Perceptual factors explain much of statistical learning given a minimal amount of exposure

Author Note

The data and analysis reported in this paper are available, at
https://osf.io/che4q/?view_only=cca901e1ecb748178ecae2b6a5f1c31c/.

The reported experiments were not preregistered.

Abstract

One of the first problems in language learning is to segment words from continuous speech. Statistical learning is a powerful mechanism for word segmentation, yet important questions such as how much statistics is required for learning, as well as the factors that explain such rapid learning are rarely studied. In this study, we find exposure to word forms only two times leads to significant learning and four times leads to better learning. The learning between different amount of exposure correlates and has large variability, suggesting that sequences have inherent learnability characteristics. We further showed that the segmentation performance of a particular sequence is associated with both how easily the rhythm of the syllable sequence is perceived and how familiar the novel word forms in the sequence are to the existing lexicon. The implications of these results to computational models of statistical learning and broader implications to language learning are discussed.

Key words: Statistical learning; artificial language; domain specificity.

Speech consists of mostly continuous utterances, most of which have multiple words in them. To understand speech, one must first segment continuous speech into words. However, due to the lack of pauses or other universal cues in continuous speech in all languages, speech segmentation is a difficult problem. Thus, the mechanisms with which learners can segment words from continuous speech are of great interest. Of the many mechanisms known to be available to language learners, statistical learning can be used to perform word segmentation by making use of distributional information in a sequence. It is also known as a powerful learning mechanism because learning occurs by mere exposure to the speech stream. For example, the seminal work on statistical learning (Saffran, Aslin & Newport, 1996) demonstrated that young infants can segment word forms in a rapid syllable stream in two minutes. In this influential study, learning only required exposure to a syllable stream, consisting of four trisyllabic words occurring 45 times each where prosodic cues to word boundaries were not present. 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.

Following this initial work showing powerful learning, there is now a large literature on how the underlying learning mechanism can be best described, as well as on how various perceptual factors can interact with this powerful learning mechanism. This fuller picture includes many factors that can influence statistical learning, including abstraction and encoding of stimuli for retention that can be language specific, language-specific stimuli properties, as well as domain general attention, processing speed and memory factors (Arciuli, 2017; Potter, Wang, Saffran, 2017; Saksida, Langus & Nespors, 2017; Wang & Saffran, 2014). For example, Wang and Saffran (2014) tested English monolingual adults, monolingual Mandarin and Mandarin–English bilingual adults in a statistical learning task using an artificial tone language, and found that the ability to

encode syllables that have tones as part of their identity played a crucial role in predicting the outcome of the statistical learning study. In many ways, the finding that statistical learning is language specific (Onnis & Thiessen, 2013; Saksida, Langus & Nespors, 2017) is broadly consistent with the idea that infants possess language specific knowledge that they can use for segmentation (e.g., Jusczyk, Cutler & Redanz, 1993; Nazzi, Bertoncini, Mehler, 1998).

The fact that statistical learning is highly language-specific indicates that one cannot simply view statistical learning as a process where the perceptual system takes in any input stream and discretizes the units into different identities, such that a “beads-on-a-string” computation can be performed. Rather, a theory of statistical learning must put perceptual constraints as a significant component. For example, in a language learning context, Potter, Wang & Saffran (2017) showed that second language learners can learn from a sequence made of syllables with tones, only after months of learning that language through extensive amount of phonetic learning. When the syllables are easy to process such as ones from the native language of the learner, only minimal amount of exposure was need for successful learning. A different way of approaching the same question is to look at the amount of input that can lead to successful learning. Gebhart, Newport and Aslin (2009) looked at statistical learning of unfamiliar sounds, and found that if the input is difficult to process, learning can take place but only after an extensive exposure period.

From this perspective, it is important to examine the amount of input learning requires for statistical learning. In this paper, we look at the question of the minimal amount of exposure for a specific type of stimuli for a specific learning population to learn successfully, which we argue can reveal important information about how the input is encoded, stored, and processed. We reason that the sequences with the least amount of exposure that leads to successful learning must possess properties that makes it easy to perceptually process the input (and vice versa), which provides a

hypothesis testing space for further experiments to figure out the specific cognitive and perceptual processes that statistical learning employs. It is also the case that the literature has a void regarding this question of the minimal amount of exposure for statistical learning. To our knowledge, only one study (Batterink, 2017) looked at this question, which used a target detection paradigm: English-speaking adult 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 found that statistical learning can be faster than previously thought: 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/la*) than to the first syllable (*tu*). This study showed that learners develop rapid sensitivity to the statistical structure of the stream given few exposures to word forms, but did not further explore the cognitive mechanisms that can explain these results.

In this study, we examine the theoretical questions outlined above, including how learning takes place as a function of exposure amount, as well as asking questions on how language specific factors influences statistical learning. Specifically, we tested adults speaking Mandarin Chinese, aiming to understand how certain language-specific factors, such as the mostly disyllabic rhythm of Mandarin Chinese (Wang, 2002) influence statistical learning given minimal amount of input for learning. Experiment 1 examined statistical learning using many syllable streams with different amount of exposure, as well as how learners treat two-syllable and three-syllable streams. This allows a comparison of the learning of the same syllable stream under different amount of exposure, as well as how streams with different rhythm patterns get learned. Given the learning results, we tested two factors that may explain the results for statistical learning in Mandarin in Experiment 2

and 3: a well-established rhythm bias that is specific to Mandarin, and linguistic entrenchment (similarity of novel words to existing words in Mandarin). Hopefully, these factors can help explain the variability in the learning results in Experiment 1.

Experiment 1

Methods

Subjects. One hundred and twenty students at the (anonymized for review) University were recruited to participate in the study. Eighty subjects participated in the four exposures condition, and forty participated in the two exposures condition.

The sample size was determined prior to data collection. Noting that there is a relatively small amount of previous research on learning with such low amount of exposure during learning, we determined the sample size by calculating the power analysis of one similar study (Batterink, 2017) and oversampling. A power analysis of Batterink (2017) was conducted (see the Online Supplementary Material for details), showing that a post-hoc power of 0.999 is achieved with 19 subjects in terms of Cohen's *d*. We purposely oversampled at 40 participants. In order to keep the number of participants who learn the same syllable stream the same between the four-exposure condition and the two-exposure condition, we used double the number of participants in the four-exposure condition.

