City Research Online

City St George's, University of London

Citation: Endress, A. (2025). The specificity of sequential statistical learning: Statistical learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory for words. Cognition, 261, 106130. doi: 10.1016/j.cognition.2025.106130

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/33398/

Link to published version: https://doi.org/10.1016/j.cognition.2025.106130

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

The specificity of sequential statistical learning: Statistical learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory for words

Learning statistical regularities from the environment is ubiquitous across domains and species. It might support the earliest stages of language acquisition, especially identifying and learning words from fluent speech (i.e., word-segmentation). But how do the statistical learning mechanisms involved in word-segmentation interact with the memory mechanisms needed to actually remember words as well as with the learning situations where words actually need to be learned? We show that, in a memory recall task after exposure to continuous, statistically structured speech sequences, participants track the statistical structure of the speech sequences and are thus sensitive to probable syllable transitions, but hardly remember any items at all. Analysis of their productions suggests that they are unable to identify probable word boundaries. As a result, they tend to produce *low-probability* items even while preferring high-probability items in a recognition test. Only discrete familiarization sequences with isolated words yield memories of actual items. Through computational modeling, we show that earlier results purportedly supporting memory-based theories of statistical learning can be reproduced by memoryless Hebbian learning mechanisms. Turning to how specific learning situations affect statistical learning, we show that it predominantly operates in continuous speech sequences like those used in earlier experiments, but not in discrete chunk sequences likely encountered during language acquisition. Taken together, these results suggest that statistical learning might be specialized to accumulate distributional information, but that it is dissociable from the (declarative) memory mechanisms needed to acquire words and does not allow learners to identify probably word boundaries.

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

1 Introduction

The ability to learn statistical regularities from the environment is remarkably widespread across species and domains (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Hauser, Newport, & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Toro, Trobalon, & Sebastián-Gallés, 2005; Turk-Browne & Scholl, 2009; Chen & Ten Cate, 2015), and might support a wide range of computations (e.g., Sherman, Graves, & Turk-Browne, 2020). Forms of statistical learning that allow learners to track sequential dependencies among sequence items might be especially important during language acquisition (Aslin & Newport, 2012; Saffran & Kirkham, 2018). However, their computational function is unclear. It is widely believed that such forms of statistical learning help learners acquire words from fluent speech

(e.g., Aslin et al., 1998; Saffran, Aslin, & Newport, 1996), and thus (presumably) store word candidates in (declarative) memory (Graf-Estes, Evans, Alibali, & Saffran, 2007; Isbilen, McCauley, Kidd, & Christiansen, 2020). However, other authors suggest that statistical learning is important for predicting events (Sherman & Turk-Browne, 2020; Turk-Browne, Scholl, Johnson, & Chun, 2010). Here, we suggest that statistical learning is critical for predicting speech material and operates predominantly under conditions where prediction is possible. However, we also suggest that statistical learning does not lead to declarative memories of words, and that separate mechanisms are required to form these memories.

We note that the label "statistical learning" has also been used for a variety of other computations, including discovering phonemic and allophonic categories (e.g., Maye, Werker, & Gerken, 2002), learning relevant locations in visual search (e.g., van Moorselaar & Slagter, 2019), compressing redundant information in visual working memory (e.g., Brady, Konkle, & Alvarez, 2009), among others (see Sherman et al., 2020 for a review). Here, we focus on forms of statistical learning that allow learners to track sequential dependencies among items in continuous sequences (and possibly also to associate simultaneously presented items in vision). We surmise that other computations referred to as "statistical learning" likely rely on different mechanisms and might well have different properties.

1.1 Statistical learning vs. declarative memory of words in fluent speech

Speech is often thought to be a continuous signal (and often perceived as such in unknown languages, but see below), and before learners can commit any words to memory, they need to learn where words start and where they end. They might rely on Transitional Probabilities (TPs) among syllables, that is, the conditional probability of a syllable σ_{i+1} given a preceding syllable σ_i, $P(\sigma_i\sigma_{i+1})/P(\sigma_i)$. Relatively predictable transitions are likely located inside words, while unpredictable ones straddle word boundaries. Early on, Shannon (1951) showed that human adults are sensitive to such distributional information. Subsequent work demonstrated that infants and non-human animals share this ability (Saffran, Aslin, & Newport, 1996; Hauser et al., 2001; Kirkham et al., 2002; Toro, Trobalon, & Sebastián-Gallés, 2005; Turk-Browne & Scholl, 2009; Chen & Ten Cate, 2015).

Statistical learning therefore supports predictive processing (Sherman & Turk-Browne, 2020; Turk-Browne et al., 2010), that is, the ability to anticipate stimuli and events based on current and past experience. This ability is critical for language (Levy, 2008; Trueswell, Sekerina, Hill, & Logrip, 1999) and other cognitive processes (Clark, 2013; Friston, 2010; Keller & Mrsic-Flogel, 2018). However, while words are clearly stored in declarative Long-Term Memory (after all, the point of knowing words is to "declare" them), statistical knowledge does not imply the formation of such memory representations. In fact, after exposure to sequences where some transitions are more likely than others, observers report greater familiarity with high-TP items than with low-TP items, even when they have never encountered either of these items and thus could not have memorized them (because the items are played backwards with respect to the familiarization sequence; Endress & Wood, 2011; Turk-Browne & Scholl, 2009; Jones & Pashler, 2007). Sometimes, observers even report greater familiarity with high-TP items they have *never* encountered than with low-TP items they have heard or seen (Endress & Langus, 2017; Endress & Mehler, 2009b; Endress, under review), suggesting that a preference for high-TP items over low-TP items does not necessarily imply that the high-TP items are encoded in declarative LTM. Further, and in line with this view, statistical learning abilities might reflect simple associative mechanisms such as Hebbian learning (Endress, 2010; Endress & Johnson, 2021; Endress, 2024): If the representation of a syllable is

still active while the next one is presented, the two syllable representations are active together and can thus form an association. These Hebbian associations will thus reflect the TPs among syllables.

While the question of whether statistical learning leads to memory for items (or chunks) is controversial (see e.g. Perruchet, 2019 vs. Endress, Slone, & Johnson, 2020 and General Discussion), statistical learning has been linked to implicit learning (e.g., Christiansen, 2018; Perruchet & Pacton, 2006; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), and is available to arguably implicit learners such as sleeping newborns, (Fló, Benjamin, Palu, & Dehaene-Lambertz, 2022). Dissociations between implicit learning and declarative memory have long been documented behaviorally (Graf & Mandler, 1984), developmentally (Finn et al., 2016), and neuropsychologically (Cohen & Squire, 1980; Knowlton, Mangels, & Squire, 1996; Poldrack et al., 2001; Squire, 1992), to the extent that statistical predictions can *impair* declarative memory encoding in healthy adults (Sherman & Turk-Browne, 2020). If statistical learning operates similarly in a word-segmentation context as in other learning situations, one would expect it to be dissociable from declarative Long-Term Memory.

That said, different memory systems can certainly interfere with each other during consolidation or support each others when the memories share a structure (see Robertson, 2022, for a review). However, given that the format of the representations created by statistical learning might differ from that used for linguistic stimuli (Endress & Langus, 2017; Fischer-Baum, Charny, & McCloskey, 2011; Miozzo, Petrova, Fischer-Baum, & Peressotti, 2016), it is at least an open question to what extent statistical learning supports declarative memories for words. In the General Discussion, we will discuss ways in which statistical learning might be useful for word learning even if it is dissociable from declarative memory.

In addition to possible dissociations between statistical learning and declarative memory, it is also unclear how continuous fluent speech really is. In fact, due to its prosodic organization, speech does not come as a continuous signal but rather as a sequence of smaller units (Cutler, Oahan, & van Donselaar, 1997; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996). This prosodic organization is perceived in unfamiliar languages (Brentari, González, Seidl, & Wilbur, 2011; Endress & Hauser, 2010; Pilon, 1981) and even by newborns (Christophe, Mehler, & Sebastian-Galles, 2001). It might affect the usefulness of statistical learning, because such speech cues tend to override statistical cues (Johnson & Jusczyk, 2001; Johnson & Seidl, 2009), and because statistical learning primarily operates *within* rather than across major prosodic boundaries (Shukla, Nespor, & Mehler, 2007; Shukla, White, & Aslin, 2011). As a result, the learner's segmentation task is not so much to integrate distributional information over long stretches of continuous speech, but rather to decide whether the correct grouping in prosodic groups such as "*thebaby*" is "*theba* + *by*" or "*the* + *baby*" (though prosodic groups are often longer than just three syllables; Nespor & Vogel, 1986).

1.2 Statistical learning in continuous sequences and discrete chunks

If statistical learning mainly supports predictive processing, it might also operate predominantly under conditions that are conducive for prediction, and associations among syllables might form more easily when the syllables are part of a continuous sequence compared to when they are packaged into discrete items (e.g., through prosodic phrasing); after all, longer, continuous sequences provide more information on which predictions can be based than shorter chunks.

Preferential statistical learning in continuous sequences would be one of numerous examples where statistical learning works better over some stimulus classes than others. The classic example is taste aversion, where rats readily associate tastes with sickness and external stimuli with pain but

cannot associate taste with pain or external stimuli with sickness (Garcia, Hankins, & Rusiniak, 1974; L. T. Martin & Alberts, 1979; Alberts & Gubernick, 1984); other examples include associations of objects with landmarks vs. boundaries (Doeller & Burgess, 2008), associations among social vs. non-social objects (Tompson, Kahn, Falk, Vettel, & Bassett, 2019), and associations among consonants vs. vowels (Bonatti, Peña, Nespor, & Mehler, 2005; Toro, Bonatti, Nespor, & Mehler, 2008).[1]

The hypothesis that statistical learning predominantly supports predictive processing thus raises the possibility that it might operate predominantly in continuous rather than discrete sequences. Conversely, discrete chunks might be more conducive for the formation of declarative memories, because such chunks have clear onsets and offsets, which appears to be a key requirement of the memory representations of linguistic stimuli (Endress & Langus, 2017; Fischer-Baum et al., 2011; Miozzo et al., 2016). The importance of discrete chunks for word learning is supported by the finding that a word-segmentation model relying just on information at the edges of discrete chunks (in the form of utterance boundaries) performed better than most other word-segmentation models (Monaghan & Christiansen, 2010), and that statistical information does not always lead to better performance when boundary information is provided (Sohail & Johnson, 2016).

In fact, statistical learning is typically explored with continuous sequences. Participants are familiarized with speech sequences consisting of random concatenations of non-sense "words" (or equivalent units in other modalities). As a result, syllables within words are more predictive of one another (and have higher TPs) than syllable combinations that straddle word boundaries. Following such a familiarization, (adult) participants typically complete a two-alternative forced-choice recognition task, where they have to choose between the words from speech stream and part-words. Part-words are tri-syllabic items that straddle a word boundary. For example, if *ABC* and *DEF* are two consecutive words, *BCD* and *CDE* are the corresponding part-words. Participants tend to choose words over part-words, suggesting that they are sensitive to the greater predictiveness (and TPs) of syllables within words. However, such results still leave open the question of whether participants can use this sensitivity to memorize words from fluent speech, and whether this sensitivity would be present in discrete sequences.

Some evidence suggests that learners might process continuous speech sequences differently from discrete ones (e.g., Endress & Bonatti, 2016; Marchetto & Bonatti, 2015; Peña, Bonatti, Nespor, & Mehler, 2002). For example, Peña et al. (2002) familiarized participants with continuous speech streams as well as with discrete, "pre-segmented" speech streams, in which each word was followed by a brief silence. The brief silences triggered additional processes such as rule-like generalizations that were unavailable after continuous familiarizations. Critically, the rule-like generalizations observed after pre-segmented familiarizations might reflect memory processes. Endress and Mehler (2009a) suggested that the role of the silences was to act as Gestalt-like grouping cues that provided learners with the location of the word edges (i.e., onsets and offsets), and thus enabled generalizations based on those word-edges (see also Glicksohn & Cohen, 2011; Morgan, Fogel, Nair, & Patel, 2019 for other perceptual grouping effects in statistical learning). Given that the grouping cues resulted in a sequence of discrete chunks, the grouping cues might also support declarative memory processing.

[1] This is not to say that statistical learning evolved *for* specific computations; statistical learning might still be a "spandrel" (Gould, Lewontin, Maynard Smith, & Holliday, 1979) that evolved as a side effect of local neural processing and might undergo positive, negative or no selection in different brain pathways.

1.3 The current experiments

Here, we explore the computational function of statistical learning in word-segmentation. In Experiment 1, we ask if statistical learning leads to declarative memory of words. We exposed (adult) participants to the speech stream from Saffran, Aslin, and Newport's (1996) classic word-segmentation experiment. The speech stream consists of four non-sense words randomly concatenated into a continuous speech sequence. As a result, TPs among syllables are higher within words than across word-boundaries. We presented the stream either as a continuous sequence (as in Saffran, Aslin, and Newport's (1996) experiments), or as a pre-segmented sequence of words, with brief silences across word boundaries. As mentioned above, these continuous vs. pre-segmented presentation modes trigger different sets of memory processes (Endress & Bonatti, 2016; Marchetto & Bonatti, 2015; Peña et al., 2002), but it is unknown if either of these processes involves declarative memory. Following this familiarization, we simply asked participants to recall what they remembered from the speech stream. In light of the finding that participants in statistical learning tasks sometimes endorse items they have never encountered (e.g., Endress & Wood, 2011; Turk-Browne & Scholl, 2009; Jones & Pashler, 2007) and can endorse them over items they *have* encountered (Endress & Langus, 2017; Endress & Mehler, 2009b; Endress, under review), we expected that participants would form declarative memories only after a pre-segmented familiarization.

To foreshadow our results, participants could recall items after a pre-segmented familiarization, but tended to produce *incorrect* items after a continuous familiarization. We then verified that a prominent statistical learning model based on memories for chunks (Perruchet & Vinter, 1998) cannot explain these data.

As these results suggest that learners do not remember items from continuous statistically structure streams and cannot even identify word boundaries, we then reconsider the strongest evidence purportedly supporting memory-based accounts of statistical learning, and report neural network simulations showing that this evidence can be explained by memory-less Hebbian learning mechanisms.

Finally, in Experiment 2, we asked whether statistical learning operates in smaller chunks such as those that might be encountered due to the prosodic organization of language, or only in longer stretches of continuous speech. Participants listened to a speech sequence of tri-syllabic non-sense words. As in Experiment 1, the words were either *pre-segmented* (i.e., with a silence after each word) or continuously concatenated.

For half of the participants, both the TPs and the chunk frequency was higher between the first two syllables of the word than between the last two syllables (TPs of 1.0 vs. .33). A statistical learner should thus split triplets like *ABC* into an initial *AB* chunk followed by a singleton *C* syllable (hereafter *AB+C* pattern). For the remaining participants, both the TPs and the chunk frequency favored an *A+BC* pattern. To make the learning task as simple as possible, the statistical pattern of the words was thus consistent for each participant. Following this familiarization, participants heard pairs of *AB* and *BC* items, and had to indicate which item was more like the familiarization items. If statistical learning predominantly operates in continuous rather than pre-segmented sequences, participants should split the triplets into their underlying chunks only after continuous but not pre-segmented familiarizations.

To preview our results, while Experiment 1 revealed that participants remember words only after listening to pre-segmented speech sequences, in Experiment 2, participants predominantly tracked TPs in continuous speech sequences, but less so in pre-segmented sequences.

2 Experiment 1: Do learners remember items in a statistical learning task?

In Experiment 1, we asked if participants would remember the items that occurred in a speech

stream. Adult participants listened to the artificial languages from Saffran, Aslin, and Newport's (1996) Experiment 2 with 8-months-old infants, except that, to increase the opportunity for learning the statistical structure of the speech stream, we doubled the exposure to 90 repetitions of each word.[2] The languages comprised four trisyllabic words, with a TP of 1.0 within words and 0.33 across word boundaries. The words were presented in a continuous stream or as a presegmented word sequence. We ran a lab-based version of the experiment (Experiment 1a) and an online replication with a larger sample size (Experiment 1b). As the results of both experiments were similar, we present them jointly.

Following a retention interval, participants had to repeat back the words they remembered from the speech stream.[3] Lab-based participants responded vocally, while online participants typed their answers into a comment field. Finally, participants completed a recognition test during which we pitted words against part-words. Part-words are tri-syllabic items that straddle a word-boundary. For example, if *ABC* and *DEF* are two consecutive words, *BCD* and *CDE* are the corresponding part-words. If participants reliably choose words over part-words, they must be sensitive to TPs (even though such a sensitivity might arise from different mechanisms).

We also asked if a prominent chunking model of word segmentation (Perruchet & Vinter, 1998) can account for the results presented here.

