Neuroscience

The speed of detection vs. segmentation from continuous sequences: Evidence for an anticipation mechanism for detection through a computational model

Meili Luo, Ran Cao, Felix Hao Wang

School of Psychology, Nanjing Normal University, Nanjing, Jiangsu, China

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 1
March 27, 2024 (this version)

Sent for peer review
January 29, 2024

Posted to preprint server
January 1, 2024

Abstract

To understand the latent structure of a language, one of the first steps in language learning is word segmentation. The rapid speed is an important feature of statistical segmentation, and exact quantifications would help us understand the underlying mechanism. In this study, we probe the speed of learning by using a novel experimental paradigm and compare them to results obtained through the traditional word segmentation paradigm. Using a novel target detection paradigm, we replicated and extended a study on when participants start to show learning effects. We successfully replicated a facilitation effect showing rapid learning, which showed that learners obtained statistical information following a single exposure. However, we also found a similar facilitation effect when the syllable sequence contained words that were uniform or mixed in length. Importantly, this contrasts with results from traditional word segmentation paradigms, where learning is significantly better in uniform-length sequences than in mixed-length sequences. Thus, even though the target detection paradigm showed robust effects, it may have required mechanisms different from those in word segmentation. To understand these mechanisms, we proposed both theoretical analyses and a computational model to simulate results from the target detection paradigm. We found that an anticipation mechanism could explain the data from target detection, and crucially, the anticipation mechanism can produce facilitation effects without performing segmentation. We discuss both the theoretical and empirical reasons why the target detection and word segmentation paradigm might engage different processes, and how these findings contribute to our understanding of statistical word segmentation.

eLife assessment

The authors present **valuable** empirical and modelling evidence that statistical learning in speech perception may contain processes like segmentation and anticipation. While the evidence for statistical learning effects is **solid**, the link between the pattern of effects (both empirical and simulated) and the theoretical concepts of segmentation and anticipation would need to be much stronger to exclude other accounts of the data. This work will be of broad interest to researchers working on, or with, statistical learning, and to any researcher interested in the challenges of whether data and modeling can effectively adjudicate between competing theoretical constructs.

For novice language learners, one of the first tasks is to understand the structure of the continuous speech streams they hear by segmenting the speech into words. In the literature, two types of information that can be used for segmentation are discussed. Prosody, such as stress, though never perfectly correlate with world boundaries in natural languages, can often provide useful information to word boundaries and has been shown to be used for word segmentation (Johnson & Jusczyk, 2001; Jusczyk, 1999; Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999). Another type of information is the distributional information of the syllables in a sequence, which was shown to be used in word segmentation as well (e.g., Saffran, Aslin, & Newport, 1996; Aslin Saffran, & Newport, 1998). The theory is that learners would track the co-occurrence information between syllables, and use this co-occurrence information to compute transitional probability, which can be a cue to word boundaries. The seminal work on statistical learning (Saffran et al., 1996) demonstrated that young infants can segment word forms in a rapid syllable stream in two minutes where the syllables in the stream formed statistical patterns to word boundaries. In this influential study, learning only required exposure to a syllable stream, consisting of four trisyllabic words occurring 45 times each where prosodic cues such as stress and co-articulation to word boundaries were not present.

Following this initial work showing powerful learning, there is now a large literature on how the underlying computational mechanism can be best described, as well as the constraints for word segmentation to be successful. To understand the underlying computational mechanism, different computational models have been proposed (e.g., Frank, Goldwater, Griffiths & Tenenbaum, 2010; Giroux & Rey, 2009; Perruchet & Vinter, 1998; Swingley, 2005). For example, different models implement ideas on boundary finding (e.g., Swingley, 2005) vs. chunking (e.g., the PARSER model from Perruchet & Vinter, 1998). Through computational modeling, concrete predictions of different theoretical approaches can be generated, which offer testable hypotheses about these different mechanisms that researchers were able to test further, using experimental methods (Endress & Mehler, 2009). In addition to computational models, leveraging the learning constraints also helps understand the computational mechanism. Elsewhere in the language acquisition literature, for example, learning constraints are an important piece in understanding why the nature of the learning problem requires a representation that's structure-dependent when studying the acquisition of syntax. In this instance, knowing when a set of learning theories succeed and fail allows us to understand the intricacies of the learning mechanism. For word segmentation, one prominent constraint is that, even though infants and adults alike have shown success segmenting syllable sequences consisting of words that were uniform in length (i.e., all words were either disyllabic; Graf Estes, Evans, Alibali, & Saffran, 2007; or trisyllabic, Aslin et al., 1998), both infants and adults have shown difficulty with syllable sequences consisting of words of mixed length (Johnson & Tyler, 2010; Johnson & Jusczyk, 2003a; 2003b; Hoch, Tyler, & Tillmann, 2013). For example, Johnson and Tyler (2010) showed that if the sequence is constructed by concatenating two trisyllabic and two disyllabic words, infants were unable to segment from such a sequence, even though the infants in the same study had no trouble

segmenting a sequence with its four words being all trisyllabic. Similarly, Hoch et al., (2013) showed that adults learned much worse with a mixed-length language than with a uniform-length language.

Another way of understanding the mechanisms for segmentation is by studying how fast learning takes place. Fast learning has always been a feature of statistical word segmentation, with the initial work showing that infants can segment words with only 2 to 3 minutes of exposure (Saffran et al., 1996; Aslin et al., 1998). It is also an important theoretical question, as the relationship between the amount of exposure and learning can be leveraged to understand the mechanism. For example, after the initial studies showed that learning was fast with relatively simple stimuli, subsequent studies testing adults with more complicated sequences or with different sounds have used longer exposure periods. In Finn and Hudson Kam (2008) for example, adults were asked to segment a sequence and the amount of exposure was manipulated during different experiments. Interestingly, even though the duration of exposure has been extended from multiple minutes to double or even quadruple the original amount, the amount of learning has not changed as a result (also see Newport & Aslin, 2004 for a similar finding). An alternative direction is to shorten the exposure amount and present very short sequences as the input, and the distributional properties that allow successful learning under such conditions can help us understand the computational processes in segmentation, though not many studies have explored this line of inquiry. Among these, Wang, Luo, and Wang (2023) showed that learners can succeed at segmentation when learners are exposed to a stream where word forms occurred only two times. Using a word segmentation paradigm, Wang et al. (2023) repeated the cycle of learning and testing for many different short sequences: Learners were first presented with a continuous syllable stream, and then asked to rate the familiarity of words and part-words, and learned the next stream and so on. This finding suggested that learners can rapidly extract word forms and remember them, and learning did not require a slow accumulation process. However, though segmentation was successful under these minimal conditions, the effect size of learning was small. Testing the same set of syllable sequences but with each word occurring four times, Wang et al. (2023) found that the effect size of learning was significantly larger in the latter condition, suggesting that, even though word forms may be extracted rapidly, the memory component of the segmentation task may require repetition. Even more impressively, Batterink (2017) found evidence that one exposure could bring about a facilitation effect consistent with word segmentation, using an online measure. This online measure involved the use of a target detection paradigm, where participants were asked to listen to syllable streams and press a key to detect a particular syllable in the stream. In each trial, twelve syllables were randomly grouped into four trisyllabic words, which were used to create a syllable sequence with all four words occurring 4 times. Batterink (2017) found that statistical learning can be faster than previously thought: After *one* exposure to a trisyllabic word (e.g., tugola), learners were able to react faster to the second (or third) syllable of that trisyllabic word (*go* or *la*) than to the first syllable (*tu*). Thus, Batterink (2017) showed that learners have sensitivity to the statistical structure of the stream after one exposure.

Understanding this effect is of great interest because it would inform the theories of statistical word segmentation and identify the computational models that can describe the effect best. Batterink (2017) discussed that a chunking model, such as the one described in PARSER (Perruchet & Vinter, 1998) is more consistent with the results than the use of conditional probabilities. She argued that computing conditional probabilities is often thought to involve multiple encounters so that a probability can be calculated. On the other hand, with a chunking model such as PARSER, exposure to the syllable sequence would result in random chunks, which are stored in memory. After a single exposure to a word form (say, ABC with different letters representing different syllables) from the syllable sequence, the random chunk may sometimes include a chunk that contains or partially contains the word form (such as ABC, or AB). Regardless of the computational framework, we believe that such an anticipation account would explain the facilitation effect: in the sense that stimuli follow, precede, or co-occur, the brain can encode such predictive relationship (e.g., Conway, 2020; Davachi & DuBrow, 2015; Summerfield & De

Lange, 2014; Turk-Browne, Scholl, Johnson, & Chun, 2010). It's even possible that this account of prediction-based facilitation may even hold without segmentation, word extraction, or chunking, per se. For example, an encounter to a sequence in which two elements co-occur (say, AB) would theoretically allow the learner to use the predictive relationship during a subsequent encounter (that A predicts B).

In the current study, we investigate statistical word segmentation with an online measure further. We aim to leverage our knowledge of the learning constraint in a typical word segmentation task, i.e., to segmentation succeed when the input sequence contained uniform-length words, but failed when the words were mixed in length, to probe the mechanisms in the online target detection task and its relationship to the offline task. If the target detection task shares the same mechanism with word segmentation, we would expect that the facilitation effect is stronger in sequences with uniform-length words compared to sequences with mixed-length words. However, as our analysis suggested above, the online target detection task may not require the learner to segment the continuous input and remember segmented forms. That is, if the facilitation effect in the target detection task is based on a general prediction mechanism as we discussed above, it would only require participants to store the sequence they hear and use that for a general prediction process. In this case, it would not matter whether the sequence contained uniform- or mixed-length words, and the size of the facilitation effect would be the same in both the uniform- and mixed-length conditions.

We report two experiments in this paper. In Experiment 1, we report a replication using the same material and the same uniform-length word design from the Batterink (2017) study (which we call the uniform condition). This serves to establish the robustness of the finding. Additionally, we conducted the replication two times, an exact replication and a conceptual replication, which allowed a comparison of a nuance variable, namely whether the sequence initial (the first and the second) or the sequence final (the 47^{th} and the 48^{th}) syllables were included in the detection task. This manipulation was included to inform us of how specific the learning condition needs to be for the effect to occur. In Experiment 2, we changed one aspect of the design, namely the lengths of the words in the sequences for target detection, while keeping all other variables the same (which we call the mixed condition). This allows us to examine the effect of learning in the mixed condition, and compare the effect size of learning in the mixed-length word condition to the uniform condition. Together, the two experiments should provide insight into the mechanisms involved in the target detection task, and its relationship to the word segmentation literature.

Experiment 1

Methods

Participants

The number of participants for the replication was determined based on a power analysis based on the data from Batterink (2017), with some over-sampling. The main effect of interest was the interaction for RTs between the first and second presentation, where the second and third syllables were predictable during the second presentation but unpredictable during the first presentation. Based on the data from Batterink (2017), this difference was −13.6ms (a standard error was 4.91). In a one-sided test, this produced a post-hoc power of 0.85 with 19 subjects, which means that the original study was well-powered. As long as we have 19 subjects in any condition in our replication, it would also ensure the power of the replication study here.

We ran the study until the end of the semester, and by the time we stopped collecting data, in the exact-replication condition, we included data from twenty-one adult participants from both the University of Nevada, Las Vegas and the University of Southern California. In the conceptual-replication condition, we included forty-eight participants from the same two institutions. IRB approval was obtained at each institution separately prior to conducting the experiment.

Stimuli

The stimuli were the same set from Batterink, who provided open materials online (retrieved from *https://osf.io/z69fs/*). Syllable sequences are constructed by concatenating syllables from two syllable inventories (from a male and a female speaker), each consisting of 24 unique syllables at a rate of 300 ms per syllable.

Design and Procedure

The study closely followed the design of Batterink (2017). To reiterate the design briefly here, each participant completed 144 iterations of the target detection task. In each iteration, 12 syllables were randomly chosen from a syllable inventory (male or female), which were used to create four trisyllabic words, exhausting all 12 syllables (i.e., one syllable occurred only in one word). Next, a syllable sequence was created by repeating the four words four times in a pseudo-random fashion, with the constraint that a word does not immediately follow itself. This meant that each syllable sequence was 48 (4*4*3) syllables long. The 144 iterations of the task included the use of 72 male- and female-voice syllable sequences, where either male or female first is counterbalanced between subjects. The experiment was self-paced and took about an hour to complete.

