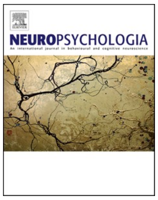
Neuropsychologia 148 (2020) 107640

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia



The influence of linguistic information on cortical tracking of words

Yan Chen [a], Peiqing Jin [a], Nai Ding [a,b,*]

[a] Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, 310027, China

[b] Research Center for Advanced Artificial Intelligence Theory, Zhejiang Lab, Hangzhou, 311121, China

ARTICLE INFO

Keywords:
Cortical tracking
Language
Learning
Speech
Words

ABSTRACT

Speech is a complex sound sequence that has rich acoustic and linguistic structures. Recent studies have suggested that low-frequency cortical activity can track linguistic units in speech, such as words and phrases, on top of low-level acoustic features. Here, with an artificial word learning paradigm, we investigate how different aspects of linguistic information, e.g., phonological, semantic, and orthographic information, modulate cortical tracking of words. Participants are randomly assigned to the experimental group or the control group. Both groups listen to speech streams composed of trisyllabic artificial words or trisyllabic real words. Participants in the experimental group explicitly learn different types of linguistic information of artificial words (phonological, phonological + semantic, or phonological + orthographic information), while participants in the control group do not explicitly learn the words. Electroencephalographic (EEG) data from the control group reveal weaker cortical tracking of artificial words than real words. However, when comparing the experimental and control groups, we find that explicit learning significantly improves neural tracking of artificial words. After explicit learning, cortical tracking of artificial words is comparable to real words, regardless of the training conditions. These results suggest training facilitates neural tracking of words and emphasize the basic role phonological information played in sequential grouping.

Author contribution

Yan Chen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Writing - original draft, Writing - review & editing. Peiqing Jin: Formal analysis, Methodology, Software, Validation, Visualization, Writing - review & editing. Nai Ding: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing - original draft, Writing - review & editing.

1. Introduction

Speech is organized with a hierarchy of units, including phonemes, syllables, words, and larger linguistic structures such as phrases and sentences (Rosen, 1992). Although the primary descriptors of lower-level speech units (such as phonemes and syllables) are acoustic features (Ding et al., 2017b; Mesgarani et al., 2008), words and higher-level linguistic structures are primarily defined by linguistic knowledge (Berwick et al., 2013; Chomsky, 1957). Only through the application of linguistic knowledge, listeners can accurately segment a continuous speech stream into discrete linguistic units. Linguistic knowledge, however, is multidimensional. For example, knowledge about a word normally includes how the word is pronounced, how the word could combine with other words, the meaning of the word, and the written form of the word. During language processing, it has been proposed that separate neural representations are built to encode phonological, syntactic, semantic, and orthographic information (Hagoort, 2005; Jackendoff, 2002), which often co-activate but can have distinct neural implementations (Hagoort and Indefrey, 2014; Rogalsky and Hickok, 2009) and behavioral consequences (Cutler, 2012).

Recent experiments showed that neural activity on different time scales could simultaneously track multiple levels of speech units (Brennan and Hale, 2019; Ding et al., 2016; Keitel et al., 2018; Martin and Doumas, 2017; Meyer et al., 2016). Critically, neural tracking of words and higher-level linguistic units reflects not just encoding of prosodic cues but encoding of linguistic units constructed based on linguistic knowledge: It has been shown that neural tracking of words and higher-level linguistic units remain after the related prosodic cues

[*] Corresponding author. Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, 310027, China.
E-mail address: ding_nai@zju.edu.cn (N. Ding).

https://doi.org/10.1016/j.neuropsychologia.2020.107640
Received 1 May 2020; Received in revised form 8 August 2020; Accepted 28 September 2020
Available online 2 October 2020

are removed (Ding et al., 2018, 2016). Furthermore, when listening to an unknown language, neural tracking of higher-level linguistic units disappears (Ding et al., 2016; Makov et al., 2017).

It remains unclear, however, which aspects of linguistic information, e.g., phonological, syntactic, or semantic, are reflected in neural activity that tracks linguistic units, even for the most basic unit of words. On the one hand, it has been hypothesized that word-tracking neural activity encodes detailed semantic/syntactic information (Broderick et al., 2018; Martin and Doumas, 2017). This hypothesis assumes that neural networks encoding different semantic/syntactic features can be resolved in macroscopic neural activity recorded by MEG or EEG. On the other hand, some studies suggest that word-tracking activity reflects phonological features. Such evidence comes from the finding of brain response which reflects the integration of phonetic information for word identification (Brodbeck et al., 2018).

Here, we investigate how phonological, semantic and orthographic features of words affect neural tracking of multisyllabic words through an artificial word learning experiment. In Experiment 1, participants were trained with three sets of trisyllabic artificial words (n = 5 in each set). In one condition, they only learned the phonological form of artificial words (set 1), while in another 2 conditions, they also learned orthographic or semantic information associated with the artificial words (sets 2 and 3). After learning, participants were exposed to a speech sequence consisting of these artificial words and their neural responses were recorded by EEG. We analyzed whether the EEG response tracking artificial words was modulated by associated phonological, orthographic or semantic information. To be noted, the EEG response during speech listening reflects not only training effect, but also statistical learning of the structured stream and physical property of audio stimuli (Batterink and Paller, 2017; Buiatti et al., 2009; Saffran et al., 1996). Therefore, we conducted Experiment 2 to control for the effects of these confounding factors. Participants in Experiment 2 listened to speech sequence consisting of the artificial words, as was played in Experiment 1. However, there was no training session prior to listening. Both experiments involve a real word condition. EEG responses to Mandarin real words were recorded and compared to the responses to artificial words.

2. Materials and methods

2.1. Participants

Totally, 46 participants were recruited through public announcement at the Zhejiang University. They were randomly assigned to the experimental group (n = 24; 20–28 years old, mean 23 years old; 12 male) or the control group (n = 22; 20–31 years old, mean 23 years old; 11 male). Two additional participants from the control group fell asleep during the experiment and their data were discarded. The experimental group took part in Experiments 1a, 1b, 1c and 1d and control group took part in Experiments 2a and 2b. All subjects were right-handed, monolingual Mandarin speakers and provided written informed consent. This study was approved by the Research Ethics Committee of the College of Medicine, Zhejiang University (2019–047).

2.2. Stimuli

2.2.1. Artificial words

Four sets of artificial words were created for participants to learn, and each set contained 5 trisyllabic artificial words. These artificial words were nonsense words, which were made up with Mandarin syllables and did not involve any semantic meaning. To construct the 5 artificial words, we first selected 5 real trisyllabic Chinese Mandarin words (e.g., "wú huā guǒ" for fig and "pí jiá kè" for jacket). In the following, the initial, middle, final syllables of a trisyllabic word were referred to as σ_1, σ_2, and σ_3 respectively. Since 4 sets of artificial words were created, 4 sets of real words were selected, which were all nouns and each syllable only appeared in one word. Within each set of words, the syllables at the same position, e.g., σ_1 of all words, were controlled so that their initial phonemes, final phonemes and tones were as diverse as possible. A set of artificial words was constructed by shuffling the syllables at the same position, e.g., σ_1 of all words. The resultant 5 artificial words had the same set of initial/middle/final syllables as the original 5 real words.

2.2.2. Speech synthesis

The sound of each artificial word or real word was created by concatenating syllables that were independently synthesized using the Neospeech synthesizer (http://www.neospeech.com/, the male voice, Liang). Each syllable was adjusted to 250 ms in duration and no acoustic gaps were inserted between the syllables.