2.1 Materials and methods

2.1.1 Participants. As we had no prior expectation about the effect size, we targeted a sample of at least 30 participants for each of the conditions (i.e., continuous vs. pre-segmented × Language 1 vs. Language 2, see below) in the (laboratory-based) Experiment 1a. This number was chosen because it is realistic in the time-frame available for a third-year honors project. In the (online) Experiment 1b, we tested 50 participants per language and segmentation condition. Participants reported to be native speakers of English, but we did not further assess their English proficiency. At least in Experiment 1a, participants were most likely exposed to English from childhood, as the experiment took place in London, UK, and the experimenters did not notice any clear non-native accents.

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 = 8 in Experiment 1a; N = 11 in Experiment 1b). Given that our aim was to assess the role of statistical learning in the formation of declarative LTM representations of words, we restricted our analysis to participants who were most likely to have engaged in the statistical learning task.

Another 11 participants were excluded from Experiment 1b 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.[4] In Ex-

[2]We doubled the exposure with respect to Saffran, Aslin, and Newport's (1996) infant studies to maximize the chance of observing successful learning, given that even the experimenters found the learning task challenging with the stimuli from Saffran, Newport, and Aslin's (1996) (adult) experiment.

[3]Given that the focus of our experiments is the potential usefulness of statistical learning for placing items into declarative memory, we introduced a brief retention interval to mimic slightly longer-term retention than in typical statistical learning studies (but see e.g. Karaman & Hay, 2018; Vlach & DeBrock, 2019).

[4]When participants produce excessively long items (e.g., *takahsakakakaratatataikokokokotatakatakatakatakatakatakataka*, *matikulatatitulapapitularimatitulaatitula*), it can take our recursive parsing algorithm (see below) a substantial amount of computing time to generate all possible matches to the speech stream. When the analysis of a single participant exceeded several days of calculations, we decided to remove this participant from analysis. Critically, and as mentioned above, the productions for

periment 1b, 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. As a result, any differences between the continuous and the pre-segmented conditions were not just a consequence of differences in statistical power. (This was not necessary in the within-participant design of Experiment 1a.) The final sample included 26 participants in the lab-based version (Experiment 1a), and 152 in the online version (Experiment 1b). Demographic information is given in Table 1. With the exception of the exclusions due to excessive computing time (which we did not anticipate), the exclusion criteria were set forth prior to analysis.

2.1.2 Materials. We re-synthesized the languages used in Saffran, Aslin, and Newport's (1996) Experiment 2. The four words in each language are given in Table 2. Each word was composed of three syllables, which were composed of two segments in turn. Stimuli were synthesized using the us3 (male American English) voice[5] of the mbrola synthesizer (Dutoit, Pagel, Pierret, Bataille, & van der Vreken, 1996), at a constant F_0 of 120 Hz and at a rate of 216 ms per syllable (108 ms per phoneme). This syllable duration is comparable to that in Saffran, Aslin, and Newport (1996) (222 ms per syllable).

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. For continuous streams, each randomization was then synthesized into a continuous speech stream (with no silences between words) using mbrola (Dutoit et al., 1996) and then converted to mp3 using ffmpeg (`https://ffmpeg.org/`). For pre-segmented streams, words were synthesized in isolation. Each randomization was then used to concatenate the words into a pre-segmented stream, with silences of 222 ms between words, 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. Syllable transitions had TPs of 1.0 within words and 0.33 across word boundaries. We created 20 versions of each stream with different random orders of words.

As the role of the silences in the pre-segmented stream was to create clearly identifiable chunks, the silence duration was chosen to result in clearly perceptible syllable groups (according to the experimenters' perception). Other investigations with pre-segmented material used shorter silences (e.g., Peña et al., 2002), longer ones (e.g., Sohail & Johnson, 2016; Endress & Mehler, 2009a) or natural prosodic phrasing (Shukla et al., 2007; Seidl & Johnson, 2008). Relatedly, other experiments mimicking the prosodic organization of speech used natural prosodic phrasing (Shukla et al., 2007; Seidl & Johnson, 2008) or grouped several "words" together using silences (Sohail & Johnson, 2016). In the light of Experiment 2, where we ask if statistical learning can be used to break up small prosodic groups such as "thebaby" into their underlying words (i.e., "the+baby"), we follow Peña et al. (2002) and present silences after each word instead of inducing longer groupings.

For the online Experiment 1b, the speech streams were combined with a silent video with no clear objects to increase attention to the stimuli. We used a panning of the Carina nebula, obtained from `https://esahubble.org/videos/heic0707g/`. The video was combined with the speech streams using the muxmovie utility.

2.1.3 Apparatus. The lab-based Experiment 1a was run using Psyscope X (`http://psy.ck.sissa.it`) in a quiet room. The online Experiment 1b was run on

which this occurred did not resemble the statistically defined words in the first place.

[5]Experiment 1 was chronologically carried out after Experiment 2, but we changed the order for readability. We chose the us3 voice because the alternative en1 (British English) voice introduced artifacts in Experiment 2a.

Table 1
Demographics of the final sample in Experiments 1 and 2. In Experiment 1a, the (lab-based) participants completed both segmentation conditions. In Experiment 2b, we conducted two independent replications with the same American English voice due to unexpected results with the British English voice in Experiment 2a.

Sequence Type	Voice	N	Females	Male	Age (*M*)	Age (range)
Experiment 1a: Lab-based recall experiment						
continuous	us3	13	13	0	19.2	18-22
pre-segmented	us3	13	13	0	19.2	18-22
Experiment 1b: Online recall experiment						
continuous	us3	76	26	50	30.7	18-71
pre-segmented	us3	76	15	61	28.9	18-62
Experiment 2a – Lab-based segmentation experiment (British English voice)						
pre-segmented	en1	30	22	8	25	18-42
continuous	en1	30	20	10	23.9	18-45
Experiment 2b – Lab-based segmentation experiment (American English voice)						
pre-segmented	us3	30	18	12	26.3	18-43
continuous	us3 (1)	32	26	6	20.1	18-44
continuous	us3 (2)	30	20	10	23.2	18-36

`https://testable.org`.

2.1.4 Procedure.

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

In the online Experiment 1b, participants watched a video with no clear objects during the familiarization.

Following the familiarization, there was a 30 s retention interval. In both Experiment 1a and 1b, participants were instructed to count backwards from 99 in time with a metronome beat at 3s per beat. Performance was not monitored. Given that our objective was to investigate the role of statistical learning in the formation of declarative LTM representations of words, we attempted to make our memory tests at least somewhat long-term by introducing this filled retention interval.

2.1.4.2 Recall test. Following the retention interval, participants completed the recall test. In Experiment 1a, participants had 45 s to repeat back the words they remembered; their vocalizations were recorded using ffmpeg and saved in mp3 format. In Experiment 1b, participants had 60 s to type their answer into a comment field, during

Table 2
Languages used Experiment 1. The words are the same as in Experiment 2 in Saffran, Aslin, and Newport (1996).

L1	L2
pabiku	bikuti
tibudo	pigola
daropi	tudaro
golatu	budopa

which they viewed a progress bar.

2.1.4.3 Recognition test. Following the recall test, participants 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.

2.1.5 Analysis strategy. As we used performance in the recognition test to restrict the analysis to those participants most likely to have engaged in statistical learning, performance in the recognition test in the final sample is not representative of the whole sample, and is thus not compared to a chance level. Therefore, we focus on the participants' recall responses.

It turned out that the written recall responses required substantial pre-processing because participants transcribed syllables using different orthographies and misperceived some phonemes, among other inconsistencies. The detailed analysis procedure is described in in Supplementary Material SM1. All analytic choices were made to maximize the correspondence between the participants' responses and the syllable sequences attested in the speech stream.

In brief, the responses were first transformed using a set of substitutions rules to allow for misperceptions (e.g., confusion between /b/ and /p/) or orthographic variability (e.g., *ea* and *ee* both reflect the sound /i/).

Second, the responses were segmented into their underlying units. This was necessary because some participants separated only words by spaces, while others separated syllables by spaces, and groups of syllables (e.g., words) by other characters (e.g., commas). For example, responses such as *bee coo tee,two da ra,bout too pa* likely reflected the words *bikuti*, *tudaro* and *budopa*.

Third, we applied another set of substitution rules to allow for other misperceptions.

Finally, we selected the best matches to the familiarization stimuli. We selected these matches by (1) maximizing the length of the match and (2) minimizing the number of substitutions with respect to the original responses.

Based on these matches, we calculate a various properties of these matches (see Table S2). For readability, we will introduce these measures in the Results section. Exclusion criteria for responses with unattested syllables are given in Supplementary Material SM1.4.

In Experiment 1a, the (lab-based) participants' verbal responses were recorded and transcribed by two independent observers. Disagreements were resolved by discussion.[6] Online participants typed their responses directly into a comment box. Analysis of these responses was fully automatic (see below).

We use likelihood ratios to provide evidence for the various null hypotheses. Following Glover and Dixon (2004), we fit the participant averages to (i) a linear model comprising only an intercept and (ii) the null model fixing the intercept to the appropriate baseline level, and evaluated the likelihood of these models after correcting for the difference in the number of parameters using the Bayesian Information Criterion.

2.2 Results

2.2.1 Analysis of the participants' productions. We present the results in three steps. First, we report some general measures of the recall items to show that participants engage in the task and track TPs in both the continuous and the pre-segmented condition. Second, we ask whether participants are more likely produce words than part-words. Third, we ask whether participants know where words start and where they end.

Descriptives, comparisons to chance levels as well as comparisons between the continuous and the pre-segmented conditions are given in Table 3.

[6] The number of disagreements can no longer be recovered.

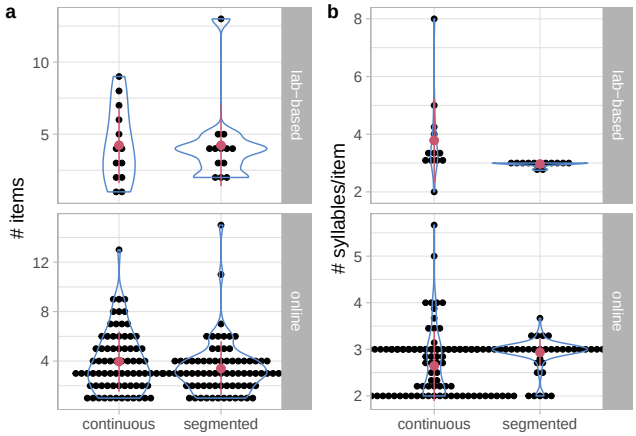

Figure 1. Number of items produced and number of syllables per item in the recall phase of Experiments 1a (top) and 1b (bottom).

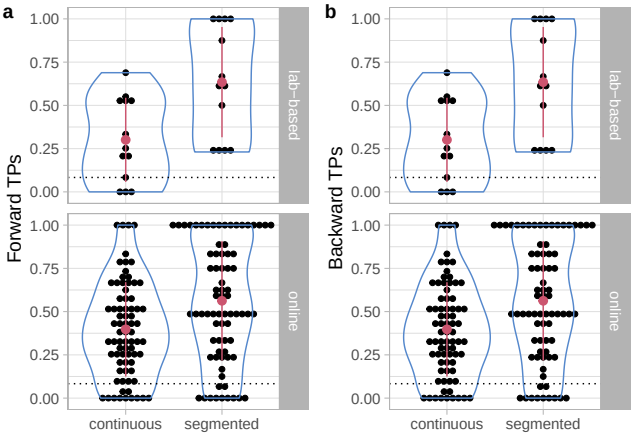

Figure 2. Forward and backward TPs in the participants' productions in the recall phase of Experiments 1a (top) and 1b (bottom). The dotted line represents the chance level for a randomly ordered syllable sequence.

2.2.1.1 General measures: Do participants engage in the task? As shown in Table 3 and Figures 1a and b, participants produced about 4 items. Neither the number of items produced nor their lengths differed across the segmentation conditions. Critically, and as shown in Table 3 and Figures 2a and b, forward and backward TPs in the participants' responses were significantly greater than the chance level of .083 in both segmentation conditions. These TPs were greater in the pre-segmented condition. These TPs likely underestimate the participants' actual performance, as we included responses with unattested syllables that might reflect misperceptions (and thus lower TPs); after removing such responses, TPs in the participants' responses were about twice as large. Participants were thus clearly sensitive to the TPs in the speech stream.

We next examined the production of two-syllable chunks. Such chunks can be either high-TP chunks (if they are part of a word) or low-TP chunks (if they straddle a word boundary). For example, with two consecutive words *ABC* and *DEF*, the high-TP chunks are *AB*, *BC*, ..., while the low-TP chunk is *CD*. As a result, two-syllable items have a 66% probability of being a high-TP chunk. As shown in Figure 3b, the proportion of high-TP among chunks high- and low-TP chunks exceeded chance in both the pre-segmented condition and the continuous condition in Experiment 1b (though not in the continuous condition of Experiment 1a), with a significantly higher proportion in the pre-segmented versions. These results thus confirm that participants are sensitive to TPs or high frequency chunks (which are confounded in the current design).

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

The traditional analysis of word segmentation experiments relies on the contrast between words and part-words. As mentioned above, part-words are tri-syllabic items that straddle a word-boundary. We thus calculated the proportion of words among words and part-words recalled by the participants. If participants faithfully produce tri-

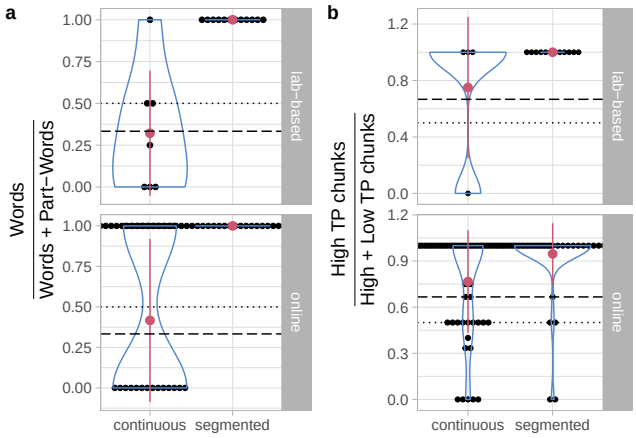

Figure 3. Analyses of the participants' productions in the recall phase of Experiments 1a (top) and 1b (bottom). (a) Proportion of words among words and part-words. The dotted line represents the chance level of 50% in a two-alternative forced-choice task, while the dashed line represents the chance 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.

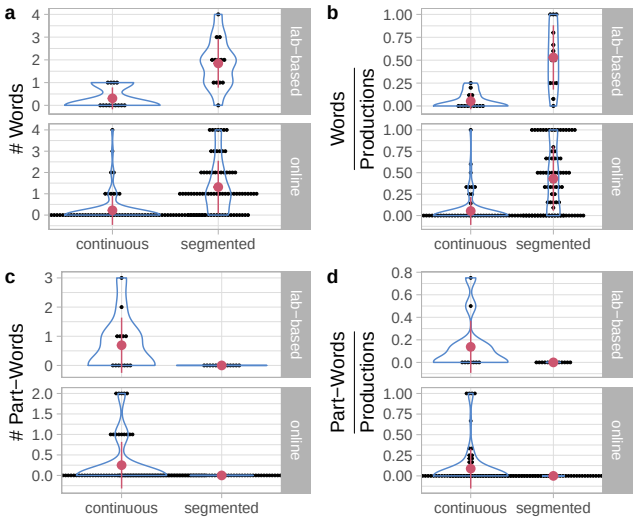

Figure 4. Number and proportion (among vocalizations) of words and part-words in the recall phase of Experiments 1a (top) and 1b (bottom).

syllabic sequences from the stream, they can start the sequences on the first, second or third syllable of a word, but only the first possibility yields a word rather than a part-word. As a result, if participants initiate their productions with a random syllable, a third of their productions should be words.

As shown in Table 3 and in Figure 3a, the proportion of words among words and part-words was close to 100% in the pre-segmented conditions, but did not differ from the chance level of 33% in the continuous conditions. This difference was statistically significant. Likelihood ratio analysis suggests that, in the continuous condition of Experiment 1b, participants were 3.5 times more likely to perform at the chance level of 33% than to perform at a level different from chance; in Experiment 1a, the likelihood ratio was 2.6. These results thus suggest that participants in the continuous condition initiate their productions at random positions in the stream, and that they do not remember any word forms.

However, inspection of Figure 3a shows that the distribution in the continuous condition is bimodal, with some participants producing only words, and others producing only part-words. Such a behavior can arise if participants pick a syllable as their starting-point, and segment the rest of the stream accordingly. If they happen to pick a word-initial syllable, they will produce only words; if they pick the second or the third syllable of a word, all subsequent items will be part-words.

