.21, *SE*= 0.723, *p*= 0.00106	*M*= 4.21, *SE*= 0.763, *p*= 0.00103	0.843
online	*M*= 4.73, *SE*= 0.315, *p*= 4.75e-12	*M*= 3.69, *SE*= 0.326, *p*= 4.13e-10	0.013
Number of syllables/item			
lab-based	*M*= 3.53, *SE*= 0.453, *p*= 0.00107	*M*= 2.8, *SE*= 0.151, *p*= 0.000546	0.053
online	*M*= 2.27, *SE*= 0.114, *p*= 4.55e-12	*M*= 2.82, *SE*= 0.0798, *p*= 7.06e-11	0.000
Proportion of words among words and part-words (or concatenations thereof)			
lab-based	*M*= 0.321, *SE*= 0.153, *p*= 0.322 (vs. 0.5); 0.798 (vs. 0.333333333333333)	*M*= 1, *SE*= 0, *p*= 0.000627 (vs. 0.5); 0.000627 (vs. 0.333333333333333)	0.034
online	*M*= 0.357, *SE*= 0.138, *p*= 0.301 (vs. 0.5); 0.649 (vs. 0.333333333333333)	*M*= 1, *SE*= 0, *p*= 9.79e-09 (vs. 0.5); 9.79e-09 (vs. 0.333333333333333)	0.000
Forward TPs			
lab-based	*M*= 0.301, *SE*= 0.0702, *p*= 0.0107	*M*= 0.634, *SE*= 0.092, *p*= 0.00159	0.006
online	*M*= 0.372, *SE*= 0.0361, *p*= 4.4e-09	*M*= 0.555, *SE*= 0.049, *p*= 8.76e-09	0.004
Backward TPs			
lab-based	*M*= 0.301, *SE*= 0.0702, *p*= 0.0107	*M*= 0.634, *SE*= 0.092, *p*= 0.00159	0.006
online	*M*= 0.372, *SE*= 0.0361, *p*= 4.4e-09	*M*= 0.555, *SE*= 0.049, *p*= 8.76e-09	0.004
Proportion of High-TP chunks among High- and Low-TP chunks			
lab-based	*M*= 0.75, *SE*= 0.289, *p*= 0.424 (vs. 0.5); 0.85 (vs. 0.666666666666667)	*M*= 1, *SE*= 0, *p*= 0.000627 (vs. 0.5); 0.000627 (vs. 0.666666666666667)	1.000
online	*M*= 0.721, *SE*= 0.0574, *p*= 0.000702 (vs. 0.5); 0.0556 (vs. 0.666666666666667)	*M*= 0.95, *SE*= 0.0353, *p*= 1.31e-08 (vs. 0.5); 6.14e-07 (vs. 0.666666666666667)	0.000
Proportion of items with correct initial syllables			
lab-based	*M*= 0.318, *SE*= 0.0984, *p*= 0.691	*M*= 0.789, *SE*= 0.0666, *p*= 0.00123	0.012
online	*M*= 0.425, *SE*= 0.0385, *p*= 0.0357	*M*= 0.72, *SE*= 0.045, *p*= 3.7e-08	0.000
Proportion of items with correct final syllables			
lab-based	*M*= 0.46, *SE*= 0.116, *p*= 0.48	*M*= 0.76, *SE*= 0.0976, *p*= 0.00275	0.045
online	*M*= 0.369, *SE*= 0.0438, *p*= 0.781	*M*= 0.694, *SE*= 0.0524, *p*= 3.56e-07	0.000