Instructions

The experiment began with a short instruction phase. The following instruction was given, and the experimenter read the instructions aloud to the participants, allowing participants to ask questions at any point of the instruction phase.

"In this study, you will be presented with a rapid succession of syllables, and your job is to detect a particular syllable in a given sequence. In each trial, a target syllable will be presented (for example, ku), both visually on the screen and aurally in the headphones. After this, you will hear the syllable sequence (for example, bakufoka…) in the headphones and your job is to press Space every time you detect the target syllable.

The key to this task is that you need to press the Space as soon as you detect the target syllable. As it would become clear to you in a moment, the syllables go by very quickly, and your job is to detect all of the target syllables as quickly and as accurately as you possibly can.

If you have understood the instructions, you may press Space to move to the next screen. If you have any questions regarding the task, please ask the experimenter now."

Syllable detection phase

After the instruction phase, the syllable detection phase began. First, the participant was given the opportunity to practice for two trials, while the experimenter was present; after the practice period and the experimenter made sure that the participant was doing the task correctly, the experimenter left the room.

Each trial in the syllable detection task began with the screen displaying "Get ready now. Press Space to start." After the participant pressed the Space bar, they saw the target syllable displayed on the screen (e.g., "target syllable: vu"). After 1.5 seconds of silence, the participant heard the

syllable from the headphones (e.g., the syllable vu), which lasted 0.3 seconds, and another 3.2 seconds of silence followed the target syllable. At this point (5 seconds after the start of the trial), the syllable stream began to play. The syllable stream lasted 14.4 seconds, during which the subjects were free to press Space to indicate that they detected the target syllable. At the end of the trial, the participant was informed as such and the next trial began ("That is the end of this trial. The next trial will begin now."). The study ended after all 144 trials were done. An illustration is shown in **Figure 1**.

There were two conditions in **Experiment 1**, though the difference between the two was minimal. In the exact-replication condition, syllables were not detection targets if they were the first two or the last two in the syllable stream, the same as in Batterink (2017). In the conceptual-replication condition, this constraint did not apply. There seemed, prima facie, no reason to exclude the detection of a syllable when it was among the first two syllables or the last two syllables of the sequence, and the conceptual-replication condition was conducted to test this effect. The conceptual-replication condition thus served to test whether this design difference would not make a difference in terms of the facilitation effect. Our null hypothesis here was whether the target syllable occurred in these arbitrary locations should not interfere with whether the learner could remember the sequence and use it for prediction.

Predictions

We re-iterate the predictions for the replication study here. The prediction is that the second syllable in a trisyllabic word is detected faster than the first syllable after one (or more) exposure, and similarly for the third syllable compared to the first syllable, because while the first syllable is unpredictable, the second and the third syllable become predictable if the participant is able to remember the trisyllabic word given one exposure.

Results and Discussion

Prior to conducting the analysis, we dropped the trials that involved the first two/last two positions to make sure that the analysis examined the same type of data for the conceptual-replication condition. This meant that all the analyses below were based on reaction time data when the syllable to be detected was in the stream position 3-46. The rest of the analysis plan closely followed the analysis described in Batterink (2017). Before the analysis, we combined the counterbalancing conditions (female voice/male voice first).

For the main analysis, the first step we took was to convert the raw reaction time data into RT data for the target syllables. This calculation included two parts, whether a target syllable was detected, and what the RT was for that syllable. A target syllable was treated as detected if there is a key press within 1200ms after the onset of the syllable. Given this criterion, participants in the exact-replication condition detected 87.7% of the syllables on average, and participants in the conceptual-replication condition detected 87.2% of the syllables on average. Thus, the detection rates of syllables in both conditions were comparable to the one reported in Batterink (2017), which is 87.4%. All subsequent analyses are conducted on these data (**Figure 2**).

Next, the crucial prediction from Batterink (2017) was examined, i.e., that after just one exposure, there is an effect of "word form extraction" where there is an interaction between syllable position and presentation order such that syllable position 2 and 3 as opposed to 1 should have a smaller reaction time in later presentations (2, 3, 4) as opposed to the first presentation. This pattern was found in both of the conditions, which showed up in the right panels (predicted values from the regression model) in **Figure 1**. For visual inspection, one easy way is to observe the slopes of the lines connecting the data points for syllable positions 1 through 3, as this slope is negative if syllables 2 and 3 are reacted to faster than syllable 1. The prediction is thus that, the slope for presentation 1 is not negative, but the slopes for presentation 2, 3 and 4 would be. Looking at **Figure 1**, we saw that the line for presentations 2 to 4 had negative slopes, whereas

Experiment	vocabulary	Target	Sequence
Exp. 1: Uniform length	yu ba vu tu ka du he lo me za so le	za	yu ba vu tu ka du he lo me za so le tu ka du he lo me yu ba vu za so le ...
Exp. 2: Mixed length	yu ba vu tu ka du he lo za so	za	yu ba vu tu ka du he lo za so tu ka du he lo yu ba vu za so ...

Figure 1

An illustration of the target detection task, for both Experiment 1 and 2.

Two sets of sample vocabulary, targets, and syllable sequences are shown. The arrows indicate where the targets are in the syllable sequence.

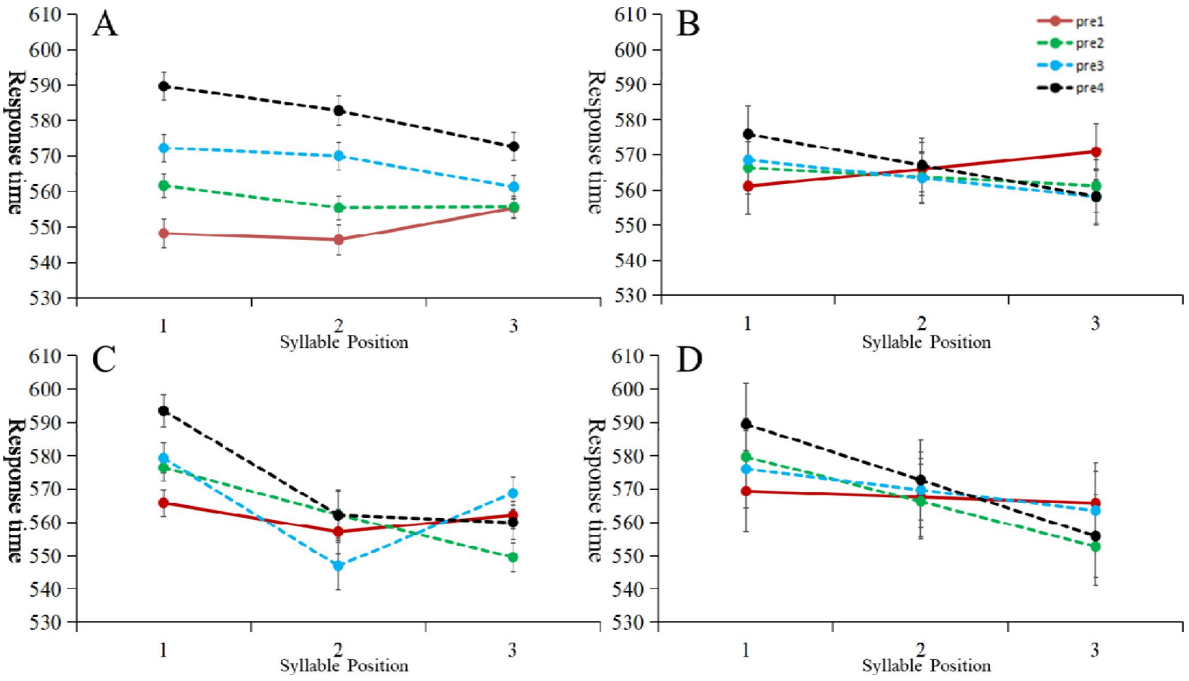

Figure 2

Reaction time (RT) data with syllable position (first, second, or third syllable in the word) on the x-axis, and word presentation (first, second, third, or fourth occurrence of the word in the stream) as different lines in the Figure.

The left panels show the raw data means and the right panels show the regression model fit. The top panels showed the data from the conceptual-replication condition and the bottom panel showed the data from the exact-replication condition. The style of the plot is similar to the ones in Batterink (2017) for ease of comparison. Error bars represent ±1 SEM.

the slope for presentation 1 was not negative. To examine this effect statistically, we ran a linear mixed effect model in which RT is the dependent variable for each condition. The independent variable included fixed effects of word presentation (1-4, categorical; the choice of the variables being categorical vs. continuous was made in Batterink, 2017), syllable position (1-3, continuous), overall stream position (3rd through 46th syllable in the syllable sequence, continuous), and the interaction between word presentation and syllable position. Note that the overall stream position was found to be a significant predictor in addition to the rest of the variables in Batterink (2017) so it was included here. Random effects included participant as a random intercept and stream position as a random slope. For each condition, we first report the statistical significance of the omnibus interaction between word presentation and syllable position, and then report the pairwise comparisons between different pairs of presentation.

Two more aspects of the data were examined following the analysis from Batterink (2017), which informs on the direction of the effect. If the effect was due to a slowdown of the unpredictable syllables for later presentations (i.e., presentation 2, 3, and 4) compared to the first presentation, this would predict the RTs for syllable position 1 to be smaller during presentation 1 compared to later presentations. This would also predict the RTs for syllable positions 2 and 3 to be the same between the presentations. On the other hand, if the effect was due to facilitation to react to the predictable syllables in later presentations, this would predict the RTs for syllable positions 2 and 3 to be smaller in later presentations compared to the first presentation, but the RTs for syllable position 1 to be similar for different presentations.

For the exact-replication condition, the omnibus interaction between word presentation and syllable position was significant ($\chi^2(3) = 14.91$, $p = 0.002$); also of note, stream position was not significant (β=0.0002, z=0.84, p=0.400); this might have been a result of a relatively small number of subjects in this condition. Next, pairwise comparisons between presentation 1 and later presentations were conducted; if the interaction coefficient is negative, it means the prediction was confirmed. The interaction between presentations 1 and 2 was negative and significant (β=-0.012, z=-2.95, p=0.003), and so was the interaction between presentations 1 and 4 (β=-0.015, z=-3.49, p< 0.001). Only the interaction between presentations 1 and 3 did not reach significance (β= -0.005, z= -1.07, p= 0.287). Thus, all of the effects were numerically in the right direction and most of the predictions were confirmed in this condition.

For the conceptual-replication condition, the omnibus interaction between word presentation and syllable position was significant ($\chi^2(3) = 16.66$, $p = 0.001$). The stream position was also significant (β=0.001, z=5.65, p<0.001), successfully replicated this effect from Battarink (2017), where syllables occurring later in the syllable stream are detected slower than syllables occurring earlier in the syllable stream. The interaction between presentation 1 and presentation 2 was negative and significant (β=-0.007, z=-2.17, p=0.030), so was the interaction between presentation 1 and 3 (β=-0.010, z=-3.03, p=0.002) and between presentation 1 and 4 (β=-0.014, z=-3.94, p<0.001). All of the effects were confirmed in the conceptual-replication condition. In sum, all of the results from the two samples showed that participants were able to react faster to the later syllables of a word compared to the first syllable following a single exposure.

Lastly, we examined the direction of the effect. Two analyses were carried out. First, we asked whether the RTs for syllable position 1 were different for presentation 1 vs. the later presentations. Secondly, we asked whether the RT for syllable positions 2 and 3 were different for presentation 1 vs. the later presentations. The results were the same between the two conditions. In the exact-replication condition, the RTs for syllable position 1 were not significantly different for presentation 1 vs. the later presentations (β=0.0001, z=0.32, p=0.747), and were significantly larger for position 2 and 3 for presentation 1 vs. the later presentations (β=-0.018, z=-3.01, p=0.003). In the conceptual-replication condition, the RTs for syllable position 1 were not significantly different for presentation 1 vs. the later presentations (β=0.008, z=1.41, p=0.160), and were significantly larger for position 2 and 3 for presentation 1 vs. the later presentations (β=-0.009,

z=-2.58, p=0.010). Thus, the effect was due to the fact that the predictable syllables (from positions 2 and 3 in the later presentations) were responded to faster, rather than unpredictable syllables were responded to slower. This analysis thus pinpoints the origin of the effect.

In sum, both the exact-replication condition and the conceptual-replication condition were successful in replicating all of the aspects from Batterink (2017). The exclusion of the detection of a syllable when it is among the first two syllables or the last two syllables of the sequence did not make a difference in generating the facilitation effect.