Assuming that the number of participants producing words vs. part-words is binomially distributed, we calculated the likelihood ratio of a model where learners identify word boundaries (and should produce words with probability 1), and a model where they track TPs and initiate

productions at random positions (and should produce words with a probability of 1/3). As shown in SM4, the likelihood ratio in favor of the first model is 3^{N_W} if participants produce no part-words (i.e., after a pre-segmented familiarization), where N_W is the number of participants producing words; otherwise, the likelihood ratio in favor of the second model is infinity. Given that the overwhelming majority of participants produce words only after a pre-segmented familiarizations, these results thus suggest that, despite their ability to track TPs, participants initiate productions at random positions in the sequence, and thus do not remember statistically defined words.

However, as shown in Figure 4, these results might be misleading because, in the continuous condition, many participants produce neither words *nor* part-words. In fact, on average, they produce only .4 words and part-words combined, respectively. (In the pre-segmented condition, most participants produce at least one word, with an average of 1.26.)

We thus turn to the question of whether participants know where words start and end, asking if participants produce correct initial and final syllables.

2.2.1.3 Do participants know where words start and where they end? If participants use statistical learning to remember words, they should know where words start and where they end. In contrast, if they just track TPs, they should initiate the responses with random syllables. As there are four words with one correct initial and final syllable each, and 12 syllables in total, $4/12 = 1/3$ of the productions should have "correct" initial syllables, and $1/3$ should have correct final syllables. Given that participants tend to produce high-TP two-syllable chunks (i.e., *AB* and *BC* rather than *CD* chunks), the actual baseline level is somewhat higher.[7] However, to evaluate the group performance, we keep the baseline of $1/3$.

As shown in Table 3 and Figure 5a and b, participants produced items with correct initial or final syllables at greater than chance level only in the pre-segmented conditions, but not in the continuous conditions. In the continuous condition of Experiment 1b, the likelihood ratio in favor of the null hypothesis was 0.785 for initial syllables and 4.06 for final syllables; in Experiment 1b, the likelihood ratios are 3.61 and 2.14, respectively. While it is possible that performance in the continuous condition might exceed the chance-level of 1/3 with more than the 78 participants currently included, the actual chance-level is somewhat higher (about 38.4%). Critically, only 42% of the productions have a correct initial syllable, which is unexpected if participants knew where words start and where they end. Together with the finding that the overwhelming majority of participants produce no word at all, these results thus suggest that TPs do not allow learners to reliably detect onsets and offsets of words.

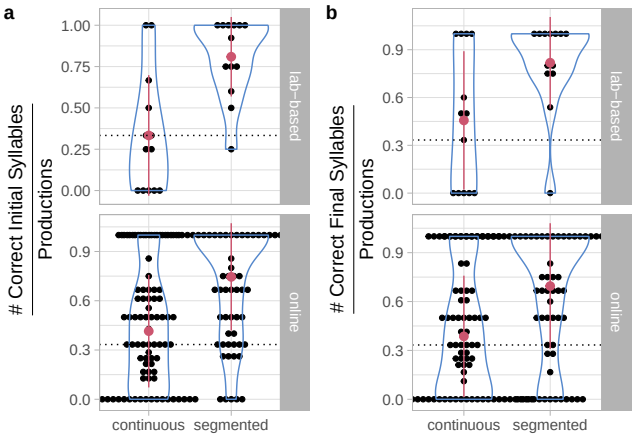

Figure 5. Analyses of the participants' productions in the recall phase of Experiments 1a (top) and 1b (bottom). (a) Proportion of productions with correct initial syllables and (b) with correct final syllables. The dotted line represents the chance level of 33%.

[7] For example, participants in the continuous condition produce about 75% high-TP chunks; if they initiate their productions with high-TP chunks, one would expect them to produce about $75\%/2 = 3/8$ rather than $1/3$ items with correct initial syllables.

Table 3

Main 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	Pre-segmented	*p*(continuous vs. pre-segmented)
Number of items			
lab-based (Exp. 1a)	*M*= 4.23, *SE*= 0.756, *p*= 0.0016	*M*= 4.23, *SE*= 0.818, *p*= 0.00152	0.812
online (Exp. 1b)	*M*= 4.03, *SE*= 0.292, *p*= 3.17e-14	*M*= 3.25, *SE*= 0.202, *p*= 2.74e-14	0.099
Number of syllables/item			
lab-based (Exp. 1a)	*M*= 3.79, *SE*= 0.421, *p*= 0.0016	*M*= 2.97, *SE*= 0.0246, *p*= 0.0007	0.026
online (Exp. 1b)	*M*= 2.65, *SE*= 0.0869, *p*= 2.29e-14	*M*= 2.93, *SE*= 0.0364, *p*= 1.04e-15	< 0.001
Forward TPs			
lab-based (Exp. 1a)	*M*= 0.301, *SE*= 0.0702, *p*= 0.0107	*M*= 0.634, *SE*= 0.092, *p*= 0.00159	0.006
online (Exp. 1b)	*M*= 0.397, *SE*= 0.0316, *p*= 6.26e-12	*M*= 0.583, *SE*= 0.04, *p*= 3.82e-13	0.001
Backward TPs			
lab-based (Exp. 1a)	*M*= 0.301, *SE*= 0.0702, *p*= 0.0107	*M*= 0.634, *SE*= 0.092, *p*= 0.00159	0.006
online (Exp. 1b)	*M*= 0.397, *SE*= 0.0316, *p*= 6.26e-12	*M*= 0.583, *SE*= 0.04, *p*= 3.82e-13	0.001
Proportion of High-TP chunks among High- and Low-TP chunks			
lab-based (Exp. 1a)	*M*= 0.75, *SE*= 0.289, *p*= 0.85 (vs. 2/3)	*M*= 1, *SE*= 0, *p*= 0.0006 (vs. 2/3)	1.000
online (Exp. 1b)	*M*= 0.767, *SE*= 0.0459, *p*= 0.00154 (vs. 2/3)	*M*= 0.97, *SE*= 0.0187, *p*= 6.75e-13 (vs. 2/3)	< 0.001
Proportion of words among words and part-words (or concatenations thereof)			
lab-based (Exp. 1a)	*M*= 0.321, *SE*= 0.153, 0.798 (vs. 1/3)	*M*= 1, *SE*= 0, *p*= 0.0006 (vs. 1/3)	0.034
online (Exp. 1b)	*M*= 0.417, *SE*= 0.105, *p*= 0.189 (vs. 1/3)	*M*= 1, *SE*= 0, *p*= 2.08e-13 (vs. 1/3)	< 0.001
Proportion of items with correct initial syllables			
lab-based (Exp. 1a)	*M*= 0.333, *SE*= 0.105, *p*= 0.856	*M*= 0.809, *SE*= 0.0694, *p*= 0.00186	0.016
online (Exp. 1b)	*M*= 0.419, *SE*= 0.0392, *p*= 0.0864	*M*= 0.738, *SE*= 0.0387, *p*= 1.58e-11	0.000
Proportion of items with correct final syllables			
lab-based (Exp. 1a)	*M*= 0.456, *SE*= 0.125, *p*= 0.5	*M*= 0.818, *SE*= 0.0829, *p*= 0.00222	0.025
online (Exp. 1b)	*M*= 0.386, *SE*= 0.043, *p*= 0.456	*M*= 0.7, *SE*= 0.0437, *p*= 4.14e-10	0.000

2.2.2 Can chunking models account for these results? Taken together, the results of Experiment 1 suggest that participants can learn statistical information from fluent speech. However, the information they retain does not allow them to learn (statistically defined) chunks that might then be encoded as word candidates in declarative long-term memory. Rather, few participants produced any words or part-words at all, and among those participants who produced such items, only one-third produced words. Further, only about a third of the participants produced items starting with word-initial syllables, while two-thirds produced items starting with word-medial or word-final syllables. Such results suggest that statistical learning does not support the very function for which it was motivated originally – to identify word boundaries in fluent speech, and thus to learn words from fluent speech.

Given the debate about whether statistical learning entails memories for chunks (see e.g. Perruchet, 2019 vs. Endress et al., 2020 and General Discussion), we illustrate the conclusion that chunking models will not produce part-words rather than words. Specifically, in SM6, we report simulations with PARSER (Perruchet & Vinter, 1998), a prominent chunking model of word segmentation, (see also Endress & Langus, 2017, for related simulations), where we attempt to bias the model to prefer part-words over words. However, despite our attempt to bias the model, it never preferred part-words to words.

Given that, in our recall experiment, the majority of those participants who produced either words or part-words produced part-words, these results suggest that chunking models (or at least at least one rather prominent chunking model) either cannot account for the current results, or, to the extent that other chunking models might account for them, that these models learn information that does not allow them to recover word boundaries from fluent speech.

Critically, such models would also need to account for the fact that participants produce part-

words even when they prefer words in a recognition test. As a result, while it might be possible to create chunking model that produce part-words (even though this would contradict their original purpose),[8] such models are unlikely to simultaneously prefer words in a recognition test. After all, the preferences of chunking models are driven by those chunks with the strongest memory representations. If these chunks happen to be words, the models will prefer words in both recognition and recall; if they are part-words, the models will prefer part-words, again in both recognition and recall. As a result, we believe that the current results are fundamentally incompatible with chunking models of statistical learning.

2.2.3 Relations between recall and recognition. The results so far suggest that the information extracted in statistical learning tasks does not allow participants to identify word boundaries. Further, the pattern of performance is unlikely to be explained by chunking models of word segmentation. As mentioned above, such models are driven by the memory strength of those chunks they happen to have memorized. As a result, even if it is possible to bias such models to prefer low-probability items, it is unclear how such models could prefer words over part-words in a recognition test (and thus have stronger memory traces of words), and simultaneously produce part-words rather than words in a recall test (and thus have stronger memory traces of part-words).

That being said, statistical learning performance (as measured in the recognition test) might still be related to memory for word candidates (as measured by the participants' productions), albeit indirectly. For example, and as mentioned above, participants might focus on particular individual syllables, and preferentially track statistics around those syllables they happen to focus on.

Given that attention affects statistical learning (e.g., Turk-Browne, Jungé, & Scholl, 2005; Toro, Sinnett, & Soto-Faraco, 2005), focusing on particular syllables might also direct the participants' attention and thus what they learn from the streams. For example, if participants happen to focus on word-medial or word-final syllables, they would also focus on statistically less cohesive syllable sequences as a result. Conversely, if participants happen to focus on word-initial syllables, they would also focus on statistically more cohesive syllables. This, in turn, which might affect recognition performance: Those participants who produced part-words might have focused on those syllables at the beginning of part-words, and those who produced words might have focused on word-initial syllables, and the syllables participants happen to focus on might be chosen randomly.

Critically, while our evidence does not allow us to decide whether participants focused on particular syllables, such views would imply that, in Experiment 1, most participants focused on other syllables than word-initial syllables, given that two thirds of the participants produced part-words rather than words. While we show in SM5 that recall performance is related to recognition performance, any memory-based views would thus still imply that statistical learning does not lead to memories of high probability sequences in most participants, and rather to memories of low-probability sequences, which would make statistical learning unsuitable for word learning in turn.

Further, and as mentioned above, most participants produced part-words even while preferring words in a recognition test. It is thus unclear how the same memory-based mechanisms can show different preferences in recall and recognition.

2.3 Discussion

Experiment 1 provided the first direct test of the contents of the participants' (episodic or semantic) declarative memory after exposure to a statistical learning task. The results suggest that, even when participants successfully track statistical informa-

[8] For example, it is possible to add an "attentional" component that forces the model to start chunks with word-medial syllables. We are grateful to a reviewer for pointing out this possibility.

tion, they remember familiarization items only when familiarized with a pre-segmented sequence. In contrast, when familiarized with a continuous sequence, their productions start with random syllables rather than actual word onsets. Given that the memory representations of linguistic items are based on their initial and final syllables (Endress & Langus, 2017; Fischer-Baum et al., 2011; Miozzo et al., 2016), these results thus suggest that statistical learning did not lead to the creation of declarative memory representations.

Contrary to this conclusion, some authors suggest that statistical learning might lead to declarative memories for chunks (Graf-Estes et al., 2007; Hay, Pelucchi, Graf Estes, & Saffran, 2011; Isbilen et al., 2020). Such experiments generally proceed in two phases. During a statistical learning phase, participants are exposed to some statistically structured sequence. Then, they are exposed to items presented in isolation, and show some processing advantage for isolated high-probability items compared to isolated low-probability items. However, we proposed that such experiments have a two-step explanation that does not involve declarative memory (Endress & Langus, 2017). First, during the statistical learning phase, participants acquire statistical knowledge without remembering any specific items. When experimenters subsequently provide participants with *isolated* chunks, the accumulated statistical knowledge facilitates processing of the experimenter-provided chunks (e.g., due to predictive processing), without these chunks having been acquired before being supplied by the experimenter. In contrast to such indirect designs, we provide a direct measure of declarative knowledge of sequence items, and show that participants do not form declarative memories of sequence items unless the sequence is pre-segmented.

Another major argument for a role of declarative memory in statistical learning comes from the observation that learners tend to recognize entire units better than sub-units (e.g., Fiser & Aslin, 2005; Giroux & Rey, 2009; Orbán, Fiser, Aslin, & Lengyel, 2008; Slone & Johnson, 2018). We now show that such results can be explained using a simple Hebbian learning model, and propose further alternative interpretations in the General Discussion.

3 Simulation 1: Does Hebbian learning provide an alternative to memory-based theories of statistical learning?

There is considerable debate about whether statistical learning leads to memory for recurring chunks (e.g., Endress et al., 2020; Goodsitt, Morgan, & Kuhl, 1993; Perruchet, 2019; Swingley, 2005; Thiessen, 2017), and some empirical results seem to support this idea.

While most of these results have alternative interpretations (see above and General Discussion), there is one research tradition that appears to provide strong evidence in favor of a memory based theory of statistical learning.

Specifically, in some studies, recognition performance is better for (statistically defined) units compared to (statistically defined) sub-units (e.g., Fiser & Aslin, 2005; Giroux & Rey, 2009; Orbán et al., 2008; Slone & Johnson, 2018). In a word recognition analogy, hearing the word *hamster* makes it difficult to recognize that the first syllable of *hamster* is a word on its own (i.e., *ham*), though, in word recognition, the reduced availability of sub-units is at least partially driven by phonetic differences between syllables that are parts of words and syllables that are words on their own (e.g., van Alphen & van Berkum, 2010; Salverda, Dahan, & McQueen, 2003; Shatzman & McQueen, 2006a, 2006b).

Similar effects are observed in statistical learning in both vision and audition. For example, the *AB* part of an *ABC* unit is harder to recognize than a complete *CD* unit, which would suggest that the entire units are stored in memory. We now provide simulation results suggesting that such results are compatible with a memory-less Hebbian learning mechanism, but discuss this issue separately for sequential, auditory sequences and simultaneously presented visual shapes as the arguments are

somewhat different.

3.1 Units vs. sub-units in audition

As mentioned above, most statistical learning results can be explained by simple Hebbian learning: If the representation of a syllable is still active while the next one is presented, the two syllable representations are active together can thus form an association. An implementation of this idea is provided in models such as Endress and Johnson's (2021). In their neural network model, neurons are connected through both excitatory and inhibitory connections, where only the excitatory connections undergo Hebbian learning. After learning, when *B* (from *ABC*) is activated, it will excite (and inhibit) both *A* and *C* in turn. Critically, the excitatory connections between *A* and *C* are weaker than those between *A* and *B* and those between *B* and *C* (since there is less temporal overlap between their activations, and thus less Hebbian learning). This idea is illustrated in Figure 6. After an (external) activation of the neuron *A* (top), excitatory connections as well as external input to *B* will activate both *B* and *C* (bottom). Depending on the balance of excitation and inhibition between *A* and *C*, the net input on from *C* to *A* might thus be inhibitory on the next time step. In contrast, in complete two-item units, there is no extra item like *C* that could reduce the activation within the unit due to inhibition.

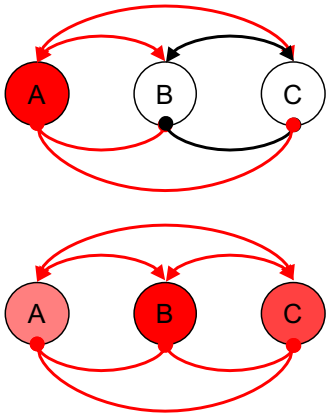

Figure 6. After an (external) activation of the neuron *A* (top), excitatory connections as well as external input to *B* will activate both *B* and *C* (bottom). Depending on the balance of excitation and inhibition between *A* and *C*, the net input on from *C* to *A* might thus be inhibitory on the next time step.

