

Statistical learning of uniform- and mixed-length artificial languages: Different computational mechanisms support different task demands

Meili Luo⁺, Ran Cao⁺, Felix Hao Wang*

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

*To whom correspondence should be addressed.

⁺These authors contributed equally to this work.

Author Note

Correspondence concerning this article should be addressed to Felix Hao Wang. Email address: haowang1@sas.upenn.edu (F. H. Wang). This work was financially supported by the National Natural Science Foundation of China (NSFC No. 32500953) and the Natural Science Foundation of Jiangsu Province of China (Grant BK20240588).

The data reported in this paper is available, at <https://osf.io/63y2g/>. The reported experiments were not preregistered.

Abstract

Statistical learning is a powerful mechanism that can support a variety of learning tasks. Many theories have assumed a single mechanism for statistical learning across different tasks, where a unitary mechanism is supposed to explain results from various studies, even across different modalities. In this study, we studied auditory statistical learning by comparing two different experimental paradigms, target detection and word segmentation, and examined if different mechanisms are required to explain results from the two paradigms. Previous work using the word segmentation paradigm suggested that learning is better with sequences containing uniform-length words than with sequences containing mixed-length words. If the same mechanism supports the target detection task, the same results are predicted. However, while learning was successful in both Experiments 1 and 2 with the target detection paradigm, the effect was larger in the mixed condition than in the uniform condition. We further replicated the uniform condition advantage in the word segmentation paradigm in Experiment 3. Thus, we hypothesized that the target detection paradigm required a different mechanism 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 a prediction mechanism, rather than clustering, could explain the data from target detection. Crucially, this 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.

Keywords: Statistical segmentation; Transitional probability; Prediction; Rapid learning

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. Prosodic information, such as stress, though never perfectly correlates with word 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 et al., 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 et al., 1996; Aslin et al., 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 et al., 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 succeeds and fails 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 et al., 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 et al., 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. A recent study showed that learners can succeed at segmentation when they are exposed to a stream where word forms occurred only two times under the word segmentation paradigm (Wang et al., 2023). 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, repetition of the word forms in the sequence may be required for robust recognition when they were later queried in the test phase. There are other experiment paradigms in statistical learning, and the target detection paradigm was better suited to understand how fast learning occurs (Batterink, 2017). In this target detection paradigm, 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, 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*). This effect was called the facilitation effect, with the idea being that learners who know a trisyllabic word can use the first syllable to predict the second and third syllables. This thus demonstrates that learners have sensitivity to the statistical structure of the stream after one exposure, under the target detection paradigm.

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. 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 in continuous input in PARSER (say, ABC with different letters representing different syllables) from the syllable sequence, the random process may sometimes produce a chunk that contains or partially contains the word form (such as ABC, or AB), and this can be used for facilitation. Notably, PARSER was originally used to explain performance with word segmentation paradigms, and if it were the case that PARSER can explain performance with target detection paradigms, it would suggest that the same mechanism would explain performance from two different paradigms.

In our view, however, there are reasons why PARSER is not well fit to explain the results from the target detection paradigms that have yet to be discussed. Of importance, forming chunks in a random manner only serves to mislead the learner whenever the chunks misalign with statistical regularities. Rather, it's possible that the facilitation effect can occur without segmentation, word extraction, or chunking, per se. Storing bigrams and tracking transitional probabilities (TPs) from continuous input may suffice. By storing all bigrams in the input per standard statistical learning theory, learners would have access to TPs for different syllable transitions, and would be faster to detect targets with high TPs than those with low TPs. Under such a view, fast learning can be explained as long as learners can use small amounts of exposure to obtain bigram counts and TP values.

Notably, this hypothesized mechanism to explain performance in target detection tasks differs from mechanisms hypothesized to explain word segmentation, an offline task. While both tasks have been hypothesized to make use of tracking TPs, word segmentation involves additional constraints, as we mentioned above. The different performance of segmenting syllable sequences

consisting of words that were uniform in length and syllable sequences consisting of words of mixed length suggests that additional processes, such as rhythm perception, are required to explain performance in word segmentation tasks (Johnson & Tyler, 2010; Wang et al., under review). Thus, while learners could potentially learn from the same syllable sequence in both a word segmentation task and a target detection task, the different task demands may tap into different learning mechanisms.

The goal of the current study is threefold. First, we aim to replicate the findings of Batterink (2017) and provide additional empirical evidence for the fast learning that was observed in that study. This replication would also provide an effect size estimate for when learners perform target detection in syllable sequences consisting of words that were uniform in length in Experiment 1. Secondly, we extend the target detection task to scenarios when learners perform target detection in syllable sequences consisting of words that were mixed in length in Experiment 2. Thirdly, we replicate the finding in word segmentation, showing a marked difference between segmenting a uniform and a mixed word-length sequence. Between these three experiments, we probe the mechanisms in the online target detection task and its relationship to the offline word segmentation task. This was done by leveraging what we know of typical word segmentation tasks, where segmentation tends to be successful when the input sequence contains uniform-length words, but fails when the words are mixed in length. 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, if the mechanism for the online target detection task is different from segmentation, it may not require the learner to segment the continuous input, and word length uniformity does not matter. In this case, target detection with sequences containing mixed-length words would also be similarly

successful as with sequences containing uniform-length words, and the mechanism involved would be the process of storing bigrams and calculating TPs. To provide further computational evidence for the proposed mechanisms, we conducted three simulations of these experiments. These simulations used either the TP tracking approach (Simulation 1) or a clustering approach (Simulations 2 and 3) to examine the learning effects in the experiments. These simulations provide information on how well these approaches using different information can approximate human performance. Together, the experiments and simulations provide insight into the mechanisms involved in the target detection task and its relationship to word segmentation.

Experiment 1

In Experiment 1, we replicate Batterink (2017) using the same material and the same uniform-length word design from the study. We conducted the replication two times, an exact replication and a conceptual replication. In addition to strengthening the robustness of the empirical finding, these different designs allowed a comparison of a nuisance variable, namely, whether the sequence initial (the first and the second) or the sequence final (the 47th and the 48th) 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.

Methods

Participants. The number of participants for the replication was determined based on a power analysis of 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 (the 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 ($4 \times 4 \times 3$) 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, the experimenter made sure that the participant was doing the task correctly and 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.





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.

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 make a difference in terms of the facilitation effect. Our null hypothesis here was whether the target syllable occurring in these arbitrary locations would not interfere with whether the learner could remember the sequence and use it for prediction.

Predictions

We reiterate 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) exposures, and similarly for the third syllable compared to the first syllable, because while the first syllable is unpredictable, the second and the third syllables 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 was 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).