Experiment 2

As we noted above, part of testing a powerful learning mechanism involves testing conditions when the learning mechanism is known to fail in specific conditions. To this end, we conducted Experiment 2, which differed from Experiment 1 in one crucial aspect. That is, we changed the lengths of the words that made up the continuous syllable sequences in Experiment 2. Rather than having them be all three syllables long, which is the case in Experiment 1, the four words making up sequences in Experiment 2 included 2 disyllabic and 2 trisyllabic words. In the word segmentation literature, using mixed-length designs leads to no segmentation (Johnson & Tyler, 2010) or significantly weaker segmentation than with uniform sequences (Hoch et al., 2013). Experiment 2 allows us to examine whether the target detection paradigm employs the same mechanism as word segmentation, which would predict that there would be a weaker facilitation effect in Experiment 2 compared to Experiment 1.

Methods

Participants

Twenty-one undergraduate students were recruited from Psychology Department subject pools at both the University of Nevada, Las Vegas and the University of Southern California.

Stimuli

The stimuli were identical to the stimuli in Experiment 1.

Design and Procedure

All aspects of the experiment were the same as Experiment 1, except for the sequences used for target detection. In Experiment 2, we generated the sequences by concatenating two disyllabic, and two trisyllabic words. In each sequence, the four words occurred 4 times, which is the same as in Experiment 1. This meant that each sequence was 40 syllables long. Target syllables could have been any position for words of any length. All the rest of the dimensions are the same as the conceptual-replication condition from Experiment 1.

Results and Discussion

We used the same analysis plan from Experiment 1, combining the counterbalancing conditions (female voice/male voice first). Under the criterion that a syllable is detected if there is a key press within the 1200ms after the onset of the syllable, participants on average detected 88.9% of the syllables. Before the analysis was run, we only kept data for stream positions 3-38, where the data for the first and last two positions in the stream were dropped.

Below, we examined the facilitation effect of disyllabic and trisyllabic words, first separately and then together. Again, the prediction for the effect is an interaction between syllable position and presentation order such that syllable positions 2 and 3 compared to 1 should have a shorter reaction time in later presentations (2, 3, 4) as opposed to the first presentation for trisyllabic words, and for disyllabic words, this was the interaction between syllable position (2 compared to 1) with presentation order. A plot of the data from Experiment 2 can be seen in **Figure 3**, and again, negative slopes are predicted for presentations 2 through 4 but not 1. To examine the effect statistically, we conducted two linear mixed effect models, for disyllabic and trisyllabic words separately. In both regressions, the RT was the dependent variable, and the independent variable included fixed effects of word presentation (1-4, categorical), position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous), overall stream position (3rd through 46th syllable in the syllable sequence, continuous), and the interaction between word presentation and syllable position. Random effects included participant as a random intercept and stream position as a random slope.

For trisyllabic words, the omnibus interaction between word presentation and syllable position was significant ($\chi^2(3) = 46.40$, $p<0.001$). Stream position was found to be significant as well ($\beta=0.002$, $z=5.00$, $p<0.001$). Next, we looked at interactions between syllable position and presentation pairs. The interaction between presentation 1 and presentation 2 was negative but not significant ($\beta= -0.007$, $z=-1.08$, $p=0.281$). The interaction between presentations 1 and 3 was negative and significant ($\beta=-0.014$, $z=-2.17$, $p=0.030$). The interaction between presentations 1 and 4 was negative and significant ($\beta=-0.043$, $z=-6.11$, $p<0.001$). For disyllabic words, the regression containing both participant as a random intercept and stream position as a random slope did not converge, so we only kept the participant as a random intercept, which converged. In this regression, the omnibus interaction between word presentation and syllable position did not reach significance ($\chi^2(3) =6.52$, $p= 0.089$). Stream position was also found not to be significant ($\beta=0.0001$, $z=0.37$, $p=0.711$). The interaction between presentation 1 and presentation 2 was negative and significant ($\beta= -0.024$, $z=-2.38$, $p=0.017$). The interaction between presentation 1 and 3 was negative and marginally significant ($\beta=-0.020$, $z=-1.84$, $p=0.066$). The interaction between presentations 1 and 4 was negative and significant ($\beta=-0.027$, $z=-2.21$, $p=0.027$). Lastly, we analyzed whether the interaction between presentation order and syllable position significantly interacted with word length. For this analysis, we added an interaction term of word length to the previous regression model, such that the fixed effect became a three-way interaction between word presentation (1-4, categorical), position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous) and word-length (2/3, categorical), as well as the overall stream position. Random effects included participant as a random intercept and stream position as a random slope. The interaction was not significant ($\chi^2(3) = 6.19$, $p= 0.103$). Together, these analyses showed that there was a robust effect for trisyllabic and disyllabic words alike, and no difference between the two types of words. A plot of the data is shown in **Figure 3**.

Lastly, we want to answer the question of whether the facilitation effect is larger in the uniform condition than in the mixed condition, which would be the prediction if the current target detection task engages the same mechanism as the word segmentation paradigm. Notably, there are some differences in terms of the structure of data in the uniform and mixed conditions. First, the mixed condition involved both disyllabic and trisyllabic words, whereas the uniform condition only had trisyllabic words. For the analysis below, we put word length in the fixed effect as a main effect, since we found the two types of words to have similar effects and no interactions, as we just discussed. Secondly, the length of syllable streams was shorter in mixed conditions compared to the uniform conditions, because half of the words were disyllabic in the mixed condition. This meant that streams were 48 syllables long in the uniform condition, but only 40 syllables long in the mixed condition. Since stream position has consistently been a significant predictor of reaction times, this is likely to affect the effects as well. Putting these two variables as main effects allowed us to observe the interaction of interest while controlling these important variables.

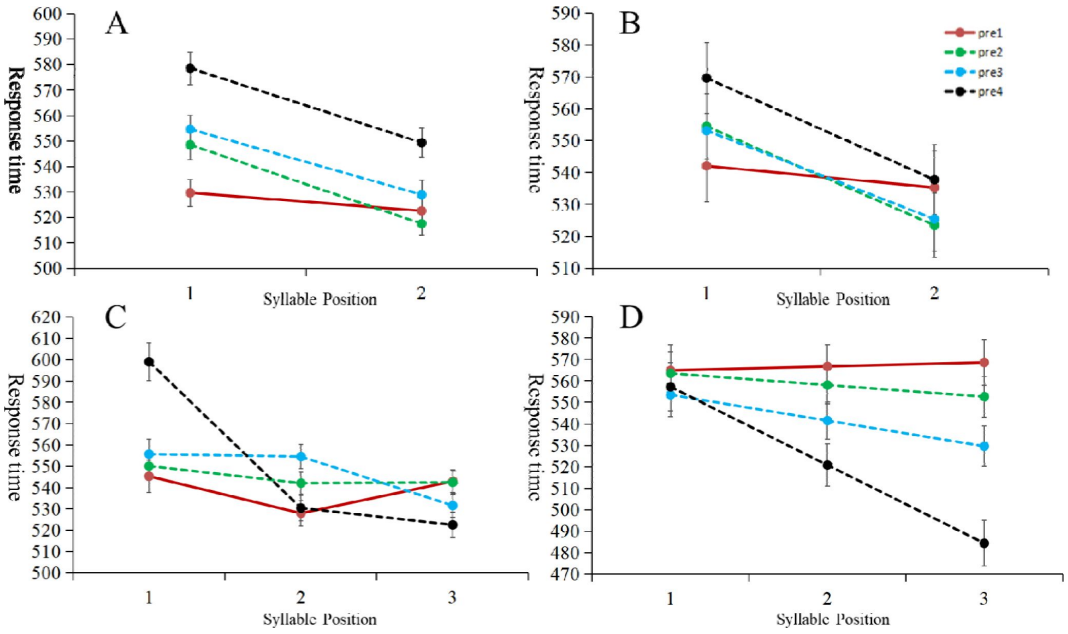

Figure 3

Reaction time (RT) data with syllable position (first, second, or third syllable in the word) on the x-axis, and word presentation (first, second, third, or fourth occurrence of the word in the stream) as different lines in the Figure.

The left panels show the raw data means and the right panels show the regression model fit. The top panels showed the data for the disyllabic words and the bottom panel showed the data for the trisyllabic words. Error bars represent ±1 SEM.

The prediction for the difference between the mixed and uniform conditions in the present target detection tasks, if they act similarly to word segmentation tasks, is that the effect is smaller in the mixed condition than in the uniform condition. For this analysis, we compared the data from **Experiment 2** to the exact-replication condition in **Experiment 1**, which had a similar number of subjects (though using data from the conceptual-replication condition yielded the same results; see **Appendix**). To examine this effect, we set up the following mixed effect regression with a three-way interaction. The RT was the dependent variable, and the independent variable included fixed effects of condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous), and the interaction between the three. Fixed effect further included overall stream position (3rd through 46th syllable in the syllable sequence in the uniform condition, 3rd through 38th syllable in the mixed condition, both continuous) and word length (disyllabic/trisyllabic, categorical). Random effects included participant as a random intercept and stream position as a random slope. The omnibus three-way interaction was significant ($\chi^2(3) = 15.79$, p=0.001), suggesting that the ways syllable position and presentation interact in the two experiments are different. To understand this three-way interaction, we looked at the three-way interaction between syllable position, condition, and pairs of presentations (i.e., 1 and 2, 1 and 3, and 1 and 4). We found that the three-way interaction for presentations 1 and 2 (β= -0.003, z=-0.52, p=0.675) was negative and not significant, became positive and not significant for presentations 1 and 3 (β=0.011, z=1.65, p=0.099), and became positive and significant for presentations 1 and 4 (β=0.021, z=2.99, p=0.003). In other words, the coefficients grow as a function of presentation in this three-way interaction. Looking at a plot of model fit (**Figure 4**), this pattern becomes clear: while the slopes (from syllable position 1 to 3) for presentation 1 were flat for both conditions, the negative slope for presentation 4 for the mixed condition was the largest in absolute value (from 570ms to 494ms) for all slopes, more than in presentation 4 for the uniform condition (from 579ms to 546ms). This was the three-way interaction we saw. We could understand this result as the mixed condition having a larger effect than the uniform condition, but as we explore in the simulation below, this statistical difference is consistent with a scenario where the facilitation effect is the same in both conditions. Importantly, these results differ from our a priori hypothesis that there is less learning in the mixed condition: the mixed condition did not generate a smaller effect than the uniform condition. This suggests that the mechanism behind the target detection task examined in this paper was different than the mechanisms involved in word segmentation.

Simulations

Having discussed an anticipation account for prediction in the introduction, the purpose of the current simulations is the implement the process computationally, which can provide insights into the nature of the computation required to produce the results we found in the experiments. We directly model RTs in this simulation, with the simple idea that syllables are either predictable or unpredictable in the input stream. RTs for predictable syllables are generated with one pattern and RTs for unpredictable syllables are generated with another pattern.