We now illustrate this point by using Endress and Johnson's (2021) model to simulate one of the first experiments showing better recognition of units compared to units Giroux and Rey's (2009) experiment. In their experiment, participants were presented with streams consisting of two three-syllable words and four two-syllable words. After such a familiarization, Giroux and Rey (2009) found better recognition for sub-units (i.e., two syllables from a three-syllable word) than for units (i.e., entire two-syllable words).

The model is a fully connected network where all neurons send both excitatory and inhibitory input to all other units. Their activations also decays over time. Critically, excitatory connections are turned using a Hebbian learning rule.

In our simulations, we randomly concatenated these words into familiarization streams with 143 occurrences of each word (matching Giroux and Rey's (2009) familiarization). We then presented the network with test items (see below) and recorded the total network activation while each item was presented, using the total activation as a measure of the network's familiarity with the test item. We tested the network for different decay rates (Λ in Endress & Johnson, 2021) and interference rates (*B* in Endress & Johnson, 2021). The cycle of familiarization and test was repeated 100 times for each parameter set, representing 100 simulated participants.

To compare the network's familiarity with two-syllable units and two-syllable sub-units, we created normalized difference scores $d = \frac{\text{Unit}-\text{Sub-unit}}{\text{Unit}+\text{Sub-unit}}$. We evaluated these difference scores against the chance level of zero using Wilcoxon tests.

As shown in Figure 7, when averaging across trials comparing two-syllable units to *AB* and *BC* sub-units, there was a significant preference for units for most parameter sets (except for some sim-

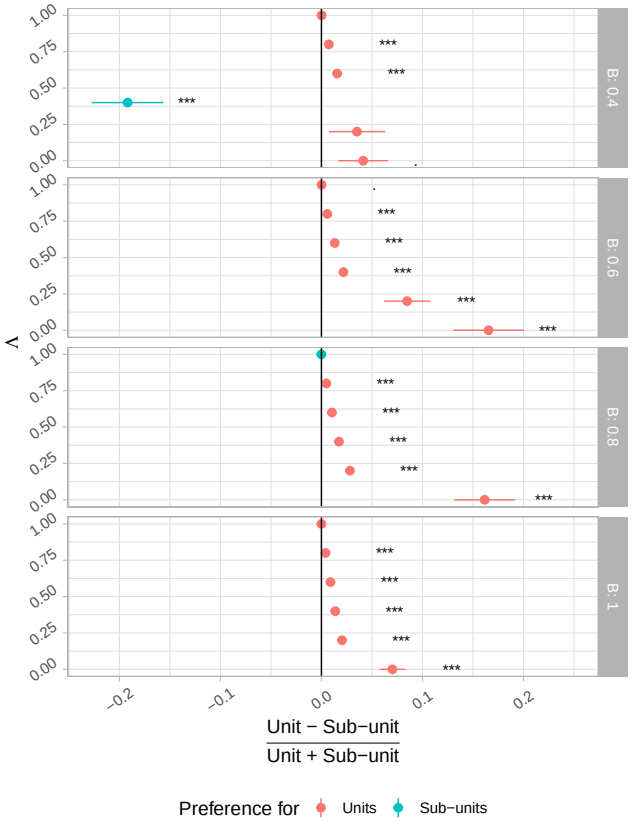

Figure 7. Normalized average difference scores of network activations after presentation of entire two-syllable units and different types of two-syllable units (i.e., AB and BC from ABC units), as a function of the forgetting rate (y axis) and the interference rate (facets in rows). As in Giroux and Rey (2009), we do not separate *AB* and *BC* sub-units. Positive values indicate stronger activations for units. Significance stars reflect a Wilcoxon test against the chance level of zero. Units generally elicit greater activation compared to the average of AB and BC sub-units. Significance labels: ***: ≤ 0.001; **: ≤ 0.01; *: ≤ 0.05; .: ≤ 0.1

ulations with low inhibition rates). A simple Hebbian network can thus account for better recognition of units compared to sub-units.

However, as shown in Figure 8, while units were systematically preferred over *AB* sub-units for most parameter values, *BC* sub-units were sometimes preferred for very low or very high interference rates. Be that as it might, the current results clearly demonstrate that a simple Hebbian network can account for the preference for units over sub-units, though the level of inhibition might need to be adequate.

To support our contention that the preference for units over sub-units might arise from the interplay between learning (and thus excitation) and inhibition, Figure 9 shows the weights between different pairs of neurons after learning. As suggested above, the connection between *A* and *C* in a three-syllable *ABC* unit is generally weaker than the other connections, and often substantially smaller than the interference rate. Depending on the parameter values, (second order) activation of *C* might thus partially suppress activation in *AB* sub-units, and activation of *A* might suppress activation in *BC* sub-units. However, the exact computational mechanisms, as well as the differences in behavior between *AB* and *BC* sub-units deserve further investigation. For the current purposes, we just conclude that the fact that a simple Hebbian learning model can account for a preference for units over sub-units demonstrates that such results do not provide evidence that units have been placed in memory, and thus do not license the conclusion that the units are stored as chunks in memory.

3.2 Units vs. sub-units in vision

The simulations reported above suggest that a simple Hebbian network can account for the preference for units over sub-units (though the level of inhibition might need to be adequate) when items are presented sequentially. As a result, such results do not provide evidence that statistical learning leads to memory for chunks.

There is also evidence that units are easier to recognize than sub-units for simultaneously presented shapes in vision (e.g., Fiser & Aslin, 2005; Orbán et al., 2008). In such experiments, shape combinations are presented simultaneously, leading to patterns of spatial statistical regularities.

However, it is unclear how reliable such effects

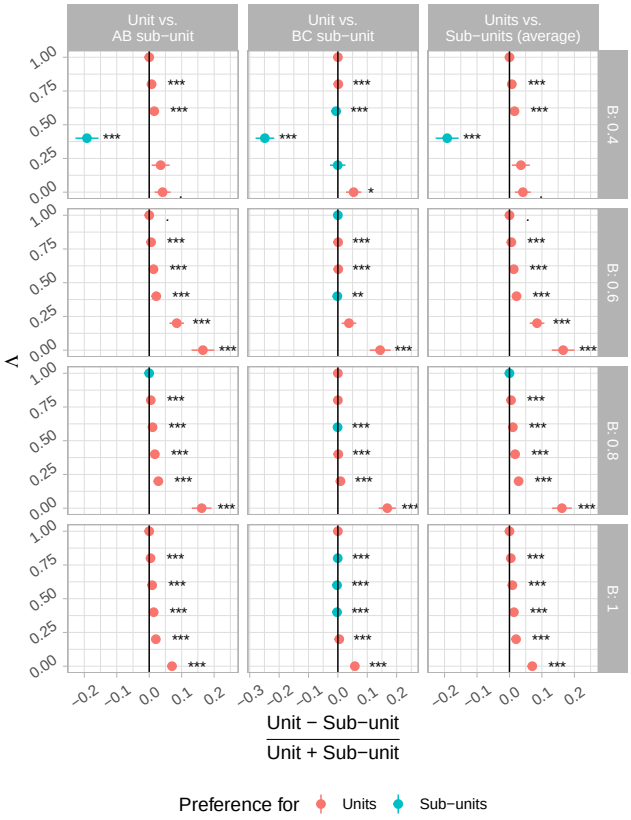

Figure 8. Normalized difference scores of network activations after presentation of entire two-syllable units and different types of two-syllable units (i.e., AB and BC from ABC units), as a function of the forgetting rate (y axis) and the interference rate (facets in rows). The rightmost column is the average of the other columns reported by Giroux and Rey (2009). Positive values indicate stronger activations for units. Significance stars reflect a Wilcoxon test against the chance level of zero. Units generally elicit greater activation compared to AB sub-units and compared to the average; when compared to BC units, the sign of the difference score depends on the parameters. Significance labels: ***: ≤ 0.001; **: ≤ 0.01; *: ≤ 0.05; .: ≤ 0.1

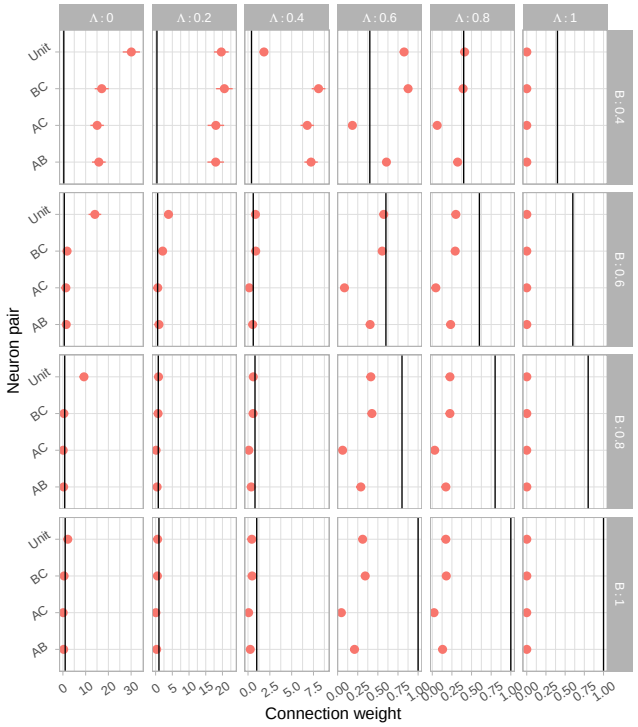

Figure 9. Connection weights between different pairs of neurons as a function of the forgetting rate (columns) and the interference rate (rows). The figure shows connection weights within a trisyllabic unit (ABC) and a bisyllabic unit (Unit). The black line represents the interference rate. The A-C connection is generally smaller than the other connections, and often substantially smaller than the interference rate.

are. For example, Fiser and Aslin (2005) observed better recognition of units in their Experiments 1 and 4, but not in their Experiment 5. Further, when presenting shapes in a sequence rather than simultaneously, Slone and Johnson (2015) also failed to find evidence for better recognition of units in their Experiment 2, where they directly contrasted the strength of representation of units vs. sub-units.

To the extent that such findings are reliable, they are consistent with a similar explanation as the sequential case above. Presumably, the strength of associations among shapes depends on their spatial distance. Further, given the ubiquity of lateral inhibitory processes in vision (Desimone & Dun-

can, 1995; Hampshire & Sharp, 2015; Kiyonaga & Egner, 2016), one would expect spatial inhibitory processes to take place in statistical learning tasks as well. As a result, one would expect a Hebbian-like model similar to the one above to reproduce better recognition of visually presented units compared to sub-units, though the temporal organization in the model above would need to be replaced with some spatial organization.

Better recognition for units compared to sub-units can thus be explained by simple Hebbian processes in the absence of the creation of memories for these units.

However, we will now suggest further alternative interpretations of a preference for units over sub-units.

3.3 Further alternative explanations of a preference for units over sub-units

In the case of *sequential* statistical learning tasks, results that units are easier to recognize than sub-units have another mutually non-exclusive alternative explanations on top of the Hebbian explanation above. This explanation is based on predictive processing. If *C* is strongly associated with *AB*, hearing an *AB* fragment *during test* might lead to a prediction error because participants expect to hear *C* (or *A* for backward predictions after hearing *BC*) even when they have no memory representation of the entire *ABC* chunk. In contrast, in entire units, there is no such prediction error. This interpretation is in line with the classic finding that tasks such as stem completion do not require declarative LTM (Graf & Mandler, 1984). *Mutatis mutandis*, participants might make predictions in test items, without any units having been placed in memory, and these predictions might affect their familiarity judgements.

In the case of *spatial* statistical learning, attentional processes provide a further alternative explanation in terms of the preference for units over sub-unit. This account relies on the spatial regions attended by participants. In unpublished results, we presented participants with simultaneously presented shape combinations, and then tested for recognition of entire units or of sub-units. We found better recognition of units than of sub-units, but only when these sub-units are located in parts of the display that do not attract attention. In contrast, when the sub-units came from salient parts of the units, recognition was as good as for units (Endress, in preparation).

Taken together, it seems reasonable to conclude that a preference for units over sub-units is not diagnostic of memory representations of the units. Rather, such results can be explained by simple and memory-less Hebbian learning mechanisms, or by the other explanations above.

4 Experiment 2: Is statistical learning available in both continuous and pre-segmented speech ?

Experiment 1 suggests that participants do not form declarative memory traces of words when the only available cues are statistical in nature. In contrast, they readily form declarative memories when items are pre-segmented.

These results do not imply that statistical learning might not play a critical role in word segmentation. As mentioned above, speech is prosodically organized (Cutler et al., 1997; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996), and 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 correct grouping in prosodic groups such as "*thebaby*" is "*theba* + *by*" or "*the* + *baby*". In principle, statistical learning might be well suited to this task. In line with the two-step explanation of Graf-Estes et al.'s (2007), Hay et al.'s (2011), Isbilen et al.'s (2020) experiments above, implicit knowledge of statistical regularities might help learners acquire words more effectively once (prosodic) segmentation cues are given (but see e.g. Ngon et al., 2013; Sohail & Johnson, 2016).

We test this issue in Experiment 2. Participants listened to a speech sequence of tri-syllabic nonsense words. For half of the participants, both the

TPs and the chunk frequency were higher between the the first two syllables of the word than between the last two syllables. We thus expected learners to split a triplet like *ABC* into an *AB+C* pattern. For the remaining participants, both the TPs and the chunk frequency favored an *A+BC* pattern. In the *pre-segmented* condition, the words were presented separated from each other and with a silence after each word. In the *continuous* condition, they were continuously concatenated. Following this familiarization, participants heard pairs of *AB* and *BC* items and had to indicate which item was more like the familiarization items. In Experiment 2a, stimuli were synthesized with the en1 (British English male) voice, though this voice turned out to produce artifacts in the continuous stream. In Experiment 2b, stimuli were synthesized using the us3 (American English male) voice.

If, as as we initially assumed, statistical learning allows learners to extract "correct" syllable groupings, they should recognize high-frequency chunks after both continuous and pre-segmented familiarizations. In contrast, if statistical learning predominantly supports predictive processing (Sherman & Turk-Browne, 2020; Turk-Browne et al., 2010), participants should extract high frequency groupings predominantly after continuous familiarizations in the *continuous* condition.

4.1 Material and Methods

We prepared two versions of Experiment 2, differing in the voice used to synthesize the stimuli. In Experiment 2a, we used a British English male (en1) voice. In Experiment 2b, we used an American English male (us3) voice. Both experiments were lab-based.

4.1.1 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). This number was chosen because it is realistic in the time-frame available for a third-year honors project. Participants reported to be native speakers of English, but we did not assess their English proficiency. However, participants were most likely exposed to English from childhood, as the experiment took place in London, UK, and the experimenters did not notice any clear non-native accents in most participants and excluded the few participants with non-native accents from analysis. The final demographic information is given in Table 1. In Experiment 2a, an additional 3 participants took part in the experiment but were not retained for analysis because they were much older than the rest of the sample ($N = 3$) or because they had a noticeable non-native accent $N = 1$. In Experiment 2b, an additional six participants were excluded from 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 ($N = 2$).

4.1.2 Design. Participants were familiarized with a sequence of tri-syllabic words. In Language 1, both the TPs and the chunk frequency were 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 4.

As a 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 and backward) TP of .33, while the second bigram has a (forward and backward) 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 4.

4.1.3 Stimuli. Stimuli in Experiment 2a were synthesized using the *en1* (British English

male) voice from mbrola (Dutoit et al., 1996). However, as discussed below, it turned out to be of relatively low quality and introduced artifacts in the data. Stimuli in Experiment 2b were synthesized using the *us3* voice (American English male) voice from mbrola (Dutoit et al., 1996).

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 (`http://sox.sourceforge.net/`) and then compressed to an mp3 file using ffmpeg (`https://ffmpeg.org/`). The stream was then presented 3 times to a participant (total familiarization duration: 7 min 17 s). The random order of the words was different for every participant.

For segmented streams, words were individually synthesized using mbrola. We then used a custom-made Perl script to randomize the words for each participant and concatenate them into a familiarization file using sox. The order of words was then randomized for each participant and concatenated into a single aiff file using sox. The silence among words was 540 ms (1.5 word durations). The total stream duration was 6 min 12s. The stream was then presented 3 times to a participant (total familiarization: 18 min 14 s).

4.1.4 Apparatus. The experiment was run using Psyscope X (`http://psy.ck.sissa.it`). Stimuli were presented over headphones in a quiet room. Responses were collected from pre-marked keys on the keyboard.

4.1.5 Procedure. Participants were informed that they would listen to a monologue by a talkative Martian, and instructed to try to remember the Martian words. Following this, they listened to three repetitions of the familiarization stream described above, for a total familiarization duration of 7 min 17 s (continuous stream) or 18 min 14 s (segmented stream).