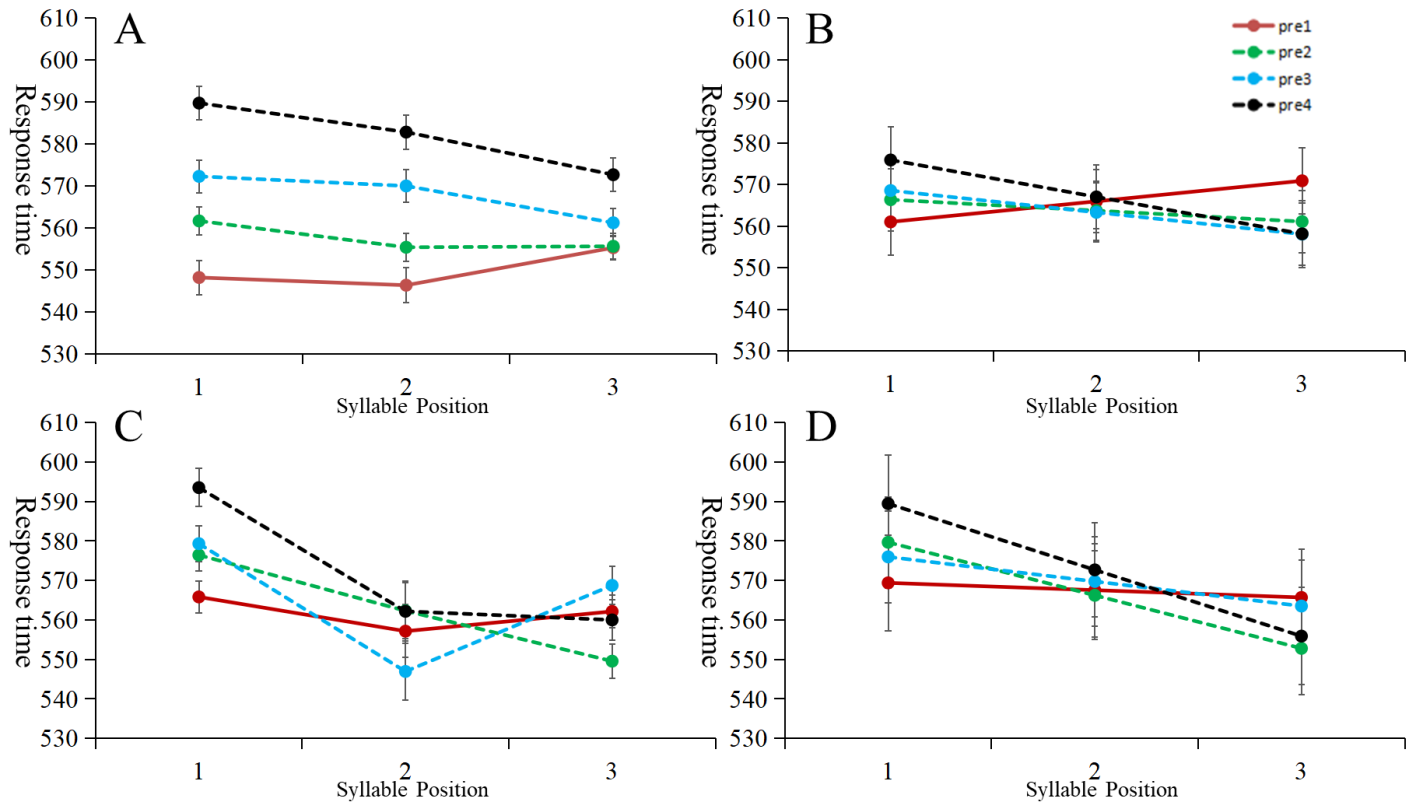


Figure 2. Reaction time (RT) data with syllable position (first, second, or third syllable in the word) on the x-axis, and word presentation (first, second, third, or fourth occurrence of the word in the stream) as different lines in the Figure. The left panels show the raw data means, and the right 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 one in Batterink (2017) for ease of comparison. Error bars represent ± 1 SEM.

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 presentations 2, 3, and 4 would be. Looking at Figure 1, we saw that the line for presentations 2 to 4 had negative slopes, whereas the slope for presentation 1 was not negative. To examine this effect statistically, we ran a linear mixed effects 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 times.

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., presentations 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 ($\beta=0.0002, z=0.84, p=0.400$); 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 ($\beta=-0.012, z=-2.95, p=0.003$), and so was the interaction between presentations 1 and 4 ($\beta=-0.015, z=-3.49, p < 0.001$). Only the interaction between presentations 1 and 3 did not reach significance ($\beta = -0.005, z = -1.07, p = 0.287$). 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 ($\beta=0.001, z=5.65, p<0.001$), successfully replicating this effect from Battarink (2017), where syllables occurring later in the syllable stream are detected more slowly than syllables occurring earlier in the syllable stream. The interaction between presentation 1 and presentation 2 was negative and significant ($\beta=-0.007, z=-2.17, p=0.030$), so was the interaction between presentation 1 and 3 ($\beta=-0.010, z=-3.03, p=0.002$) and between presentation 1 and 4 ($\beta=-0.014, z=-3.94, p<0.001$). 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 facilitation effect. This analysis mainly regards whether RTs for syllable position 1 in later presentations are faster or slower than in the first presentation. To conduct this analysis, we combined the data from the conceptual-replication condition and the exact-replication condition to achieve better statistical power. In the first mixed effects regression, the fixed effects included presentation and stream position, and the random effects included a by-subject intercept and a random slope of stream position. Results showed that later presentations took significantly longer to respond to compared to the first presentation ($\chi^2(3) = 10.70, p=0.014$), where the effect grew larger with each presentation (second presentation: $\beta=0.011, z=1.82, p=0.069$; third presentation: $\beta=0.019, z=2.40, p=0.016$; fourth presentation: $\beta=0.034, z=3.23, p=0.001$). Thus, while the predictable syllables (from positions 2 and 3 in the later presentations) were responded to faster, unpredictable syllables were also responded to more slowly, starting from the third presentation. This finding was different from the results from Batterink (2017), who did not find the effect of presentation on syllable position 1 to be significant, possibly due to a power issue¹.

In sum, both the exact-replication condition and the conceptual-replication condition were successful in replicating all of the aspects of 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

¹ If we ran the same analysis with data from either the exact or the conceptual condition, this effect was not significant either, which indicates the same power issue. We thank our reviewer for making this suggestion for running this analysis by combining data from the two conditions for better statistical power.

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; also see Experiment 3). 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 in 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

Under the criterion that a syllable is detected if there is a key press within 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.

The main analysis for Experiment 2 involved a linear mixed effects model that involved all factors of interest. In this regression (which we call regression 1), 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), and word length (disyllabic/trisyllabic) and their three-way interaction. One additional fixed effect was included, which was overall stream position (3rd through 46th syllable in the syllable sequence, continuous). Random effects included participant as a random intercept and stream position as a random slope. The main analysis showed that the three-way interaction was not significant ($\chi^2(3) = 6.19, p = 0.103$), suggesting that word length did not significantly interact with the main interaction of interest (between presentation and syllable position). Thus, we dropped the three-way interaction, and the new regression included word length as a main effect, the two-way interaction between presentation and syllable position, and a fixed effect of overall stream position as well as the same random effects as above (which we call regression 2). In this regression, the omnibus interaction between word presentation and syllable position was significant ($\chi^2(3) = 49.77, p < 0.001$), suggesting that the interaction between presentation and syllable position was significant for both disyllabic and trisyllabic words. This was in addition to a significant fixed effect of word length ($\beta = 0.018, z = 6.19, p < 0.001$). This result thus shows that there is an overall effect of word length in addition to the interaction between presentation and syllable position, so next, we turn to pairwise comparisons in different word length conditions.