\end{table}

\begin{figure}

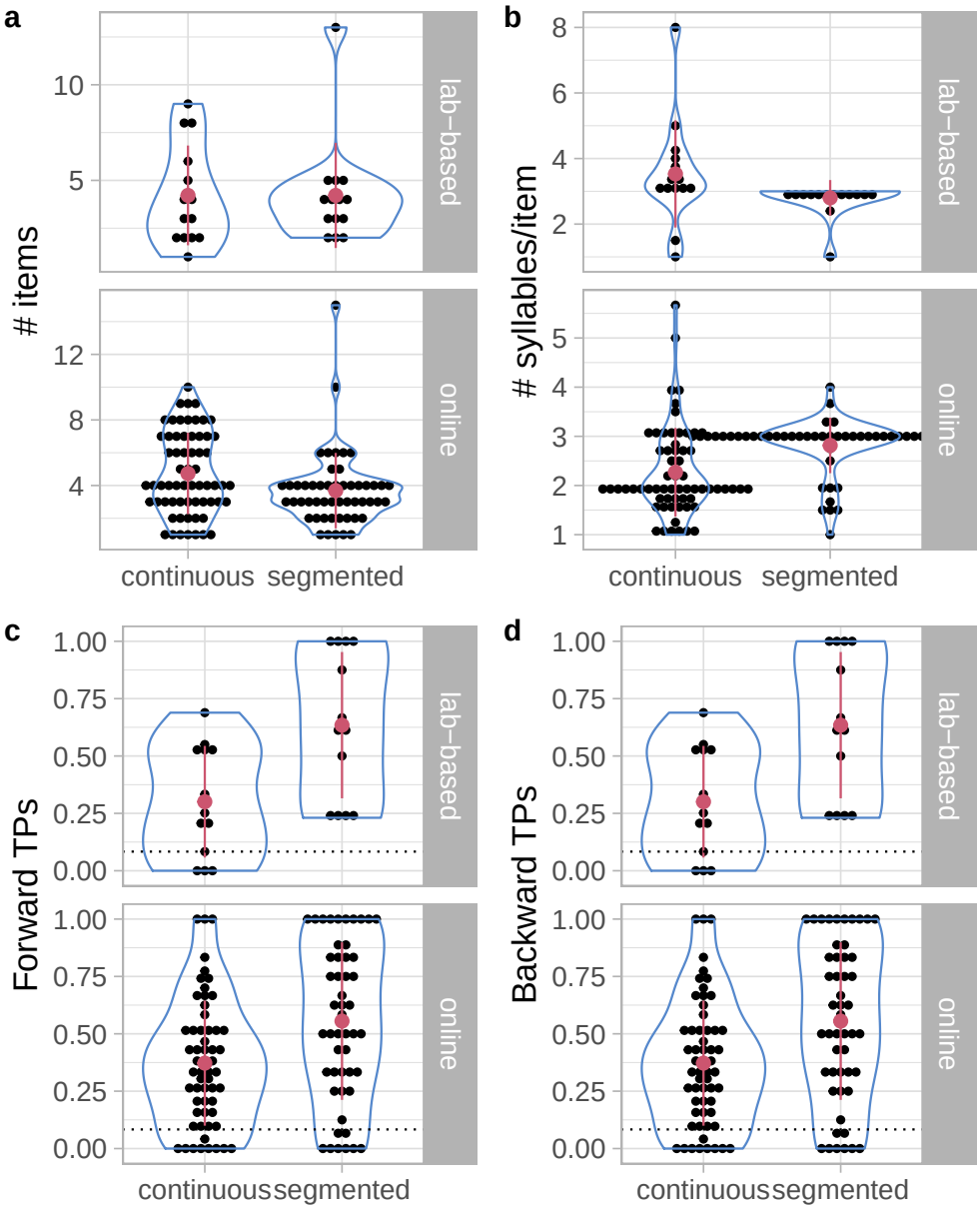