Following this familiarization, participants were presented with pairs of items with an inter-stimulus interval of 500 ms, and had to choose which items was more like what they heard during familiarization. One item comprised the first two syllables of a word, and was a correct choice for Language 1. The other 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.

4.1.6 Analysis strategy. 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

Table 4
Design of 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	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:-nA:
FGD	FGD			faIz@-vOI	faI-g@-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:

for participants, correct items and foils. Following Baayen, Davidson, and Bates (2008), random factors were removed from the model when they did not contribute to the model likelihood.

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

4.2 Results

4.2.1 Experiment 2a (British English voice). We first report the results from Experiment 2a, using a British English voice. When the familiarization stream was pre-segmented, participants failed to split smaller utterances into their underlying components. As shown in Figure 10 (top), the average performance did not differ significantly from the chance level of 50% when the stream was synthesized with the *en1* voice (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 3.55 after correction with the Bayesian Information Criterion. Further, as shown in Table 5, performance did not depend on the language condition.

In contrast to the common finding that humans and other animals are sensitive to TPs, our participants failed to use TPs to split pre-segmented utterances into their underlying units. We thus asked if, in line with previous research, they can track TPs units are embedded into a *continuous* speech stream. That is, participants in the continuous condition listened to the very same artificial speech stream as in the pre-segmented condition, except that the stream was continuous and had no silences between words.

Participants also failed to use TPs to segment words when the speech stream was continuous. Specifically, and as shown in Figure 10 (top), 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 5.22 times more likely than the alternative hypothesis after a correction with the Bayesian Information Criterion. However, as shown in Table 5, 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 likely affected grouping, and prevented participants from using statistical learning. We thus decided to replicate Experiment 2a with a different, American English voice.

4.2.2 Experiment 2b (American English voice). When the familiarization stream was pre-segmented, participants failed to split smaller utterances into their underlying components. As shown in Figure 10 (bottom), the average performance did not differ significantly from the chance level of 50% when the stream was synthesized with the *us3* voice (M = 51.67, SD = 15.17), V = 216, p = 0.307. Likelihood ratio analysis favored the null hypothesis by a factor of 4.57 after correction with the Bayesian Information Criterion. As shown in Table 5, performance did not depend on the language condition. However, Figure 10 also shows a clearly defined outlier. In Supplementary Information SM7, we remove participants for Experiments 2a and 2b who differ by more than 2.5 standard deviations from the condition mean. This analysis yields similar results to the unfiltered analyses.

The failure to use statistical learning to split pre-segmented units was conceptually replicated in a pilot experiment with Spanish/Catalan speakers using chunk frequency and backwards TPs as the primary cues (SM8).

As in Experiment 2a, and in contrast to the common finding that humans and other animals are sensitive to TPs, our participants failed to use TPs to split pre-segmented utterances into their under-

Table 5

Performance differences across familiarization conditions in Experiment 2. 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			Odds 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]	-0.22	0.826
Continuous familiarization, British English voice (Exp. 2a)									
language = L2	en1	-1.024	0.410	[-1.83, -0.22]	0.359	0.147	[0.161, 0.803]	-2.50	0.013
Pre-segmented vs. continuous familiarization, British English voice (Exp. 2a)									
language = L2	en1	-1.061	0.382	[-1.81, -0.313]	0.346	0.132	[0.164, 0.732]	-2.779	0.005
stream type = segmented	en1	-0.242	0.360	[-0.949, 0.464]	0.785	0.283	[0.387, 1.59]	-0.673	0.501
language = L2 × stream type = segmented	en1	0.967	0.508	[-0.0292, 1.96]	2.631	1.338	[0.971, 7.13]	1.903	0.057
Pre-segmented familiarization, American English voice (Exp. 2b)									
language = L2	us3	0.114	0.673	[-1.2, 1.43]	1.121	0.754	[0.3, 4.19]	0.170	0.865
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, American English voice (Exp. 2b, 1)									
language = L2	us3	-0.019	0.558	[-1.11, 1.07]	0.982	0.547	[0.329, 2.93]	-0.033	0.973
stream type = segmented	us3	-0.328	0.188	[-0.696, 0.0391]	0.720	0.135	[0.499, 1.04]	-1.752	0.080
Pre-segmented vs. continuous familiarization, American English voice (Exp. 2b, 2)									
language = L2	us3	0.215	0.657	[-1.07, 1.5]	1.240	0.814	[0.342, 4.49]	0.327	0.743
stream type = segmented	us3	-0.608	0.244	[-1.09, -0.13]	0.544	0.133	[0.337, 0.878]	-2.493	0.013

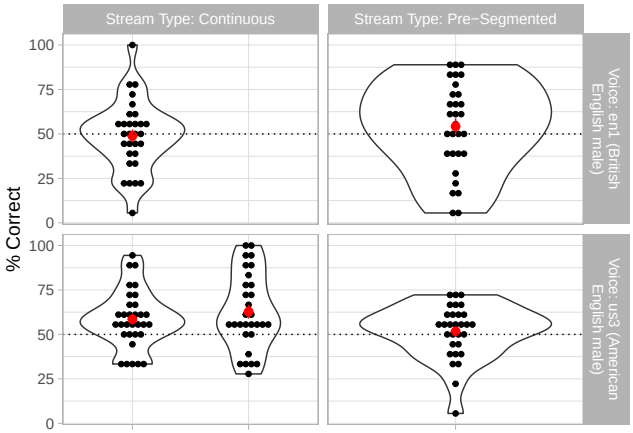

Figure 10. Results of Experiment 2. Each dot represents a participant. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a continuous familiarization stream or (right) a pre-segmented familiarization stream, with a British English voice (en1, top) or an American English voice (us3, bottom). The two continuous conditions with the American English voice are replications of one another.

lying units. We thus asked if they could track TPs units that are embedded into a *continuous* speech stream. As in Experiment 2a, participants in the continuous condition listened to the very same artificial speech stream as in the pre-segmented condition, except that the stream was continuous and had no silences between words.

As shown in Figure 10 (bottom), when the speech stream was synthesized with the *us3* voice, the average performance differed significantly from the chance level of 50%, (M = 58.51, SD = 16.21), Cohen's $d = 0.52$, $CI_{.95} = 52.66, 64.35$, V = 306.5, $p = 0.02$. As shown in Table 5, performance did not depend on the language condition, and was marginally better than in the pre-segmented condition (p = .08).

Given the likely confound introduced by the voice used in Experiment 2a, we sought to ensure that the results of Experiment 2b would be reliable, and replicated the successful tracking of statistical information using a new sample of participants, still with the *us3* voice. As shown in Figure 10 (bottom), the average performance differed significantly from the chance level of 50%, (M = 62.78, $SD = 21.35$), Cohen's $d = 0.6$, $CI_{.95} = 54.81$, 70.75, V = 320, $p = 0.008$. As shown in Table 5, performance did not depend on the language condition, and was significantly better than in the pre-segmented condition (p = .013).

Taken together, these results thus suggest that statistical learning mechanisms predominantly operate in continuous sequences, but less so in pre-segmented sequences (see also Shukla et al., 2007, 2011). Such a result is compatible with the view that statistical learning is important for predictive processing, given that continuous sequences are more conducive for prediction. In contrast, it raises doubts as to whether participants can use statistical learning mechanisms to memorize words, given that they do not seem to be able to do so in pre-segmented streams.

4.3 Discussion

In Experiment 2, participants tracked statistical dependencies predominantly when they were embedded in a continuous speech stream, but not across pre-segmented chunk sequences. This finding does not contradict the results from the Experiment 1 above, where TPs were somewhat higher in the pre-segmented condition; after all, if participants faithfully recall familiarization items, the resulting TPs will be high as well.

This result is also consistent with earlier findings that statistical learning predominantly occurs within major prosodic groups, and, within these groups, predominantly at the edges of those groups (Shukla et al., 2007; Seidl & Johnson, 2008). We show that, with shorter and better separated groups, statistical learning can be weakened further, to the extent that it is no longer detectable (at least in the current experiment).

In line with results from conditioning experiments (Alberts & Gubernick, 1984; Garcia et al., 1974; Gubernick & Alberts, 1984; L. T. Martin & Alberts, 1979), statistical learning, and maybe associative learning in general, can thus be enhanced or suppressed depending on the learning situation. The enhanced statistical learning in continuous sequences is consistent with the view that statistical learning is important for predictive processing (Turk-Browne et al., 2010; Sherman & Turk-Browne, 2020), given that prediction is arguably more useful in lengthy chunks. It is also consistent with the view that statistical learning may be less important for memorizing words (or at least to break up utterances so that the underlying words can be memorized), especially given that, due to its prosodic organization, speech tends to be pre-segmented into smaller groups (Cutler et al., 1997; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996; Brentari et al., 2011; Endress & Hauser, 2010; Pilon, 1981; Christophe et al., 2001).

A possible alternative interpretation is that, in the continuous streams of Experiment 2, repeated bisyllabic items pop out (and are thus remembered), while, in the pre-segmented streams, chunking cues (in the form of silences) prevent sub-chunks from popping out. However, if repeated bisyllabic items pop out in Experiment 2's continuous streams, repeated *trisyllabic* items (i.e., words) should pop out in Experiment 1 as well, and participants should be able to recall them as a result. As this prediction is falsified, a reasonable conclusion is that statistical learning does not make repeating elements pop out. Conversely, the availability of chunks might make statistical learning of within-chunk regularities more difficult, especially if chunks are memorized as whole units. This possibility would also confirm that statistical learning is separable from the (declarative) mechanisms involved in memorizing chunks.

Further, while our trisyllabic items are relatively short, so are utterances in infant-directed speech. For example, infant-directed utterances have a typical duration of about 1 s (with some cross-language variability; see e.g., Fernald et al., 1989; Grieser & Kuhl, 1988), with a mean utterance length of about 4 (e.g., Snow, 1977; Smolak & Weinraub, 1983; see also A. Martin, Igarashi, Jincho, & Mazuka, 2016). As a result, if statistical learning is difficult in shorter utterances, the utility of statistical learning for language acquisition might be reduced.

This is not to say that statistical learning can never occur in pre-segmented units. While the available statistical information does not always improve performance when chunking information is available (e.g., Sohail & Johnson, 2016), Shukla et al. (2007) showed that, when adults learners are exposed to 10-syllables chunks (defined by intonational contours), they have some sensitivity to statistical information within the chunks, though they might also use declarative memory mechanisms to remember sub-chunks (see also Endress & Bonatti, 2007; Endress & Mehler, 2009a; Endress & Wood, 2011 for additional results suggesting that statistical learning is possible within chunks, at least when the structure of the test items made the TP contrast rather salient). However, Shukla et al.

(2007) also found that participants predominantly retain information at chunk edges rather than at chunk medial positions. At minimum, it is thus an empirical question to what extent statistical learning is useful for word segmentation in the short utterances infants are faced with.

5 General Discussion

In the current experiments, we explored to what extent statistical learning can fulfill the function that is often attributed to it: Identifying word boundaries in fluent speech so that participants can learn words and, ultimately, commit them to declarative LTM.[9] In Experiment 1, we exposed (adult) participants to the speech streams from Saffran, Aslin, and Newport's (1996) classic word-segmentation experiment with infants, and asked whether they would be able to recall the words contained in these speech streams. When the speech streams were continuous, participants clearly tracked TPs in the speech streams, but we found no evidence that they had remembered any words at all. The overwhelming majority produced neither words nor part-words, and, even among those who produced word or part-words, two thirds produced part-words. Further, only about a third of the participants produced items starting with word-initial syllables, while two-third produced items starting with word-medial or word-final syllables. Statistical learning thus does not appear to provide participants with the ability to identify word boundaries in fluent speech nor to remember the words to which they have been exposed. Through simulations with a prominent chunking model (Perruchet & Vinter, 1998), we confirmed that these results cannot be explained by chunking models of word segmentation. Further, and as mentioned above, the fact that participants produce part-words even when they prefer words in a recognition test is fundamentally incompatible with such models, given that the models' preferences are driven by those chunks with the strongest memory representations, in *both* recall and recognition. As a result, they should show the same preferences in both recall and recognition. In contrast, when brief silences were inserted at word boundaries, mimicking the prosodic organization of speech, participants reliably produced words.

In Experiment 2, we asked whether statistical learning operates in smaller chunks, such as those that might be encountered due to the prosodic organization of language, or only in longer stretches of continuous speech. Participants listened to a speech sequence of tri-syllabic non-sense words. As in Experiment 1, the words were either *pre-segmented* (i.e., with a silence after each word) or continuously concatenated. We found that participants preferred high probability sequences only after exposure to continuous but not to pre-segmented streams, suggesting that statistical learning might be much less effective in the short and prosodically structured sequences that are typical of language acquisition (e.g., Fernald et al., 1989; Grieser & Kuhl, 1988; A. Martin et al., 2016; Snow, 1977; Smolak & Weinraub, 1983).[10]

Taken together, Experiments 1 and 2 suggest that statistical learning does not lead to declarative LTM representations of words, does not allow learners to identify word boundaries, and might not even operate under those conditions likely encountered during language acquisition. As a result, statistical learning and (declarative) memory might fulfill different computational functions in the process of word segmentation.

These results echo dissociations between associative learning and declarative memory (Cohen & Squire, 1980; Graf & Mandler, 1984; Finn et al., 2016; Knowlton et al., 1996; Poldrack et al.,

[9] As mentioned above, we focus on forms of statistical learning that allow learners to track sequential dependencies among items in continuous sequences and possibly also to associate simultaneously presented items in vision. Other forms of statistical learning might well have different properties.

[10] As mentioned above, we do not propose that statistical learning is impossible within chunks, and there is evidence that statistical learning can occur within chunks under some conditions.

2001; Squire, 1992), suggesting that the (cortical) declarative memory system might be independent of a (neostriatal) system for associative learning (Knowlton et al., 1996; Poldrack et al., 2001; Squire, 1992), though other authors propose that both types of memory involve the hippocampus (Ellis et al., 2021; Schendan, Searl, Melrose, & Stern, 2003; Sherman & Turk-Browne, 2020) and different memory systems can interact during consolidation (Robertson, 2022). In line with earlier proposals (Turk-Browne et al., 2010; Sherman & Turk-Browne, 2020), we thus suggest that the computational function of statistical learning might be distinct from that of (declarative) memory encoding, and that statistical learning might be more important for predictive processing. The relative salience of these mechanisms might depend on how useful and adaptive they are for the learning problem at hand.

5.1 Can chunking models account for word-segmentation data?

As mentioned above, there is considerable debate about whether statistical learning leads to memory for recurring chunks (e.g., Endress et al., 2020; Goodsitt et al., 1993; Perruchet, 2019; Swingley, 2005; Thiessen, 2017). However, and as also mentioned above, there are a number of results that seem incompatible with a declarative memory theory of statistical learning.

For example, observers sometimes report greater familiarity with high-TP items than with low-TP items when they have never encountered either of them (because the items are played backwards with respect to the familiarization sequence; Endress & Wood, 2011; Turk-Browne & Scholl, 2009; Jones & Pashler, 2007). Further, observers sometimes report greater familiarity with high-TP items they have *never* encountered than with low-TP items they have heard or seen (Endress & Langus, 2017; Endress & Mehler, 2009b; Endress, under review), a result that has been indirectly replicated even in findings that purportedly challenge these conclusions (Perruchet & Poulin-Charronnat, 2012).[11] Such results clearly demonstrate that a sensitivity to statistical structure does not imply that the statistically favored items have been encoded in LTM. In line with this view, many statistical learning results can be explained by purely correlational, memory-less Hebbian learning mechanisms (e.g., Endress & Johnson, 2021, 2023; Endress, 2024; Verosky & Morgan, 2021).

In our view, the main evidence in favor of memory-based models of statistical learning comes in three flavors (see Endress et al., 2020, for a critical review of other evidence). First, different authors suggested that statistically favored items are preferentially encoded in memory (e.g., Graf-Estes et al., 2007; Hay et al., 2011; Isbilen et al., 2020). However, as mentioned above, such results are compatible with a two-step explanation: First, during the statistical learning phase, participants acquire statistical knowledge without remembering any specific items. When experimenters *subsequently* provide participants with *isolated* chunks, the accumulated statistical knowledge facilitates processing of the experimenter-provided chunks, without these chunks having been acquired before being supplied by the experimenter. According to this explanation, there would not be any declarative memory of these chunks due to statistical learning.

The second major source of evidence that is compatible with a memory-based model for statistical learning is the observation that statistically structured sequences can elicit periodic electrophysiological activity with rhythms corresponding to word durations. For example, if words are three syllables long, a neural rhythm with a periodicity of three syllables can arise (e.g., Batterink & Paller, 2017; Buiatti, Peña, & Dehaene-Lambertz, 2009; Fló et al., 2022; Kabdebon, Pena,

[11] In Perruchet and Poulin-Charronnat's (2012), as in Endress and Langus's (2017) and Endress and Mehler's (2009b) experiments, it was much harder to choose between words and unattested high-TP items than to choose between words and part-words, a result that is incompatible with current chunking models.