Stimuli. To create the syllable streams used in the study, we used the NeoSpeech synthesizer (<http://www.neospeech.com/>, the male voice, Liang) to synthesize all of the syllables independently, before concatenating them into syllable streams. The syllables were real Mandarin Chinese syllables (with legal onsets, codas and tones), and the tones of the syllables were from 1 through 4 (syllables with a light tone were not used, because light tones were used in very restricted situations mostly having to do with functors or end of utterances). Each syllable took slightly under

250ms with small and variable amount of silence at the end of each syllable such that the syllable streams were presented precisely at 4Hz. The list of the syllables as well as the syllable combinations can be found in the Online Supplementary Material.

Design. We chose to use a traditional word segmentation paradigm, rather than the reaction time study design in Batterink (2017), because we wanted to show that participants have word form knowledge extracted from the continuous sequence¹. To this end, we create syllable streams and the corresponding word/part-word sequences under the unbalanced frequency design from Aslin et al. (1998)². To begin, we chose three-syllable combinations made of legal Mandarin syllables (which made words under statistical learning definitions) that are not real Mandarin words, in the sense that any of the two consecutive syllables in these units are not real Mandarin words (nor the three-word combination). A complete list of these three-syllable combinations can be found in the Online Supplementary Material. After we created 160 sets three-syllable combinations, every 4 of these trisyllables were taken together to make a set of words in statistical learning syllable streams, making 40 different syllable sets. For each syllable set, we created two trisyllabic syllable streams, where one is a counterbalancing trisyllabic syllable stream of the other. The counterbalancing stream was created by switching syllables such that words and partwords were swapped between these two streams (mirroring the designs in Saffran et al., 1996 and Aslin et al., 1998). This makes a total of 80 trisyllabic streams. In addition, we created the first 40 disyllabic streams, by taking the first two syllables of the three-syllable groups. For each of these

¹ The reaction time task from Batterink (2017) can show that participants reacting faster to later syllables in the word compared to the first syllable, which means that they have knowledge that the first syllable can predict the later syllables, but this does not show that participants know which syllables make a word.

² This is another place where our design differed from the Batterink (2017) design. In Batterink (2017), all the words in syllable streams have equal frequency (a characteristic from Saffran et al., 1996), which makes it difficult to disentangle transitional probability and frequency as the source of the learning effect: these factors can only be disentangled if the words in the sequence have lower frequency than part-words, as Aslin et al. (1998) explained. We designed all of our sequences with an unbalanced frequency design from Aslin et al. (1998).

streams, a counterbalancing disyllabic stream was created as well, so there was a total of 80 disyllabic streams. In sum, we used a total of 160 syllable streams, with 40 pairs of counterbalanced trisyllabic streams and 40 pairs of counterbalanced disyllabic streams.

Each syllable stream was designed in a similar way as Aslin et al. (1998), where there were high frequency words and low frequency words. As explained from Aslin et al. (1998), this enables the fact that part-words can be constructed such that they would have the same frequency in the syllable stream as the low-frequency words. In the two exposures condition, syllable streams were created by concatenating two high frequency words that occurred 4 times and two low frequency words that occurred 2 times. In the four exposures condition, high frequency words occurred 8 times and the low frequency words occurred 4 times. Sequencing of these words was random with two constraints: that a word does not follow itself in the sequence, and that the syllable stream contains the two partwords exactly the same number of times as the low frequency words.

The test items were constructed in the same way as in Aslin et al. (1998): In trisyllabic sequences, part-words were constructed with the last syllable of a high frequency word and the first two syllable of the other high frequency word. In disyllabic sequences, part-words were constructed with the last syllable of a high frequency word and the first syllable of the other high frequency word, as illustrated in Figure 1.

Figure 1. Schematic illustration of the design of Experiment 1. A) The design of a disyllabic syllable sequences, in which different numbers represent different syllables, and different colors represent different words. In this sequence, four disyllabic words are used, where two words had high frequency (occurring two times, [1, 2] and [7, 8]) and two words had low frequency (occurring two times, [3, 4] and [5, 6]). The sequence was created randomly with the constraints that no word follows itself, and in addition, that the transition between the high frequency words [8,1], [2 7] making up the part-words also occurs two times. Note that trisyllabic sequences are similarly constructed, and other than there were three syllables per word, all the other distributional aspects of disyllabic sequences were held constant. B) The design of the two conditions in the study. In both conditions, there were disyllabic sequences and trisyllabic sequences, and they were randomly presented to participants. The four exposures condition differed from the two exposures condition only in terms of the number of times each word occurred in the sequence. In the four

exposures condition, low frequency words occurred four times and high frequency words occurred eight times.

The study in both conditions lasted for roughly the same amount of time, and it took all participants under an hour to complete. Since the syllable streams in the four exposures condition are twice as long as those in the two exposures condition, participants in the four exposures condition learned a subset of 20 syllable streams consisting of 10 trisyllabic and 10 disyllabic streams each, and participants in the two exposures condition learned a subset of 40 syllable streams consisting of 20 trisyllabic and 20 disyllabic streams. For all participants, the streams within subject were drawn from entirely different three-syllable groups (i.e., any particular three-syllable group was learned once, either as a trisyllabic stream or a disyllabic stream with the first two syllables). Between subjects, the streams were counterbalanced by having 8 lists such that the lists covered all of the syllable streams in equal frequency. For each participant, the presentation order of the different streams was randomized.

Procedure. In the beginning of the study, participants were informed that they will be learning from multiple syllable streams and answering questions about each syllable stream after learning. As such, the study alternated between two types of trials, learning trials and testing trials. In learning trials, participants listened to a syllable stream passively through the headphones while the screen was blank. Participants were asked to pay attention to the syllable stream.

After each syllable stream is played, a test phase with 8 test trials would start immediately. The instructions for the test phase were displayed on the screen, indicating that a number of syllable sequences were about to be played, and participants would be asked to make familiarity ratings about these syllable sequences, given the syllable stream they just heard. The 8 test items

were created by repeating 4 unique test items, which included two *words*, which were low frequency words from the language stream, and two *part-words*, made from parts of two high frequency words from the language stream. The presentation order of these items was randomized for each participant, under the constraint that all 4 unique items were presented before they were presented again, in a new random order and the same item was not tested twice in a row.

In each test trial, we first played the syllable sequence (either a word or a part-word) twice, with a 1 second pause in between (e.g., xian2mo2...xian2mo2)³. Participants were then asked to rate the familiarity of the item: “Given the new language you just learned, how familiar is this word?” Their responses were marked on a scale containing four response items: “Very familiar”, “Somewhat familiar”, “Somewhat unfamiliar”, and “Very unfamiliar”. 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. After all the 8 test items were rated, participants were told that they are done learning this syllable stream, with one remaining question, which is whether they have heard any words in the language stream that is a real Mandarin word. If they heard anything that resembles a real word, participants were asked to write them down on an answer sheet. After they are done writing, participants were informed that the next syllable stream would start, which is the beginning of the learning phase for the next syllable stream. This repeated until all 40 (the two exposures condition) or all 20 (the four exposures condition) syllable streams for that participant have been learned and tested.