2.2.3. Speech sequences

In the experiments, artificial words or real words were presented in sequences. In each of the four conditions of Experiment 1, a sequence of 2400 syllables (corresponding to 800 artificial words or real words) were presented at a rate of 250 ms per syllable, i.e. 4 Hz. The same artificial word/real word was not allowed to repeat immediately. We added 7–8 random syllables to the beginning and the end of the speech sequence, so that the listeners could not simply grouped every 3 syllables into a chunk from the beginning of the speech sequence but instead have to rely on their memory to find the word onsets. The random syllables at the beginning and the end of a sequence were randomly drawn from syllables in the artificial words/real words. The total duration of the speech sequence in each condition of Experiment 1 was 10 min and 3.75 s. Ten syllables in each speech sequence were manipulated to sound like a different voice (formant shift ratio = 1.5). In Experiment 2, the real word sequence was the same as the real word sequence in Experiment 1. The artificial word sequence, however, was 4 times longer. It contained 9600 syllables (corresponding to 3200 artificial words) and lasted 40 min and 3.75 s. Ten syllables in the real word sequence and forty syllables in the artificial word sequences were with a different voice (formant shift ratio = 1.5).

2.3. Experimental design

The procedures of the experiments are shown in Fig. 1A. All participants were tested in a silent room, wearing earphones. Each experiment was run in a separate session and EEG was recorded throughout the experiments (see 2.4. EEG Recording section). In Experiment 1, the 4 sets of artificial words were randomly assigned to the Experiment 1a, 1b, 1c and 1d, and the order of these four conditions was counterbalanced over subjects. In the real word condition, i.e., Experiment 1d, the artificial words were replaced with the real words used to generate the artificial words. In Experiment 2, one set of artificial words was randomly assigned to one participant in the control condition (2a), and the real word condition contained the real words used to generate these artificial words (2b).

2.3.1. Experiment 1a: phonological condition

During a training session, participants learned the phonological form of a set of artificial words. Each time an artificial word was auditorily presented once and at the same time the spelling, i.e., the *pinyin*, was shown on the screen, which both lasted 750 ms (Fig. 1B). Two successive artificial words were separated by a 250-ms silence and black screen. Sixty repetitions for each of the five artificial words yielded a total of 300 repetitions, resulting in a 5-min training session. Participants were instructed to remember the pronunciation the artificial words and prepare to take a test afterwards.

A two-alternative forced-choice (2AFC) test was then conducted to assess phonological learning. In each of the 5 trials, participants heard an artificial word they learned (e.g., shēn luò tái) and a part-word foil which consisted of a syllable pair from the artificial word plus an

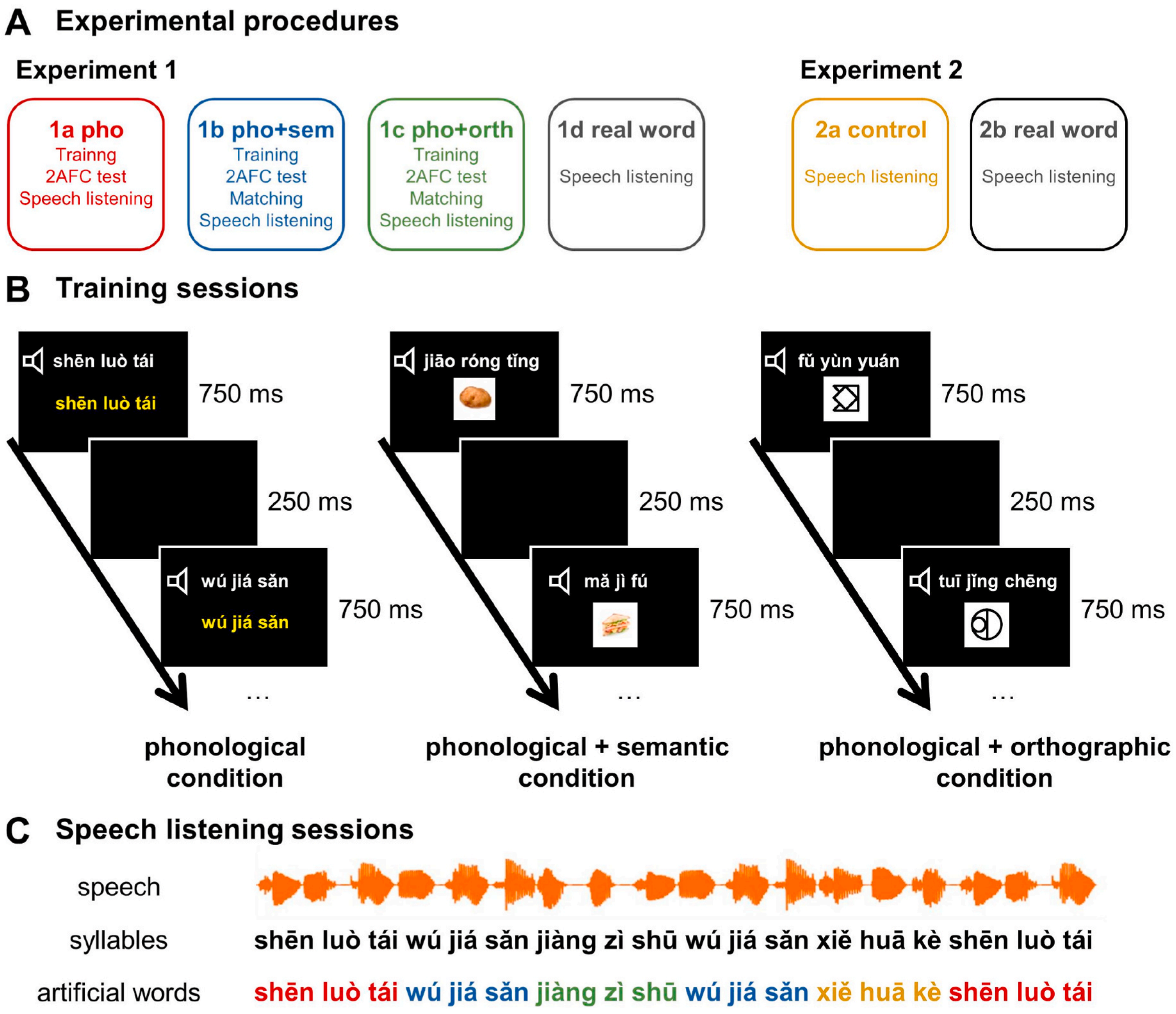


Fig. 1. Experimental design. (**A**) Experimental procedures of the two experiments. The full names of Experiments 1a, 1b and 1c are given in Fig. 1B. 2AFC test = two-alternative forced-choice test. (**B**) Training sessions for three linguistic learning conditions in Experiment 1. Each artificial word was auditorily presented and after a pause another artificial word was presented. During the presentation of each word, the participants could see the spelling of the word (*pinyin*), a picture that represents the meaning of the word, or a symbol that represents the writing form of the word. (**C**) Speech listening sessions. The stimulus consisted of syllables presented at a constant rate of 4 Hz. Every 3 syllables constructed an artificial word. Without learning, participants did not know which combinations of syllables constituted artificial words. After learning, however, the brain could segment the syllable sequence into artificial words.

additional syllable (e.g., luò tái wú). The artificial word and part-word foil were presented in a random order and separated by a 500 ms silence. Participants were asked to choose which word was learned by a button press. The next trial would begin after the participant gave a response on one trial.

To test the neural tracking of artificial words after phonological learning, participants listened to a 10 min 3.75 s-speech sequence composed by the 5 artificial words they just learned in the training session (see 2.2. Stimuli and Fig. 1C). Participants were instructed to perform a target detection task during listening, i.e., they had to press the key if they detected any syllable with a higher pitch than other syllables. Responses that did not occur within 1.25 s from pitch deviant onset were considered as false alarms.