Buiatti, & Dehaene-Lambertz, 2015; Moser et al., 2021). At first sight, such results seem to suggest that participants must track (and thus remember) words, though not all of these authors espoused a memory-based perspective of statistical learning. However, it turns out that this periodic activity can also result from Hebbian learning mechanisms that do not place any items in memory (Endress, 2024). After all, in each word, the final syllable is maximally predictive, and thus receives more associative input from other syllables than word-initial and word-medial syllables. As a result, one would expect an activation peak on word-final syllables, and thus a rhythm with a periodicity of a word duration.

The third major source of evidence for a memory-based model of statistical learning comes from studies revealing better recognition of (statistically defined) units compared to (statistically defined) units (e.g., Fiser & Aslin, 2005; Giroux & Rey, 2009; Orbán et al., 2008; Slone & Johnson, 2018). In the word recognition analogy used above, hearing the word *hamster* makes it difficult to recognize that the first syllable of *hamster* is a word on its own (i.e., *ham*; leaving aside phonetic differences between syllables that are parts of words and syllables that are words on their own; e.g., van Alphen & van Berkum, 2010; Salverda et al., 2003; Shatzman & McQueen, 2006a, 2006b). In actual statistical learning tasks, the *AB* part of an *ABC* unit is harder to recognize than a complete *CD* unit, which would suggest that the entire units are stored in memory.

However, the simulations reported here suggest that such results are compatible with memory-less Hebbian learning mechanisms, due to the interplay between excitation and inhibition. We also provided additional alternative explanations, which suggest that the evidence for chunk-based memory due to statistical learning is much weaker than commonly believed.

Taken together, these results suggest there are several alternative explanations for better recognition of units than of sub-units that do not involve declarative memory representations of the units. Given that the relatively direct memory test presented here revealed no evidence that statistical learning leads to memory representation for recurring units, a plausible conclusion is that it does not. Potentially, statistical learning might reflect simple Hebbian learning as in Endress and Johnson's (2021) model.[12]

The conclusion that statistical learning does not lead to declarative memories of words does not imply that statistical learning has no role in word learning. For example, and as mentioned above, prior associations among syllables (or other phonological units) might facilitate the subsequent establishment of declarative memory representations for words once suitable cues become available. Pre-existing associations might be particularly useful for word learning if the initial (phonological) representations of word sounds are not yet integrated in the mental lexicon, and if this integration requires additional exposure to these words (e.g., Gaskell & Dumay, 2003; see also Viviani & Crepaldi, 2022, for evidence that lexica are acquired gradually in second language acquisition). However, most words are exceedingly rare (Yang, 2013), which, in turn, raises the question of whether sufficient exposure would be available to learners to acquire all but the most frequent

[12]This conclusion does not imply that there are no explicit components to statistical learning. In fact, statistical learning is sensitive to attentional manipulations (Turk-Browne et al., 2005; Toro, Sinnett, & Soto-Faraco, 2005), and recognition performance in statistical learning tasks tends to be better when participants are more confident in their responses (e.g., Batterink, Reber, Neville, & Paller, 2015; Smalle, Daikoku, Szmalec, Duyck, & Möttönen, 2022). However, such results do not imply that statistical learning leads to declarative memory for words. For example, after familiarization with an episode of Looney Tunes, participants would presumably be highly confident in the association between Bugs Bunny and a carrot. However, this association does not imply that the Bugs Bunny-carrot combination is stored as a chunk in LTM.

words. Conversely, when potential meanings are available, people can learn words from just one or a few exposures (e.g., Aravind et al., 2018; Carey & Bartlett, 1978; Stevens, Gleitman, Trueswell, & Yang, 2017; Trueswell, Medina, Hafri, & Gleitman, 2013), suggesting that considerable exposure is not required for all forms of word learning.

Be that as it may, the current results also demonstrate that statistical learning does not allow learners to identify the beginnings and endings of words in the absence of other cues. While statistical learning might lead to helpful prior associations among syllables, other cues seem to be required to identify the (phonological) word forms that can later be consolidated.

5.2 Cues to word boundaries

These current results have implications for how words can be learned from fluent speech. If learners cannot use statistical learning to encode word candidates in (declarative) memory, they need to use other cues. Possible cues include using known words as delimiters for other words (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Brent & Siskind, 2001; Mersad & Nazzi, 2012), attentional allocation to beginnings and ends of utterances (Monaghan & Christiansen, 2010; Seidl & Johnson, 2008; Shukla et al., 2007), legal sound sequences (McQueen, 1998) and universal aspects of prosody (Brentari et al., 2011; Christophe et al., 2001; Endress & Hauser, 2010; Pilon, 1981). Such cues might plausibly support declarative memories of words because they (but not transition-based associative information) are consistent with how linguistic sequences are encoded in declarative long-term memory: Linguistic sequences are encoded with reference to their first and their last element (Endress & Langus, 2017; Fischer-Baum et al., 2011; Miozzo et al., 2016). Moreover, even a fairly simple computational model attending to utterance edges yielded excellent segmentation and word-learning performance (Monaghan & Christiansen, 2010), suggesting that such cues might be useful for actual language learners as well.

5.3 Potential roles of statistical learning

This is no to say that statistical learning might play no implicit role in word learning even when it is not sufficient to produce memories that can be recalled. For example, and as mentioned above, associations among syllables might facilitate the establishment of declarative memories once suitable (and explicit) segmentation cues become available (Endress & Langus, 2017), and, once words are acquired, word processing is not immune to unconscious stimuli such as masked primes (e.g., Forster, 1998; Kouider & Dupoux, 2005). Statistical learning might also facilitate word learning indirectly, for example through the acquisition of phonotactic constraint that might affect word learning in turn (e.g., Friederici & Wessels, 1993; Mattys, Jusczyk, Luce, & Morgan, 1999; McQueen, 1998). However, the extent to which statistical learning supports such computations remains to be established. For example, the phonotactic regularities above can be learned by keeping track of material at utterance boundaries (Monaghan & Christiansen, 2010), and thus just using the type of cues we introduced in the pre-segmented conditions. However, given that the current results suggest that statistical learning and declarative memory might have separable functions, and that statistical learning does not lead to memory for words nor to knowledge of word boundaries, we believe that it is an important topic for further research to determine the role statistical learning plays in word acquisition.

References

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

Aravind, A., de Villiers, J., Pace, A., Valentine, H., Golinkoff, R., Hirsh-Pasek, K., ... Sweig Wilson, M. (2018). Fast mapping word meanings across trials: Young children forget all but their first guess. *Cognition*, *177*, 177–188. doi: 10.1016/j.cognition.2018.04.008

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

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

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

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. doi: 10.1016/j.cortex.2017.02.004

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of memory and language*, *83*, 62–78. doi: 10.1016/j.jml.2015.04.004

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

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

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of experimental psychology. General*, *138*, 487–502. doi: 10.1037/a0016797

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

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

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage*, *44*(2), 509–519. doi: 10.1016/j.neuroimage.2008.09.015

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. In *Proceedings of the stanford child language conference* (Vol. 15, pp. 17–29).

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

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

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

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

Cohen, N., & Squire, L. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, *210*(4466), 207–210. doi: 10.1126/science.7414331

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

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Doeller, C. F., & Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(15), 5909-14. doi: 10.1073/pnas.0711433105

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

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

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica*, *135*(2), 182–190. doi: 10.1016/j.actpsy.2010.06.005

Endress, A. D. (2024). Hebbian learning can explain rhythmic neural entrainment to statistical regularities. *Developmental Science*. doi: 10.1111/desc.13487

Endress, A. D. (under review). Transitional probabilities outweigh frequency of occurrence in statistical learning of simultaneously presented visual shapes.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299. doi: 10.1016/j.cognition.2006.09.010

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

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

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

Endress, A. D., & Johnson, S. P. (2023). Hebbian, correlational learning provides a memory-less mechanism for statistical learning irrespective of implementational choices. *Cognition*, *230*, 105290. doi: 10.1016/j.cognition.2022.105290

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

Endress, A. D., & Mehler, J. (2009a). Primitive computations in speech processing. *Quarterly Journal of Experimental Psychology*, *62*(11), 2187–2209. doi: 10.1080/17470210902783646

Endress, A. D., & Mehler, J. (2009b). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351-367. doi: 10.1016/j.jml.2008.10.003

Endress, A. D., Slone, L. K., & Johnson, S. P. (2020). Statistical learning and memory. *Cognition*, *204*, 104346. doi: 10.1016/j.cognition.2020.104346

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology*, *63*(3), 141–171. doi: 10.1016/j.cogpsych.2011.07.001

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477-501.

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

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

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology. General*, *134*(4), 521-37. doi: 10.1037/0096-3445.134.4.521

Fló, A., Benjamin, L., Palu, M., & Dehaene-Lambertz, G. (2022). Sleeping neonates track transitional probabilities in speech but only retain the first syllable of words. *Scientific reports*, *12*, 4391. doi: 25865749

Forster, K. I. (1998). The pros and cons of masked priming. *Journal of psycholinguistic research*, *27*, 203–233. doi: 10.1023/a:1023202116609

Friederici, A., & Wessels, J. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, *54*(3), 287-95.

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

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

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*, 105–132. doi: 10.1016/s0010-0277(03)00070-2

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive science*, *33*, 260–272. doi: 10.1111/j.1551-6709.2009.01012.x

Glicksohn, A., & Cohen, A. (2011). The role of gestalt grouping principles in visual statistical learning. *Attention, Perception and Psychophysics*, *73*(3), 708–713. doi: 10.3758/s13414-010-0084-4

Glover, S., & Dixon, P. (2004). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin and Review*, *11*(5), 791–806.

Goodsitt, J., Morgan, J. L., & Kuhl, P. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language*, *20*(2), 229-52.

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

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

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

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support

for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20. doi: 10.1037/0012-1649.24.1.14

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

Hampshire, A., & Sharp, D. J. (2015). Contrasting network and modular perspectives on inhibitory control. *Trends in cognitive sciences*, *19*, 445–452. doi: 10.1016/j.tics.2015.06.006

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

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. doi: 10.1016/j.cogpsych.2011.06.002

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

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

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

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

Kabdebon, C., Pena, M., Buiatti, M., & Dehaene-Lambertz, G. (2015). Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain and language*, *148*, 25–36. doi: 10.1016/j.bandl.2015.03.005

Karaman, F., & Hay, J. F. (2018). The longevity of statistical learning: When infant memory decays, isolated words come to the rescue. *J. Exp. Psychol. Learn. Mem. Cogn.*, *44*(2), 221–232. doi: 10.1037/xlm0000448

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

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

Kiyonaga, A., & Egner, T. (2016). Center-surround inhibition in working memory. *Current biology : CB*, *26*, 64–68. doi: 10.1016/j.cub.2015.11.013

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

Kouider, S., & Dupoux, E. (2005). Subliminal speech priming. *Psychological science*, *16*, 617–625. doi: 10.1111/j.1467-9280.2005.01584.x

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

Marchetto, E., & Bonatti, L. L. (2015). Finding words and word structure in artificial speech: the development of infants' sensitivity to morphosyntactic regularities. *Journal of Child Language*, *42*(4), 873–902. doi: 10.1017/S0305000914000452

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Maternal speech to infants in a tonal

language: Support for universal prosodic features in motherese. *Cognition*, *156*, 52–59. doi: 10.1016/j.cognition.2016.07.015

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

Mattys, S. L., Jusczyk, P. W., Luce, P., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465-94.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-11.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46.

Mersad, K., & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, *8*(3), 303-315. doi: 10.1080/15475441.2011.609106

Miozzo, M., Petrova, A., Fischer-Baum, S., & Peressotti, F. (2016). Serial position encoding of signs. *Cognition*, *154*, 69–80. doi: 10.1016/j.cognition.2016.05.008

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

Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and gestalt-like principles predict melodic expectations. *Cognition*, *189*, 23–34. doi: 10.1016/j.cognition.2018.12.015

Moser, J., Batterink, L. J., Li Hegner, Y., Schleger, F., Braun, C., Paller, K. A., & Preissl, H. (2021). Dynamics of nonlinguistic statistical learning: From neural entrainment to the emergence of explicit knowledge. *NeuroImage*, *240*, 118378. doi: 10.1016/j.neuroimage.2021.118378

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

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34. doi: 10.1111/j.1467-7687.2012.01189.x

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745–2750. doi: 10.1073/pnas.0708424105

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

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? transitional probabilities versus chunks in language learning. *Topics in cognitive science*, *11*, 520–535. doi: 10.1111/tops.12403

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in cognitive sciences*, *10*, 233–238. doi: 10.1016/j.tics.2006.03.006

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. doi: 10.1016/j.jml.2012.02.010

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–63.

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

122.

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

Robertson, E. M. (2022). Memory leaks: information shared across memory systems. *Trends in cognitive sciences*, *26*, 544–554. doi: 10.1016/j.tics.2022.03.010

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

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual review of psychology*, *69*, 181–203. doi: 10.1146/annurev-psych-122216-011805

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–21.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the comer of your ear. *Psychological Science*, *8*(2), 101-105. doi: 10.1111/j.1467-9280.1997.tb00690.x

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*(1), 51-89.

Schendan, H. E., Searl, M. M., Melrose, R. J., & Stern, C. E. (2003). An fmri study of the role of the medial temporal lobe in implicit and explicit sequence learning. *Neuron*, *37*, 1013–1025. doi: 10.1016/s0896-6273(03)00123-5

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

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

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193-247.

Shatzman, K. B., & McQueen, J. M. (2006a). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, *17*(5), 372-7. doi: 10.1111/j.1467-9280.2006.01714.x

Shatzman, K. B., & McQueen, J. M. (2006b). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics*, *68*(1), 1-16.

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current opinion in behavioral sciences*, *32*, 15–20. doi: 10.1016/j.cobeha.2020.01.015

Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 22760–22770. doi: 10.1073/pnas.2013291117

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

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6038–6043. doi: 10.1073/pnas.1017617108

Slone, L. K., & Johnson, S. (2015). Statistical and chunking processes in adults' visual sequence learning. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science*

society (pp. 2218–2223). Austin, TX: Cognitive Science Society. Paper presented at the annual meeting of the cognitive science society.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics: Infants represent visual sequences in terms of chunks. *Cognition*, *178*, 92–102. doi: 10.1016/j.cognition.2018.05.016

Smalle, E. H. M., Daikoku, T., Szmalec, A., Duyck, W., & Möttönen, R. (2022). Unlocking adults' implicit statistical learning by cognitive depletion. *Proceedings of the National Academy of Sciences of the United States of America*, *119*. doi: 10.1073/pnas.2026011119

Smolak, L., & Weinraub, M. (1983). Maternal speech: strategy or response? *Journal of Child Language*, *10*(2), 369-380. doi: 10.1017/S0305000900007820

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*, 1–22.

Sohail, J., & Johnson, E. K. (2016). How transitional probabilities and the edge effect contribute to listeners' phonological bootstrapping success. *Language Learning and Development*, 1-11. doi: 10.1080/15475441.2015.1073153

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

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*(S4), 638-676. doi: 10.1111/cogs.12416

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*(1), 86-132. doi: 10.1016/j.cogpsych.2004.06.001

Thiessen, E. D. (2017). 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*. doi: 10.1098/rstb.2016.0056

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

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

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25-34. doi: 10.1016/j.cognition.2005.01.006

Toro, J. M., Trobalon, J. B., & Sebastián-Gallés, N. (2005). Effects of backward speech and speaker variability in language discrimination by rats. *Journal of Experimental Psychology. Animal Behavior Processes*, *31*(1), 95-100. doi: 10.1037/0097-7403.31.1.95

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*, 126–156. doi: 10.1016/j.cogpsych.2012.10.001

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

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552-64. doi: 10.1037/0096-3445.134.4.552

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

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

10.1523/JNEUROSCI.0858-10.2010

van Alphen, P. M., & van Berkum, J. J. A. (2010). Is there pain in champagne? Semantic involvement of words within words during sense-making. *Journal of Cognitive Neuroscience*, *22*, 2618–2626. doi: 10.1162/jocn.2009.21336

van Moorselaar, D., & Slagter, H. A. (2019). Learning what is irrelevant or relevant: Expectations facilitate distractor inhibition and target facilitation through distinct neural mechanisms. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *39*, 6953–6967. doi: 10.1523/JNEUROSCI.0593-19.2019

Verosky, N. J., & Morgan, E. (2021). Pitches that wire together fire together: Scale degree associations across time predict melodic expectations. *Cognitive science*, *45*, e13037. doi: 10.1111/cogs.13037

Viviani, E., & Crepaldi, D. (2022). Masked morphological priming and sensitivity to the statistical structure of form-to-meaning mapping in l2. *Journal of cognition*, *5*, 30. doi: 10.5334/joc.221

Vlach, H. A., & DeBrock, C. A. (2019). Statistics learned are statistics forgotten: Children's retention and retrieval of cross-situational word learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(4), 700 - 711.