³ For all the syllables used in the study, we use a pinyin representation where the syllables are followed by a number 1 through 4, such that the number represents the tone the syllable is in. In this example, both syllables xian and mo are in the second tone.

Results

Overall Learning Results. We asked whether learners can successfully extract word forms after only two or four exposures by asking participants to rate words and part-words, and if words are rated more familiar than part-words, it would be taken as evidence for learning. In the experiment, participants rated words and part-words with a familiarity scale, and we coded the scale of “Very familiar”, “Somewhat familiar”, “Somewhat unfamiliar”, and “Very unfamiliar” into the numeric values 1, 2, 3, and 4, respectively. The prediction for learning is that participants would rate words as more familiar than part-words. The first question was whether learners can successfully extract word forms only after two exposures. Taking all the rating data in the two exposures condition, we conducted our statistical analysis with mixed effect linear regressions. In the regression, the rating was the independent variable, and the fixed effect was item (word/part-word), with a random structure of subject intercepts nested in individual language stream intercepts. The result was significant, showing that words were rated more familiar than part-words ($\beta=0.061$, $z=4.36$, $p<0.001$).

Next, we asked whether learners can successfully extract word forms after four exposures, and whether the learning is better than in the two exposures condition. First, we took all the rating data in the four exposures condition, we conducted our statistical analysis with a mixed effect linear regression, where the fixed effect was item (word/part-word), and a random structure was subject intercepts nested in individual language stream intercepts. The result was significant, showing that words were rated more familiar than part-words ($\beta=0.285$, $z=19.78$, $p<0.001$). Next, we combined the data between the two exposures condition and the four exposures condition and conducted a mixed effect linear regression. In this regression, the fixed effect was the interaction and the main effects of both item (word/part-word) and number of exposure (2/4). We also used a

random structure of subject intercepts nested in individual language stream intercepts. The interaction was significant ($\beta=0.224$, $z=11.12$, $p<0.001$), showing that the learning in the four exposures condition was significantly better than in the two exposures condition. The results of Experiment 1 are shown in Figure 2.

Figure 2. The rating results from Experiment 1, showing the mean of the ratings (the error bars represent 95% confidence interval). On the y-axis the mean rating is plotted, where the smaller values correspond with more familiarity.

To visualize the learning result in such a way that the effect size of learning can be compared directly between conditions and between languages, we converted the results from the regressions to an effect size estimate, similar to a Cohen's d (Brysbaert & Stevens, 2018). Note that the β that the regressions report is the difference of rating between item types, and this allows us to directly obtain the effect size from the mixed effect regression estimates by dividing the model estimate (the β s) by the sum of random effect variance and residual variance. The results

can be seen in Figure 3, showing that word segmentation is successful after just two exposures, and learners have knowledge of the word forms, as shown from a segmentation paradigm. This is consistent with the findings from Batterink (2017), who used a target detection paradigm to show participants' knowledge of the statistical structure in the syllable stream, though our result is more directly interpretable as word segmentation performance. Furthermore, the amount of learning resulting from just two exposures is quite small, and there is a big difference in terms of the amount of learning between the two exposures and four exposures conditions such that even a small amount of increase in terms of exposure can bring about significant more knowledge of word forms in statistical learning.

It is worth noting that in previous adult studies on statistical learning, the exposure period is typically much longer, and the effect size of learning can be either higher or comparable to these estimates. In a study where the design was very close to the current study, Weiss, Gerfen and Mitchel (2009) found an effect size of 2.02, 1.74 and 3.75 with the same syllable sequence but with three different male and female voices after 12 minutes of exposure in Experiment 1. In Saffran et al. (1997), the syllable sequence was designed by concatenating 6 different trisyllabic words, and learners were exposed to the sequence for 21 minutes (Experiment 1) or 42 minutes (Experiment 2), which generated effect sizes of 0.69 and 1.84, respectively. In similar work on adjacent dependency learning, Misyak and Christiansen (2012) had an effect size of 0.85 after a 25-minute exposure period, and Wang, Trueswell, Zevin & Mintz (2020) had an effect size of 0.75 after 2.4 minutes of exposure to the artificial language. Note, however, that the results here are the combined results of many syllable sequences not yet considering syllable stream variability, a likely contributor to the effect size estimate, which we discuss next.

Before we analyze potential perceptual factors that contribute to the variability of learning between different syllable streams, we turn to analyze the variability of learning of different syllable streams, both in terms of the range of learning results by different syllable streams, as well as the learnability of a particular syllable stream given different amounts of exposures.

Figure 3. The effect sizes of statistical learning of different languages on average in units of Cohen’s d , after learning from a sequence in which words were presented either two or four times. Error bars represent the 95% confidence interval of the effect size estimate.

Syllable Stream Variability Ranges. In this analysis, we look at the range that different syllable sequences would generate, given different amount of learning. Recall that the design of the study includes 160 different syllable sequences, and this allows us to examine the variance of the learning effect given different syllable sequences. If it were the case that different syllable sequences generate roughly the same amount of learning, we would expect that the syllable

sequences in the two exposures condition all have roughly a small amount of learning, whereas the syllable sequences in the four exposures condition all have roughly a large amount of learning.

In order to conduct this analysis, we ran an analysis on rating for each language stream separately, under the two- and four-exposure conditions respectively. In each regression we included rating as the independent variable, with item type (word vs. part-word) as the fixed effect and by-subject random intercepts. After the estimates were generated for each regression, we computed an effect size (obtained by dividing the regression model estimate (the β) by the residual variance) for each regression. To visualize all the different syllable sequences, we sorted all of the effect size estimates of different syllable streams from low to high and plotted the sorted result in Figure 3. The sorted figure shows that the variance in the four exposures condition (s.d. = 0.42) is more than in the two exposures condition (s.d. = 0.27), and a Levene's F statistic on variance showed a significant difference ($F(1, 318) = 25.402, p < 0.001$). A bar and whisker plot showing all observations is also shown in Supplementary materials.