This model is to process syllable sequences online, and to generate a RT for each syllable that is processed. At the beginning of processing a syllable sequence, the model assumes the learner to detect the target with a baseline amount of time, RT0, which is a constant. From this point on, the model stores each bigram it encounters. Based on the bigrams that are stored at any point, the next syllable is either predictable or unpredictable. The core assumptions are that 1) predictable syllables get a facilitation effect when it is reacted to, and 2) unpredictable syllables do not. As such, we propose a simple relation between the RT of a syllable occurring for the n^{th} time and the $n+1^{th}$ time, which is:

$$\mathrm{RT}(n+1) = \begin{cases} \mathrm{RT(n)} + \mathrm{stream\_pos} * \mathrm{stream\_inc} & if\ unpredictable \\ \mathrm{RT(n)} + \mathrm{stream\_pos} * \mathrm{stream\_inc} + \mathrm{occ\_inc} * \mathrm{occurrence} & if\ predictable \end{cases}$$

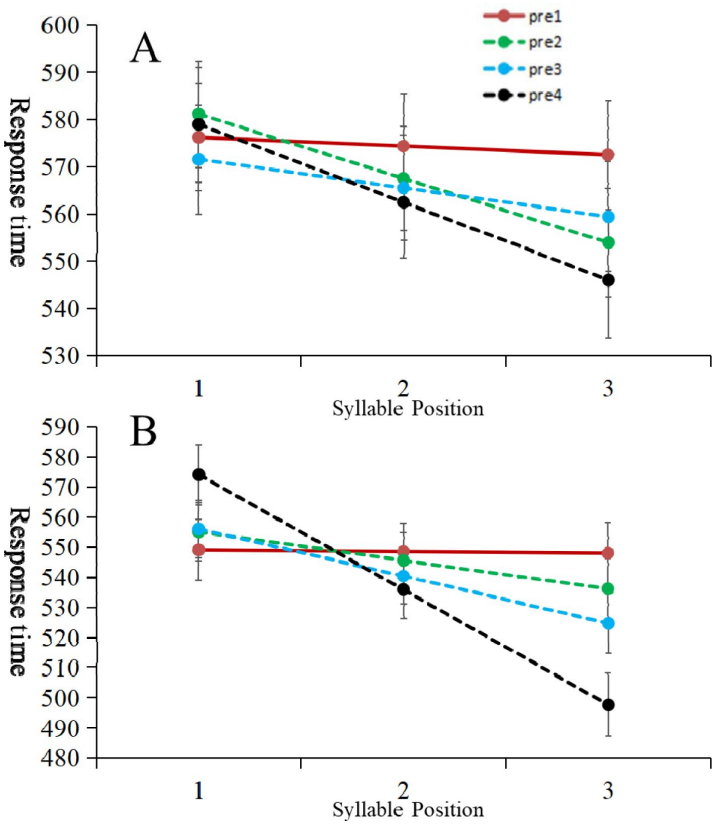

Figure 4

Regression model fit from the three-way interaction between condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous).

Figure 4A showed results from the uniform condition from **Experiment 1**, and Figure 4B showed results from the mixed condition from **Experiment 2**. Error bars represent ±1 SEM.

and

RT(1) = RT0 + stream_pos * stream_inc, where the n in RT(n) represents the RT for the n^{th} presentation of the target syllable, stream_pos is the position (3-46) in the stream, and occurrence is the number of occurrences that the syllable has occurred so far in the stream.

This process applies to the rest of the syllables in the sequence, until the end of the syllable stream. At this point, each syllable in the sequence will have a corresponding RT. To simulate the process of a participant reacting to a single target syllable, we will output the RTs corresponding to a random target syllable, such that the only data left for a syllable sequence are 4 RT values for the 4 occurrences of the target syllable.

Here is a more in-depth discussion of the assumptions behind this simple model. First, if a syllable occurs for the first time, we expect the learner to detect the target with a baseline amount of time. Theoretically, we take this to mean that it would take a certain amount of time to recognize and react to a syllable for the first time. Secondly, the next time the target syllable occurs, the amount of time it takes to react to the target syllable depends on whether this syllable is predictable or not. If it is unpredictable, the amount of time it takes to react is the same amount of time as the last time it was reacted to. If the target syllable is predictable, the amount of time it takes to react is different from the last time, by a constant (occ_inc) times the number of times this syllable has occurred so far. Theoretically, if it takes a certain amount of time to react to the target syllable the last time, this time, the reaction to the target syllable is facilitated by its predictability, where the amount is proportional to the number of times this target syllable has occurred so far. The assumption that the facilitation amount is proportional to the number of times the target syllable already occurred is based on the empirical finding that the more the target syllable was detected, the faster the RT is. The constant (occ_inc) represents the amount of facilitation effect due to predictability. In addition to the predictability factors, one more (positive) number needs to be added to each RT, which is a stream-position effect: the later the syllable is in the stream, the slower the RT is. This is also based on empirical findings from the task. For a discussion of the specifics of setting these parameters, see **Appendix**.

There are three parameters in our set of equations. The first, the baseline RT (RT0), does not factor into the pattern of data results later, as all RTs share this component equally. We set this RT0 to be the constant from the regression coefficient, from previous regressions. The second constant is the stream_inc, the increment amount for stream position. Again, it is common to all RTs. We set it as a small, positive number, which represents the general trend that RTs are larger the later the target is in the stream. The third constant is occ_inc, the increment for the number of targets that already occurred. We know this number to be negative (i.e., more occurrences would mean smaller RTs). We took a small, negative number from the corresponding regression coefficient. Notably, though we took the estimates from the regressions, this by no means would mean that the resulting RT distribution would resemble the RT distributions from the humans. The point of this simulation is to consider the properties of the model when we only consider very few factors (predictability/structure of the syllable sequence), and see if RT distributions based on these factors can be similar to the RT distributions from the human data.

To implement this model computationally, we went through a few steps. First, we constructed the syllable sequences, in the same way as we did in the experiments. Note that, during this step, there is randomness in constructing the syllable sequences, as different words can be concatenated in different orders while maintaining the constraints for the order (i.e., no words can follow itself). Next, we implement the target detection section of the task, randomly picking a target syllable in the syllable stream. We generated RTs for all syllables based on the formula described above, though, for the data from this simulation, only the RTs associated with the targets were saved in the data. To do this, in an online fashion, the model stores the bigrams that it has encountered so far, and calculates the RT of the next syllable based on the bigrams from the collection of bigrams

that are remembered, and the RT of the syllable from the last occurrence. Simply put, the RTs for the unpredictable syllables only include the baseline RT plus positive change as a function of the stream position. The RTs for the predictable syllables are a function of how predictable they are, on top of initial conditions. Again, note that no "word extraction" is required: the model only requires exposure to the input and stores the bigrams it encounters; There are bigrams that are predictable and unpredictable, and there is no need to make inferences over where the word boundaries are in the input sequence for the model to operate.

Given this model, we conducted two simulations, a uniform condition simulation, and a mixed condition simulation. These two simulations mirrored the structure of **Experiments 1** and **2** above, in terms of how the syllable sequences were set up. In each simulation, we generated the data for the same number of subjects (19, from Batterink 2017) and the same number of trials (144). For each trial, we generated the RT values according to the formula described above. Notably, the same parameters are used in both conditions. The simulations thus represent learners with the same learning characteristics: by using the same set of parameters going into the two conditions, we are assuming these learners behave the same for the two conditions.

Running the model generates simulated data for each condition. With the simulated data, we ran the same set of regressions as we did in the experiments. First, for each simulation, we looked at the (fixed) effect of syllable position (1-3), presentation (1-4), and their interaction, in addition to stream position (1-48). Next, we conducted a three-way interaction for syllable position (1-3), presentation (1-4), and condition (uniform/mixed). All these regressions included by-subject random intercepts and a random slope of stream position, the same as the regressions we ran for experiments.

The results for the model mirrored the qualitative pattern of data from human experiments. First, we found that the slope for the first presentation in the fitted model across three syllable presentations is the same, flat slope as we observed in the human data, for both the uniform and mixed conditions. Second, we found that there was a three-way interaction, the same way as the human results: The slope for the fourth presentation of the mixed condition is larger than in the uniform condition, given the same slopes for the first presentations in both conditions (**Figure 5**).

The fact that such a simple model can capture the same patterns from the human results is remarkable. The simplicity is based on the number of assumptions that went into the model, which are simply that predictable targets get shorter RTs, the amount of which is based on the number of times this particular target has occurred so far. This means that no other assumptions are required for the facilitation effect to occur. If one compares this model to other models for segmentation (e.g., the ones listed in Bernard et al., 2020), this model would have the least number of assumptions built in. More importantly perhaps, when we set the same parameter for the uniform and the mixed condition in the simulations, that is, setting the amount of change to be the same for predictable items in two conditions, we find that the same difference as we found in the human experiments, which is that the mixed condition showed a larger effect than the uniform condition. This may provide an explanation for our behavioral result, namely, that the larger effect in the mixed condition does not suggest that people reacted more quickly in the mixed condition, but is a reflection of mean lengths of the words in the syllable sequence – that is, the effect may be a result of total stream length difference between the two conditions (for a more thorough exploration of this effect, see the additional simulations in the Appendix). Importantly, for the current discussion on the origin of the difference, the same amount of facilitation effect from previous occurrences in our computational model provides a good fit for the human data.

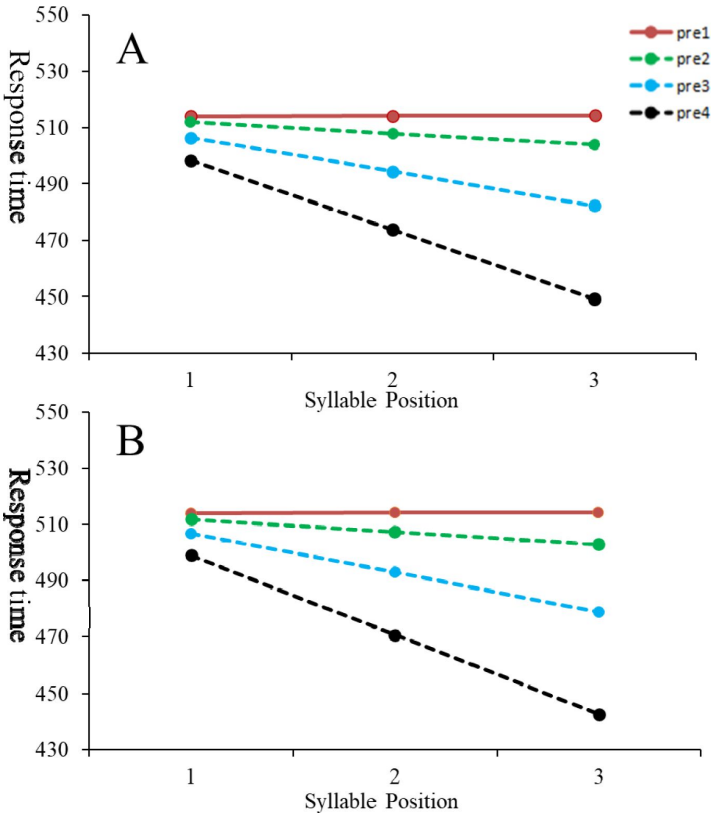

Figure 5

Regression model fit from the three-way interaction between condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous) for the simulated data.

Figure 5A showed results from the simulation for the uniform condition, and Figure 5B showed results from the simulation for the mixed condition.

General Discussion

This paper investigated the mechanisms involved in statistical word segmentation, reporting two experiments using the target detection task and comparing them to studies from the word segmentation literature. In Experiment 1, we reported a successful replication of Batterink (2017), including both a conceptual replication and an exact replication. The facilitation effect in question was successfully replicated, where the reaction time was shorter for predictable syllables (syllable position 2 and 3 in a triplet) compared to unpredictable syllables (syllable position 1) in later presentations (2, 3, 4) as opposed to the first presentation. In Experiment 2, we changed the structure of the syllable sequences in the study, where instead of using words of uniform length (which was the case for Experiment 1), we used sequences with mixed-length words. Such a change has been shown to generate a smaller amount of learning in the word segmentation literature, under the segmentation paradigm. However, with the target detection task, we found a similar facilitation effect in the mixed-length condition in Experiment 2, with the same speed that a single exposure was enough for this effect. Contrary to our prediction based on the segmentation literature that uniform-length sequences are learned better than mixed-length sequences, we found that the effect in the mixed condition (Experiment 2) was larger in the uniform condition (Experiment 1). To explain these results, we computationally modeled a prediction process, where the only assumptions in the model involved changes to the RT based on predictability. Simulations provided evidence that the same computational processes and parameters generated similar effects for both the uniform and mixed conditions. In fact, with the same facilitation parameter in both conditions, we found a larger effect in the mixed condition, and this is consistent with our data. We took this as evidence supporting the hypothesis that humans employed the same processes for mixed and uniform conditions. Taken together, these results suggest that a simple prediction-based anticipation mechanism can explain the results from the target detection task, and the mechanisms involved in this task may be different from the ones employed in word segmentation.

Mechanisms in the target detection tasks

What are the mechanisms behind the target detection task, if the mechanisms are not the same set of mechanisms involved in word segmentation? In the paragraphs below, we will discuss the following points. First, we discuss the mechanism behind the target detection task, which we argue to be a prediction-based anticipation mechanism. Notably, under such a mechanism, one exposure suffices for the prediction to occur. Secondly, we will provide a theoretical analysis of why segmentation based on distributional evidence would require more evidence. We argue that the prediction-based anticipation mechanism involves processes at word-internal locations, which is different from mechanisms for segmentation, which involves decisions at word boundaries. As such, whereas one exposure enables prediction-based anticipation, minimally two occurrences in a sequence are required for a set of syllables to form a word statistically.