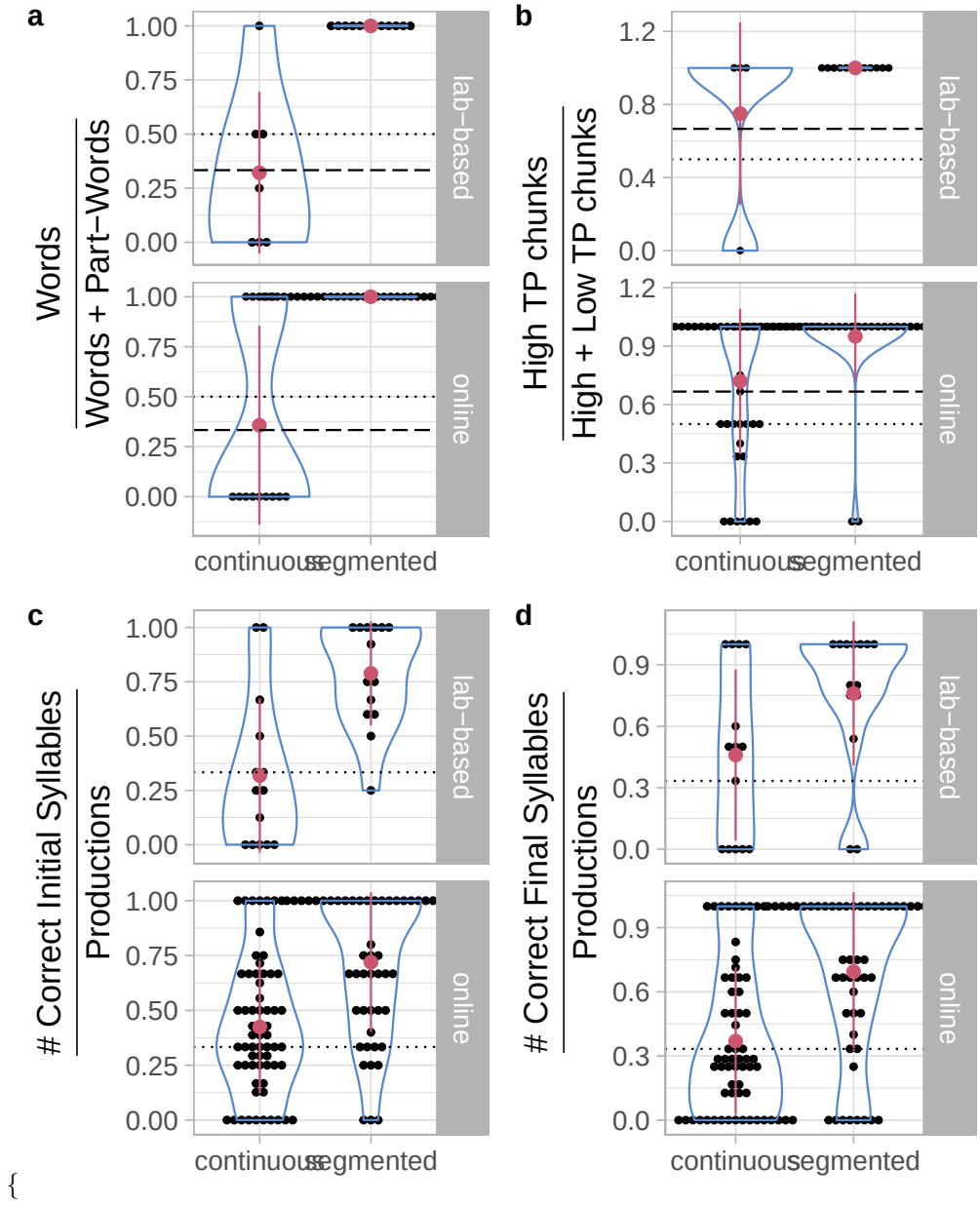