Yang, C. D. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6324–6327. doi: 10.1073/pnas.1216803110

Supplementary Online Materials

SM1 Analysis of the productions in Experiment 1

SM1.1 Pre-processing of the responses

Each recall response was analyzed in five steps. First, we applied pre-segmented substitution rules to make the transcriptions more consistent (see Table S1, "before segmentation", for a complete list of substitution rules). 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. This was necessary because some participants separated only words by spaces, while others separated syllables by spaces, and groups of syllables (e.g., words) by other characters (e.g., commas). If the response did not contain any commas (,) or semicolons (;), any spaces in the response were used to delineate units. For example, if the response was "*tudaro pigola*", *tudaro* and *pigola* would be accepted as 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 S1). For example, we treated the voiced and unvoiced varieties 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.

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.

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 syllables, one syllable change was allowed with respect to the speech streams. If no match was 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 substrings.

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

1. The winning underlying form had the maximal *number of attested syllables* among candidate underlying forms;

Table S1

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 intended to make transcriptions more consistent, and were not counted when calculating the derivation length. Substitutions after segmentation allowed for misperceptions, and were counted when calculating derivation length. These substitution rules were motivated by three observations: (1) /O/ might be perceived as /A/. (2) Voiced and unvoiced consonants can be confused; that is /g/ can be confused with /k/, /d/ with /t/ and /b/ and /p/. (3) /b/ might be perceived as /v/.

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			

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.

SM1.2 Measures of interest

We computed various properties for each underlying form, given the "target" language the participants had been exposed to. All measures provided in the raw data are described in Table S2. For each underlying form, we calculated:

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 was 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.

SM1.3 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, there is only one (out 12) non-zero TP continuation with a TP of 1.0, and the 11 other continuations have a TP of zero. As a result, the random TP is $1.0 \times 1/12 + 0.0 \times 11/12 = 1/12$. For a C syllable, there are 3 (out of 12) possible continuations with a TP of 1/3; the other 9 continuations have a TP of zero. As a result, the random TP is $1/3 \times 3/12 + 0.0 \times 9/12 = 1/12$. On average, the random TP is thus $(1/12 + 1/12 + 1/12)/3 = 1/12 \approx .083$.

SM1.4 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.

SM2 Measures and column names in the supplementary data file for Experiment 1

Table S2
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.

SM3 Additional results for Experiment 1

Table S3

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.

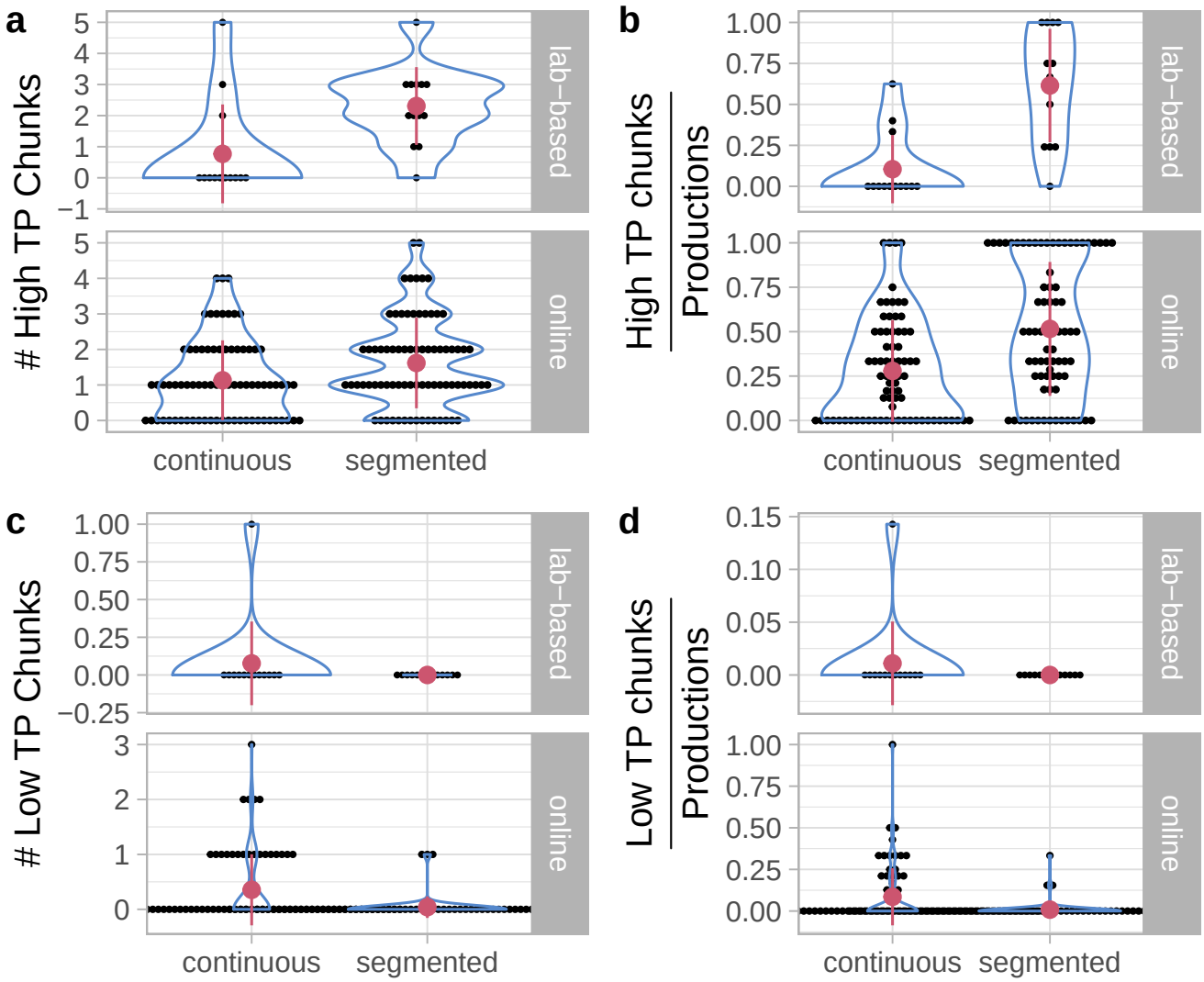

Figure S1. Plot of High and Low TP chunks.

SM4 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:

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

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.

Table S4
Counts of participants producing exclusively words, exclusively part-words, neither words nor part-words, or a mixture of both. To compare the recognition performance of participants who produced part-words to that of participants producing words, we excluded participants who produced neither of these item types or a mixture thereof.

	Participants producing			
Segmentation Condition	**Part-words**	**Words**	**Neither** (excluded)	**Mixture** (excluded)
Lab-based				
Continuous	3	1	6	3
Pre-segmented	0	12	1	0
Online				
Continuous	14	10	52	0
Pre-segmented	0	52	24	0

SM5 Relations between recall and recognition performance in Experiment 1

In this section, we seek to compare recognition and recall performance in Experiment 1. This is somewhat problematic, because our data resulting from the recall phase of Experiment 1 is discrete rather than continuous. We will thus link recognition and recall performance through two analyses. First, and as mentioned above, two-thirds of the participants in the continuous condition produced part-words, while only one-third produced words. We will compare performance in the recognition phase between those participants producing part-words and those producing words.

Second, it turned out that, during the recall phase, the proportion of productions with "correct" initial or final syllables was reasonably continuous (see Figure 5). We will thus correlate these proportions as well as the TPs in the strings *produced* by the participants with their performance in the recognition phase.

SM5.1 Recognition performance in word-producers vs. part-word producers

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

Further, we excluded participants who produced neither words nor part-words. The counts are shown in Table S4. Finally, since no participants in the pre-segmented conditions produced part-words, some statistical comparisons are not available for the pre-segmented condition.

As shown in Table S5 and Figure S2, participants in the continuous condition of the online experiment who produced words performed statistically better in the recognition test than participants who produced part-words. (In the lab-based experiments, there were only 4 participants in total who produced either words or part-words, making statistical comparisons unreliable.)

SM5.2 Production of correct initial and final syllables vs. recognition performance

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

Table S5

Recognition performance as a function of whether participants produced words or part-words. The p value reflects a Wilcoxon test comparing participants producing words and participants producing part-words, respectively. The statistical comparisons are available only in the continuous conditions because no participant in the pre-segmented condition produced part-words.

			Recognition performance		
Segmentation Condition	**Productions**	***N***	***M***	***SE***	***p***
Lab-based					
Continuous	Words	1	50.0	—	0.637
Continuous	Part-Words	3	75.0	17.68	
Pre-segmented	Words	12	91.7	4.91	—
Online					
Continuous	Words	10	90.0	5.83	< 0.001
Continuous	Part-Words	14	42.9	5.04	
Pre-segmented	Words	52	93.3	2.09	—

As shown in Figure S3, recognition performance in the continuous condition was correlated both with the proportion of correct initial syllables in the participants' productions and with the production of correct final syllables. However, it was not correlated with the average TPs in the participants' productions. These correlations were not significant in the pre-segmentation conditions, presumably because of the very high level of performance.

While these results suggest that recognition and recall performance are related, the underlying causal pathway is unclear. On the one hand, and as mentioned above, participants who happen to focus on those syllables corresponding to words rather than part-words would also focus on more statistically cohesive syllable sequences, which, in turn, would lead to better recognition performance as well. Alternatively, recall and production performance might be linked directly.

Critically, under either hypothesis, statistical learning would not allow help participants with the problem that motivated sequential statistical learning approaches to word segmentation in the first instance, namely to identify word boundaries in fluent speech. According to the first interpretation, participants do not remember any words to begin with. According to the second interpretation, participants might use statistical learning to recover word boundaries, but recover *incorrect* word boundaries in two-thirds of the cases even in a highly simplified learning situation, and thus cannot rely on statistical learning for word learning either.

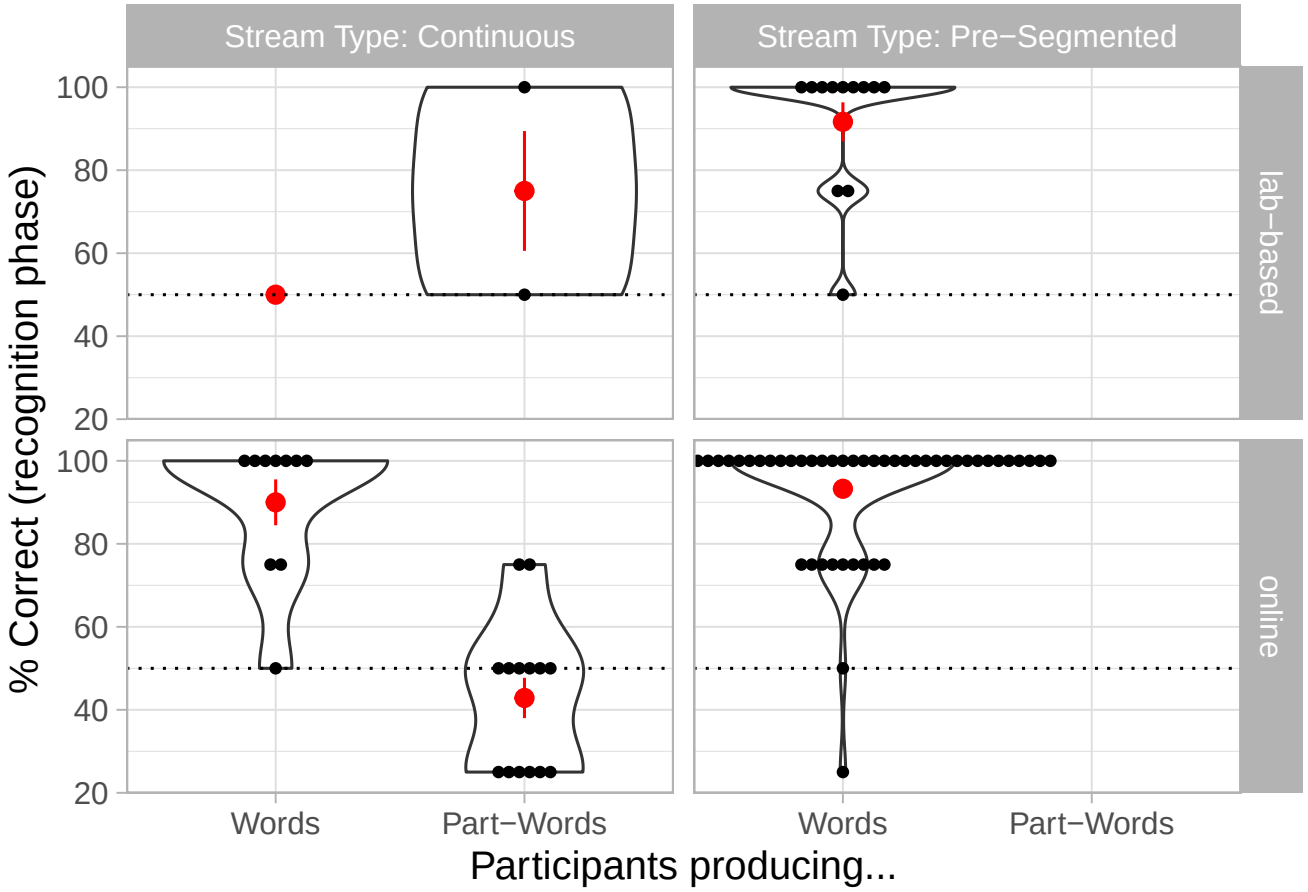

Figure S2. Recognition performance in Experiment 1 as a function of whether a participant produces words or part-words. Each dot represents a participant. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) a continuous familiarization stream or (right) a pre-segmented familiarization stream, in the lab-based version of the experiment (top) or in the online version (bottom). Participants producing words performed better in the recognition test than participants producing part-words.

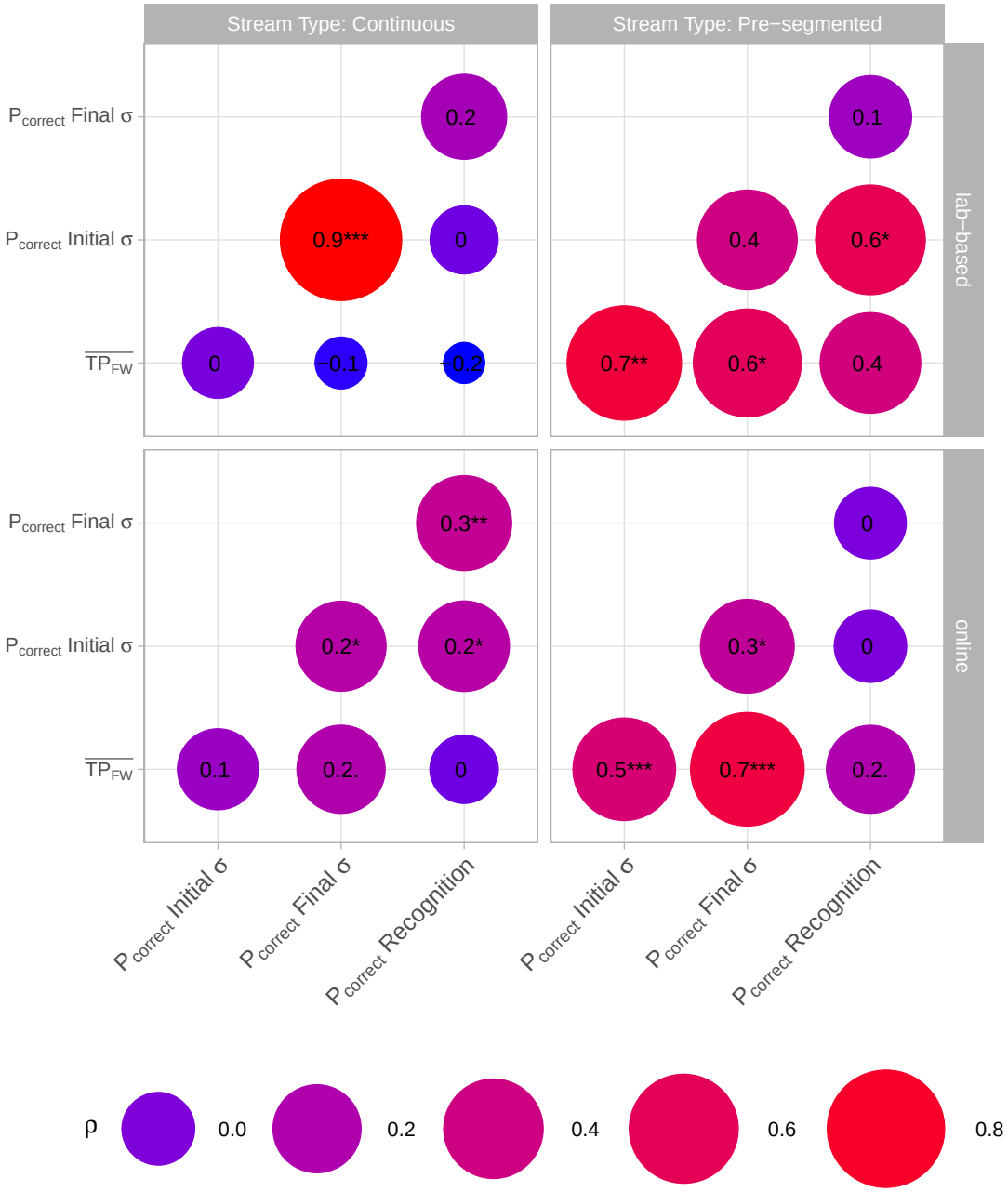

Figure S3. Spearman correlations between the performance in the recognition test ($P_{correct}$ Recognition) and three measures of the participants' productions: The proportion of correct initial syllables ($P_{correct}$ Initial σ) and of final syllables ($P_{correct}$ Final σ) as well as the average forward TPs in the participants' productions ($T\bar{P}_{FW}$). Correlations were calculated separately after familiarization with (left) a continuous familiarization stream or (right) a pre-segmented familiarization stream, in the lab-based version of the experiment (top) or in the online version (bottom). Significance labels: ***: ≤ 0.001; **: ≤ 0.01; *: ≤ 0.05; .: ≤ 0.1

SM6 Simulations with PARSER (Perruchet & Vinter, 1998)

PARSER segments continuous streams by recursively chunking units in the stream. These units are syllables or syllable combinations the model has encountered and retained in the speech stream. Units are built up recursively. For example, if a unit *A* is followed by a unit *B*, the model can create a new and larger unit *AB* that it can recognize later on. As a result, if this new unit *AB* is later followed by *C*, a new and still larger unit *ABC* might be created. These units are stored in a lexicon and have some memory weight. The weight of recurring units is strengthened, while spurious units are eliminated through decay and interference.