First, we begin with a theoretical analysis that can shed light on the difference between the two tasks. As a general statement that has implications for all the discussion below, the two tasks require learners to use different information, which is used at different locations in the sequence. In terms of the information required in the target detection task, let's consider the following sequence: GHIABCDEFABCGHI, where the word "ABC" is preceded and followed by different words. By the second time "ABC" occurs (underlined in the sequence), the syllable B is preceded by the syllable A, and this is predictable because the AB transition occurred prior. This simple analysis suggests two things. First, the location of the information enabling the facilitation effect is word-internal – rather than word boundaries. Secondly, only one co-occurrence was enough for

the facilitation effect to occur during the second encounter, because the prediction occurs word-internally. Together, this means that one prior occurrence can enable learners to generate a prediction.

In the literature, this has been discussed as a prediction-based anticipation mechanism for statistical learning (e.g., Barakat, Seitz, & Shams, 2013; Davachi & DuBrow, 2015; Summerfield & De Lange, 2014; Turk-Browne et al., 2010). Under such a mechanism, the brain can encode the co-occurrences of stimuli from the past. In statistical learning terms, such a prediction-based anticipation mechanism can be viewed as a simpler version of a conditional probability model, where the conditional probability becomes 1 for two elements. That is to say, the conditional probability for two elements, after an initial encounter, can be calculated as 1 (i.e., $p(B|A) = 1$ when there is only a single encounter to AB). Notably, this fact is contrary to a specific claim in Batterink (2017), where it was argued that the calculation of conditional probabilities could not support the facilitation effect given a single exposure, because "the computation of conditional probabilities depends on accruing statistical data across a sample of input and cannot occur instantly after only a single exposure to an underlying pattern (Batterink, p. 926)". However, a single exposure does provide information about the transitions within the single exposure, and the probability of B given A can indeed be calculated from a single occurrence of AB. In our model, for example, the second time a predictable syllable occurs, it is marked predictable because it occurred one time prior, and another syllable can predict it.

This brings us to the discussion of the difference between the mechanisms that one needs to explain word segmentation and target detection. As we just discussed, one single occurrence of AB is enough for the prediction of B the next time A appears. However, for statistical segmentation, a single occurrence is not enough. Let's consider the example sequence above one more time, but this time only the section prior to the second occurrence of ABC (i.e., "GHIABCDEF"). Even if a learner can remember these syllables perfectly, there is no information for segmentation. That is to say, since all the syllables have occurred exactly once, there is no distributional information for segmentation. Only after ABC occurs the second time, word boundaries defined by distributional information begin to emerge: The fact that syllable A is preceded by different syllables (I and F) makes the forward transitional probability of I or F going to A to be 1/2, and the fact that syllable C is followed by different syllables (D and G) makes the forward transitional probability of C going to D or G to be 1/2. Thus, in this example, having had two exposures would enable the segmentation of ABC from this sequence (using a similar measure, such as backward transitional probability or mutual information, would require the same information). Going back to the location of the information for segmentation, it's clear that the decisions to segment require information at word boundaries; and concretely, prior to A, and after C. This also marks the difference to the prediction-based anticipation mechanism, where the critical information for the effect is word-internal. In sum, this theoretical analysis suggests that multiple exposures are required to make segmentation possible, whereas a single exposure could allow predictions between syllables to occur[1].

Thus, both the difference in terms of the location of the information and the information requirements for segmentation means that there is a disassociation of the mechanisms between the facilitation effect in the target detection task and the word segmentation task. In target detection, the decision to react to the target is local to positions that involve specific transitions in the sequence. In this sense, no segmentation is required; remembering bigrams, as we demonstrated in our model, would suffice for this task. However, the segmentation task requires the learner not only to segment the sequence, but also to remember the segmented subsequences in memory. To segment a single word requires two decisions for word boundary, and then, the segmentation task requires the segmented sequence to be remembered (i.e., only representing where the word boundaries are would not be enough.) These differences mean that, detecting targets from sequences with uniform-length and mixed-length words would generate a similar amount of learning (as evidenced by our experiments above). However, segmenting words from

the two types of sequences is differentially difficult, because sequences with mixed-length words are more complex (Johnson & Tyler, 2010; Wang, Trueswell, Zevin, & Mintz, under review). In sum, the two tasks may both require the learner to use co-occurrence information from the sequence, the two tasks require the learner to process different information to accomplish, and thus require different task-demands and mechanisms.

Time course for the facilitation and other similar effects

Through empirical work and a computational model, we provided evidence that the facilitation effect happened only after one exposure. However, for a complete theory for the time course of word segmentation, it's not the case that an exposure or two should be considered the whole picture. For example, even though learning was successful within the word segmentation paradigm with only two exposures, four exposures produced significantly more robust learning (Wang et al., 2023). At the same time, it's not the case that more exposure equals more learning. Other than the examples in the introduction, Bulgarelli and Weiss (2016) conducted a study looking at the time course of learning. Participants were presented with multiple 67-second syllable sequences (which contained hundreds of syllables), and tested between the presentation of each syllable sequence. Learning plateaued after a single block of learning, where the effect size of learning never changed following the first block or after several blocks of learning. In sum, the relationship between exposure and learning is complicated, requiring an examination of the cognitive mechanisms involved in segmentation as a function of time and complexity of the learning materials (e.g., sequences with uniform- vs. mixed-length words), a topic for future work.

The timing characteristics of target detection may be unique in the literature, as most tasks cannot detect learning so quickly. For example, even though serial reaction time (SRT) tasks have also been used to examine the learning of statistical dependencies (e.g., Howard & Howard, 1997; Hunt & Aslin, 2001; Wang & Kaiser, 2022), the effect emerges much slower. The difference between SRT tasks and the target detection task is that, in SRT tasks, participants make a key press for every stimulus, whereas the target detection task requires key presses only for a single target. The slow emergence of the learning effect may have to do with the fact that making a key press for every stimulus requires the learner to pay constant attention to the upcoming stimulus in order for an action (making a key press). In contrast, in target detection tasks, there is no action required for most of the stimulus, so that the participants may plan their action while processing the stimuli. For example, in Hunt and Aslin (2001), participants completed 70-word sessions, and completed 8 sessions a day for 6 consecutive days. While the question of how many sessions are required to produce a reliable effect was not explored directly in that study, the data showed participants took multiple sessions to show a learning effect in many experiments. Notably, even though the target detection task has been used in other studies (Bertels, Boursain, Destrebecqz, & Gaillard, 2014; Bertels, Demoulin, Franco, & Destrebecqz, 2013; Bertels, Franco, & Destrebecqz, 2012; Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015; Kim, Seitz, Feenstra, & Shams, 2009; Turk-Browne et al., 2010), Batterink (2017) was the first study to demonstrate that learners can show a facilitation effect after a single exposure to our knowledge. Most of the other studies using the target detection task (e.g., Franco et al., 2015) provided an exposure phase to the participants before the target detection task began, making it unclear when the effect arose. Our current study provides further empirical evidence that the facilitation effect for predictable syllables emerges given a single exposure, demonstrating that the effect is equally applicable when learning from uniform-length and mixed-length sequences.

Target Detection and PARSER

Zooming in on the facilitation effect in the target detection task specifically, one of the candidates for explaining the effect involves clustering (Batterink, 2017). Batterink (2017) discussed that clustering may explain the data better than the use of conditional probabilities, because obtaining conditional probabilities may require more than a single exposure, citing a computational model known as PARSER (Perruchet & Vinter, 1998). PARSER accomplishes segmentation in two

iterative steps. In the first step, PARSER randomly picks a number from 1 through n (typically 3), and clusters this random number of syllables as a chunk. This step creates chunks and stores them in memory with certain weights associated with each one (termed Perceptual Shaper). In the second step, PARSER either strengthens the weight or decreases the weight of items in the Perceptual Shaper: If the incoming chunk matches an existing chunk, the weight of the existing chunk (and its components) is increased. However, if the incoming chunk is completely new, it is added to the Perceptual Shaper, but at the same time, the weights of all the previous chunks are decreased. The updating of the weights occurs in time steps, and the two steps occur during each time step. With this iterative process, PARSER can successfully segment a syllable sequence into its component words, because these words (and their components) are more likely to repeatedly occur, much more likely than part-words.

So, can PARSER explain learning after a single exposure? To answer this question empirically, we created a simulation. In this simulation, we used the U-Learn program (Perruchet, Robinet, & Lemaire, 2014) to examine the learning of short sequences (see Appendix). Notably, there is a lack of a linking assumption translating the weights of different chunks to RT differences in the target detection. Here, we asked PARSER to evaluate words vs. part-words, which is a function built-in to the U-Learn program, and made a linking assumption: if the words and part-words are differentially weighted, this is equivalent to the facilitation effect in target detection (i.e., we take the learning effect from PARSER to indicate learning). The U-Learn program reports the rate words are preferred over part-words in 10 time-steps (which corresponds to 1/10 of the learning sequence, however long the learning sequence is). Thus, if we use 4 different words each of which occurs 10 times (modifying the existing "ready-to-use configurations"), 1/10 of the syllable sequence is 4 words long and 2/10 of the syllable sequence is 8 words long. Running the simulation 50 times to represent running 50 subjects on this task, we find that, out of 50 times, the percentage of time where words were segmented but part-words were not during the first 1/10 of the sequence was 0 times, and this percentage became 1% after 2/10 of the sequence. Thus, it's not the case that PARSER can successfully segment words following a single exposure. In this instance, it would appear that humans are better learners than PARSER. In a second simulation, we created the training sequences from Johnson and Tyler (2010) which contained both sequences with uniform-length and mix-length words. The result of the simulation was that PARSER was equally successful with the uniform- and mixed-length conditions (see Appendix). In this instance, PARSER is perhaps more powerful than humans in terms of segmentation. Thus, it would appear that PARSER cannot account for the kind of results that humans produce, where it is not as sensitive to statistical regularities as humans in a target detection task, and too powerfully equipped to learn when humans would have trouble.

Conclusions

In summary, the current study found that the facilitation effect from the target detection task is empirically robust, and can be shown with sequences with uniform-length words or mixed-length words alike. The speed for a facilitation effect following a predictable sequence to appear is indeed at its theoretical limit of just one prior encounter. Through empirical findings and a computational model, we provided a possible mechanism to explain this facilitation effect from the target detection task. Furthermore, by comparing the current results in the target detection task to work from the word segmentation literature, we argued that the mechanisms involved in the target detection task are different from the word segmentation task. Future exploration is needed to understand the relationship between the amount of exposure and learning in statistical word segmentation, as well as a characterization of the memory mechanisms that are involved during the segmentation process.

Note

This reviewed preprint has been updated to correct the corresponding author's name.

Appendix

Before we present our simulations in the Appendix, here is a summary of the simulations below. In Part 1, we simulated with the computational model described in the paper. The purpose of Simulation 1 is to show the robustness of the results given a range of parameters, and that the model behavior is the same regardless of the parameter values. Next, we ask a bigger question, which is why the effect sizes in the mixed condition is larger than in the uniform condition. We test the hypothesis that, the effect sizes were a function of mean word length. By manipulating 5 different possible types of sequences, the purpose of Simulation 2 is to show that given the same structure and parameters of the model, the effect size of the facilitation effect is correlated with the mean word length.

In Part 2, we present a different set of simulations with PARSER. In Simulation 1, we want to track the relationship between the amount of exposure and learning. By giving PARSER a small amount of data, the model can show us how much data was needed to get PARSER started with segmentation. This simulation shows that, given a single exposure, there is no learning from PARSER; in fact, with two exposures to a word form, there is still no learning. In Simulation 2, we simulate the Johnson and Tyler (2010) experiment, asking whether PARSER would be sensitive to the type of word length (uniform/mixed) during segmentation. We found that PARSER can segment both conditions equally well, and this is different from findings from humans.

In Part 3, we present an additional analysis of the difference between the uniform and mixed conditions. In this analysis, we used the data from the conceptual replication. The results showed the same pattern as the analysis in the main text.