With data from both disyllabic and trisyllabic words, we constructed a regression model, each looking at the interactions between syllable position and presentation pairs. For trisyllabic

words, the interaction between presentation 1 and presentation 2 was negative but not significant ($\beta = -0.007$, $z = -1.09$, $p = 0.277$). The interaction between presentations 1 and 3 was negative and significant ($\beta = -0.014$, $z = -2.07$, $p = 0.039$). The interaction between presentations 1 and 4 was negative and significant ($\beta = -0.043$, $z = -6.21$, $p < 0.001$). For disyllabic words, the regression containing both participant as a random intercept and stream position as a random slope did not converge. Thus, the regression for disyllabic words only included a by-subject random intercept as the random effect. The interaction between presentation 1 and presentation 2 was negative and significant ($\beta = -0.024$, $z = -2.20$, $p = 0.028$). The interaction between presentation 1 and 3 was negative and marginally significant ($\beta = -0.020$, $z = -1.85$, $p = 0.065$). The interaction between presentations 1 and 4 was negative and significant ($\beta = -0.027$, $z = -2.16$, $p = 0.031$). Together, these analyses showed that there was a robust effect for trisyllabic and disyllabic words alike. A plot of the data is shown in Figure 3.

Notably, we wondered if there is evidence that word length made no difference for the interaction between presentation and syllable position, and the frequentist approach (regression 1) only yielded a non-significant p-value, which cannot be considered as evidence for the null. To this end, we conducted a Bayesian analysis, using the approach outlined in Harms and Lakens (2018). The Bayes factor for the interaction can be obtained by computing Bayesian Information Criteria (BIC) for the null model and the alternative model (which are regression 2 and regression 1, respectively). The Bayes Factor is then given by $e^{(BIC_{alt} - BIC_{null})/2}$. The Bayes Factor was found to be $e^{25.65}$ which is more than 10^{11} , providing strong evidence that word length made no difference for the interaction between presentation and syllable position.

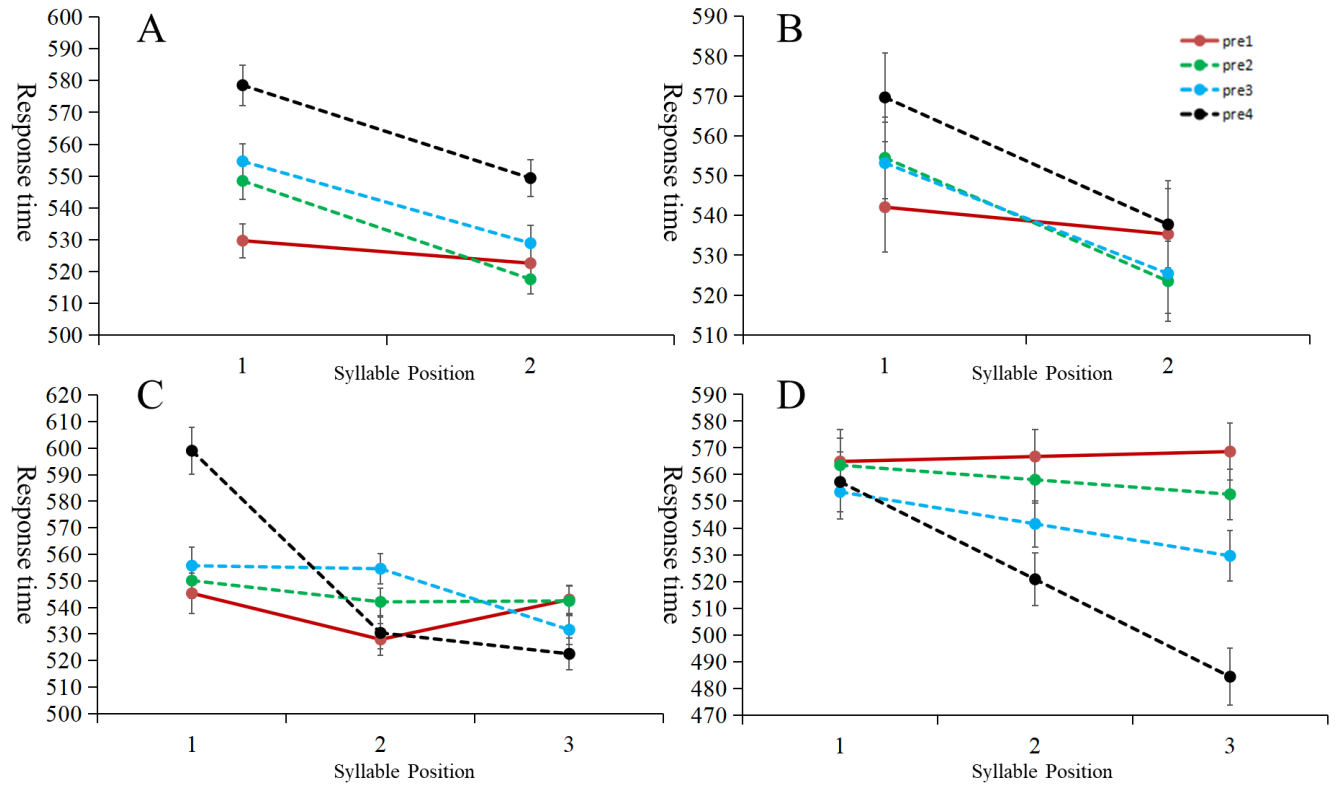


Figure 3. Reaction time (RT) data with syllable position (first, second, or third syllable in the word) on the x-axis, and word presentation (first, second, third, or fourth occurrence of the word in the stream) as different lines in the Figure. The left panels show the raw data means, and the right 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.

Lastly, we want to answer the question of whether the facilitation effect is larger in the uniform condition in Experiment 1 than in the mixed condition in Experiment 2, 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 (Experiment 2) involved both disyllabic and trisyllabic words, whereas the uniform condition (Experiment 1) only had trisyllabic words.

For the analysis below, we put the experiment number in the fixed effect (with potential interactions with other variables). Secondly, the length of syllable streams was shorter in Experiment 2 compared to in Experiment 1, because half of the words were disyllabic in Experiment 2. 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 for these variables.

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 effects regression with a three-way interaction. The RT was the dependent variable, and the independent variable included fixed effects of experiment (Experiment 1/2, which correspond with uniform/mixed conditions, coded categorically), word presentation (1-4, coded categorically), position (1-3 for trisyllabic words and 1-2 for disyllabic words, coded continuously), 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 ($\beta=-0.003$, $z=-0.52$, $p=0.600$) was negative and not significant, became positive and not significant for presentations 1 and 3 ($\beta=0.011$, $z=1.65$, $p=0.099$), and became positive and significant for presentations 1 and 4 ($\beta=0.021$, $z=2.99$, $p=0.003$). 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.

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 finding suggests that the mechanism behind the target detection task examined in this paper was different than the mechanisms involved in word segmentation. As the computational model below would suggest, the larger interaction effect in the mixed condition compared to the uniform condition is consistent with a facilitation effect that is compared in two conditions. Importantly, however, these effects are not expected if there is less learning in the mixed condition compared to the uniform condition, which has been the finding from segmentation tasks.