Figure 2: Number of items produced, number of syllables per item and forward and backward TPs. The dotted line represents the chance level for a randomly ordered syllable sequence.



{

}

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

6 Appendix

6.1 Additional results for the recall experiments

\begin{table}

\caption{Various supplementary analyses pertaining to the productions as well as test against their chances levels. Number of items produced, their numbers of syllables, number of words, number of part-words (chance level: 0), proportion of words among productions, proportion of part-words among productions, proportion of words among words and part-words (chance level 50%), average forward TPs (chance level: 1/12), difference between positionally expected and actual TPs, average backward TPS. CHUNKS }

	Continuous	Segmented	*p* (Continuous vs. Segmented)
Number of words			
lab-based	*M*= 0.286, *SE*= 0.13, *p*= 0.0719	*M*= 1.71, *SE*= 0.316, *p*= 0.00224	0.005
online	*M*= 0.143, *SE*= 0.0752, *p*= 0.0545	*M*= 1.24, *SE*= 0.18, *p*= 3.63e-07	0.000
Proportion of words among productions			
lab-based	*M*= 0.286, *SE*= 0.13, *p*= 0.0719	*M*= 1.71, *SE*= 0.316, *p*= 0.00224	0.005
online	*M*= 0.143, *SE*= 0.0752, *p*= 0.0545	*M*= 1.24, *SE*= 0.18, *p*= 3.63e-07	0.000
Number of part-words			
lab-based	*M*= 0.643, *SE*= 0.258, *p*= 0.031	*M*= 0, *SE*= 0, *p*= NaN	0.031
online	*M*= 0.19, *SE*= 0.0639, *p*= 0.00693	*M*= 0, *SE*= 0, *p*= NaN	0.005
Proportion of part-words among productions			
lab-based	*M*= 0.643, *SE*= 0.258, *p*= 0.031	*M*= 0, *SE*= 0, *p*= NaN	0.031
online	*M*= 0.19, *SE*= 0.0639, *p*= 0.00693	*M*= 0, *SE*= 0, *p*= NaN	0.005
Actual vs. expected forward TPs			
lab-based	*M*= -0.513, *SE*= 0.0814, *p*= 0.000244	*M*= -0.328, *SE*= 0.0823, *p*= 0.00909	0.126
online	*M*= -0.462, *SE*= 0.0385, *p*= 2.57e-10	*M*= -0.38, *SE*= 0.043, *p*= 3.66e-08	0.156
Number of High-TP chunks			
lab-based	*M*= 0.714, *SE*= 0.427, *p*= 0.181	*M*= 2.14, *SE*= 0.375, *p*= 0.00224	0.022
online	*M*= 1.02, *SE*= 0.14, *p*= 6.57e-08	*M*= 1.53, *SE*= 0.193, *p*= 5.53e-08	0.039
Proportion of High-TP chunks among productions			
lab-based	*M*= 0.097, *SE*= 0.056, *p*= 0.181	*M*= 0.561, *SE*= 0.102, *p*= 0.00241	0.003
online	*M*= 0.208, *SE*= 0.0311, *p*= 1.12e-07	*M*= 0.432, *SE*= 0.0485, *p*= 7.19e-08	0.000
Number of Low-TP chunks			
lab-based	*M*= 0.0714, *SE*= 0.0741, *p*= 1	*M*= 0, *SE*= 0, *p*= NaN	1.000
online	*M*= 0.365, *SE*= 0.0832, *p*= 8.81e-05	*M*= 0.0588, *SE*= 0.0439, *p*= 0.371	0.001
Number of Low-TP chunks among productions			
lab-based	*M*= 0.00893, *SE*= 0.00927, *p*= 1	*M*= 0, *SE*= 0, *p*= NaN	1.000
online	*M*= 0.0692, *SE*= 0.0165, *p*= 0.000206	*M*= 0.00915, *SE*= 0.00706, *p*= 0.371	0.001