We first familiarized the model with one of the speech streams used in Experiment 1 (i.e., one of the speech streams from Saffran, Aslin, and Newport's (1996) Experiment 2). Following this, we recorded the memory strength of words and part-words. Specifically, we created 4 test trials pitting the two words against the two part-words, and, in each test trial, we compared the weight of the word and that of the part-word. We assigned a score of 1 to a trial if the weight of the word in the lexicon was higher than that of the part-word, a score of 0 with the weight of the part-word was higher, and a score of 0.5 if the two weights were the same. We then averaged these scores for all trials, and used this average as the performance of a simulated participant (see below).

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

PARSER has five parameters: the maximal number of units considered, the increment in memory strength upon encountering a unit, the weight threshold for an item to be removed from the lexicon, the initial weights of the syllables, the forgetting rate, and the interference rate. We varied the forgetting and interference rates and kept the original values of the other variables. We used forgetting rates from 0 to 0.1 and interference rates from 0 to 0.01, both in 101 equidistant steps. (In the original model, the forgetting rate was 0.05 and the interference rate 0.005.) These parameter combinations thus yielded $101 \times 101 = 10,201$ simulated "experiments." Each experiment was run with 50 random initializations, representing 50 participants. We created 40 different "speech streams" with different random orderings of the words. For each simulated participant, we then randomly chose one of those speech streams.

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

Further, all 10,201 simulated experiments showed a statistically significant preference for words. Across experiments, the average effect size (Cohen's *d*) was 1.616 (range: 0.344, 3.672), with the smaller effect sizes mainly occurring for high forgetting rates (see Figure S4). With Perruchet and Vinter's (1998) original parameters, the effect size was 1.833.

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

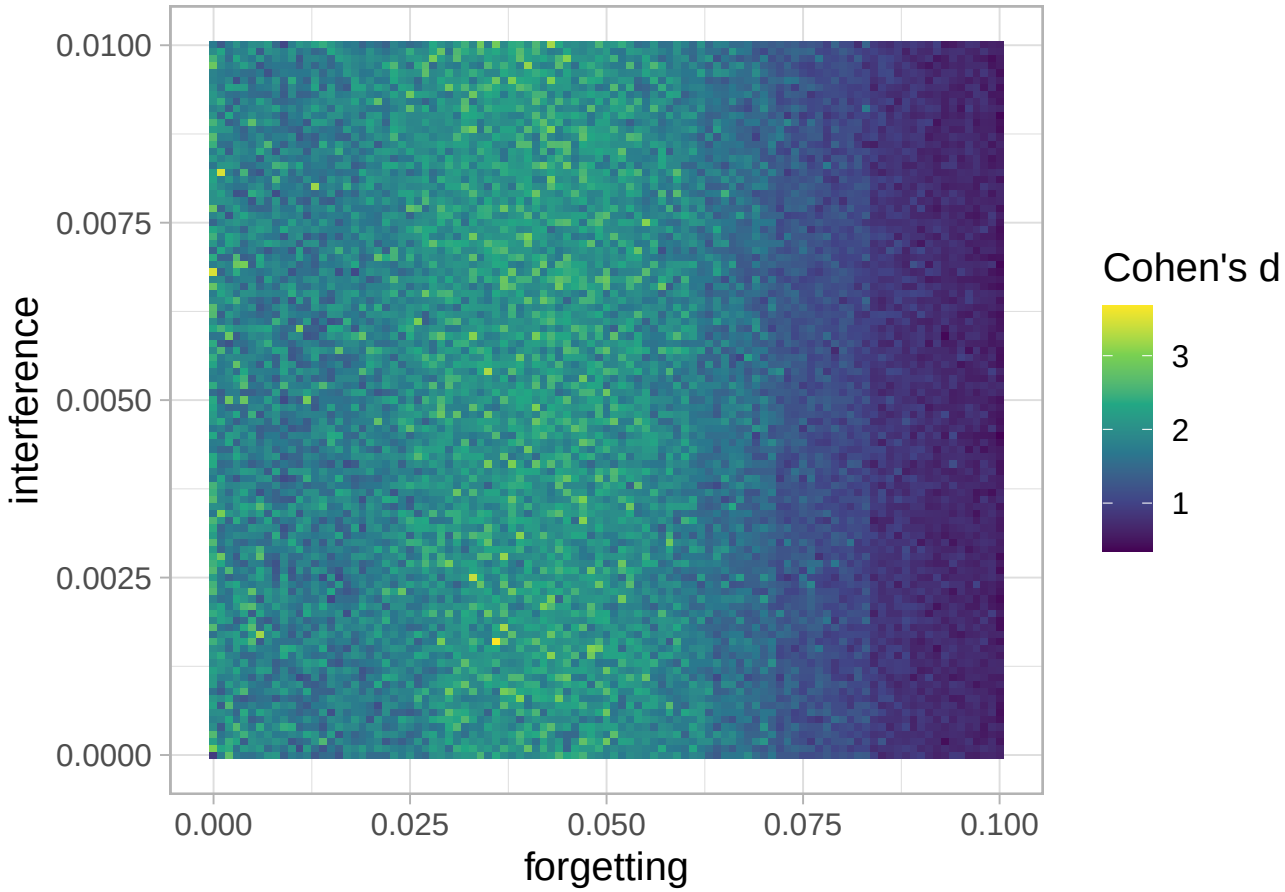

Figure S4. Effect sizes (Cohen's *d*) of the preference for words over part-words in simulations with PARSER as a function of the forgetting rate and the interference rate. All simulated experiments yielded a significant preference for words.

SM7 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.

SM7.1 Experiment 2a (British English voice)

Figure S5 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 S6, 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 S5 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 S6, 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.

SM7.1.1 Experiment 2b (American English voice). Figure S5 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 S6, 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 S5, 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 S6, 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 S5, 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 S6, 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.

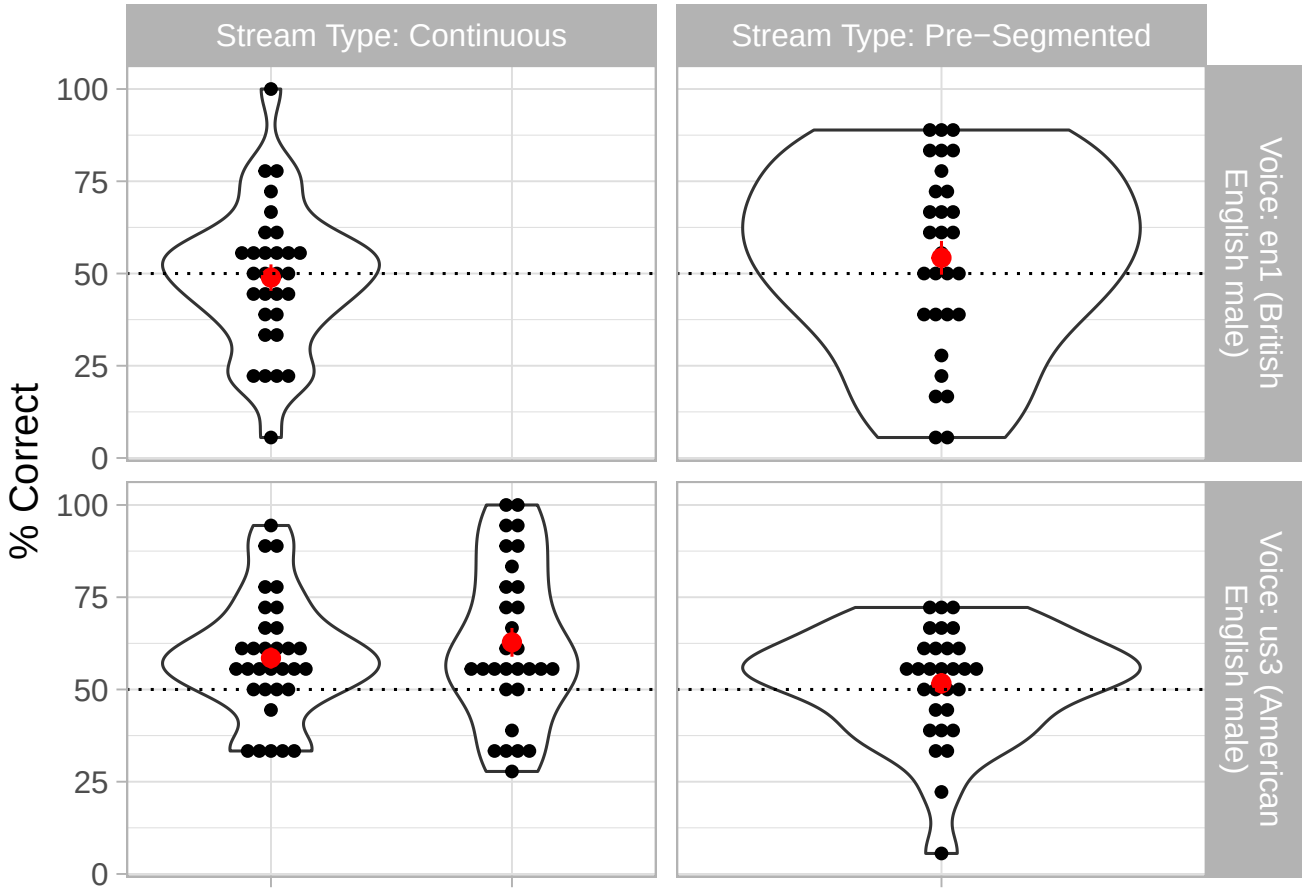

Figure S5. Results of Experiment 1 after outliers of more than 2.5 standard deviations from each condition mean were excluded. Each dot represents a participant. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 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 S6

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

SM8 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.

SM8.1 Methods

Table S7
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

SM8.1.1 Participants. Demographic information of the pilot experiment is given in Table S7. Participants were native speakers of Spanish and Catalan and were recruited from the Universitat Pompeu Fabra community.

SM8.1.2 Stimuli. Stimuli transcriptions are given in Table S8. 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.

SM8.1.3 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/`).

SM8.1.4 Familiarization. The design of the pilot experiment is shown in Table S8. 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 4 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.

SM8.1.5 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 4 were combined into four test pairs, which were presented twice with different item orders. A new trial started 100 ms after a participant response.

Table S8

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

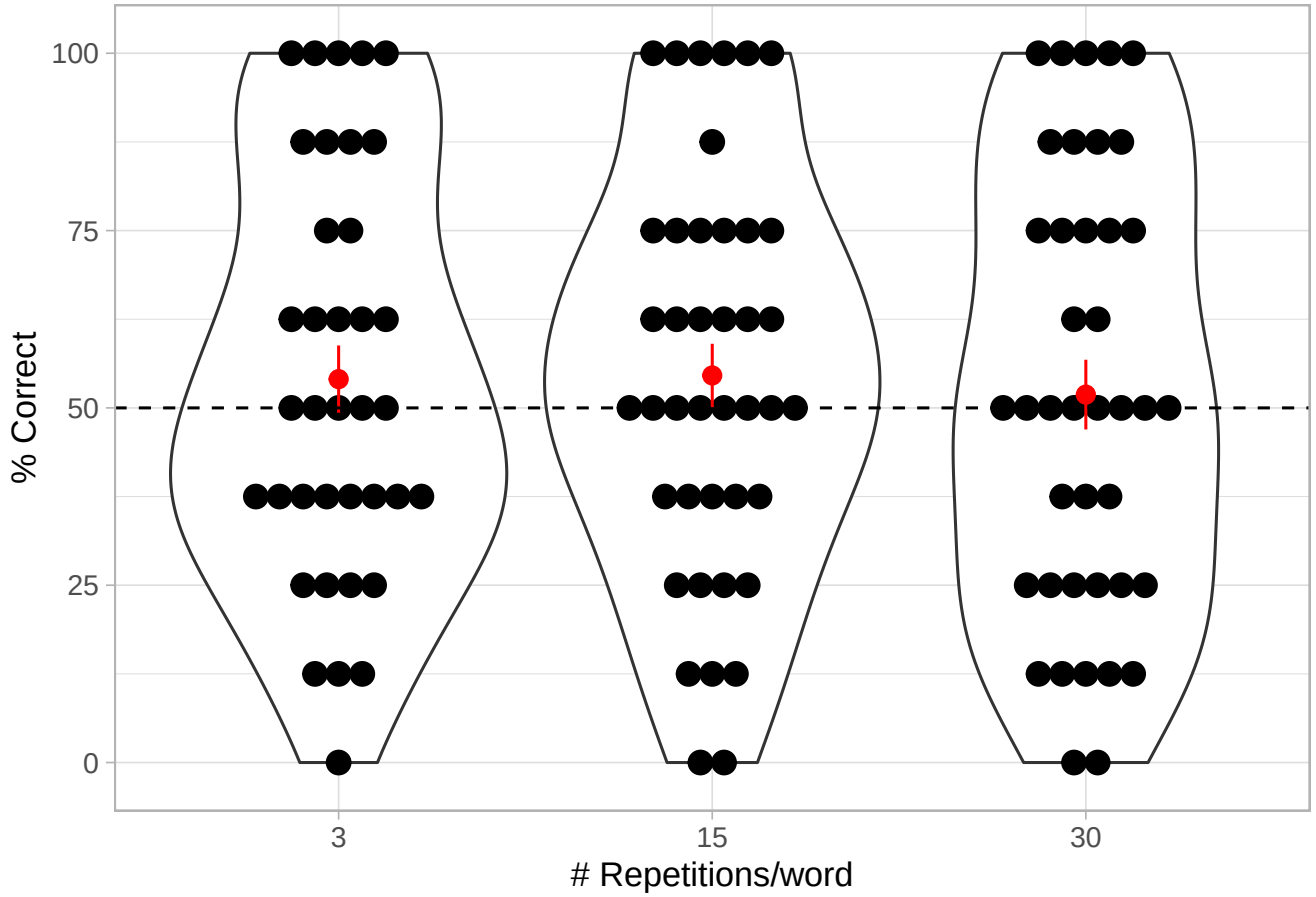

Figure S6. Results of the pilot experiment. Each dot represents a participants. The central red dot is the sample mean; error bars represent standard errors from the mean. The results show the percentage of correct choices in the recognition test after familiarization with (left) 3, (middle) 15 or (right) 30 repetitions per word.

Table S9

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

SM8.2 Results

As shown Table S9, a generalized linear model revealed that performance depended neither on the amount of familiarization nor on the familiarization language. As shown in Figure S6, 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.