2.3.2. *Experiment 1b: phonological + semantic condition*

In the training session, each artificial word was paired with a picture, e.g. a picture of potato or sandwich. Each time an artificial word was auditorily presented once and at the same time the picture was shown on the screen, which both lasted 750 ms (Fig. 1B). After training, participants received the 2AFC test described in Experiment 1a. Then, they did another test, in which they needed to match the words with corresponding pictures. In the test, the five words and five pictures were shown in two rows on a piece of paper and participants had to connect the associated pairs. Other procedures were the same as Experiment 1a.

2.3.3. *Experiment 1c: phonological + orthographic condition*

In the training session, each artificial word was paired with a meaningless symbol (Fig. 1B; Song et al., 2010), instead of the pictures in Experiment 1b. Other procedures were the same as Experiment 1b.

2.3.4. *Experiment 1d: real word condition*

Participants listened to the real word sequences described in the 2.2. Stimuli section, and had to respond to syllables with a higher pitch by a key press.

2.3.5. *Experiment 2a: control condition*

Experiment 2 did not involve any training session, and the participants were not aware that each artificial word had 3 syllables. In the artificial word sequence, the transitional probability between neighboring syllables was clearly 1 within each artificial word. Since there were 5 artificial words and each artificial word would not immediately repeat, the transitional probability between neighboring syllables was 0.25 across artificial word boundaries.

Before the experiment, the participants were instructed to listen to a speech of a “new language” they’ve never heard before for 40 min and press the key if they detect any syllable with a higher pitch than other syllables. Responses that did not occur within 1.25 s from a target onset were considered to be false alarms. After the 40 min of exposure, participants completed the 2AFC test to distinguish artificial words and part-word foils, like in Experiments 1a. Participants were instructed to indicate which syllable sequence sounded more like what they heard during the 40-min exposure.

2.3.6. *Experiment 2b: real word condition*

This condition was identical to Experiment 1d.

2.4. *EEG recording*

EEG responses were continuously recorded using a 64-channel Biosemi ActiveTwo system. Four additional electrodes placed at the outer canthi of both eyes and above and below the right eye were used to record horizontal and vertical EOGs. Two electrodes placed on the left and right mastoid were used as the reference. The EEG recordings were low-pass filtered below 400 Hz and sampled at 2048 Hz (default in Biosemi ActiveTwo system). Since the study focused on word-rate and syllable-rate neural responses (1.33 Hz and 4 Hz respectively), the EEG recordings were high-pass filtered above 0.5 Hz with a linear-phase finite impulse response (FIR) filter (10 s Hamming window). To remove EOG artifacts in EEG, the horizontal and vertical EOG signals were regressed out. The EEG recordings were referenced to the averaged mastoid recording. The EEG signals were resampled to 32 Hz using the *resample* function of MATLAB, in which an FIR anti-aliasing low-pass filter was applied to prevent aliasing in the lower frequencies.

2.5. *Frequency-domain analysis*

During speech listening sessions, words and syllables were presented at constant rates and the neural tracking of these linguistic structures was analyzed in the frequency domain. To avoid onset/offset effects and focus on steady-state response, neural activity for the random syllables at both ends of each session was not analyzed.

The 40-min control condition was divided into four 10-min blocks (0–10 min, 10–20 min, 20–30 min, 30–40 min). Then, the neural recordings during each block of the control condition and neural recordings of other conditions in Experiments 1 and 2 were all segmented into 50 12-s epochs, with each epoch corresponding to duration of 16 trisyllabic words. Finally, for each control block and other condition, the neural responses were averaged over all epochs and transformed into the frequency domain using the Discrete Fourier transform (DFT). The response amplitude at the word frequency (1.33 Hz) and the syllable frequency (4 Hz) were extracted.

As has been stated by the previous research, the word identification component, indexed by the neural tracking at the word frequency relative to that at the syllable frequency, could be an indicator of learning efficiency (Batterink and Paller, 2017). In other words, if participants learned one specific set of words more efficiently, they would show relative more preference in the concurrent tracking of underlying words relative to individual syllables. The word identification component was anticipated to reach the highest in the real word condition, in which the words were supposed to be sufficiently learned on various aspects before the experiment. Within each condition, we quantified the word identification component by *Word Learning Index* (WLI) using the following formula:

$$\mathrm{WLI} = \frac{\text{Response amplitude}_{\text{word frequency}}}{\text{Response amplitude}_{\text{syllable frequency}}}$$

The WLI was computed across 8 electrodes where response amplitude at the word and syllable frequencies showed the strongest values (Fz, F1, F2, F3, F4, FCz, FC1, FC2). The EEG spectrum was averaged over these 8 channels and subjects (Figs. 3 and 4). The response topography showed the amplitude of the DFT coefficients at the word and the syllable frequency (Figs. 3 and 4).

2.6. *Statistical analysis*

For behavioral data, percentage scores of the 2AFC test were computed for each participant. Statistical significance was tested using one-sample *t*-test against the chance level (50%).

For spectral peaks (Figs. 3 and 4), a paired *t*-test was used to test if the neural response in a frequency bin was significantly stronger than the average of the neighboring four frequency bins (two bins on each side). Such a test was applied to the word frequency, the syllable frequency, and the frequency corresponding to the second harmonic of the word, and a threshold of $q < 0.05$ with false discovery rate (FDR) was adopted to correct for multiple comparisons.

For Experiments 1 and 2, we tested whether there was an effect of training. We tested the difference of WLI between 1) each linguistic learning condition, 2) the average of three linguistic learning conditions, with 1) the average of four blocks and 2) the first block of control condition. In order to correct for differences across participants, the WLI of each participant was corrected by subtracting his or her WLI in the real word condition. Independent-sample *t*-tests were performed.

For Experiment 1, we tested whether there was a difference of the WLI between different linguistic learning conditions and the real word condition. A repeated-measures ANOVA was performed with condition (phonological condition, phonological + semantic condition, phonological + orthographic condition, and real word condition) as a within-subject factor. Planned comparisons were conducted between conditions using paired *t*-tests.

For Experiment 2, we tested whether there was a difference of the WLI in different control blocks and the real word condition. Four control blocks and the real word condition were regarded as five conditions. We computed the WLI within each condition. A repeated-measures ANOVA was conducted with condition as a within-subject factor. Planned comparisons were conducted between conditions using paired *t*-tests.

3. Results

3.1. *Behavioral performance of participants*

The behavioral performance of the participants is presented in Table 1. The accuracies (>95.5%) and the false alarms (<1.5) of the target detection tasks during trisyllabic words listening reflect the wakefulness of the participants. The experimental group performed

Table 1
Behavioral performance.

	Target detection		2AFC test	Matching
	Accuracy	False alarm	Accuracy	Accuracy
Experimental group				
Phonological condition	99.2% (2.8%)	0.1 (0.3)	98.3% (5.6%)	–
Phonological + semantic condition	99.6% (2.1%)	0.0 (0.2)	98.3% (8.2%)	100.0% (0.0%)
Phonological + orthographic condition	99.2% (2.8%)	0.3 (0.9)	99.2% (4.1%)	90.0% (17.7%)
Real word condition	98.8% (6.1%)	0.1 (0.3)	–	–
Control group				
Control condition	95.5% (7.0%)	1.5 (3.6)	65.5% (20.6%)	–
Real word condition	99.5% (2.1%)	0.0 (0.2)	–	–

Note: 2AFC = two-alternative forced-choice. The values in table are mean (standard deviation).

almost at ceiling on the 2AFC test (>98.3%; assessing phonological learning) and matching tasks (>90.0%; assessing semantic and orthographic learning). The accuracy of 2AFC test for the control group was significantly above chance level (65.5%; $t = 3.5$, $p < 0.01$), indicative of statistical learning process during speech listening. The experimental group obtained significantly higher scores of 2AFC test than the control group (98.6% vs. 65.5%; $t = -7.4$, $p < 0.001$). This suggests that participants in the experimental group have acquired stronger the word-level representations after training.