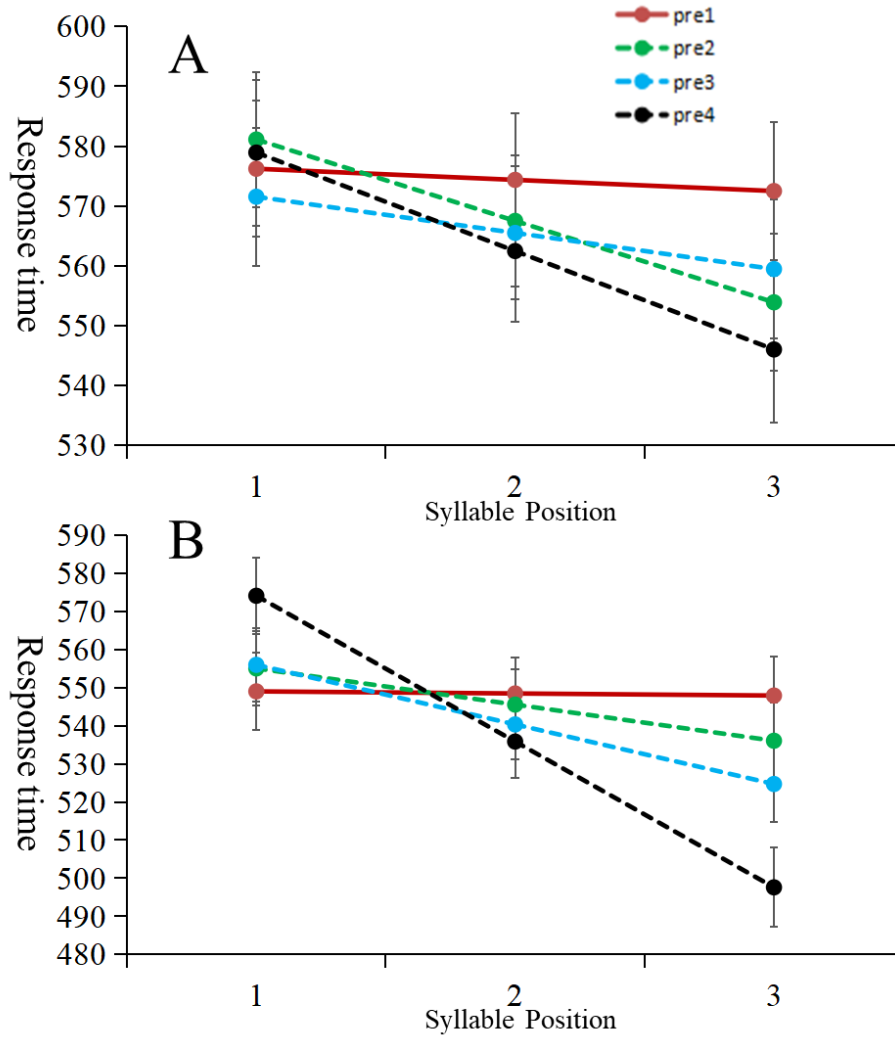


Figure 4. Regression model fit from the three-way interaction between condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2 for disyllabic words, continuous). Figure 4A showed results from the uniform condition from Experiment 1, and Figure 4B showed results from the mixed condition from Experiment 2. Error bars represent ± 1 SEM.

Before we report our computational model for these target detection results, we report an additional experiment testing word segmentation with both uniform and mixed length sequences. Even though we have cited evidence in the literature that participants are worse at segmenting

mixed-length sequences than uniform-length sequences, this result is worth replicating in the same paper.

Experiment 3

Experiment 3 was designed to provide evidence that word lengths make a difference for a different experiment paradigm, i.e., word segmentation. In the experiment presented below, we conducted a replication study of Johnson and Tyler (2010) with adult participants, asking whether adults rely on rhythm for segmentation tasks, as we propose infants do. Notably, if the conclusion is that syllable sequences with uniform and mixed word lengths are similarly learnable for adults, it demonstrates that infants alone have difficulty segmenting sequences with mixed word lengths given the comparison to Johnson and Tyler (2010). Conversely, if adults also have difficulty learning sequences with mixed word lengths but no problem with uniform word lengths, this provides evidence that infants and adults alike have difficulty segmenting continuous sequences without rhythm, and with transitional probability alone, and this would be in contrast to findings from target detection paradigms in Experiments 1 and 2.

Methods

Subjects. A total of sixty undergraduate students at the University of Pennsylvania were recruited from the Psychology Department subject pool. Thirty subjects each were run in the uniform word length (UWL) condition and the mixed word length (MWL) condition.

Stimuli. We used the mbrola software package (Dutoit et al., 1996) to synthesize the speech stimuli in Experiment 1, in order to be consistent with the quality of the speech in Johnson and Tyler (2010). For our English-speaking adults, we used the us1 diphone set. The syllables used were (in SAMPA notation that mbrola uses): [p A], [b I], [t I], [b u], [g @U], [l A], [t u], [d A], [d

@U], [p I]. Consonants were 49ms long and vowels were 173ms long, so that all syllables were 222ms long. The fundamental frequency was 220Hz.

Design and procedure. Both the UWL and the MWL conditions in Experiment 1 had two phases: a learning phase followed by a testing phase. In the learning phase, participants were told to listen to a syllable sequence passively through the headphones while the screen was blank. They were asked to pay attention to the syllable sequence as they were to be tested on it later.

There were two conditions: a uniform word length (UWL) condition and a mixed word length (MWL) condition. In the UWL condition, four disyllabic words were randomly concatenated, two of which occurred with high frequency (90 times) and the other two occurred with low frequency (45 times). In the MWL condition, the two high frequency words were trisyllabic (each occurring 90 times) and the two low frequency words were disyllabic (each occurring 45 times). These two conditions were designed such that the part-words in both the UWL and the MWL conditions are the same and have the same bigram frequency (45 times). Each subject received a different random concatenation of the syllables during training (similar to Johnson & Tyler, 2010), and the test items were also individually generated, given the randomization in the training sequence.

Immediately after the learning phase, we displayed instructions for the test phase on the screen. The instructions indicated that participants would hear a number of sound sequences and make judgments about the sequences. There was a total of 4 test items, two words and two part-words. The presentation sequence of test trials was randomized for each participant. Participants initiated each test trial. For each test item, a disyllabic sequence (either a word or a part-word) was played 3 times, with a 1-second pause between each sequence. After the presentation of the test item, participants were asked to rate the familiarity of the item: “Do you think that you heard this

sequence in the previous section?” Their responses were marked on a scale containing five response items: “Definitely”, “Maybe”, “Not Sure”, “Maybe Not”, and “Definitely Not”. Once the participants made their choice, the screen went blank and the trial ended. After a one-second inter-trial interval, the next trial began.

Results and Discussion

We coded the scale of “Definitely”, “Maybe”, “Not Sure”, “Maybe Not”, and “Definitely Not” into the numeric values 4, 3, 2, 1, and 0, respectively (Figure 5). To compare these ratings statistically, we ran mixed effects linear regressions on the data. First, we ran an analysis on rating for both conditions separately, with item type (word vs. part-word) as the fixed effect and by-subject random intercepts. In the MWL condition, item type was not significant, indicating that there was no learning ($\beta = -0.333$, $z = -1.65$, $p = 0.100$). We calculated a Bayes Factor for this condition with the same procedure as in Experiment 2, and found that the Bayes Factor was 2.87. According to Jeffreys (1998), a Bayes Factor between 1.6 and 3.3 is considered to be substantial evidence for the null. In the UWL condition, item type was significant, indicating learning ($\beta = -0.933$, $z = -4.52$, $p < 0.001$). Second, we ran an analysis on data from both conditions, using the interaction and main effects of both conditions (UWL vs. MWL) and item type (word vs. part word). The interaction was significant ($\beta = -0.600$, $z = -2.07$, $p = 0.038$), suggesting that learning was different between the two conditions.

A strong interpretation of this result is that adults, like 5.5-month-old and 8-month-old infants in Johnson and Tyler (2010), are able to learn from a uniform-length word sequence but not from a mixed-length word sequence. A weaker interpretation is that adults are able to learn from a uniform-length word sequence, and learn it much better than the mixed-length word sequence. However, it is clear that both interpretations are consistent with our prediction that a

crucial factor for infants and adults to perform word segmentation in these artificial language studies is word length uniformity (and as such, mechanisms other than tracking transitional probabilities may have explanatory power).