Not only is the variance in four exposures is more than in two exposures, but we also see disyllabic streams are learned better than trisyllabic streams, replicating Yu et al. (2021). We examine this statistically as an interaction between item type (words/part-words) and the syllable length per word in the stream (2/3) and find that this is the case in the four exposures condition ($\beta = -0.574, z = -20.27, p < 0.001$) as well as in the two exposures condition ($\beta = -0.406, z = -14.60, p < 0.001$), an effect we will explore in later experiments.

Figure 4. Sorted effect sizes of statistical learning of different languages on average, after learning from a sequence in which words were presented either two or four times. The error bars represent the 95% confidence interval of the effect size estimate. The horizontal line at 0 represents chance. Note that the x-axis shows the rank order of the different syllable streams, which can be different for streams in four exposures and in two exposures, so there is no correspondence between the streams in two and four exposures along the x-axis; the relationship between the same stream with different amount of learning is shown in Figure 5.

Syllable Stream Learnability. As we discussed, one of the questions that remain unclear in the literature is regarding the learnability of different streams that consists of different syllable combinations. For different syllable combinations, does the amount of learning scale to the number of exposures? Note that the streams in Figure 4 have been sorted by the effect size, so the learning of the same syllable stream in the two exposures and four exposures condition is not yet clear. To

answer the learnability question, we examine the correlation between the learning effects of the same stream after 2 and 4 exposures, looking at a scatterplot to show the learning of the same syllable stream in the two exposures and four exposures condition (Figure 5).

Figure 5. The amount of learning effect sizes of statistical learning of different languages on average, after learning from a sequence in which words were presented either two or four times.

We observe a strong correlation between the effect sizes of learning in the two exposures and four exposures condition ($r = 0.589$, $p < 0.001$). If we separate the data by syllable sequence type (disyllabic vs. trisyllabic) and look at the correlations separately, the correlation is significant for disyllabic streams ($r = 0.412$, $p < 0.001$) as well as for trisyllabic streams ($r = 0.331$, $p = 0.003$). This indicates that the learnability of a particular syllable stream is inherent to the specific syllable combinations, where the learning of words in a particular syllable stream after two exposures will

get a certain amount better on average after two more exposures.

For completeness, we analyzed the number of real words that were perceived during the study per syllable stream, following the work from Palmer, Hutson, White and Mattys (2019) showing that lexical knowledge of word forms in the syllable sequence can boost the statistical segmentation of other words in the sequence. To be clear, the reason that participants in our study would perceive real words in the syllable streams could be due to a variety of factors, such as the syllables produced by the synthesizer could be somewhat ambiguous between two easily confusable syllables, producing a percept of some syllable combinations make up real words. The number of real words per subject per syllable stream was 0.74 (s.d. = 0.37) in the four exposures condition, and the number of real words per subject per syllable stream was 0.55 (s.d. = 0.31) in the two exposures condition, and this number were significantly different between the two conditions in a two-sample t-test assuming unequal variance ($t(308.7) = 4.985$, $p < 0.001$). We looked at the correlation between the learning effect sizes of each syllable stream and the number of real words and the correlation was neither significant in the four exposures condition ($r = 0.038$, $p = 0.630$) nor in the two exposures condition ($r = 0.106$, $p = 0.181$). This indicates that the learnability of syllable streams is not correlated with the number real words one perceives during learning with respect to different syllable streams.

What explains the variability in learning of different syllable sequences in Experiment 1? Below, we explore two perceptual factors, rhythm biases in Mandarin Chinese, and the idea of linguistic entrenchment, the idea that even though the words making up syllable sequences are novel, their similarity to real words of the learners can still contribute to segmentation performance. These perceptual factors hold the potential of explaining the variability of learning between different individual syllable streams.

Experiment 2

The fact that most statistical learning sequences were designed by concatenating words with the same number of syllables makes rhythm a salient percept (Wang, Zevin & Mintz, 2019). In the current experiment, we directly test the extent to which learners can perceive rhythms in statistical learning streams. As we mentioned in the introduction, there is a rhythmic expectation for two-syllable streams in Mandarin Chinese (Wang, 2002). We also see from Experiment 1 that two-syllable streams are learned better with a novel syllable sequence. We thus hypothesize that, two-syllable rhythms are perceived more accurately than three-syllable streams, and if so, this supports statistical learning by providing a segmentation cue that the two-syllable streams can be cut up every two syllables. When streams were made of all three-syllable words, perceiving this rhythm can be more difficult, leading to worse learning.

Methods

Subjects. A total of forty students at the (anonymized for review) University were recruited to participate in the study. These subjects were split into eight groups of five, where each group learned 40 different language streams (described below). This particular procedure is made possible by mirroring the distribution of syllables to participants in the two-exposure condition in Experiment 1, which determines the sample size in Experiment 2.

Stimuli. The stimuli used in this study are the same ones from the four exposures condition in Experiment 1, where statistical learning of these streams was studied.

Design. Given the 160 different syllable streams from Experiment 1, we distributed the streams to participants by creating eight different lists, a process that is the same as the one the four-exposure condition of the Experiment 1. Each list included 40 streams, half of which were disyllabic, and half were trisyllabic. Each participant in Experiment 2 was given the 40 streams to

rate their rhythms using a simple procedure, described below. In total, 40 participants rated 40 streams each, and each stream was rated by 10 different participants.

Procedure. The study began with an instruction phase that explained the experimental task in detail. In the instruction phase, participants were told that they will be listening to syllable streams, and introduced to two concepts, that some syllable streams have a *two-syllable rhythm*, and some syllable streams have a *three-syllable rhythm*. Their job in the study is to figure out whether a particular syllable stream has a *two-syllable rhythm* or *three-syllable rhythm*. The instruction phase proceeded by giving syllable stream examples. Participants were told that they would be hearing a stream with a *two-syllable rhythm*, and a *two-syllable rhythm* stream was presented to them. This two-syllable stream contained disyllabic non-words only and lasted 8 seconds. Participants were then asked whether they wanted to hear the stream again, and if they wanted to, they could play the same syllable stream again until they indicated to proceed. Then, participants were told that they would be hearing a stream with a three-syllable rhythm. This three-syllable stream contained trisyllabic words only and also lasted 8 seconds. Participant could also play the same trisyllabic syllable stream again until they indicated to move on.

Given these initial experiences rating syllable streams, participants were then informed that if they could figure out the rhythm before the syllable stream is done playing, they could press Space to end the playing of the stream, and answer the rating question. We asked participants to spend as little time as possible on each stream, and ending the playing of the streams as early as possible as long as they have figured out the rhythm stream. To illustrate how to end the playing of the streams, we repeated playing the two-syllable example stream, followed by the three-syllable example stream, each time letting participants to end the streams before they were done playing. This concludes the instruction phase.