Part 1. Simulations with the present computational model

Simulation 1

In this first simulation, we test a range of parameter values and show that the simulation is robust to the choice of the parameters. As we said in the paper, the parameters include the baseline RT (RT0), stream_inc, the increment amount for stream position, and occ_inc, the increment for the number of targets that already occurred. The first two parameters do not factor into the behavior of the model (i.e., anything that leads to the facilitation effect), as the model is mostly concerned with the facilitation effect, which is a result of the interaction between syllable position and presentation. Neither the baseline RT nor the increment for stream position would influence this interaction. For the simulations reported in the paper and the simulations in the rest of this Appendix, we set the baseline RT value (RT0) to 500ms, and the increment amount for stream position to 0.72ms, all positive numbers. Note that, we could have drawn these values as a random number from a distribution, but again, such choices would not influence the interaction of interest. A priori, for the purpose of this simulation, we only considered the third parameter, the increment for the number of targets that already occurred, to hold any potential to influence the interaction.

From the outset, we knew this increment to be a negative number, because predictable syllables were reacted to faster than unpredictable syllables. To see how this parameter influenced the interaction between syllable position and presentation, we wanted to manipulate this parameter within a range. On the larger side of the range, the value can be a negative number close to 0. For our purpose, -0.1ms was a number close to 0 that is still meaningful for RT values. On the smaller side of the range, we did the following calculation. If this increment was a large number with a negative sign (say -1000), it would mean that the RT in presentation 4 would be smaller than 0 (i.e., RT0 plus 3 times this negative number), which was impossible. Thus, by setting predictable RTs in presentation 4 to be a positive number near 0, we calculated that the increment number was close to -70ms. Thus, -70ms was used as the larger end of the range.

Thus, we took 5 different values ranging from -0.1ms to -70ms, and ran the model to generate the simulated data. For data from each simulation, we conducted the same regression, and looked at the plots of the regression predictions. The regression estimates for shown in Table A1. We saw that for all of the different instances of occ_inc, the interaction between syllable position and presentation went the same direction, where the slope for presentation 1 was flat, and became more negative as the presentation number increased. The slope for presentation 4 was the most negative in all instances. Thus, we concluded that the parameter values, with a range where the values were reasonable, did not qualitatively change the results of the simulations reported in the paper.

Simulation 2

In this simulation, we test 5 different ways of constructing a sequence in the target detection paradigm, and processed the data in the same way to explain why the mixed condition generated a larger effect than in the uniform condition. Our hypothesis is that the shorter the component words are (in terms of the number of syllables), the more negative the slope of the later presentations. To test this hypothesis, we created 5 different ways of constructing a sequence, which are listed in Table A2.

As is shown in Table A2, we created 5 different ways of constructing a sequence, manipulating the mean lengths of words. All these conditions included 4 different words in them with unique syllables. The mean lengths of 2.5 and 3 are the same as the mixed and uniform conditions, respectively.

After we generated the RT values from the model, we conducted the same set of regressions for each model. The RT was the dependent variable, and the independent variable included fixed effects of word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous), and the interaction between the two. Random effects included participant as a random intercept and stream position as a random slope. Of particular interest are the beta estimates for the two-way interactions, specifically, the estimates between presentations 1 and 2, presentations 1 and 3, and presentations 1 and 4. The prediction is that, the smaller the mean length of words is, the larger the changes are for pairs of presentations, and the largest for presentations 1 and 4.

The results from the regressions are shown in Table A3.

From these regression estimates, we see that our prediction is confirmed. We see that these effects grow linearly with respect to the mean length of words for the five conditions we tested. In addition, whether the syllable sequence was uniform or mixed in length did not matter, and only the mean length was predictive of the beta coefficients.

occ_inc values / pairs of presentation	-0.1	-17.6	-35.1	-52.6	-70
Presentation 1 and 2	-.046	-8.352	-16.362	-24.020	-32.663
Presentation 1 and 3	-.137	-23.872	-48.712	-72.215	-97.366
Presentation 1 and 4	-.285	-50.091	-99.800	-147.916	-198.262

Table A1

The beta estimates for regressions for different simulations. On the top, we show the different occ_inc values we used in each simulation. On the left, we list the different pairs of presentation that interact with syllable position, and in the table, the beta coefficient for these interactions are shown.

mean length	content (words by number)
2	4 disyllabic words
2.5	2 disyllabic words, 2 trisyllabic words
3	4 trisyllabic words
3.5	2 trisyllabic words, 2 quadruple-syllabic words
4	4 quadruple-syllabic words

Table A2

The content of the present simulation.

Mean word length	2	2.5	3	3.5	4
Presentation 1 and 2	-8.145	-4.634	-4.132	-2.653	-2.491
Presentation 1 and 3	-23.947	-14.043	-12.178	-8.440	-7.449
Presentation 1 and 4	-49.316	-28.181	-24.678	-16.775	-14.909

Table A3

The beta estimates for regressions for different simulations.

Part 2. Simulations with PARSER

Simulation 1

In this simulation, we used the U-Learn program (Perruchet, Robinet, & Lemaire, 2014) to examine the learning of short sequences. Here, we asked PARSER to evaluate words vs. part-words, which is a function built-in to the U-Learn program. The U-Learn program reports the rate words are preferred over part-words in 10 time steps (which corresponds to 1/10 of the learning sequence, however long the learning sequence is). Thus, if we use 4 different words each of which occurs 10 times (modifying the existing Aslin et al. 1998 "ready-to-use configurations"), 1/10 of the syllable sequence is 4 words long and 2/10 of the syllable sequence is 8 words long. We created the sequence and the test items by modifying the ready-to-se configurations (**Figure A1**).

Using this setup, we ran the simulation 50 times to represent running 50 subjects on this task. The results are shown in **Figure A2**. We find that, PARSER is successful after finishing running the 40-word sequence most of the time in simulation. The crucial question for the current simulation is whether there is any learning after $2/10^{th}$ of the sequence. Observing both the percentage and weight changes in the learning curve, and we see that there is no learning.

Simulation 2

In this simulation, we used the U-Learn program to examine the learning of uniform and mixed conditions in Johnson and Tyler, 2010. The uniform condition was the same as the existing Aslin et al. 1998 configurations, and the mixed condition was modified to include two disyllabic and two trisyllabic words, with all test items being disyllabic. In both conditions, the four words had an unbalanced frequency profile (45, 45, 90, 90). The mixed condition setup configuration is shown in **Figure A3**.

Using this set-up, we ran the simulation 50 times each in the uniform and mixed conditions to represent running 50 subjects on this task. The results are shown in **Figure A4**. We find that, PARSER is successful in both conditions. In the uniform condition, words were successfully segmented 79% of the time on average, and part-words were segmented 4% of the time on average. In the mixed condition, words were successfully segmented 82% of the time on average, and part-words were segmented 1% of the time on average. Thus, we see that PARSER is capable of learning in both the mixed and uniform conditions.

Part 3. Additional analyses between the uniform and mixed conditions

The main text mentioned that the difference between the uniform and mixed condition is the same, whether the conceptual or exact replication data was used to compare to the mixed condition. The following section shows this analysis.

For this analysis, we compared the data from **Experiment 2** to the conceptual-replication condition in **Experiment 1**. The same set of regression setup was used. In this mixed effect regression with a three-way interaction, the RT was the dependent variable, and the independent variable included fixed effects of condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous), and the interaction between the three. Fixed effect further included overall stream position (3rd through 46th syllable in the syllable sequence in the uniform condition, 3rd through 38th syllable in the mixed condition, both continuous) and word length (disyllabic/trisyllabic, categorical). Random effects included participant as a random intercept and stream position as a random slope.

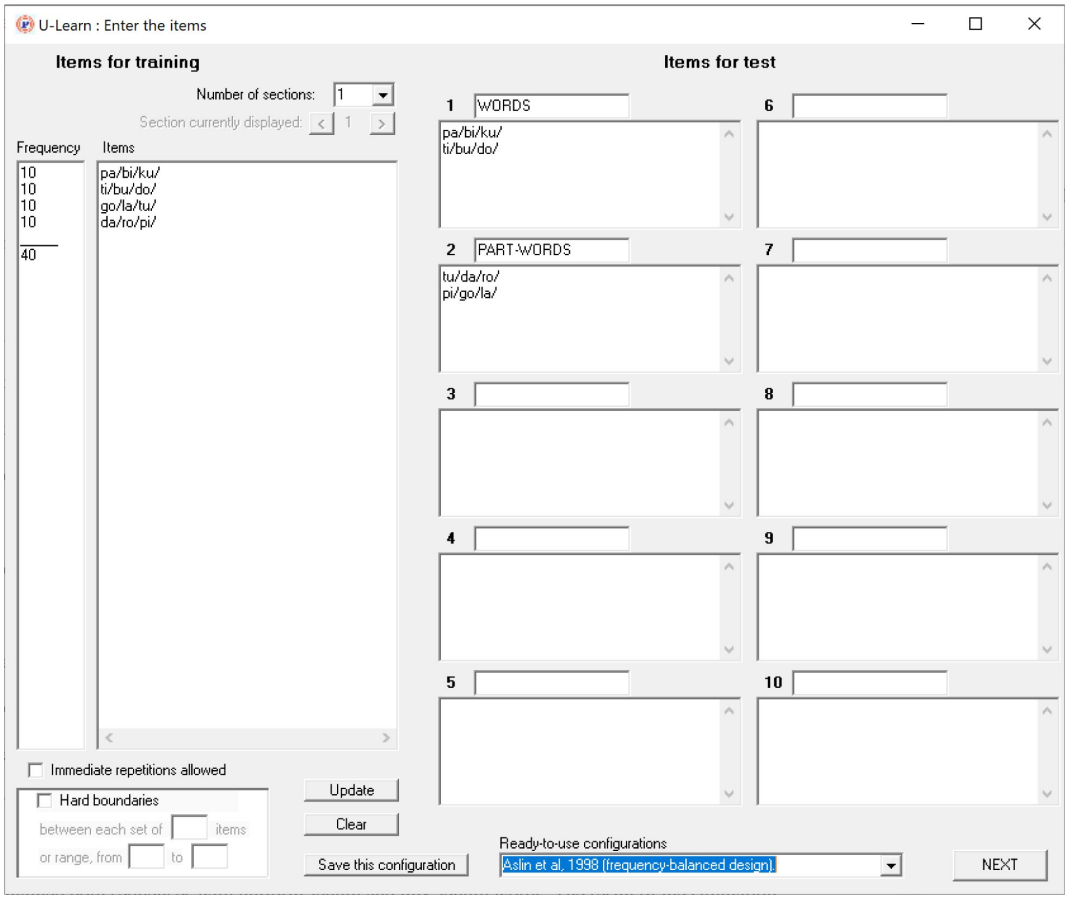

Figure A1

The set-up for simulation one, where four trisyllabic words occur 10 times each.

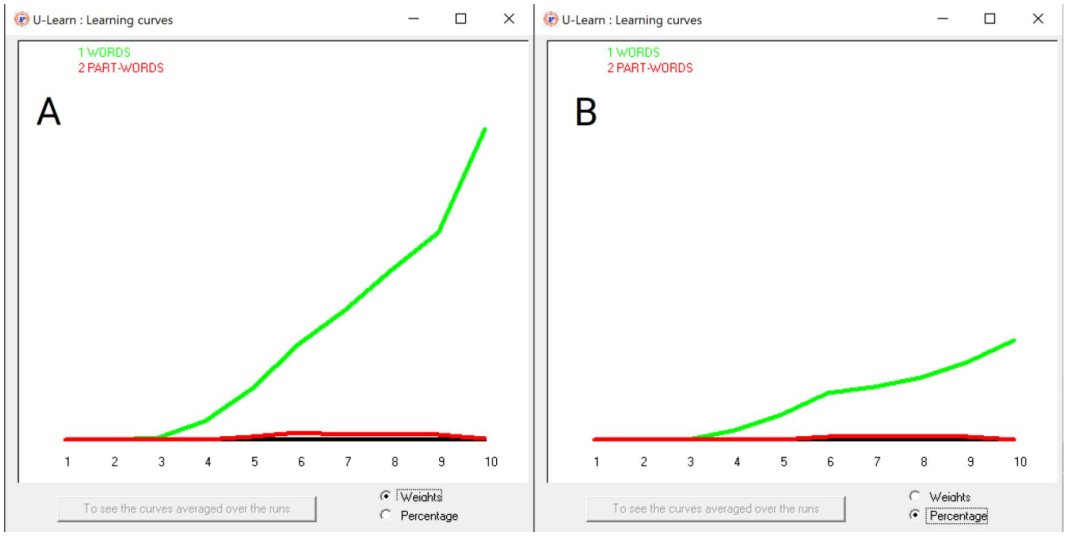

Figure A2

The results for the first simulation, where Figure 2A shows the weight changes of words and part-words over the course of learning and Figure 2B shows the percentage of words and part-words discovered over the course of learning.