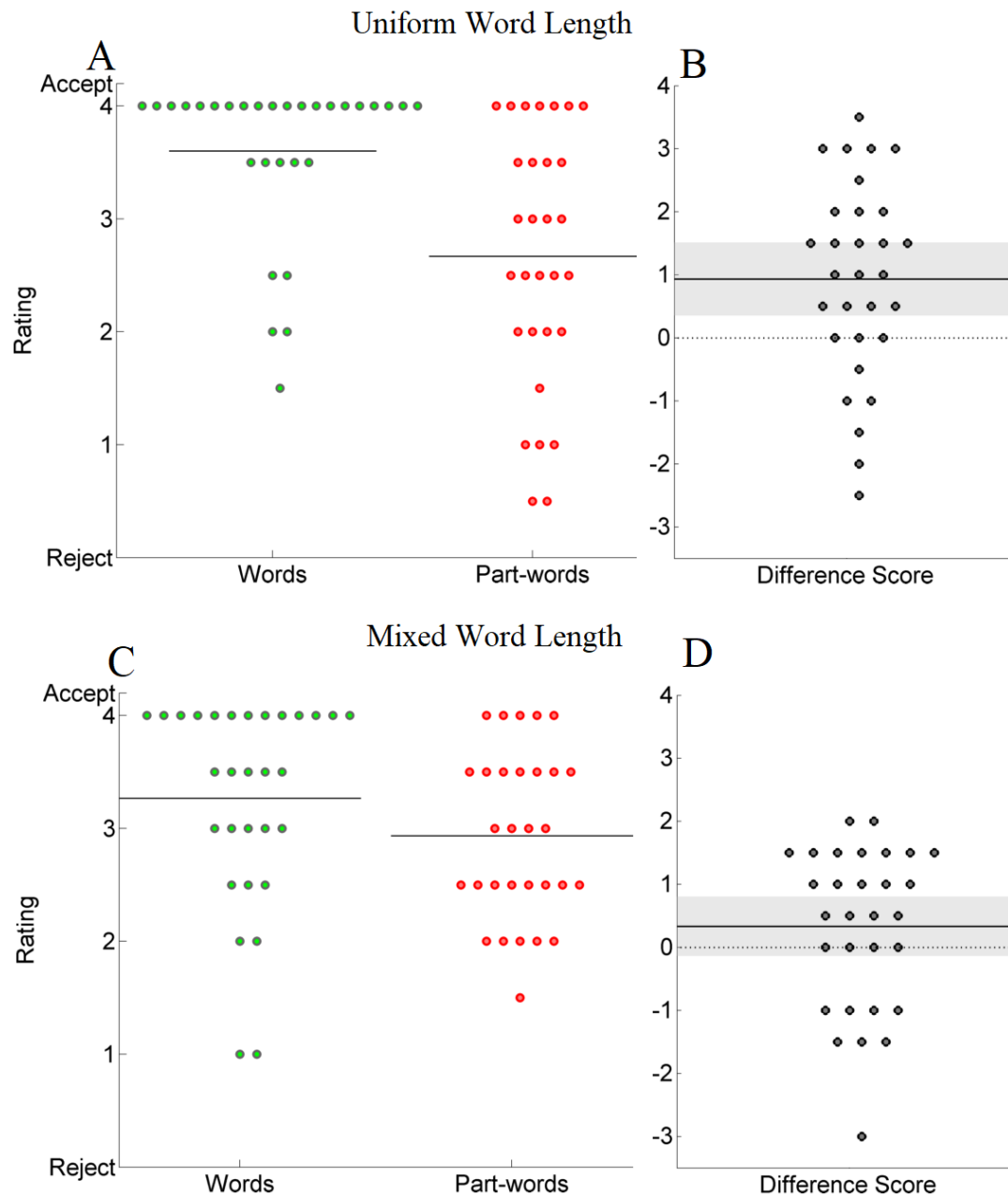


Figure 5. Means for words and part-words, and difference scores for each subject in Experiment 1, collapsed over counterbalancing conditions. In Figures 5A and 5C, each circle represents the mean rating of a subject for all words and part-words in the uniform and mixed word length

conditions, with a solid line indicating the mean value for each item type. In Figure 5B and 5D, each circle represents the difference between mean ratings (words - part-words) for each subject in the uniform and mixed word length conditions, with the solid line showing the mean difference, and shadows showing 95% confidence intervals around the mean. The dotted line at 0 represents chance.

Simulation 1

In Simulation 1, we hypothesized a specific computational process that can potentially explain human performance in the target detection paradigm in Experiments 1 and 2. The purpose of Simulation 1 is to provide a simple computational process that instantiates the assumption of what we believe contributes to the facilitation effect in target detection. In this model, 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 an 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, RT_0 , 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 they are reacted to, and 2) unpredictable syllables do not. As such, we propose a simple recursive relation between the RT of a syllable occurring for the n^{th} time and the $n+1^{\text{th}}$ time, which is:

$$RT(n + 1) = \begin{cases} RT(n) + stream_pos * stream_inc & \text{if unpredictable} \\ RT(n) + stream_pos * stream_inc + occ_inc * occurrence & \text{if predictable} \end{cases}$$

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

has already occurred is based on the empirical finding that the more the target syllable is 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 (RT_0), does not factor into the pattern of data results later, as all RTs share this component equally. We set this RT_0 to be the constant from the regression coefficient from previous regressions. The second constant is the stream_inc , the increment amount for the 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 have 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 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 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 word 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 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 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 6).

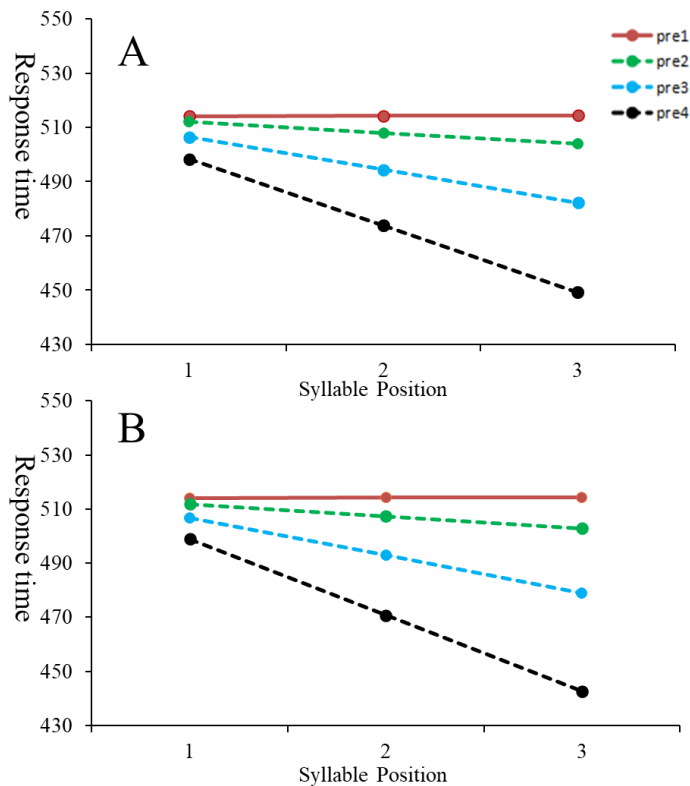


Figure 6. Regression model fit from the three-way interaction between condition (mixed/uniform, categorical), word presentation (1-4, categorical), and position (1-3 for trisyllabic words and 1-2

for disyllabic words, continuous) for the simulated data. Figure 6A showed results from the simulation for the uniform condition, and Figure 6B showed results from the simulation for the mixed condition.