After the instruction phase, the main phase began where participants rated the rhythms of syllable streams. In the main phase, syllable streams were presented one after another: Each time a stream was presented, the rating question came up (described below). Note that given the statistical design of the syllable streams, disyllabic streams had a duration of 12 seconds and trisyllabic streams had a duration of 18 seconds, though the actual playing time for streams could be shorter than the total duration. For each syllable stream, participants first played the syllable stream (while capable of stopping the stream playing any time after 1 second of playing). The one second threshold was to prevent from erroneous key presses that ends playing prematurely (it is also impossible to perceive the rhythms under 1 second, when syllables were presented at 4Hz). Once the streams stopped playing (either stopped by the participant or after it finished playing), participants were asked to rate the rhythm of the stream: “Did the language stream have a two-syllable rhythm or a three-syllable rhythm?” Their responses were marked on a scale containing four response items: “Definitely a two-syllable stream”, “Possibly a two-syllable stream”, “Possibly a three-syllable stream”, and “Definitely a three-syllable stream”. Once the participants made their choice, the screen alerted them that they could start the next stream. This repeated until all of the 40 syllable streams were over for the participant. The whole experiment lasted around 40 minutes for each participant.

Results

We analyze the rating results in two ways, treating the responses as either categorically correct/incorrect, or viewing it as a response on a continuous scale. First, we took on the view that the rating scale is categorical, where the expected result for the rhythm rating task is that disyllabic streams would be rated as “Definitely a two-syllable stream” or “Possibly a two-syllable stream” and trisyllabic streams would be rated as “Definitely a three-syllable stream” or “Possibly a three-

syllable stream”. The choice of “possibly” indicates uncertainty from the participants, but it is not related to the correctness of the ratings. Thus, we coded “Definitely a two-syllable stream”, “Possibly a two-syllable stream” to correct (coded as a 1) and “Possibly a three-syllable stream”, and “Definitely a three-syllable stream” to incorrect for responses to two-syllable streams, and vice versa for responses to incorrect (coded as a 0). Given this coding, we find that the ratings for two-syllable streams are more likely to be correct than three-syllable streams by a large margin. The mean accuracy for two-syllable streams was 73.4% (SE = 0.6%) and the mean accuracy for three-syllable streams was 26.0% (SE = 0.5%). To compare the accuracy for these two types of streams, we conducted a mixed-effect logistic regression where rating was the dependent variable, and the fixed effect was stream type (disyllabic stream vs. trisyllabic stream) as the fixed effect. By-subject random intercepts and a random slope of stream number (coding for different streams) was used as the random effect structure. The result was that the ratings were highly different for the two types of streams ($\beta=-2.518$, $z=-28.94$, $p<0.001$),

The accuracy analysis shows a big difference between the rhythm rating of syllable sequences, but it cannot show the gradient differences of ratings for different syllable streams. For example, it cannot demonstrate that the certainty or the lack of for a particular syllable stream. In order to capture the gradient nature of these rhythm ratings, we coded the same rating data by treating the items on the same to be continuous. For two-syllable streams, we coded “Definitely a two-syllable stream”, “Possibly a two-syllable stream”, “Possibly a three-syllable stream”, and “Definitely a three-syllable stream” are coded into the numeric values 1, 2, 3, and 4, respectively, representing four continuous states (correct with certainty, correct with uncertainty, incorrect with uncertainty, and incorrect with certainty). For three-syllable streams, the reverse order was used where for example, “Definitely a three-syllable stream” was coded as 1, and “Definitely a two-

syllable stream” was coded as 4, etc. Under this coding scheme, the mean for two-syllable streams is 1.867 (SE = 0.012, CI= [1.843, 1.891]), and the mean for trisyllabic streams is 3.071 (SE =0.012, CI= [3.048, 3.094]). To compare these ratings statistically (Figure 6), we ran mixed effect linear regressions on the data, where the rating was the dependent variable, and the fixed effect was stream type (disyllabic stream vs. trisyllabic stream) as the fixed effect. By-subject random intercepts and a random slope of stream number (coding for different streams) was used as the random effect structure. The result was that the ratings were highly different for the two types of streams ($\beta=1.204$, $z=33.59$, $p<0.001$, CI= [1.134, 1.274]). This result confirms the hypothesis that participants rated disyllabic streams (compared to trisyllabic streams) are more likely to have a two-syllable stream than three-syllable stream. The distribution of the results, tallied by mean ratings to individual syllable streams, is shown in Figure 6.

Figure 6. The distribution of rhythm ratings. Each observation is the mean rating of a particular syllable stream. The x-axis shows the mean rating (from 1, correct with certainty to 4, incorrect

with certainty), and the y-axis shows the count. The tallies are broken down by the type of syllable stream (disyllabic vs. trisyllabic).

For completeness, we also analyzed the duration of the syllable streams participants played. To do this, we looked at the distribution of the duration for both the disyllabic and trisyllabic streams separately, because while the maximum duration of a disyllabic stream is 12 seconds, the maximum duration for a trisyllabic stream is 18 seconds, due to the fact that the same number of words were in all streams. Looking at the durations for disyllabic streams and trisyllabic streams, we find that a large number of streams were played until the sound finished playing, generating a large number of 12-second durations for disyllabic streams (482 out of 800) and 18-second durations for trisyllabic streams (402 out of 800). Dropping these trials, the mean duration for disyllabic streams is 7.251 seconds ($SE = 0.143$, 95% $CI = [6.970, 7.533]$) and the mean duration for trisyllabic streams is 9.667 seconds ($SE = 0.207$, 95% $CI = [9.260, 10.074]$). The mean duration for the disyllabic and trisyllabic streams was different in a mixed effect regression, where the fixed effect is the type of stream, and the random effect was a by-subject intercept only ($\beta = 1.892$, $z = 9.31$, $p < 0.001$), due to the fact that the model fails to converge if a by-stream random intercept is included.