3.2. Effect of training on cortical tracking of words

The effect of training on cortical tracking of words was examined. For Experiments 1 and 2, we tested the difference of Word Learning Index (WLI) between 1) each linguistic learning condition, 2) the average of three linguistic learning conditions, with 1) the average of four blocks and 2) the first block of control condition. The WLI was computed by the response amplitude at the word frequency relative to that at the syllable frequency. The WLI of each subject was corrected by subtracting his or her WLI of the real word condition to correct for individual difference. Independent-sample *t*-tests revealed higher WLI in the phonological condition, phonological + semantic condition, phonological + orthographic condition, and the average of them than WLI of the average of four blocks of control condition (*t*-values > 2.06, *p*-values < 0.05; Fig. 2A). The WLI of the phonological condition and the average of three linguistic learning conditions were significant higher than the WLI of the first block of control condition ($t = 2.67$, $p = 0.01$; $t = 2.40$, $p = 0.03$; Fig. 2B). The phonological + semantic condition and the phonological + orthographic condition have marginally significantly higher or a trend of higher WLI than the first block of statistical learning ($t = 1.99$, $p = 0.053$; $t = 1.50$, $p = 0.14$; Fig. 2B).

3.3. Effect of linguistic learning on neural tracking of words

In Experiment 1, we sought to determine the effects of linguistic knowledge (i.e. phonological, semantic and orthographic features) on neural tracking of trisyllabic words. Fig. 3A showed the EEG response spectrums in four conditions (phonological, phonological + semantic, phonological + orthographic, and real word conditions) as a function of frequency. As expected, the response spectrums exhibited clear peaks at the syllabic rate (4 Hz: *p*-values $< 3.6 \times 10^{-5}$, paired *t*-test, FDR corrected $q < 0.05$), word rate (1.33 Hz: *p*-values $< 7.6 \times 10^{-5}$, paired *t*-test, FDR corrected $q < 0.05$), as well as the rate corresponding to the second harmonic of the word (2.67 Hz: *p*-values $< 1.8 \times 10^{-4}$, paired *t*-test, FDR corrected $q < 0.05$). The Word Learning Index (WLI), computed by the response amplitude at the word frequency relative to that at the syllable frequency, was not significantly different among conditions ($F = 1.9$, $p = 0.15$, one-way repeated-measures ANOVA; Fig. 3B). Planned comparisons showed that the WLI was not significantly different between real word and each linguistic learning conditions (0.1 < *p*-values < 0.8, paired *t*-test) and was not significantly different among the explicit learning conditions (0.1 < *p*-values < 0.3, paired *t*-test). To be noted, a failure to find the difference between the conditions is not positive evidence that there is no difference between them. To test the hypothesis that there was no difference between the conditions, we used a Bayesian approach to compute the odds on the null with SPSS 25.0. Briefly, we specified both a null and alternative prior distribution of WLI and computed posterior probabilities for each hypothesis. The Bayes factor was calculated as the ratio of the likelihoods of two hypotheses. If the Bayes factor is bigger than 1, we find evidence for the null. If the Bayes factor is smaller than 1, we find evidence against the null. In our data, The Bayes factors for the null were 1.5–6.3 for paired *t*-tests between every two conditions, indicating that there was no difference of WLI between phonological, phonological + semantic, phonological + orthographic, and real word conditions. These results suggest that

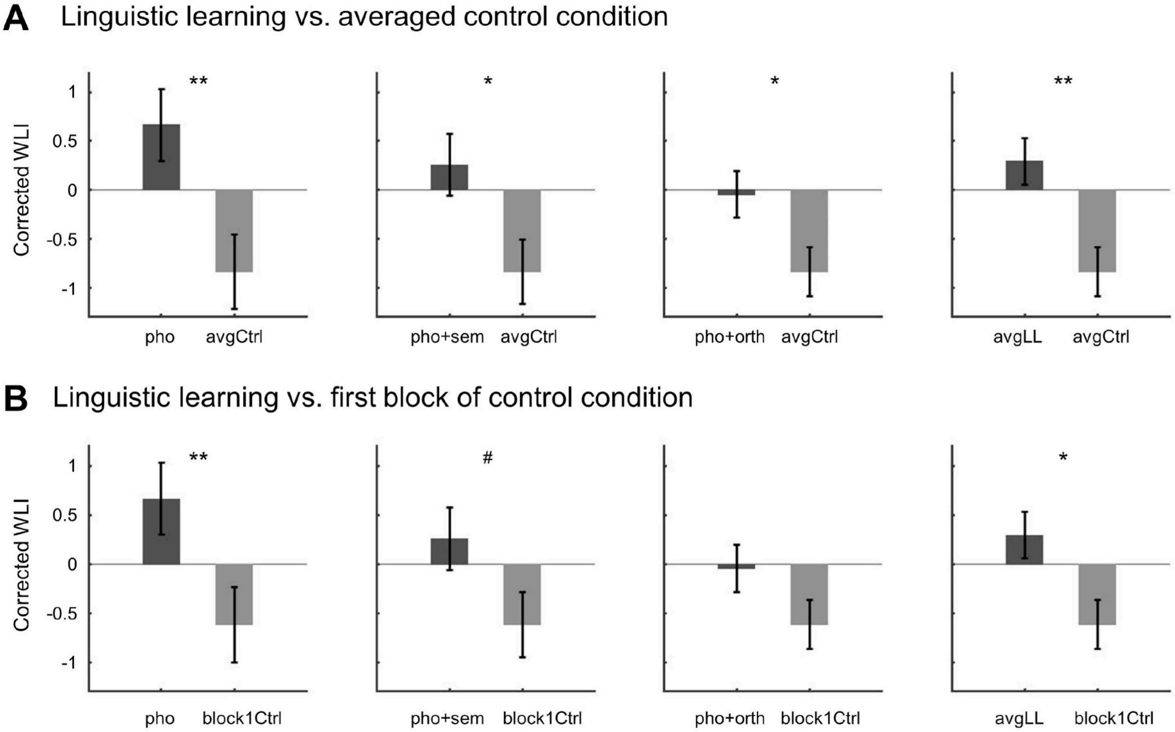


Fig. 2. The comparisons of corrected Word Learning Index (WLI) between linguistic learning conditions (**A**) with the average of four blocks and (**B**) with the first block of control condition. Pho = phonological condition; pho + sem = phonological + semantic condition; pho + orth = phonological + orthographic condition; avgLL = the average of three linguistic learning conditions; avgCtrl = the average of four blocks of control condition; block1Ctrl = the first block of control condition. #*p* < 0.10, **p* < 0.05, ***p* < 0.01.