Importantly, at time 2 on the x-axis, there is no learning of words. The y-axis represents the weights/percentages, but since these units are arbitrary and only meaningful when comparing two curves, the units are not displayed in the plotting function of U-Learn.

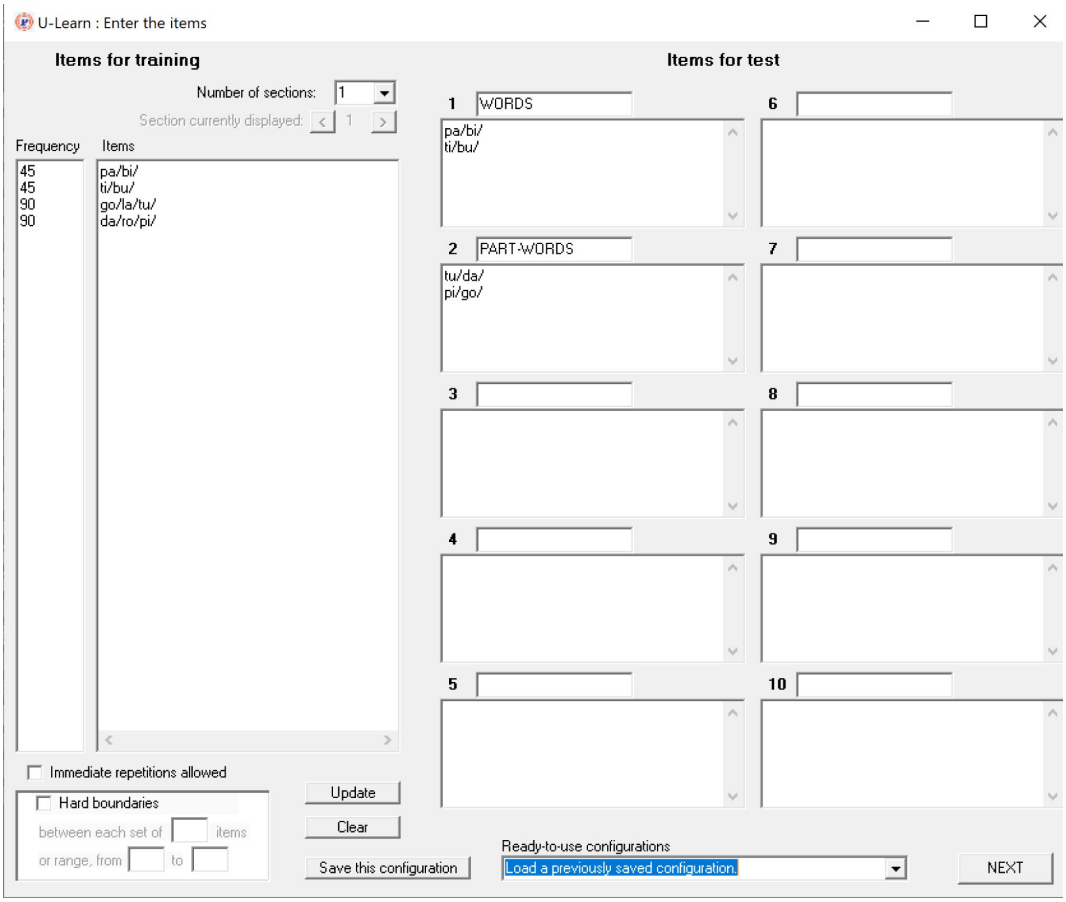

Figure A3

The set-up for the mixed condition in Simulation 2.

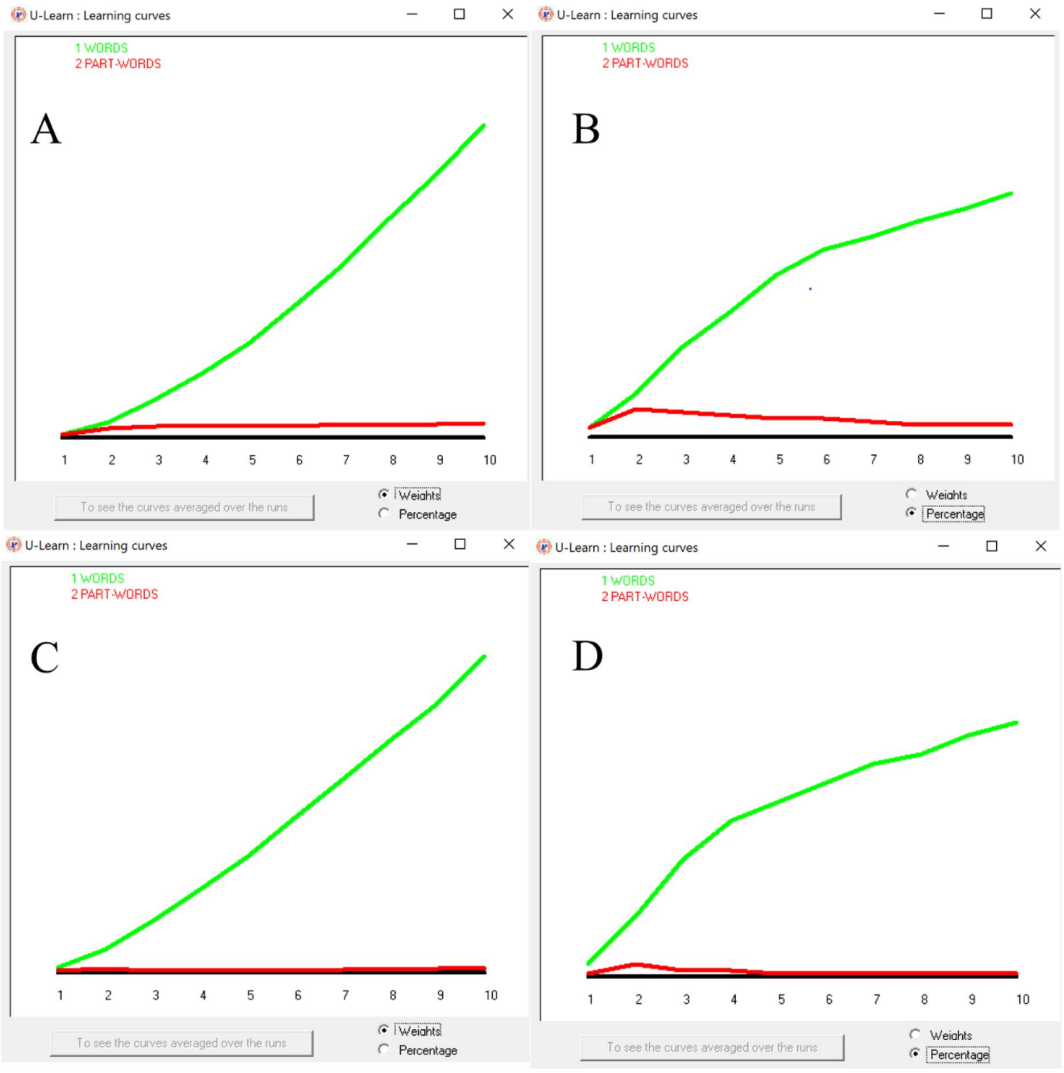

Figure A4

The results for the second simulation, where Figures 4A and 4B show the weight and percentage changes of words and part-words over the course of learning in the uniform condition, and Figures 4C and 4D show the weight and percentage changes of words and part-words over the course of learning in the mixed condition.

On the x-axis, each number represents 1/10 of the sequence. The y-axis represents the weights/percentages, but since these units are arbitrary and only meaningful when comparing two curves, the units are not displayed in the plotting function of U-Learn.

The omnibus three-way interaction was significant ($\chi^2(3) = 16.06$, p=0.001), suggesting that the ways syllable position and presentation interact in the two experiments are different. To understand this three-way interaction, we looked at the three-way interaction between syllable position, condition, and pairs of presentations (i.e., presentations 1 and 2, 1 and 3, and 1 and 4). We found that the three-way interactions for presentations 1 and 2 (β=0.001, z=0.26, p=0.797) was positive and not significant, and positive and not significant for presentations 1 and 3 (β=0.005, z=0.85, p=0.396), and positive and significant for presentation 1 and 4 (β=0.022, z=3.52, p<0.001). This is the same pattern as the analyses in the main text, where the coefficients grow as a function of presentation in this three-way interaction, same as the exact-replication condition. Looking at a plot of model fit (**Figure A5**), this pattern becomes clear: while the slopes (from syllable position 1 to 3) for presentation 1 were similarly non-negative for both conditions, the negative slope for presentation 4 for the mixed condition was the largest slope (from 569ms to 493ms) for all slopes, more than in presentation 4 for the uniform condition (from 572ms to 554ms).

References

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

Barakat B. K., Seitz A. R., Shams L. (2013) **The effect of statistical learning on internal stimulus representations: Predictable items are enhanced even when not predicted** *Cognition* **129**:205–211

Batterink L. J. (2017) **Rapid statistical learning supporting word extraction from continuous speech** *Psychological Science* **28**:921–928

Bernard M., Thiolliere R., Saksida A., Loukatou G. R., Larsen E., Johnson M., Cristia A. (2020) **WordSeg: Standardizing unsupervised word form segmentation from text** *Behavior research methods* **52**:264–278

Bertels J., Boursain E., Destrebecqz A., Gaillard V. (2015) **Visual statistical learning in children and young adults: how implicit?** *Frontiers in Psychology* **5**

Bertels J., Demoulin C., Franco A., Destrebecqz A. (2013) **Side effects of being blue: influence of sad mood on visual statistical learning** *PloS one* **8**

Bertels J., Franco A., Destrebecqz A. (2012) **How implicit is visual statistical learning?** *Journal of Experimental Psychology: Learning, Memory, and Cognition* **38**

Conway C. M. (2020) **How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning** *Neuroscience & Biobehavioral Reviews* **112**:279–299

Davachi L., DuBrow S. (2015) **How the hippocampus preserves order: the role of prediction and context** *Trends in cognitive sciences* **19**:92–99

Endress A. D., Mehler J. (2009) **The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words** *Journal of Memory and Language* **60**:351–367

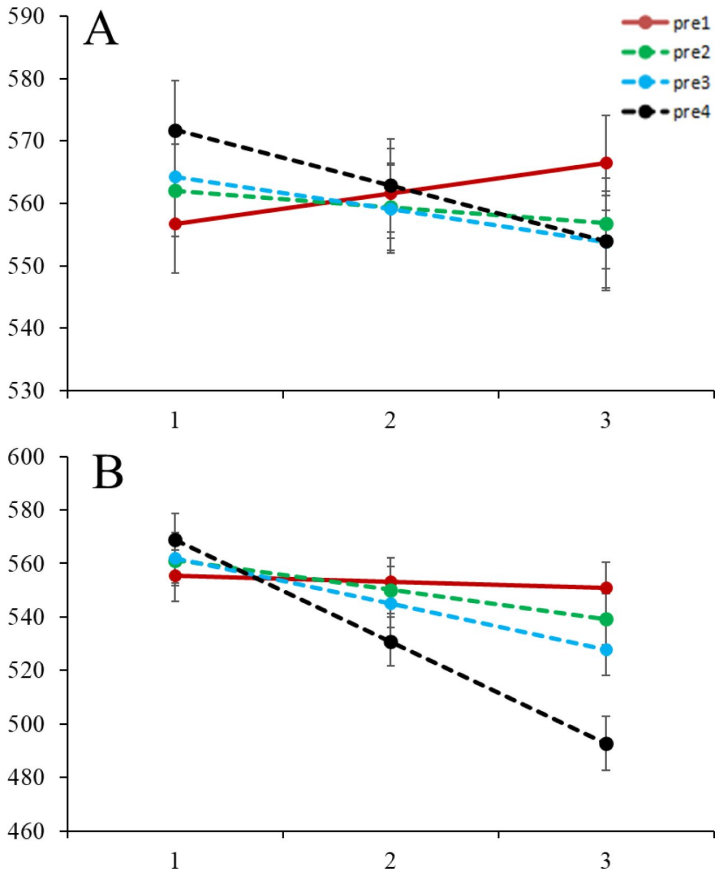

Figure A5

Regression model fit from the three-way interaction between condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous).