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 fewest number of assumptions built in. More importantly, 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.

Simulation 2

An alternative computational model to explain the facilitation effect in the target detection task is PARSER (Perruchet & Vinter, 1998), which was proposed as a potential explanation for the facilitation effect in Batterink (2017). PARSER processes continuous syllable sequences 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 the facilitation effect, given that the facilitation effect occurs after a single exposure in both Batterink (2017) and Experiment 1 in the present study? To answer this question empirically, we created the following simulations using the U-Learn program (Perruchet et al., 2014), a program developed by the authors of the original PARSER paper. Whereas the PARSER model was created to model word segmentation of continuous syllable sequences, there is a lack of a linking assumption between the weights of different chunks and data from target detection experiments. We made the following linking assumption: after asking PARSER to output weights for words vs. part-words, 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). As the weights of syllable sequences are the main output of the PARSER model, we felt that this was the most reasonable approach to adopt.

One other feature of the U-Learn program is that it reports its weight changes in 10 time-steps (which corresponds to 1/10 of the learning sequence, however long the learning sequence is). We made use of this feature to examine the relationship between the amount of exposure and learning. That is, 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 (which should contain a word form once, on average) and 2/10 of the syllable sequence is 8 words long (which should contain the same word form twice, on average). The idea is that, if the weights for words became significantly more than the weights for part-words after 1 or 2 time points in this simulation, it would be evidence that there is learning after 1 or 2 exposures to a word form. We created the sequence and the test items by modifying the ready-to-se configurations (Figure 7).

U-Learn : Enter the items

Items for training

Number of sections: 1

Section currently displayed: 1

Frequency	Items
10	pa/bi/ku/
10	ti/bu/do/
10	go/la/tu/
10	da/ro/pi/
40	

☐ Immediate repetitions allowed

☐ Hard boundaries

between each set of items

or range, from to

Update

Clear

Save this configuration

Items for test

1 WORDS

pa/bi/ku/

ti/bu/do/

2 PART-WORDS

tu/da/ro/

pi/go/la/

3

4

5

6

7

8

9

10

Ready-to-use configurations

Aslin et al. 1998 (frequency-balanced design)

NEXT

Figure 7. The set-up for simulation one, where four trisyllabic words occur 10 times each.

Using this setup, we ran the simulation 50 times to represent running 50 subjects on this task. The results are shown in Figure 8. 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 $1/10^{\text{th}}$ or $2/10^{\text{th}}$ of the sequence. Observing the weight changes in the learning curve, we see that there is no learning (Figure 8) during this period. The model also produced the weights for a potential statistical comparison, but in this instance, all weights of words and partwords for $1/10^{\text{th}}$ and $2/10^{\text{th}}$ of the sequences were 0s (and thus a statistical

test is not needed). Thus, PARSER failed to segment/show a facilitation effect after either one or even two exposures. In fact, it can be seen from Figure 8 that out of 50 runs of the model in each condition, only 1 run produced a non-zero weight for words during 3/10th of the sequence.

Simulation 2 thus shows that PARSER cannot learn after a small amount of exposure, and thus cannot explain the facilitation in the target detection tasks. In both scenarios, humans are better learners than PARSER as they were able to learn from sequences much faster than the algorithm using clustering as the mechanism for learning.

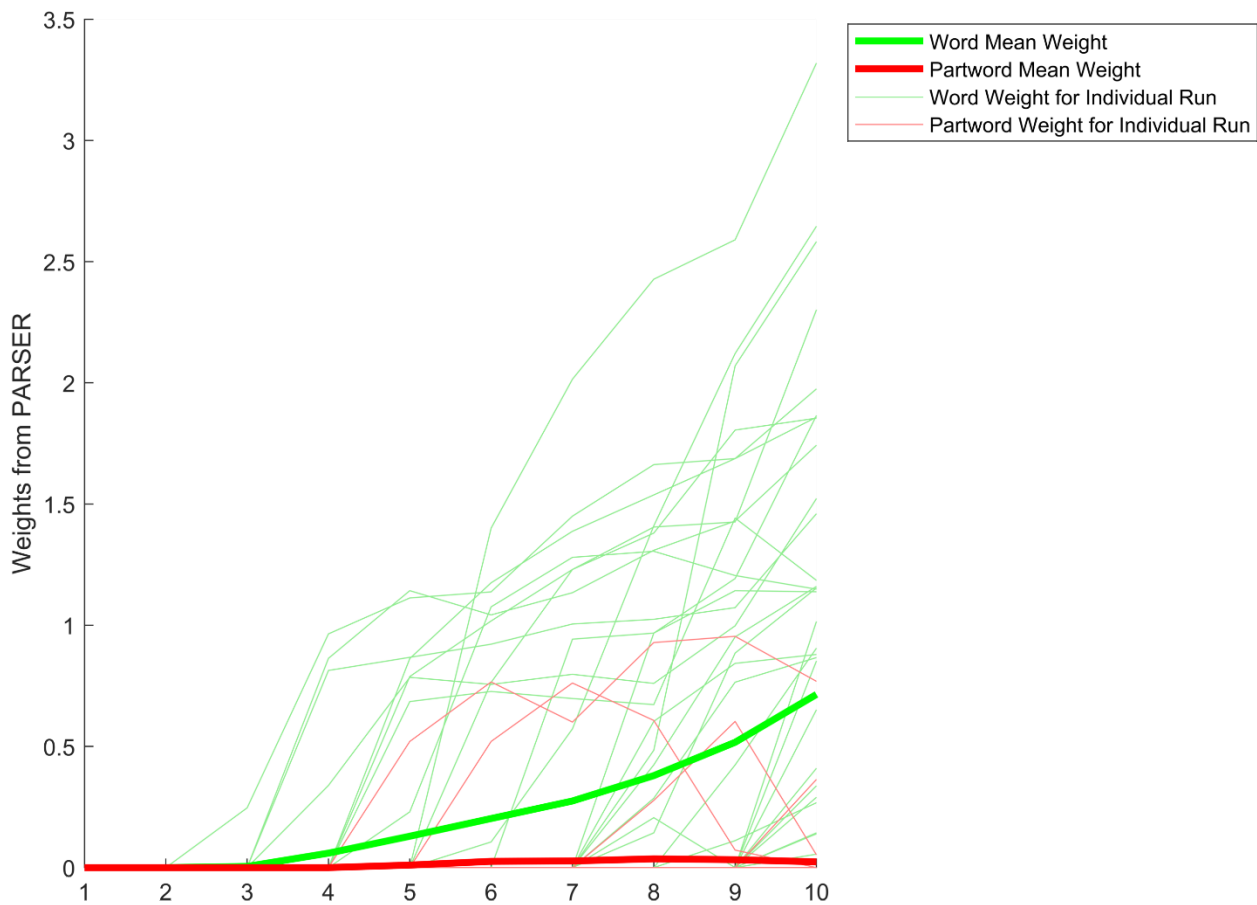


Figure 8. The results for Simulation 2, where the thin lines show the weight changes of words and part-words over the course of learning in individual runs, the thick lines show the mean weight of words and part-words across runs over the course of learning. On the x-axis, each number represents the model having processed 1/10 of the sequence (i.e., 3 means the model has processed

3/10 of the sequence so far). The y-axis represents weights, which are in arbitrary units (as described in PARSER documentation).