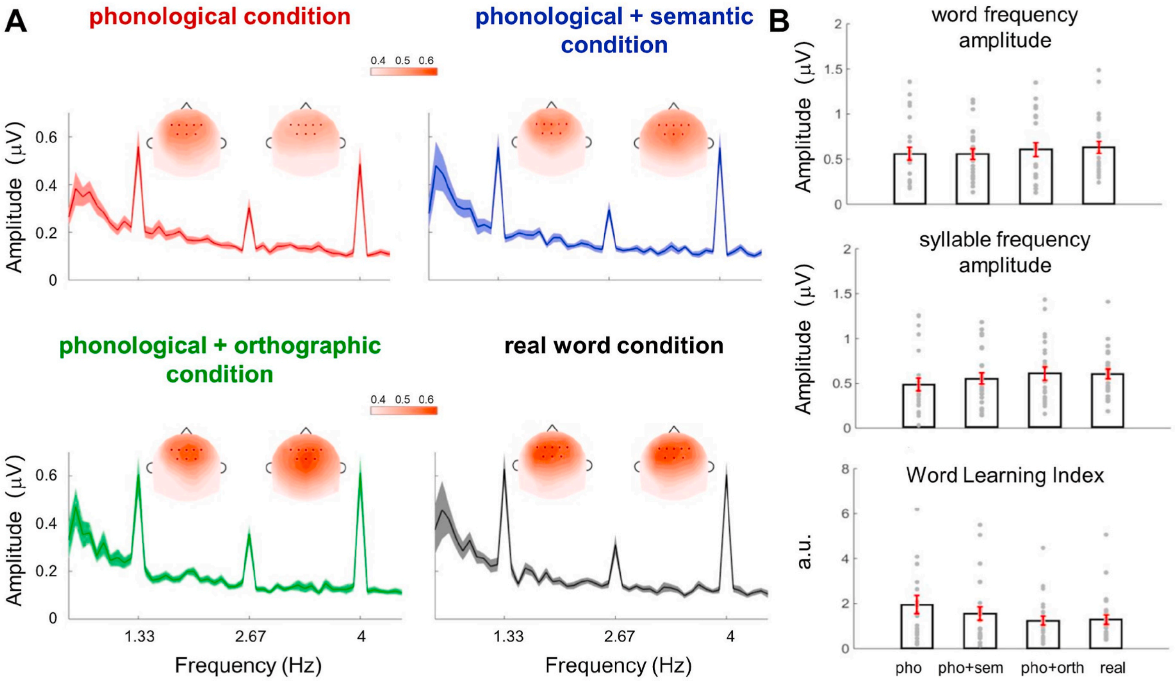


Fig. 3. EEG responses to words in the experimental group. (**A**) Spectral peaks are observed at word and syllable rates (1.33 Hz and 4 Hz respectively) in all 4 conditions. In the spectrum, the shaded area covers 2 standard errors across participants (standard error = standard deviation/$\sqrt{n}$; n, number of participants). The topographical distribution of response is shown for the responses at word and syllable rates. The eight black dots on the topography denote the locations of the eight electrodes used for the spectrum analysis. (**B**) Response amplitude at the word and syllable frequencies and the Word Learning Index. The Word Learning Index is the ratio between word- and syllable-frequency responses. Error bars represent standard error and each gray dot represents data from 1 participant. The full names of the conditions are given in Fig. 3A. No significant difference was observed between conditions, suggesting that phonological learning is sufficient to drive word-tracking responses.

learning of phonological information alone is sufficient to elicit neural tracking of artificial words, and the amplitude of this response is comparable to the neural response tracking real words. Therefore, the neural tracking of words might primarily reflect the segmentation of a continuous speech stream into phonological units.

3.4. *Neural tracking of words in the control condition*

In Experiment 2, participants listened to sequence of artificial without training. We explored the neural tracking of artificial words in the control condition. The neural tracking reflects both statistical learning and the physical property of audio stimuli (**Supplementary material**). Fig. 4A shows the EEG response of five conditions: four 10-min control blocks and real word condition. For all the conditions, there were significant peaks at the word (p-values $< 1.9 \times 10^{-3}$, paired t-test, FDR corrected $q < 0.05$) and syllable frequencies (p-values $< 3.6 \times 10^{-5}$, paired t-test, FDR corrected $q < 0.05$), as well as the frequency corresponding to the second harmonic of the word (p-values $< 4.3 \times 10^{-3}$, paired t-test, FDR corrected $q < 0.05$). There was a main effect of conditions ($F = 4.0$, $p = 0.04$, one-way repeated-measures ANOVA) on the WLI. Further comparisons showed the WLI of real word was significantly larger than four blocks in the control condition (p-values < 0.05, paired t-test; Fig. 4C). WLI was not significantly different among the four control blocks ($0.1 < p$-values < 0.7, paired t-test).

4. Discussion

When listening to speech, cortical activity can track the rhythm of words. The current study investigates how the word-tracking response is modulated by learning and reflects different aspects of linguistic processing. We revealed significantly higher word-tracking response in the experimental group than the control group. Moreover, we found the neural tracking of learned artificial words was comparable to real words in the experimental group, regardless of learning conditions (phonological, phonological + semantic, phonological + orthographic conditions). Higher word-tracking response in the experimental group than the control group reflects a facilitation of training in the segmentation of continuous speech streams into discrete units. Comparable word-rate response of phonological condition with the other conditions (phonological + semantic, phonological + orthographic, real word conditions) emphasizes the primary role of phonological information in the neural tracking of words.

In the last decade, neural oscillation has been recognized to play an important role in auditory coding and speech coding at phoneme or syllable level (Di Liberto et al., 2019, 2015; Giraud and Poeppel, 2012; Peelle et al., 2013; Teoh et al., 2019). Researchers have found, when listening to speech, cortical activity tracks the temporal envelope of speech (Kerlin et al., 2010; Lalor and Foxe, 2010; Luo and Poeppel, 2007), which carries the acoustic rhythm of speech. Neural tracking of the speech envelope occurs both for native language, unfamiliar language, and non-speech (Ding et al., 2016; Lalor and Foxe, 2010; Zou et al., 2019) and therefore to some extent reflects auditory encoding. Recent studies, however, have demonstrated neural oscillations could track the rhythms of linguistic units, e.g., words (Buiatti et al., 2009; Ding et al., 2017a, 2016; Farthouat et al., 2017; Getz et al., 2018) in the absence of relevant acoustic cues.

The current study used only 5-min training of artificial words, and achieved a significant word-tracking response. Results show significantly stronger word-tracking response for learning conditions than the control condition, confirming the importance of learning/training on sequential grouping (Fig. 2). Previous ERP studies have also observed an effect of learning on speech segmentation. For instance, Sanders et al. (2002) recorded ERPs while participants listened to continuous speech consisting of nonwords and compared the responses to nonword onsets before and after they explicitly learned the nonwords. They found that word onsets elicited a larger N100 after than before training. Moreover,