Figure A5A showed results from the conceptual-replication condition from **Experiment 1**, and Figure A5B showed results from the mixed condition from **Experiment 2**. Error bars represent ±1 SEM.

Estes K. G., 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**:254–260

Finn A. S., Kam C. L. H. (2008) **The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation** *Cognition* **108**:477–499

Franco A., Eberlen J., Destrebecqz A., Cleeremans A., Bertels J. (2015) **Rapid serial auditory presentation** *Experimental psychology* **62**:346–351

Frank M. C., Goldwater S., Griffiths T. L., Tenenbaum J. B. (2010) **Modeling human performance in statistical word segmentation** *Cognition* **117**:107–125

Giroux I., Rey A. (2009) **Lexical and sublexical units in speech perception** *Cognitive Science* **33**:260–272

Hoch L., Tyler M. D., Tillmann B. (2013) **Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages** *Psychonomic Bulletin & Review* **20**:142–147 https://doi.org/10.3758/s13423-012-0309-8

Howard J. H., Howard D. V. (1997) **Age differences in implicit learning of higher order dependencies in serial patterns** *Psychology and Aging* **12**:634–656 https://doi.org/10.1037/0882-7974.12.4.634

Hunt R. H., Aslin R. N. (2001) **Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners** *Journal of Experimental Psychology: General* **130**

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**:548–567

Johnson E. K., Jusczyk P. W. (2003) **Exploring possible effects of language-specific knowledge on infants' segmentation of an artificial language** *Jusczyk Lab Final Report* :141–148

Johnson E. K., Jusczyk P. W. (2003) **Exploring statistical learning by 8-month-olds: The role of complexity and variation** *Jusczyk Lab Final Report* :141–148

Johnson E. K., Tyler M. D. (2010) **Testing the limits of statistical learning for word segmentation** *Developmental Science* **13**:339–345

Jusczyk P. W. (1999) **How infants begin to extract words from speech** *Trends in cognitive sciences* **3**:323–328

Jusczyk P. W., Aslin R. N. (1995) **Infants' detection of the sound patterns of words in fluent speech** *Cognitive psychology* **29**:1–23

Jusczyk P. W., Houston D. M., Newsome M. (1999) **The beginnings of word segmentation in English-learning infants** *Cognitive psychology* **39**:159–207

Kim R., Seitz A., Feenstra H., Shams L. (2009) **Testing assumptions of statistical learning: is it long-term and implicit?** *Neuroscience letters* **461**:145–149

Newport E. L., Aslin R. N. (2004) **Learning at a distance I. Statistical learning of non-adjacent dependencies** *Cognitive psychology* **48**:127–162

Perruchet P., Vinter A. (1998) **PARSER: A model for word segmentation** *Journal of Memory and Language* **39**:246–263

Perruchet P., Robinet V., Lemaire B. (2014) **U-Learn: Finding optimal coding units from unsegmented sequential databases**

Saffran J. R., Aslin R. N., Newport E. L. (1996) **Statistical learning by 8-month-old infants** *Science* **274**:1926–1928

Summerfield C., De Lange F. P. (2014) **Expectation in perceptual decision making: neural and computational mechanisms** *Nature Reviews Neuroscience* **15**:745–756

Swingley D. (2005) **Statistical clustering and the contents of the infant vocabulary** *Cognitive Psychology* **50**:86–132

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

Wang F. H., Kaiser E. (2022) **Linguistic Priming and Learning Adjacent and Nonadjacent Dependencies in Serial Reaction Time Tasks** *Language Learning* **72**:695–727

Wang F. H., Luo M., Wang S. (2023) **Statistical word segmentation succeeds given the minimal amount of exposure** *Psychonomic Bulletin & Review* :1–9

Wang F. H., Trueswell J., Zevin J., Mintz T. H. **Wang, F. H., Trueswell, J., Zevin, J., & Mintz, T. H. (under review). Repetition induced rhythm as an alternative account to statistical word segmentation: A model and meta-analysis.**

Article and author information

Meili Luo
School of Psychology, Nanjing Normal University, Nanjing, Jiangsu, China

Ran Cao
School of Psychology, Nanjing Normal University, Nanjing, Jiangsu, China

Felix Hao Wang
School of Psychology, Nanjing Normal University, Nanjing, Jiangsu, China
For correspondence: haowang1@sas.upenn.edu

Copyright

Editors

Reviewing Editor
Andrea Martin
Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Senior Editor
Floris de Lange
Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

Reviewer #1 (Public Review):

Summary:

This paper presents two experiments, both of which use a target detection paradigm to investigate the speed of statistical learning. The first experiment is a replication of Batterink, 2017, in which participants are presented with streams of uniform-length, trisyllabic nonsense words and asked to detect a target syllable. The results replicate previous findings, showing that learning (in the form of response time facilitation to later-occurring syllables within a nonsense word) occurs after a single exposure to a word. In the second experiment, participants are presented with streams of variable-length nonsense words (two trisyllabic words and two disyllabic words) and perform the same task. A similar facilitation effect was observed as in Experiment 1. The authors interpret these findings as evidence that target detection requires mechanisms different from segmentation. They present results of a computational model to simulate results from the target detection task and find that an "anticipation mechanism" can produce facilitation effects, without performing segmentation. The authors conclude that the mechanisms involved in the target detection task are different from those involved in the word segmentation task.

Strengths:

The paper presents multiple experiments that provide internal replication of a key experimental finding, in which response times are facilitated after a single exposure to an embedded pseudoword. Both experimental data and results from a computational model are presented, providing converging approaches for understanding and interpreting the main results. The data are analyzed very thoroughly using mixed effects models with multiple explanatory factors.

Weaknesses:

In my view, the main weaknesses of this study relate to the theoretical interpretation of the results.

(1) The key conclusion from these findings is that the facilitation effect observed in the target detection paradigm is driven by a different mechanism (or mechanisms) than those involved in word segmentation. The argument here I think is somewhat unclear and weak, for several reasons:

First, there appears to be some blurring in what exactly is meant by the term "segmentation" with some confusion between segmentation as a concept and segmentation as a paradigm. Conceptually, segmentation refers to the segmenting of continuous speech into words. However, this conceptual understanding of segmentation (as a theoretical mechanism) is not necessarily what is directly measured by "traditional" studies of statistical learning, which typically (at least in adults) involve exposure to a continuous speech stream followed by a forced-choice recognition task of words versus recombined foil items (part-words or

nonwords). To take the example provided by the authors, a participant presented with the sequence GHIABCDEFABCGHI may endorse ABC as being more familiar than BCG, because ABC is presented more frequently together and the learned association between A and B is stronger than between C and G. However, endorsement of ABC over BCG does not necessarily mean that the participant has "segmented" ABC from the speech stream, just as faster reaction times in responding to syllable C versus A do not necessarily indicate successful segmentation. As the authors argue on page 7, "an encounter to a sequence in which two elements co-occur (say, AB) would theoretically allow the learner to use the predictive relationship during a subsequent encounter (that A predicts B)." By the same logic, encoding the relationship between A and B could also allow for the above-chance endorsement of items that contain AB over items containing a weaker relationship.

Both recognition performance and facilitation through target detection reflect different outcomes of statistical learning. While they may reflect different aspects of the learning process and/or dissociable forms of memory, they may best be viewed as measures of statistical learning, rather than mechanisms in and of themselves.

(2) The key manipulation between experiments 1 and 2 is the length of the words in the syllable sequences, with words either constant in length (experiment 1) or mixed in length (experiment 2). The authors show that similar facilitation levels are observed across this manipulation in the current experiments. By contrast, they argue that previous findings have found that performance is impaired for mixed-length conditions compared to fixed-length conditions. Thus, a central aspect of the theoretical interpretation of the results rests on prior evidence suggesting that statistical learning is impaired in mixed-length conditions. However, it is not clear how strong this prior evidence is. There is only one published paper cited by the authors - the paper by Hoch and colleagues - that supports this conclusion in adults (other mentioned studies are all in infants, which use very different measures of learning). Other papers not cited by the authors do suggest that statistical learning can occur to stimuli of mixed lengths (Thiessen et al., 2005, using infant-directed speech; Frank et al., 2010 in adults). I think this theoretical argument would be much stronger if the dissociation between recognition and facilitation through RTs as a function of word length variability was demonstrated within the same experiment and ideally within the same group of participants.

(3) The authors argue for an "anticipation" mechanism in explaining the facilitation effect observed in the experiments. The term anticipation would generally be understood to imply some kind of active prediction process, related to generating the representation of an upcoming stimulus prior to its occurrence. However, the computational model proposed by the authors (page 24) does not encode anything related to anticipation per se. While it demonstrates facilitation based on prior occurrences of a stimulus, that facilitation does not necessarily depend on active anticipation of the stimulus. It is not clear that it is necessary to invoke the concept of anticipation to explain the results, or indeed that there is any evidence in the current study for anticipation, as opposed to just general facilitation due to associative learning.

In addition, related to the model, given that only bigrams are stored in the model, could the authors clarify how the model is able to account for the additional facilitation at the 3rd position of a trigram compared to the 2nd position?

(4) In the discussion of transitional probabilities (page 31), the authors suggest that "a single exposure does provide information about the transitions within the single exposure, and the probability of B given A can indeed be calculated from a single occurrence of AB." Although this may be technically true in that a calculation for a single exposure is possible from this formula, it is not consistent with the conceptual framework for calculating transitional probabilities, as first introduced by Saffran and colleagues. For example, Saffran et al. (1996, Science) describe that "over a corpus of speech there are measurable statistical regularities that distinguish recurring sound sequences that comprise words from the more accidental

sound sequences that occur across word boundaries. Within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another within a word, whereas transitional probabilities spanning a word boundary will be relatively low." This makes it clear that the computation of transitional probabilities (i.e., Y | X) is conceptualized to reflect the frequency of XY / frequency of X, over a given language inventory, not just a single pair. Phrased another way, a single exposure to pair AB would not provide a reliable estimate of the raw frequencies with which A and AB occur across a given sample of language.

(5) In experiment 2, the authors argue that there is robust facilitation for trisyllabic and disyllabic words alike. I am not sure about the strength of the evidence for this claim, as it appears that there are some conflicting results relevant to this conclusion. Notably, in the regression model for disyllabic words, the omnibus interaction between word presentation and syllable position did not reach significance (p= 0.089). At face value, this result indicates that there was no significant facilitation for disyllabic words. The additional pairwise comparisons are thus not justified given the lack of omnibus interaction. The finding that there is no significant interaction between word presentation, word position, and word length is taken to support the idea that there is no difference between the two types of words, but could also be due to a lack of power, especially given the p-value (p = 0.010).

(6) The results plotted in Figure 2 seem to suggest that RTs to the first syllable of a trisyllabic item slow down with additional word presentations, while RTs to the final position speed up. If anything, in this figure, the magnitude of the effect seems to be greater for 1st syllable positions (e.g., the RT difference between presentation 1 and 4 for syllable position 1 seems to be numerically larger than for syllable position 3, Figure 2D). Thus, it was quite surprising to see in the results (p. 16) that RTs for syllable position 1 were not significantly different for presentation 1 vs. the later presentations (but that they were significant for positions 2 and 3 given the same comparison). Is this possibly a power issue? Would there be a significant slowdown to 1st syllables if results from both the exact replication and conceptual replication conditions were combined in the same analysis?

(7) It is difficult to evaluate the description of the PARSER simulation on page 36. Perhaps this simulation should be introduced earlier in the methods and results rather than in the discussion only.

https://doi.org/10.7554/eLife.95761.1.sa1

Reviewer #2 (Public Review):

Summary:

This valuable study investigates how statistical learning may facilitate a target detection task and whether the facilitation effect is related to statistical learning of word boundaries. Solid evidence is provided that target detection and word segmentation rely on different statistical learning mechanisms.

Strengths:

The study is well designed, using the contrast between the learning of words of uniform length and words of variable length to dissociate general statistical learning effects and effects related to word segmentation.

Weaknesses:

The study relies on the contrast between word length effects on target detection and word learning. However, the study only tested the target detection condition and did not attempt to

replicate the word segmentation effect. It is true that the word segmentation effect has been replicated before but it is still worth reviewing the effect size of previous studies.

The paper seems to distinguish prediction, anticipation, and statistical learning, but it is not entirely clear what each term refers to.

https://doi.org/10.7554/eLife.95761.1.sa0