\end{table}

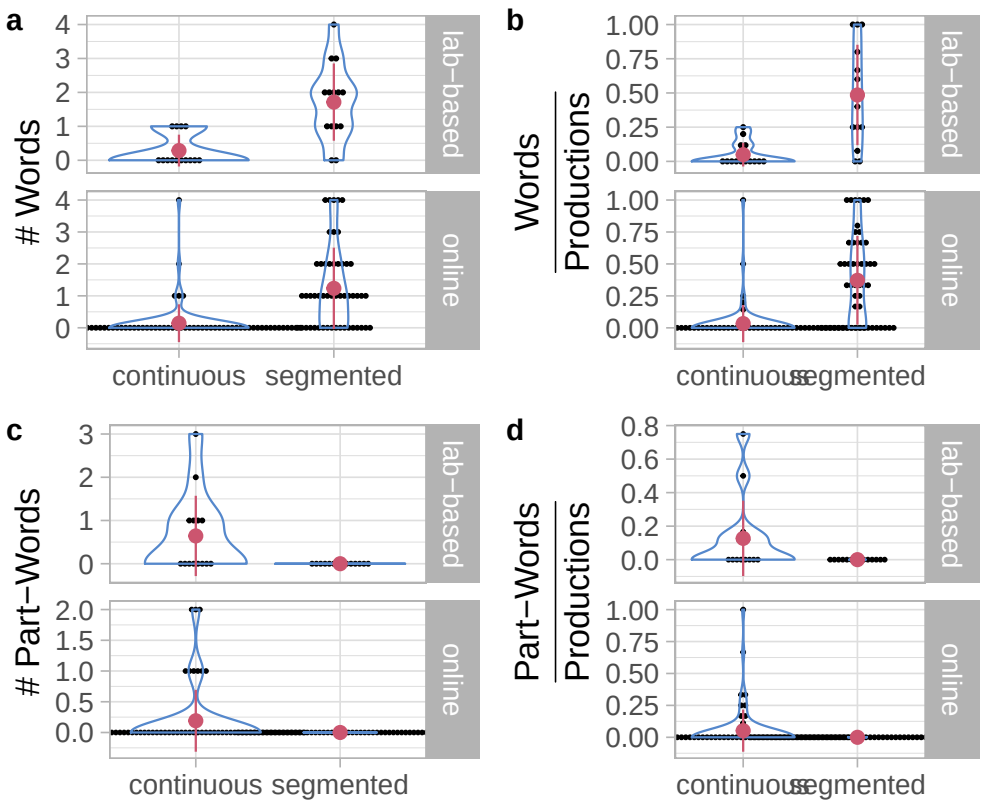


Figure 3: Number and proportion (among vocalizations) of words and part-words.

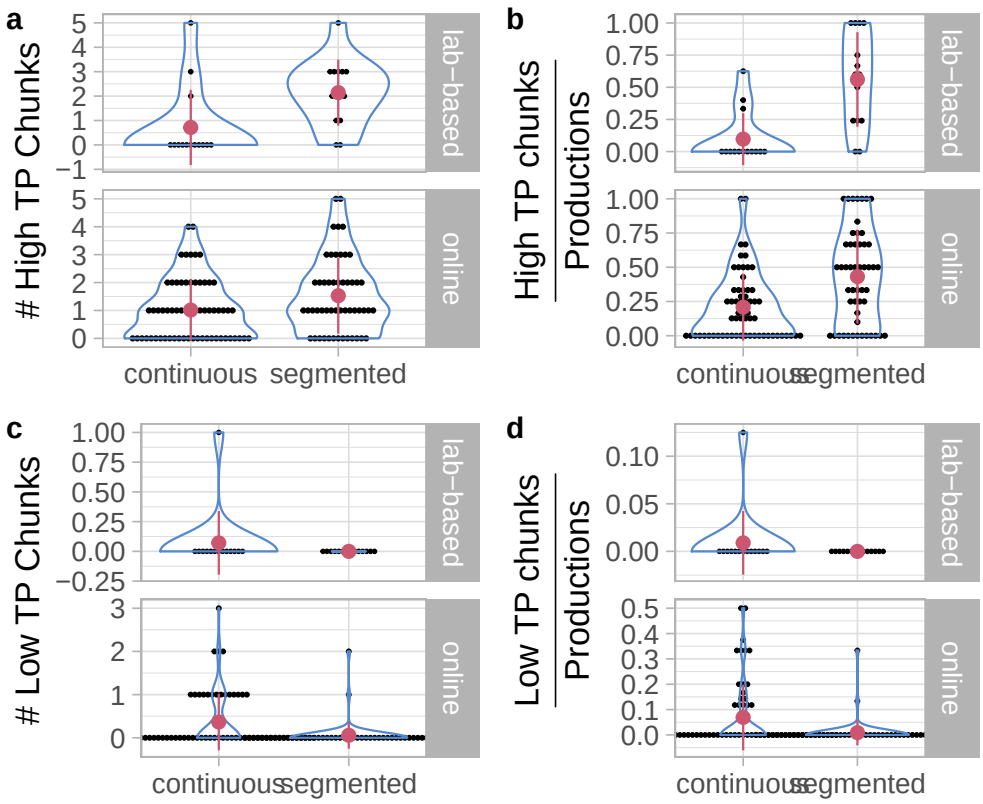


Figure 4: Plot of High and Low TP chunks.