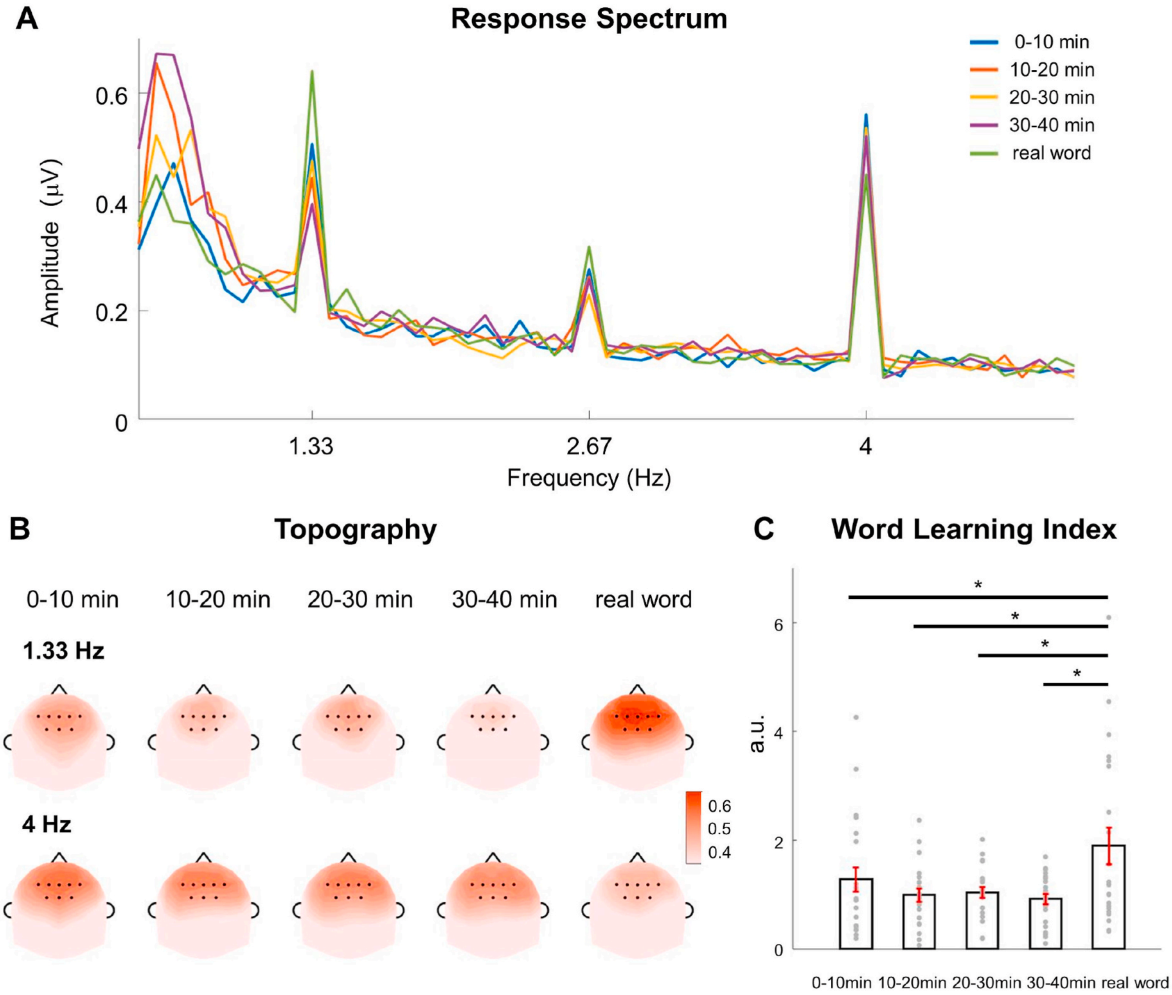


Fig. 4. EEG responses to words in the control group. (**A**) The EEG spectrum is separately shown for four 10-min blocks in the control condition. The EEG spectrum in the real word condition is also shown. (**B**) Response topography. (**C**) Word Learning Index. Error bars represent standard error and each gray dot represents data from 1 participant (standard error = standard deviation/$\sqrt{n}$; n, number of participants). $*p < 0.05$.

the N100 word-onset effect was also observed in subjects listening to real words in their native language. In addition to N100, N400 component was also observed in studies of speech learning. It was observed that word onset elicited larger N400 amplitudes in high learners than low learners (Abla and Okanoya, 2009; Sanders et al., 2002). Here, we explored the temporal dynamics of the EEG response in the current study. As illustrated in Fig. 5B, the real word condition differed from artificial words of the first block of control condition in the 300–500 ms range, showing a significant larger N400 effect (independent-sample *t*-test, cluster-based permutation test, $p < 0.05$). From visual inspection, there were also trend of larger N400 effect in three explicit learning conditions compared with first block of control condition. The N400 component is a robust ERP index of speech segmentation (Cunillera et al., 2009; De Diego Balaguer et al., 2007). It might be one of the bases for the word-rate response in the frequency domain analyses. The higher amplitude near the peak for the second syllable in two explicit learning conditions versus the first block of control condition (independent-sample *t*-test, cluster-based permutation test, $p <$ 0.05) might indicate attention to high probability transitions—perhaps in anticipation of the completion of an intact word. Both the ERP results and the tracking results in the current study provide evidence for the importance of learning/training on the grouping of syllables into words.

The current study extends previous studies by showing that neural tracking of word primarily reflects the segmentation of a continuous speech stream into phonological unit. The word-rate response is comparable in the phonological, phonological + semantic, and phonological + orthographic learning conditions. In all 3 conditions, the participants learned the phonological form of artificial words but in the latter 2 conditions participants additionally learned semantic and orthographic information. Cortical tracking of artificial words with phonological training is comparable to the latter 2 conditions and comparable to neural tracking of real words (Fig. 3). These results suggest that the word-rate response mainly reflects neural encoding of phonological words. It is likely that the word-tracking EEG response primarily reflects the encoding of word boundaries.

Notably, the modulation effect of semantic or orthographic information is not clearly revealed by the current results. However, such effect has been discovered in previous studies on speech processing (Broderick et al., 2020, 2018). In the current study, the stimulus is a sequence of unrelated words and the task is to detect a voice change. It is clear that the stimulus and task do not motivate semantic or orthographic processing. Nevertheless, the task does not encourage the participants to segment speech into phonological words either, since the voice change can be detected without lexical segmentation. In fact, since the voice change is applied to a random syllable, the task cannot be facilitated by lexical segmentation. It is possible that the segmentation of speech into phonological words is a process that occurs more automatically than semantic/orthographic processing. Studies have found semantic processing can occur for words that presented at the attended location and the N400 component of the event-related potentials (ERPs)