Simulation 3

In Simulation 3, we used PARSER to perform a simulation of Experiment 3. In Experiment 3, there was a uniform-length condition and a mixed-length condition, so Simulation 3 can potentially inform us of whether clustering-based approaches can explain human performance in the word segmentation task. To set this up, we modified the existing Aslin et al. 1998 simulation in PARSER. Specifically, the uniform-length condition was the same as the existing Aslin et al. 1998 simulation. The mixed-length condition was modified in the way shown in Figure 9. Specifically, the two low frequency words were changed to disyllabic, while maintaining the other two high frequency words to be trisyllabic. The test items were changed to disyllabic words, as was done in other experiments.

U-Learn : Enter the items

Number of sections: 1

Section currently displayed: 1

Frequency	Items
45	pa/bi/
45	ti/bu/
90	go/la/tu/
90	da/ro/pi/

☐ Immediate repetitions allowed

☐ Hard boundaries

between each set of items

or range, from to

Update

Clear

Save this configuration

Ready-to-use configurations

Load a previously saved configuration

NEXT

Items for training

Items for test

1 WORDS

2 PART-WORDS

3

4

5

6

7

8

9

10

Figure 9. The set-up for the mixed condition in Simulation 2.

Next, we ran the simulations for the two conditions 50 times each. This corresponds to running 50 subjects in each condition. The results are shown in Figure 10. From these simulations, we see that PARSER is similarly successful in both conditions. With the weights from the simulations, we performed two statistical comparisons to answer two questions. The first question was whether the weight difference between words and part-words is different between conditions. For this analysis, we gathered the final weights (the weights after all the sequences have been learned, which correspond to the 10th segment in the figures) from each type of item in the two

conditions. With the two factors (word type and condition), we performed a two-way anova, which showed that there was no significant interaction ($F(1) = 1.78$, $p=0.184$). This suggests that, following the exposure to the entire sequence in each condition (which simulates human learning performance in segmentation), there was no difference in terms of PARSER performance in the two conditions. The second question was whether there was any point during the time course of the exposure to the sequence where learning performance diverged. For this analysis, we gathered the weights in each of the 10 segments and performed the same two-way anova, asking if there was a significant interaction between word type and condition. Even without multiple comparison correction, which lowers the alpha values to a value smaller than 0.05, none of the 10 segments had a significant interaction, as all the p-values were larger than 0.14. Thus, we find that there was no point at which there was a significant interaction between word type and condition.

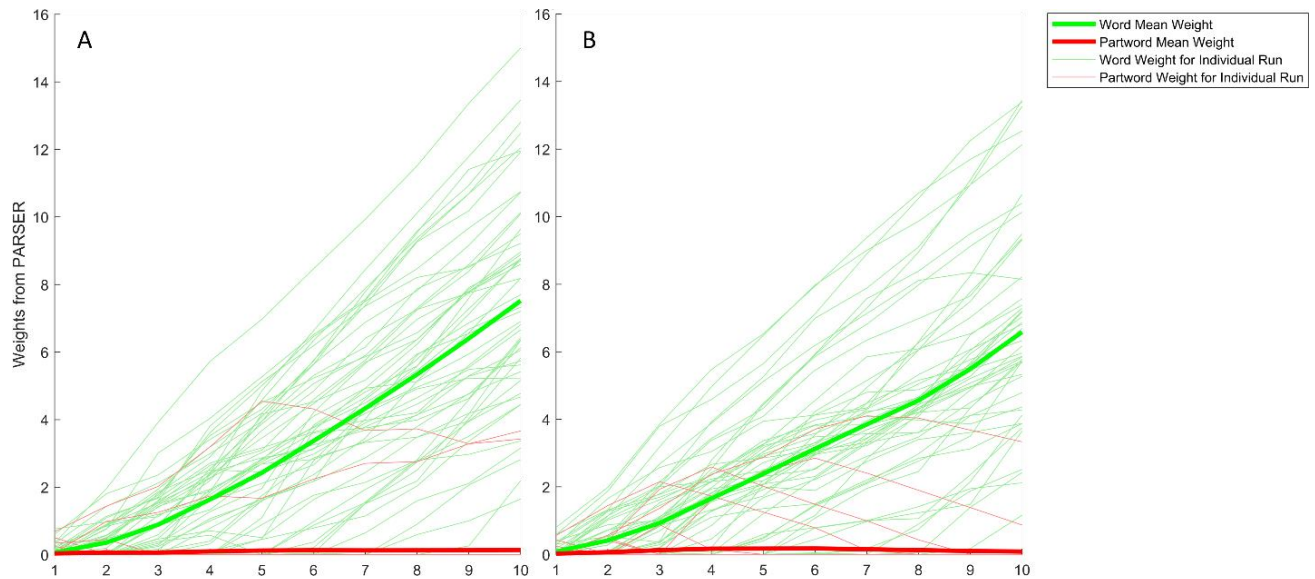


Figure 10. The results for Simulation 3, where Figure 10A shows the weight changes of words and part-words in the uniform condition, and Figure 10B shows the weight changes of words and part-words in the mixed condition. On the x-axis, each number represents the model having processed 1/10 of the sequence (i.e., 3 means the model has processed 3/10 of the sequence so far).

The y-axis represents weights, which are in arbitrary units (as described in PARSER documentation).

Given these results from Simulation 3, we see that PARSER is capable of learning in both mixed and uniform conditions. We also see that the rate of learning as well as final learning outcomes, are qualitatively similar. When compared with human performance in the word segmentation task given a relatively large amount of exposure, PARSER represents a more powerful learner than humans, as PARSER has no trouble with segmenting a mixed-length sequence.

Taken together, we find that results from PARSER and human performance differ in important ways. When learning from a small amount of exposure, which is the case in a target detection task, PARSER cannot learn, but humans can. When learning after a large amount of exposure, which is the case in word segmentation tasks, PARSER learns uniform and mixed word length sequences equally well, which also differ from human performance. Thus, we find that PARSER is not as sensitive to statistical regularities as humans during the initial encounters with statistical regularities, and is too powerfully equipped to learn complicated sequences when humans would have trouble.

General Discussion

This paper investigated the mechanisms involved in statistical word segmentation and target detection, reporting three experiments and three simulations. In Experiment 1, we reported a replication of Batterink (2017), including both a conceptual replication and an exact replication. The facilitation effect was successfully replicated in both cases, where the reaction time was shorter for predictable syllables (syllable positions 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 was shown to generate a smaller amount of learning in the word segmentation literature using the word segmentation paradigm. Contrary to the 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). In Experiment 3, the finding that adults are better at segmenting a sequence containing uniform-length words than a sequence containing mixed-length words using the word segmentation task was replicated.

To explain the computational mechanisms for our experimental results, we conducted three simulations. In Simulation 1, we computationally modeled the facilitation effect in the target detection task. The assumption built into the model involved changes to the RT based on predictability, where we assumed a recursive relation between RTs for consecutive items, and predictive and unpredictable targets had different relations to the RT of the previous item. The simulation produced data patterns similar to data from Experiments 1 and 2, and to the uniform-length conditions and mixed-length conditions alike. Simulation 1 thus 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. In Simulation 2, we explored the viability of using clustering-based approaches (specifically, the computational model PARSER) as it was assumed to account for target detection performance in the previous literature (Batterink, 2017). Simulation 2 showed that, while PARSER can eventually produce different weights for words vs. part-words, which were akin to the facilitation effect, this effect did not appear unless

the model encountered more than one or two exposures. Compared to human performance, this computational model required more exposure to learn statistical regularities in the input. In Simulation 3, we used PARSER again to simulate Experiment 3, which was a word segmentation task. We found that PARSER was able to learn from uniform-length and mixed-length sequences in a similar fashion, which differed from how humans learned the two types of sequences. Together, these results suggest that a simple prediction-based mechanism can explain the results from the target detection task, and clustering-based approaches such as PARSER cannot, contrary to previous claims.