6.2 Experiments with the *en1* diphone base

6.2.1 Segmented stream, 3 repetitions of the stream, en1 diphone based

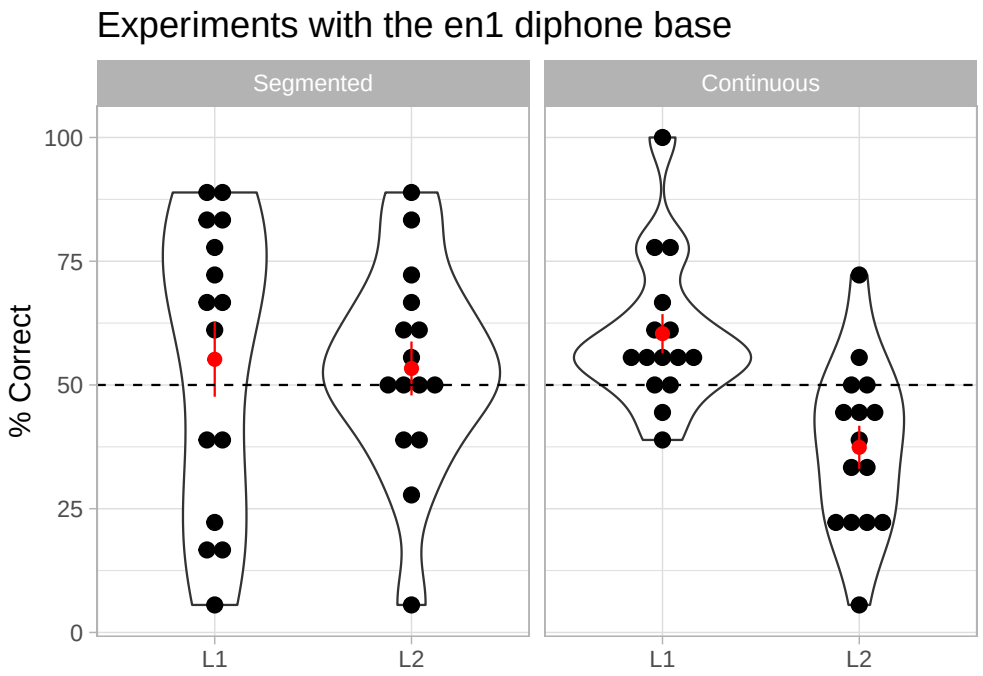


Figure 5: Results for a pre-segmented presentation of the stream (540 ms silences, left) and continuous presentation of the stream (right). Each word was repeated 45 times. The diphone based was *en1*.

As shown in Figure 5, 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 12, performance did not depend on the language condition.

6.2.2 Continuous stream, 3 repetitions of the stream, en1 diphone based

As shown in Figure 5, 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 12, 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.

6.3 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.

Table 12: 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

Effect	Estimate	Std. Error	CI	t	p
stats.3x.en.segm					
langL2	-0.097	0.441	-0.96, 0.767	-0.22	0.826
stats.3x.en.cont					
langL2	-1.024	0.410	-1.83, -0.22	-2.50	0.013

Table 13: 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

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

6.3.1 Methods

6.3.1.1 Participants

Demographic information of Pilot Experiment 1 is given in Table 13. Participants were native speakers of Spanish and Catalan and were recruited from the Universitat Pompeu Fabra community.

6.3.2 Stimuli

Stimuli transcriptions are given in Table 14. They were synthesized using the *es2* (Spanish male) diphone base of the mbrola [Dutoit et al., 1996] speech synthesized, using a segment duration of 225 ms and an fundamental frequency of 120 Hz.

6.3.2.1 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/).

6.3.2.2 Familiarization

The design of Pilot Experiment 1 is shown in Table 14. The languages comprise trisyllabic items. All foward 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 sequency 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 14: 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

Table 15: 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

6.3.2.3 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.

6.3.3 Results

As shown Table 15, a generalized linear model revealed that performance depended neither on the amount of familiarization nor on the familiarization language. As shown in Figure 6, a Wilcoxon test did not detect any deviation from the chance level of 50%, neither for all amounts of familiarization combined, M= 53.5, SE= 2.71, p= 0.182, nor for the individual familiarization conditions (3 repetitions per word: M= 54.1, SE= 4.81, p= 0.416; 15 repetitions per word: M= 54.6, SE= 4.52, p= 0.325; 30 repetitions per word: M= 51.9, SE= 4.98, p= 0.63). Following Glover and Dixon [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.