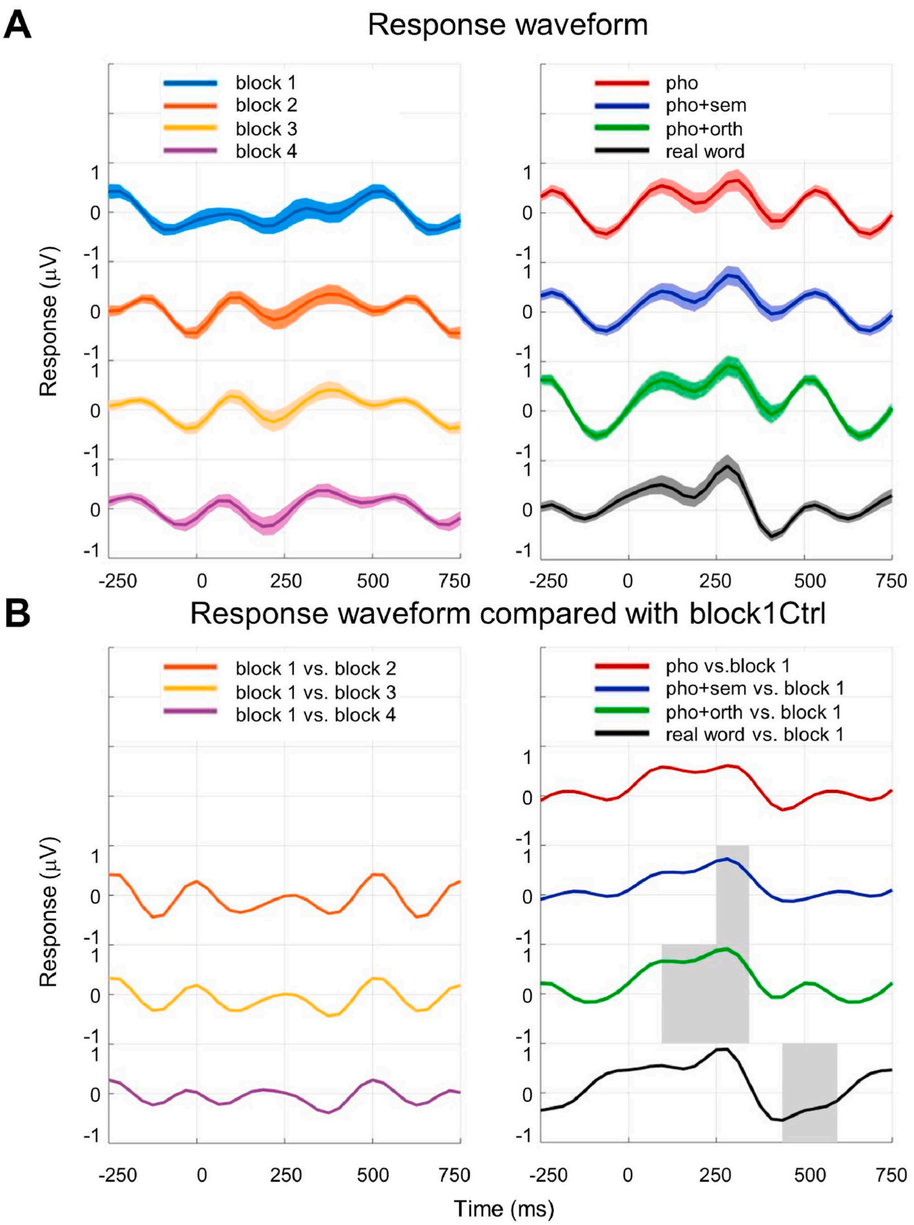


Fig. 5. Temporal dynamics of the EEG response. (A) Original EEG response of Experiment 2 and Experiment 1. Real word or artificial word onset is shown by 0 ms. Left: response time course for the four blocks of control condition. Right: response time course for three linguistic learning conditions and the real word condition. (B) EEG response as compared with the first block of control condition. Left: comparison between the block 1 with block 2, 3, 4 of control condition. Right: comparison between each condition with the block 1 of control condition. Significant differences are marked by gray bars (independent-sample *t*-test, cluster-based permutation test, $p < 0.05$). Block1Ctrl = the first block of control condition.

is elicited by attended words (Bentin et al., 1995; Luck et al., 1996; Naccache et al., 2002; Nobre and McCarthy, 1995). However, the N400 response disappears when the words are unattended (Bentin et al., 1995; Nobre and McCarthy, 1995). Broderick et al. (2018) conducted an electroencephalographic study in which they found that when comprehending natural speech, the brain responds to the contextual semantic content of each word in a time-locked fashion. However, the semantic responses only emerged for attended but not for unattended speech. These studies indicate that semantic processing of words is attention and task dependent. Therefore, if the task motivates semantic or orthographic processing, it is possible such processes will modulate the word-tracking response. A second reason for the null effect between the conditions might be due to the lack of higher-level processing. Previous studies indicated that words might be tracked more strongly when they can be integrated into phrases; phrases might be tracked more strongly when they can be integrated into sentences (Kaufeld et al., 2020; Keitel et al., 2018; Martin, 2020, 2016). However, the current study only focused on word and syllable level processing and did not encourage other higher-level processing, such as processing of phrases and sentences. This might result in the absence of difference between the linguistic learning conditions. The third reason for the null effect of semantic and orthographic knowledge might be due to the relatively short time of training. Given that each training session in our study only took 5 min, the learning might not be sufficient. The neural tracking could possibly be modulated by semantic and orthographic information for longer training and/or more consolidated knowledge at these levels.

In the control condition, the word-rate response emerged within the first 10 min, peaking at approximately 5–7 min of exposure, and plateaued after that (**Supplementary material**). This is consistent with previous evidence for rapid statistical learning (Farthouat et al., 2017; Getz et al., 2018; Saffran et al., 1996). For example, a previous study on statistical learning of using non-linguistic tone sequences as material observed tritone frequency-related responses after 3–5 min of exposure (Farthouat et al., 2017). These suggest the statistical learning is very rapid and is a remarkable ability of humans to learn about structured temporal patterns existing in the environment (Abla et al., 2008; Saffran et al., 1999). Prior studies have also investigated the neural correlates of statistical learning. For instance, Karuza et al. (2013) conducted a functional magnetic resonance imaging study in which they found significant activation in the pars opercularis and pars triangularis regions of the left inferior frontal gyrus (LIFG) during statistical learning. These findings suggest that the LIFG may mediate statistical learning and are involved in extracting temporally ordered patter information in speech processing (Abla et al., 2008; Turk-Browne et al., 2009). Dale et al. (2012) used a sequential learning task to promote predictive behaviors in participants as they responded to a sequence of stimulus events along a continuum of regularity. They found that explicit awareness measures correlated strongly with predictive behavior. Participants reported low awareness of sequence pattern also showed modulated reaction times relative to the regularity of the structure they received. These are suggestive of both explicit and implicit learning in the exposure. Therefore, the authors considered a "two system" hypothesis related to statistical learning: When implicit learning system extracts sufficient structure, the brain can seek out a forthcoming stimulus, thus producing an error signal and then the explicit awareness and learning system kicks in.

The current study has some caveats. First, in Experiment 1b, we trained the participants to associate the artificial words with semantic concepts, by showing them the pictures of the objects. The problem is that when the pictures of the objects elicit semantic representation, it might elicit the name of the object as well. Second, it is worth noting that the stimuli in phonological + orthographic condition only carried visual symbolic characteristics and didn't include enough orthographic information. Further work will be necessary to use words from unknown languages or artificial words which have orthographic information as training materials.

In sum, using explicit learning paradigms, we tested the difference of neural tracking between explicit learning and control condition and investigated the effects of linguistic knowledge (phonology, orthography and semantic) on the cortical tracking of lexical units. The results indicate that three linguistic learning conditions yield comparable cortical tracking of learned words, all of which also achieve similar effect with the real word condition. These results indicate that neural tracking of word primarily reflects the segmentation of a continuous speech stream into phonological units. For the control condition, though neural tracking to artificial words can also be observed, it is much weaker than the real word and explicitly learned words, which suggests explicit learning of words is crucial for sequential grouping. The study emphasizes the critical role of explicit learning and phonological information on word-tracking, deepening our understanding of the neural processing of lexical units in the human brain.

Declaration of competing interest

None.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number 31771248]; the Major Scientific Research Project of Zhejiang Lab [grant number 2019KB0AC02]; and the Fundamental Research Funds for the Central Universities. We would like to thank Jianfeng Zhang for comments on the manuscript and Naifei Su for help with imaging analyses. We are also grateful to all research participants for their patience.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2020.107640.

References

Abla, D., Katahira, K., Okanoya, K., 2008. On-line assessment of statistical learning by event-related potentials. J. Cognit. Neurosci. 20, 952–964. https://doi.org/10.1162/jocn.2008.20058.

Abla, D., Okanoya, K., 2009. Visual statistical learning of shape sequences: an ERP study. Neurosci. Res. 64, 185–190. https://doi.org/10.1016/j.neures.2009.02.013.

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

Bentin, S., Kutas, M., Hillyard, S.A., 1995. Semantic processing and memory for attended and unattended words in dichotic listening: behavioral and electrophysiological evidence. J. Exp. Psychol. Hum. Percept. Perform. 21, 54–67. https://doi.org/10.1037/0096-1523.21.1.54.

Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J.J., 2013. Evolution, brain, and the nature of language. Trends Cognit. Sci. 17, 89–98. https://doi.org/10.1016/j.tics.2012.12.002.

Brennan, J.R., Hale, J.T., 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. PloS One 14 (1), e0207741. https://doi.org/10.1371/journal.pone.0207741.

Brodbeck, C., Hong, L.E., Simon, J.Z., 2018. Rapid transformation from auditory to linguistic representations of continuous speech. Curr. Biol. 28, 3976–3983. https://doi.org/10.1016/j.cub.2018.10.042.

Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biol. 28, 803–809. https://doi.org/10.1016/j.cub.2018.01.080.

Broderick, M.P., Di Liberto, G.M., Anderson, A.J., Rofes, A., Lalor, E.C., 2020. Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. bioRxiv. https://doi.org/10.1101/2020.04.17.046201, 2020.04.17.046201.

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

Chomsky, N., 1957. Syntactic Structures. Mouton de Gruyter, The Hague.

Cunillera, T., Càmara, E., Toro, J.M., Marco-Pallares, J., Sebastián-Galles, N., Ortiz, H., Pujol, J., Rodríguez-Fornells, A., 2009. Time course and functional neuroanatomy of speech segmentation in adults. Neuroimage 48, 541–553. https://doi.org/10.1016/j.neuroimage.2009.06.069.

Cutler, A., 2012. Native Listening: Language Experience and the Recognition of Spoken Words. MIT Press, Cambridge.

Dale, R., Duran, N.D., Morehead, J.R., 2012. Prediction during statistical learning, and implications for the implicit/explicit divide. Adv. Cognit. Psychol. 8, 196–209. https://doi.org/10.2478/v10053-008-0115-z.

De Diego Balaguer, R., Toro, J.M., Rodriguez-Fornells, A., Bachoud-Lévi, A., 2007. Different neurophysiological mechanisms underlying word and rule extraction from speech. PloS One 2, e1175. https://doi.org/10.1371/journal.pone.0001175.

Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C., 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. 25, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030.

Di Liberto, G.M., Wong, D., Melnik, G.A., de Cheveigné, A., 2019. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage 196, 237–247. https://doi.org/10.1016/j.neuroimage.2019.04.037.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., Poeppel, D., 2017a. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front. Hum. Neurosci. 11, 1–9. https://doi.org/10.3389/fnhum.2017.00481.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164. https://doi.org/10.1038/nn.4186.

Ding, N., Pan, X., Luo, C., Su, N., Zhang, W., Zhang, J., 2018. Attention is required for knowledge-based sequential grouping: insights from the integration of syllables into words. J. Neurosci. 38, 1178–1188. https://doi.org/10.1523/JNEUROSCI.2606-17.2017.

Ding, N., Patel, A.D., Chen, L., Butler, H., Luo, C., Poeppel, D., 2017b. Temporal modulations in speech and music. Neurosci. Biobehav. Rev. 81, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011.

Farthouat, J., Franco, A., Mary, A., Delpouve, J., Wens, V., Op de Beeck, M., De Tiège, X., Peigneux, P., 2017. Auditory magnetoencephalographic frequency-tagged responses mirror the ongoing segmentation processes underlying statistical learning. Brain Topogr. 30, 220–232. https://doi.org/10.1007/s10548-016-0518-y.

Getz, H., Ding, N., Newport, E.L., Poeppel, D., 2018. Cortical tracking of constituent structure in language acquisition. Cognition 181, 135–140. https://doi.org/10.1016/j.cognition.2018.08.019.

Giraud, A.L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517. https://doi.org/10.1038/nn.3063.

Hagoort, P., 2005. On Broca, brain, and binding: a new framework. Trends Cognit. Sci. 9, 416–423. https://doi.org/10.1016/j.tics.2005.07.004.

Hagoort, P., Indefrey, P., 2014. The neurobiology of language beyond single words. Annu. Rev. Neurosci. 37, 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847.

Jackendoff, R., 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press, Oxford.

Karuza, E.A., Newport, E.L., Aslin, R.N., Starling, S.J., Tivarus, M.E., Bavelier, D., 2013. The neural correlates of statistical learning in a word segmentation task: an fMRI study. Brain Lang. 127, 46–54. https://doi.org/10.1016/j.bandl.2012.11.007.

Kaufeld, G., Bosker, H.R., Alday, P.M., Meyer, A.S., Martin, A.E., 2020. Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. bioRxiv. https://doi.org/10.1101/2020.02.05.935676, 2020.02.05.935676.

Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol. 16, e2004473 https://doi.org/10.1371/journal.pbio.2004473.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party. J. Neurosci. 30, 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010.

Lalor, E.C., Foxe, J.J., 2010. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. 31, 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x.

Luck, S.J., Vogel, E.K., Shapiro, K.L., 1996. Word meanings can be accessed but not reported during the attentional blink. Nature 383, 616–618. https://doi.org/10.1038/383616a0.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004.

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., Golumbic, E.Z., 2017. Sleep disrupts high-level speech parsing despite significant basic auditory processing. J. Neurosci. 37, 7772–7781. https://doi.org/10.1523/JNEUROSCI.0168-17.2017.

Martin, A.E., 2020. A compositional neural architecture for language. J. Cognit. Neurosci. 32, 1407–1427. https://doi.org/10.1162/jocn_a_01552.

Martin, A.E., 2016. Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. Front. Psychol. 7, 120. https://doi.org/10.3389/fpsyg.2016.00120.

Martin, A.E., Doumas, L.A.A., 2017. A mechanism for the cortical computation of hierarchical linguistic structure. PLoS Biol. 15 (3), e200663 https://doi.org/10.1371/journal.pbio.2000663.

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am. 123, 899–909. https://doi.org/10.1121/1.2816572.

Meyer, L., Henry, M.J., Gaston, P., Schmuck, N., Friederici, A.D., 2016. Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cerebr. Cortex 27, 4293–4302. https://doi.org/10.1093/cercor/bhw228.

Naccache, L., Blandin, E., Dehaene, S., 2002. Unconscious masked priming depends on temporal attention. Psychol. Sci. 13, 416–424. https://doi.org/10.1111/1467-9280.00474.

Nobre, A.C., McCarthy, G., 1995. Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. J. Neurosci. 15, 1090–1098. https://doi.org/10.1523/JNEUROSCI.15-02-01090.1995.

Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cerebr. Cortex 23, 1378–1387. https://doi.org/10.1093/cercor/bhs118.

Rogalsky, C., Hickok, G., 2009. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. Cerebr. Cortex 19, 786–796. https://doi.org/10.1093/cercor/bhn126.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B Biol. Sci. 336, 367–373. https://doi.org/10.1098/rstb.1992.0070.

Saffran, J.R., Aslin, R.N., Newport, E.L., 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928. https://doi.org/10.1126/science.274.5294.1926.

Saffran, J.R., Johnson, E.K., Aslin, R.N., Newport, E.L., 1999. Statistical learning of tone sequences by human infants and adults. Cognition 70, 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4.

Sanders, L.D., Newport, E.L., Neville, H.J., 2002. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. Nat. Neurosci. 5, 700–703. https://doi.org/10.1038/nn873.

Song, Y., Bu, Y., Hu, S., Luo, Y., Liu, J., 2010. Short-term language experience shapes the plasticity of the visual word form area. Brain Res. 1316, 83–91. https://doi.org/10.1016/j.brainres.2009.11.086.

Teoh, E.S., Cappelloni, M.S., Lalor, E.C., 2019. Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. Eur. J. Neurosci. 50, 3831–3842. https://doi.org/10.1111/ejn.14510.

Turk-Browne, N.B., Scholl, B.J., Chun, M.M., Johnson, M.K., 2009. Neural evidence of statistical learning: efficient detection of visual regularities without awareness. J. Cognit. Neurosci. 21, 1934–1945. https://doi.org/10.1162/jocn.2009.21131.

Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., Pan, X., Chen, F., Zheng, J., Ding, N., 2019. Auditory and language contributions to neural encoding of speech features in noisy environments. Neuroimage 192, 66–75. https://doi.org/10.1016/j.neuroimage.2019.02.047.