Rhythm rating of syllable streams is of interest not only because it shows that learners are sensitive to rhythm in statistical learning streams, but also could potentially explain the statistical learning results from Experiment 1. To this end, we examined the relationship between the learning outcome from the word segmentation task (the four exposures condition in Experiment 1), and rhythm ratings of syllable streams, in two separate regressions looking at disyllabic and trisyllabic sequences. This can potentially answer the question of whether the ease with which the rhythm of

a particular syllable stream is perceived predict the learning of that syllable stream. In terms of the learning scores for syllable sequences, the data from the four-exposure condition in Experiment 1 was used to generate an effect size estimate for each sequence (these were used to generate Figure 4 and 5), where we ran a regression for each sequence looking at the difference between words and part-words and obtained its effect size. For rhythm ratings, the means of syllable streams were used. First, we ran the interaction between syllable sequence type (disyllabic/trisyllabic) and effect size of the learning per syllable sequence. The interaction was significant ($\beta=0.334$, $t=2.19$, $p=0.030$), showing that the relationship between rhythm rating and learning was different for disyllabic sequences and disyllabic sequences, so we ran two separate regressions looking at both types. For the disyllabic sequences, the relationship between rhythm rating and learning was significant ($\beta= -0.368$, $t(79)=-3.13$, $p=0.002$), such that the better the rhythm of the sequence is judged, the better the learning is. However, this was not the case for trisyllabic sequences ($\beta= -0.034$, $t(79) = -0.35$, $p=0.726$). This pattern of data could emerge if learners apply their two-syllable stream rhythm bias to segment all the syllable sequences: for disyllabic sequences, the success to find a two-syllable rhythm predicts how well segmentation happens, but for trisyllabic sequences, this rhythmic bias will make segmentation fail.

Figure 7. The correlation between the rhythm rating score and the effect sizes of streams in the statistical learning tasks, as broken down by disyllabic (Figure 7A) or trisyllabic (Figure 7B) sequences. The y-axis plots the mean rhythm rating for each syllable streams from the current experiment, and the x-axis plots the effect sizes of the learning outcome from Experiment 1 in the four-exposure condition.

Experiment 3

In Experiment 3, we explore the idea of linguistic entrenchment (Siegelman et al., 2018; Seidenberg & Zevin, 2006), namely, that even though the syllable sequences contained only non-words from the perspective of participants' native language, the similarities of these non-words to real words in Mandarin can be variable and could attribute to the ease with which non-words can be segmented. We take the same set of stimuli from Experiment 1 to use in Experiment 3 to assess the similarities of the non-words to real words in Mandarin in a different group of participants. The similarity data will be used to explain the variance between different learning sequences.

Methods

Subjects. Forty students at the (anonymized for review) University were recruited to participate in the study, and none of them participated in Experiment 1 or 2. This particular procedure is made possible by mirroring the distribution of syllables to participants in the two-exposure condition in Experiment 1, which determines the sample size in Experiment 3.

Stimuli and Design. We took the test items from Experiment 1, with the same distribution in terms of which participants get which test items. In Experiment 1, each participant learned from 40 syllable streams and rated 160 items, since each syllable stream had 4 test words associated with them (2 words, 2 part-words). In Experiment 3, each participant rated 160 words, and the counterbalancing of all items is the same as in Experiment 1. The order of the 160 words was randomized for each participant.

Procedure. Experiment 3 began with instructions. The instruction informs participants that they are about to rate the similarity of novel words to real words in Mandarin. The instruction proceeded showing the question participants will answer in the study, which is “How similar is the artificial word you just heard to words in Chinese?” They were also shown the four options to choose from: “Sounds to be a real word”, “Sounds somewhat similar to a real word”, “Sounds somewhat dissimilar to a real word”, and “Sounds very dissimilar to a real word”. After informing participants of how to use the scale, two examples (one real word and one novel word) were given, and instructions informed participants to select “Sounds to be a real word” for the real word item, and used their own judgment for the other example test item.

After the instructions, participants started on the rating task. Novel words (which were words and part-words from Experiment 1) were presented one by one in a randomized fashion. Each word was played twice, with a second of pause in between (so that participants can listen to

the same word twice). After the non-word was played, a question appeared on the screen: “How similar is the artificial word you just heard to words in Chinese?” To answer the question, participants were given four options: “Sounds to be a real word”, “Sounds somewhat similar to a real word”, “Sounds somewhat dissimilar to a real word”, and “Sounds very dissimilar to a real word”.

Results

Distribution of the similarity ratings of words and part-words. For all items that were rated in Experiment 3, we computed an average rating of that word. The frequency of the mean ratings of novel words is plotted in Figure 8, broken down by syllable sequence type (disyllabic/trisyllabic sequences). We see a large effect of syllable sequence type such that participants rated disyllabic words significantly more similar to real words than trisyllabic words. To analyze this statistically, we conducted a mixed effect linear regression looking at the rating of all words, with a fixed effect of syllable sequence type (disyllabic/trisyllabic), and a by-subject random intercept. The effect was significant ($\beta=0.625$, $z=26.12$, $p<0.001$). This is consistent with our understanding that, because most Mandarin words are disyllabic, the words in disyllabic sequences would appear more similar to real words. This effect predicts that disyllabic sequences are learned better than trisyllabic sequences as a matter of linguistic entrenchment, where similarity ratings can explain the variances in the learning result of test items, which we explore next.

Figure 8. A histogram of the distribution of the real-word similarity ratings for all the novel words in Experiment 3, which shows that participants rate disyllabic words as more real-word like than trisyllabic novel words.

Relationship between novel words to real words similarity and their learnability in the word segmentation task. To examine the relationship between novel words (i.e., “word” or “part-word” in statistical learning terms) to real words similarity and their learnability in the word segmentation task more fully, we conducted several regression analyses, looking at the relationship between the learning score from the word segmentation task (the four exposures condition), and the similarity score of novel words to real words. This relationship has a two-by-two structure, syllable sequence type (word/part-word) by counterbalance status of the sequence (sequence/counterbalance sequence), shown in Figure 9. The unit of analysis is for a test item for Experiment

3, unlike Experiment 2, where the unit was a syllable stream that was used for learning. For a particular test item, we regress the mean of the statistical learning scores from Experiment 1 (the four exposures condition), by the mean of similarity rating from Experiment 3 for that item. We see positive relationships between the statistical learning scores and the similarity rating only for “words”, and zero relationships for “part-words” in linear regressions. For “words” in sequences, the slope is significantly positive ($\beta=0.544$, $t(158)=5.90$, $p<0.001$) but the slope for “part-words” in sequences is not ($\beta=-0.086$, $t(158)=-0.78$, $p=0.436$). The slope for “words” in counterbalanced sequences is significantly positive ($\beta=0.625$, $t(158)=5.89$, $p<0.001$) but the slope for “part-words” in counterbalanced sequences is not ($\beta=-0.074$, $t(158)=-0.65$, $p=0.518$). In sum, the similarity of “words” to real words in Mandarin is helpful for segmentation, but the similarity to real words of “part-words” has no effect. These relationships are shown in Figure 9, where learning scores are on the x-axis, and the similarity data are on the y axis. For the visualization, we fit a least-square line through all the data for each figure. Notably, these relationships if one is to believe that the first option, “Sounds to be a real word” is not a continuous part of the scale along the other options. After removing all observations where participants selected “Sounds to be a real word” (which included 17% of the data), the same results are observed. This analysis is reported in the Online Supplementary Material.