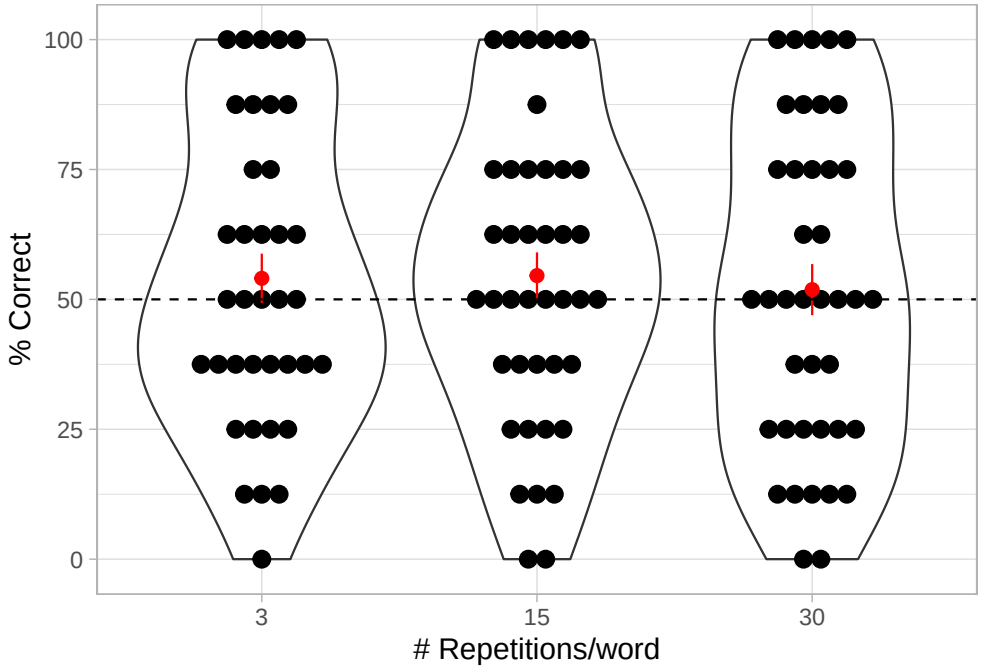


Figure 6: Results of Pilot Experiment 1. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right) 30 repetitions per word.

References

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

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

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

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

M Beckman and J B Pierrehumbert. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:15–70, 1986.

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

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

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

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

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

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

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

Christian F Doeller, John A King, and Neil Burgess. Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proc Natl Acad Sci U S A*, 105(15):5915–20, Apr 2008. doi: 10.1073/pnas.0801489105.

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

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

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

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

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

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

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

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

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

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

J Garcia, W G Hankins, and K W Rusiniak. Flavor aversion studies. *Science*, 192:265–267, 1976.

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

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

Ariel Gout, Anne Christophe, and James L. Morgan. Phonological phrase boundaries constrain lexical access ii. infant data. *J Mem Lang*, 51(4):548–567, 2004.

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

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

K. Hirsh-Pasek, D. G. Kemler Nelson, P. W. Jusczyk, K. W. Cassidy, B. Druss, and L. Kennedy. Clauses are perceptual units for young infants. *Cognition*, 26(3):269–86, Aug 1987.

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

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

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

Céline Ngon, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat, and Sharon Peperkamp. (non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Dev Sci*, 16 (1):24–34, Jan 2013. doi: 10.1111/j.1467-7687.2012.01189.x.

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

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

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

Martin E. Seligman. On the generality of the laws of learning. *Pyschol Rev*, 77(5):406–418, 1970.

E. Selkirk. On derived domains in sentence phonology. *Phonology Yearbook*, 3:371–405, 1986.

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

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

Rushen Shi and Mélanie Lepage. The effect of functional morphemes on word segmentation in preverbal infants. *Developmental science*, 11:407–413, May 2008. ISSN 1467-7687. doi: 10.1111/j.1467-7687.2008.00685.x.

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

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

Noam Siegelman and Ram Frost. Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *J Mem Lang*, 81:105–120, May 2015. doi: 10.1016/j.jml.2015.02.001.

Erik D. Thiessen. What's statistical about learning? insights from modelling statistical learning as a set of memory processes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 372, January 2017. ISSN 1471-2970. doi: 10.1098/rstb.2016.0056.

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

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

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

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

Nicholas B Turk-Browne, Justin Jungé, and Brian J Scholl. The automaticity of visual statistical learning. *J Exp Psychol Gen*, 134(4):552–64, Nov 2005. doi: 10.1037/0096-3445.134.4.552.