The experimental data and the simulations provide evidence for our hypothesis that statistical word segmentation and target detection tasks employ different mechanisms. Notably, a null hypothesis would be that a single mechanism would explain performance in the two tasks. Following the discussion above, if this singular mechanism were the tracking of transitional probabilities, it could explain performance in target detection tasks, but could not explain performance differences in learning sequences with uniform-length and mixed-length words in word segmentation tasks. If this single mechanism were rhythm perception, it could explain performance in statistical word segmentation tasks, but could not explain performance in target detection tasks. If this single mechanism were clustering, it could neither explain the better performance in performance differences in learning sequences with uniform-length and mixed-length words in word segmentation tasks, nor the rapid learning in target detection tasks, as we showed in our simulations. As such, different mechanisms need to be posited to explain performance in different tasks, even though the tasks share the feature of requiring learners to learn from statistical regularities.

Different mechanisms in the target detection and word segmentation tasks

Having shown that statistical word segmentation and target detection tasks employ different mechanisms, we will discuss potential reasons why different tasks require different mechanisms, which include different computations involved in the two tasks as well as the representations required to support the computations. To summarize, we believe that there are potentially shared computational processes for the two tasks, but the learning effects require different computations at different locations in the syllable sequence, and the representations supporting the two tasks are also different.

While both tasks potentially require learners to store bigrams and larger n-grams during learning, the two tasks require learners to use this information differently. In the target detection task, learners need to store the bigrams and compute transitional probabilities for syllable transitions, and this knowledge of syllable transitions with different TPs is all that is required. Whereas the syllable transitions with low TPs are responded to with a baseline level of RTs in the target detection task, the main effect of interest, i.e., the facilitation effect, occurs at word internal locations. This differs from the word segmentation task, where learners need to make decisions about where the word boundaries are. As an example, we consider the following sequence: GHIABCDEFABCGHI, where the word "ABC" is preceded and followed by different words. In the target detection task, if the target is syllable B, it is theoretically possible to predict B from the preceding syllable A after one occurrence of AB, and experimental evidence supports this theoretical analysis. This differs from word segmentation, where learners need to decide where the word boundaries are. For each word, two word boundaries are required. Given the example sequence in this paragraph, after the learner processed the ABCGHI section, they would understand that A can be preceded by I or F, C is followed by D or G, both of which are required

for segmentation. As such, multiple exposures (at least two) are required to make segmentation possible² Whereas a single exposure could allow predictions between syllables in target detection to occur.

The two tasks also require different representations. In target detection, the detection speed for reacting to the target is only influenced by the knowledge of the specific transitions in the sequence (i.e., whether the current syllable is predictable or unpredictable given the previous one), as we showed in our model. In this sense, no segmentation is required; remembering bigrams, as we demonstrated in our model, would suffice for this task. It is possible that the only representations required for target detection are the bigrams. This differs from the word segmentation tasks in several aspects. First, the segmentation task requires the learner not only to segment the sequence, but also to remember the segmented subsequences in memory. Representing only where the word boundaries are would not be enough. Secondly, in word segmentation, learners use all possible information that they can use, including prosodic information (e.g., Jusczyk et al., 1999). As such, it requires learners to represent other information, such as prosodic information. Specifically in sequences used for statistical word segmentation tasks, while prosodic information such as stress is typically not present, sequence with uniform-length words produces a rhythm percept that learners can use for segmentation. This representation of rhythm is required to explain the different levels of learning performance for uniform- and mixed-length sequences. In contrast, learners performing target detection in these different types of sequences had similar performance, suggesting that rhythm did not play a role in target detection.

² There are cases exceptional to these discussions. If the learner can segment through subtraction, it would only require a single exposure to the novel word. In this instance, the linguistic materials surrounding it need to be known. For example, consider the English sentence “I bought a dax yesterday”, where one knows all the other words in the sentence. In this case, the novel word can be segmented distributionally through subtraction (Lignos & Yang, 2010). But most discussions on word segmentation regard scenarios where most if not all word forms are unknown, so the example in this footnote counts more as an exception.

In sum, the two tasks may both require the learner to use co-occurrence information from the sequence, they require the learner to process different information to accomplish and thus require different task-demands and mechanisms.

Time courses 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 in the target detection task, and we mentioned that two exposures may suffice for the word segmentation task (Wang et al., 2023). While these results nicely demonstrate the rapidity with which statistical learning occurs, there is a fuller picture of the relationship between learning and the amount of exposure when we consider studies that provide more exposure to learners. 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. For example, 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 (similar results have been reported in Finn & Hudson Kam, 2008; Newport and Aslin, 2004). 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.

There is another line of studies that provides insight into the relationship between exposure and online learning, similar to the target detection task. In serial reaction time (SRT) tasks,

participants make a key press for every stimulus, unlike the target detection task, which requires key presses only for a single target. These studies have also been used to examine the learning of statistical dependencies (e.g., Howard & Howard, 1997; Hunt & Aslin, 2001; Wang & Kaiser, 2022), though the learning effects are typically observed in blocks, rather than for the first few observations. 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. 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 to take an action (making a key press). In contrast, in target detection tasks, there is no action required for most of the stimuli, so that the participants may plan their action while processing the stimuli. It is of interest to understand the specific time course of when the learning effect takes place in SRT tasks in future studies.

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, both when the syllable sequence contains mixed-length words and only uniform-length words. Through computational modeling, we provided a possible mechanism to explain this facilitation effect from the target detection task and ruled out a computational model that was previously theorized to account for the learning process. Importantly, these results provide evidence for our claim that the mechanisms involved

in the target detection task are different from those in 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.

References

- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4), 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(2), 205-211.
- Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science*, 28(7), 921-928.
- Bernard, M., Thiollie, 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, 1541.
- 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(3), e59832.
- Bertels, J., Franco, A., & Destrebecqz, A. (2012). How implicit is visual statistical learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5), 1425.
- 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(2), 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(3), 351-367.
- 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(3), 254-260.
- Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, 108(2), 477-499.
- Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation. *Experimental psychology*, 62(5), 346-351.
- Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117(2), 107-125.
- Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, 33(2), 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.
<http://dx.doi.org/10.3758/s13423-012-0309-8>
- Howard, J. H., Jr., & 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(4), 658.
- Jeffreys, H. (1998). *The theory of probability*. Oxford, UK: Oxford University Press.
- Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of memory and language*, 44(4), 548-567.
- Johnson, E. K., & Jusczyk, P. W. (2003a). 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. (2003b). 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(2), 339–345.
- Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in cognitive sciences*, 3(9), 323-328.
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1), 1-23.
- Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3-4), 159-207.
- Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3-4), 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(2), 145-149.

- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2), 127-162.
- Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39(2), 246-263.
- Perruchet, P., Robinet, V., & Lemaire, B. (2014). U-Learn: Finding optimal coding units from unsegmented sequential databases. Unpublished manuscript.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928.
- Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11), 745-756.
- Swingle, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50(1), 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(3), 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. (under review). Repetition induced rhythm as an alternative account to statistical word segmentation: A model and meta-analysis.