Figure 9. The relationships of statistical learning scores of “words” and “part-words” (y-axis) and their real-word similarity score, broken down by “word”/”part-word” and whether they are in counterbalancing sequences. In the figure, each dot represents a particular test item. Note that Figure 9A and 9D has the same items, and 9B and 9C has the same items because they are counterbalanced conditions.

Taken together, results from Experiment 3 show that the similarity of novel words to real words can explain variability in the learning outcomes in statistical learning. The finding is that when the novel words in syllable sequences resemble real words, they are much more easily

extracted even when they appear in continuous syllable sequences. The fact that statistical learning scores for words and the similarity of these novel words were obtained from a different group of participants makes a strong point that it is the characteristics of the novel words that determines how well they get segmented from continuous sequences.

Lastly, we looked at the relationship between learning, non-word to real word similarity rating and rhythm rating of different syllable streams as a function of disyllabic and trisyllabic streams. Note that this analysis is conducted with data from separate participants, making a strong case that the result can indicate the learnability of the individual syllable streams, rather than an indication of learning patterns of the specific participants we ran. In a mixed regression, the learning outcome of statistical learning (taken from the four-exposure condition from Experiment 1) was the dependent variable, and the four-way interaction between 1) item type (words vs. part-words, coded given its counterbalancing condition⁴), 2) language type (disyllabic/trisyllabic), 3) mean similarity rating of test items from Experiment 3 (a continuous variable) and 4) mean rhythm rating of syllable sequences (with continuous coding from Experiment 2, a continuous variable) were the fixed effect, with both a by-subject and by-stream random intercept.

The four-way interaction was significant ($\beta = -0.406$, $z = -2.03$, $p = 0.042$), suggesting that similarity rating and rhythm rating contributed significantly to explain the variance in learning but did so differently for disyllabic and trisyllabic streams for words/part-words. To understand this interaction, we plotted contour plots of predicted statistical learning ratings from Experiment 1 for words vs. part-words for both disyllabic and trisyllabic sequences given the other factors (Figure 10). Different colors in the contour plots represent the predicted value of the rating given in Experiment 1 at different non-word to real word similarity rating from Experiment 3 (shown on

⁴ This was done such that words and part-words are reversed in the counterbalancing condition.

the x-axis), and the rhythm rating of different streams in Experiment 2 (shown on the y-axis). To understand these effects beyond a visual estimation, we conducted two separate mixed effect linear regressions looking at disyllabic and trisyllabic streams separately. Both regressions used the learning outcome of statistical learning as the dependent variable and we used the following procedure to determine the appropriate fixed effects: We started with a three-way interaction between item type (words/part-words), the similarity rating of test items and the rhythm rating of syllable sequences with continuous coding. If the highest-order interaction was not significant, it was dropped, and all the next order interactions were used. This process was done iteratively to select the model that explains the data best for both the disyllabic and trisyllabic streams separately with the least number of parameters. Given this process, we find that for disyllabic sequences, there was a two-way interaction between rhythm ratings and item type ($\beta=-0.334$, $z=-6.12$, $p<0.001$) and a main effect of similarity rating ($\beta=0.119$, $z=5.14$, $p<0.001$). This also means that all other interactions are not significant. This result indicates that, for disyllabic syllable streams, rhythm perception made words easier to learn but part-words harder to learn (something we expect given the hypothesis that rhythm perception facilitates segmentation). Similarity rating of test items had a main effect such that the more familiar the novel word is to known words, which did not interact with other variables for disyllabic words. Looking at Figure 10, the distribution of words and part-words for disyllabic streams are left-right reversed, roughly speaking: if one flips the figure for disyllabic part-words horizontally, the distribution becomes the same. For trisyllabic sequences, the three-way interaction was significant ($\beta= -0.423$, $z= -2.90$, $p=0.004$). In this regression, two additional two-way interactions are significant: item type (words/part-words) interacts with similarity ratings ($\beta=1.053$, $z= 3.07$, $p=0.002$) as well as rhythm ratings ($\beta= 1.275$, $z=2.79$, $p=0.005$). In other words, for trisyllabic syllable streams, all three factors interact and

contribute to the variance of how well an item was learned. Looking at the two figures for trisyllabic words and part-words, we see that only word ratings were impacted by rhythm perception and similarity, but part-word ratings were only impacted by similarity ratings. In other words, the learning pattern for trisyllabic words is such that, if factors such as linguistic entrenchment and rhythm perception can help with identifying words from a syllable stream, learners will remember them, but they have no knowledge of part-words, such that the ratings for part-words are entirely based on their similarities to existing words in the lexicon.

Figure 10. Contour plots showing the relationship between the statistical learning rating of words/part-words (from the four-exposure condition of Experiment 1, shown as colored areas),

rhythm rating (from Experiment 2) and similarity rating (from Experiment 3), broken down by the status of the test item (word/part-word) and syllable sequence type (disyllabic/trisyllabic). The values of the statistical learning rating are the predicted values from the above-reported four-way regression, where the color represent the margin estimates for words and part-words, where smaller values represent a response that this item was more familiar given the coding of responses from Experiment 1. Also, smaller values for rhythm rating represent a correct identification of the rhythm with certainty (from Experiment 2) and smaller values for similarity rating represent more similar to real words (from Experiment 3).

General Discussion

The results of the present study answered a variety of learnability questions in statistical learning, contributing to our understanding of the factors at play during the processing of short syllable streams. In terms of learnability as a function of exposure, we found that a word form in a continuous sequence can be learned after just two exposures, and furthermore, this learning became more robust after two more exposures. This difference showed that even a small amount of increase in exposure could bring about significant more knowledge of word forms. The fact that we used a large number of different syllable sequences consisting of different words (set of syllables) allowed us to observe the variability between the learning result of these different syllable sequences. This is similar to studying individual differences, except that we learned about the differences in learnability of these different syllable streams. We found that sequences consisting of different syllable combinations were drastically different in terms of learnability: the learnability of a particular syllable stream is inherent to the specific syllable combinations, such that having four exposures linearly increases the knowledge of the extracted word form compared

to having two exposures. Lastly, we explored two factors that can explain the differences between the learnability of different syllable streams: linguistic entrenchment, the idea that the similarity of these novel words to real words in the language contributes to how easy the word forms of novel words can be extracted, and rhythm perception, the idea that learners can perceive rhythm inherent to the syllable streams to segment the continuous syllable stream.

These learnability issues have to do with a fuller description of statistical learning. The findings regarding the minimal amount of exposure that leads to robust learning demonstrate that statistical learning is a robust learning mechanism. The variability in learning for different syllable streams, especially the fact that the same syllable stream has a certain learnability profile (see section *Syllable Stream Learnability*), demonstrates that the outcome of statistical learning is related to facts beyond the conditional probabilities of the syllables in the sequence⁵, and warrants an explanation. In Experiment 2 and 3, we examined how well listeners can identify the rhythm of the syllable streams and how similar test items are to real words of the native language. We find that both the rhythm property of the syllable streams and the similarity of the novel words to real words play a significant role in the learning of novel syllable sequences from continuous input, highlighting the idea that the perceptual process of statistical learning engages prior knowledge of the language, which in turn heavily influences the processing of syllable streams, even given minimal amount of exposure.

The current study highlights the fact that the knowledge of the lexicon influences statistical learning both at the scale of individual speech streams (as a function of the ability to perceive a certain rhythm in a manner that's language specific), as well as individual novel words (as a

⁵ This is also rarely discussed, because the prior studies almost exclusively used materials of CV syllables, the learning of which are assumed to be representative of learning of all syllable combinations. But as we discuss in this paper, this is almost certainly untrue.

function of its similarities to existing words). We believe that these findings are hold implications for a complete description of the mechanisms that support successful segmentation. As we mentioned, the transitional probability model by itself is a poor fit to the data, because it only explains at the coarse level that learning performance is above chance collectively⁶, but cannot explain any of the variability between the learning outcome of different syllable sequences. In the same vein, other computational models that only considers the identity of syllables as either the same or different would also fail, because again, the distributional information of all sequences are the same. However, all of many aforementioned factors, whether it is the expectation of word lengths or the effect of linguistic entrenchment on segmentation generates a different prediction for each syllable sequence, either by setting a prior on how many syllables to segment each time, or generate word candidates of different saliency for different syllable streams, could explain the variability between different syllable sequences. As such, this work calls for a computational model that takes into consideration the linguistic knowledge of a learner and how that knowledge biases the segmentation process, a topic for future work.

In sum, this study shows that even though statistical learning can take place after a minimal number of exposures, increasing the amount of exposure can improve learning. The variability in the learning outcomes have large variability between different syllable sequences, which factors like rhythmic biases and linguistic entrenchment can explain. This variability is striking in the

⁶ To be clear, with only two exposures, a model computing transitional probabilities can explain the above chance learning result, contrary to the claims in Batterink (2017). Consider the sequence in Figure 1A. In the sequence, we see that the word “3 4” have the syllable 3 always precede syllable 4, but the part-word “2 7” have the syllable 2 is followed by syllables 3, 5, and 7, which makes the transitional probability of syllable 2 to syllable 7 to be less than 1, suggesting a boundary between 8 and 1. Models such as PARSER, on the other hand, may or may not work with such a sequence. While it may accidently segment the right words, it is also unconstrained during the initial segmentation process, and if it segments part-words by chance, the reuse mechanism is not helpful in reducing the plausibility of part-words due to the short sequence in total, as well as matched frequency of part-words compared to the low-frequency words.

sense that its explanatory power lies in factors that are domain specific. Even in a simple task such as word segmentation, because it involves language, the process is far different from counting of beads on a string.

References

- Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science*, 28(7), 921-928.
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1), 1-20. <https://doi.org/10.5334/joc.10>
- Emberson, L. L., Liu, R., & Zevin, J. D. (2013). Is statistical learning constrained by lower level perceptual organization? *Cognition*, 128(1), 82-102.
- 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. <https://doi.org/10.1016/j.jml.2008.10.003>
- Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117(2), 107–125. <https://doi.org/10.1016/j.cognition.2010.07.005>
- Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning. *Language Learning and Development*, 9(1), 1–24. <https://doi.org/10.1080/15475441.2012.707101>
- Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*, 16(3), 486-490.
- Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, 33(2), 260–272. <https://doi.org/10.1111/j.1551-6709.2009.01012.x>

- 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.
<https://doi.org/10.1006/jmla.2000.2755>
- 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.
- Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, 122(2), 241–246.
<https://doi.org/10.1016/j.cognition.2011.10.007>
- Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, 62(1), 302-331.
- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2), 127–162.
[https://doi.org/10.1016/S0010-0285\(03\)00128-2](https://doi.org/10.1016/S0010-0285(03)00128-2)
- Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298(5593), 604–607. <https://doi.org/10.1126/science.1072901>
- Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39(2), 246-263.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>

- Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and Learning) out of the Corner of Your Ear. *Psychological Science*, 8(2), 101–105. <https://doi.org/10.1111/j.1467-9280.1997.tb00690.x>
- Seidenberg, M. S., & Zevin, J. D. (2006). Connectionist models in developmental cognitive neuroscience: Critical periods and the paradox of success. *Processes of change in brain and cognitive development: Attention and performance*, 21, 315-347.
- Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, 177, 198–213. <https://doi.org/10.1016/j.cognition.2018.04.011>
- Swingle, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50(1), 86–132. <https://doi.org/10.1016/j.cogpsych.2004.06.001>
- Velan, H., & Frost, R. (2009). transposition effects are not universal: The impact of transposing letters in Hebrew. *Journal of Memory and Language*, 61(3), 285-302.
- Wang, F. H., Hutton, E. A., & Zevin, J. D. (2019). Statistical learning of unfamiliar sounds as trajectories through a perceptual similarity space. *Cognitive Science*, 43(8), e12740.
- Wang, F. H., Zevin, J. D., Trueswell, J. C., & Mintz, T. H. (2020). Top-down grouping affects adjacent dependency learning. *Psychonomic Bulletin & Review*, 27(5), 1052-1058.
- Wang, H.. (2002). How the metrical boundary relates to metrical pattern, syntax, pragmatics in Chinese Mandarin. *Linguistics Collection*, 26, 279-300.
- Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech Segmentation in a Simulated Bilingual Environment: A Challenge for Statistical Learning? *Language Learning and Development*, 5(1), 30–49. <https://doi.org/10.1080/15475440802340101>

Yu, W., Wang, L., Qu X., Wang T., Zhang J., Liang D. (2021). Transitional probabilities and expectation for word length impact verbal statistical learning. *Acta Psychologica Sinica*, 53(6), 565-574.