Brain & Language 261 (2025) 105532

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Low-frequency neural activity tracks syntactic information through semantic mediation

Yuan Xie [a,b], Peng Zhou [c,*], Likan Zhan [d], Yanan Xue [d]

[a] School of Engineering, Westlake University, Hangzhou, Zhejiang 310030, China
[b] Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou, Zhejiang 310024, China
[c] Department of Linguistics, School of International Studies, Zhejiang University, Hangzhou 310058, China
[d] School of Communication Sciences, Beijing Language and Culture University, Beijing 100083, China

ARTICLE INFO

Keywords:
Low-frequency neural activity
Semantics
Syntax
Human language processing mechanism

ABSTRACT

How our brain integrates single words into larger linguistic units is a central focus in neurolinguistic studies. Previous studies mainly explored this topic at the semantic or syntactic level, with few looking at how cortical activities track word sequences with different levels of semantic correlations. In addition, prior research did not tease apart the semantic factors from the syntactic ones in the word sequences. The current study addressed these issues by conducting a speech perception EEG experiment using the frequency-tagging paradigm. Participants ($N = 25$, $Mean_{age} = 23;4$, 16 girls) were asked to listen to different types of sequences and their neural activity was recorded by EEG. We also constructed a model simulation based on surprisal values of GPT-2. Both the EEG results and the model prediction show that low-frequency neural activity tracks syntactic information through semantic mediation. Implications of the findings were discussed in relation to the language processing mechanism.

1. Introduction

How our brain processes different levels of linguistic units (e.g. syllables, phrases, sentences) is a central focus in neurolinguistic studies. Many recent studies investigated this phenomenon by looking at the neural basis of processing the basic two-word constructions, e.g., “the boy”, “he cried” (Kang et al., 1999; Sinai & Pratt, 2002; Burton et al., 2009; Herrmann et al., 2012; Tsigka et al., 2014; Schell, et al., 2017; Maran et al., 2022a, 2022b; Rafferty et al., 2023; Li et al., 2024), because the two-word paradigm allows for a detailed observation of a compositional operation that mirrors the nature of combinatory operations described in theoretical linguistics, and reduces additional processes needed for the computation of long and complex constructions, thereby making it ideal for testing a variety of linguistic hypotheses that require multi-dimensional comparisons (Bemis & Pylkkänen, 2011; Iwabuchi et al., 2019; Maran et al., 2022a). Adopting neuroimaging methods, the two-word paradigm studies mainly investigated the processing mechanisms of syntactic phenomena such as categorical and agreement violations, phrase and wordlist contrasts, noun and verb contrasts, etc. (Maran et al., 2022a). For example, Event-Related Potential (ERP) studies revealed that categorical and agreement violations in the two-word constructions evoke ERP components such as an Early Syntactic Negativity (ESN) (Hasting & Kotz, 2008), an increased negativity N400 (Münte et al., 1993; Barber & Carreiras, 2003, 2005), a prolonged negativity lasting until 500 ms (Maran et al., 2022b), and an increased late positivity at 500 ms (Hasting & Kotz, 2008). Electroencephalography/Magnetoencephalography (EEG/MEG) studies found that delta, alpha and beta oscillations are associated with the syntactic computations in the two-word constructions (Lu et al., 2022). Functional Magnetic Resonance Imaging (fMRI) studies demonstrated that the left Brodmann Area 44 (BA 44) is involved in categorical analysis (Herrmann et al., 2012), and the left Inferior Frontal Gyrus (IFG) is engaged in the processing of syntactic agreement (Carreiras et al., 2010) and hierarchical phrases (Zaccarella & Friederici, 2015).

Prior ERP studies on two-word constructions mainly focused on when and what specific components can be evoked during the processing of linguistic units, failing to capture how different linguistic units are dynamically assembled and implemented in the brain. Previous fMRI studies on two-word constructions enabled us to understand what cortical regions are involved in the processing of specific linguistic units without telling us how different linguistic units are dynamically integrated to be finally comprehensible. Prior EEG/MEG research on two-

[*] Corresponding author.
E-mail address: zhoupeng1892@zju.edu.cn (P. Zhou).

https://doi.org/10.1016/j.bandl.2025.105532
Received 30 January 2024; Received in revised form 17 December 2024; Accepted 2 January 2025
Available online 8 January 2025
0093-934X/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

word constructions explored the oscillatory dynamics of linguistic units and revealed the potential of capturing the dynamic process in which different levels of linguistic units are rapidly and constantly assembled and integrated in the brain. To build upon previous EEG studies and to further explore the nature of this dynamic process, the current study conducted an EEG experiment to explore how neural activity tracks and integrates different levels of linguistic units in two-word constructions (i.e., 'noun + noun' pairs versus 'noun + verb' pairs) using the frequency-tagging paradigm. Before presenting our study, we briefly review previous oscillation studies on the processing of different levels of linguistic units in two-word constructions, particularly focusing on the processing of 'noun + noun' and 'noun + verb' pairs that are directly relevant to the current study.

It is widely acknowledged that delta-band (0.5–4 Hz) neural activity can track different levels of linguistic units, such as syllables (4 Hz, Ding et al., 2016), words (2 Hz, Jin et al., 2020), and sentences (1 Hz, Bai et al., 2022; Lu et al., 2022, 2023). Yet, how human brain integrates lower levels of linguistic units (e.g. syllables) into higher level units (e.g. phrases) remains controversial. One spectrum of the existing studies argues that our brain combines words into phrases and sentences by solely relying on the semantic information of words without recourse to the syntactic information (Frank & Yang, 2018); whereas the other spectrum claims that neural activity tracks the syntactic structural information instead of the semantic properties of words (Ding et al. 2016, 2017). Recent EEG studies seem to provide evidence for the latter point of view (Martin & Doumas, 2017; Lo, et al., 2022; Lu et al., 2022, 2023). In these studies, word sequences either with a particular syntactic structure or with a certain specific semantic correlation were presented auditorily to the participants to see what neural responses were evoked while they were processing these sequences. The results showed that the power evoked in the processing of word sequences with a particular syntactic structure was significantly stronger than the power evoked in the processing of word sequences with a specific semantic correlation. They interpreted the findings as evidence that neural activity tracks the syntactic structural information instead of the semantic properties of words.

But we wish to point out that the experimental stimuli used in some of these prior studies appear to be problematic. Consider Lu et al. (2022), for example. The study used EEG to track Mandarin-speaking adults' neural activity in response to two types of sequences containing synthesized continuous isochronous disyllabic Mandarin words, with each sequence consisting of 24 disyllabic words and each word lasting for 500 ms. One type of sequence was composed of two disyllabic nouns (noun + noun) from two semantic categories (i.e., living and nonliving) in an alternating order (e.g. '牡蛎茶杯熊猫尺子……' 'oyster teacup panda ruler…'), referred to as the 'semantic condition'; and the other type of sequence consisted of a four-syllabic 'noun + verb' structure (e.g. '父母回来兔子逃跑……', 'parents returned rabbits escaped…'), referred to as the 'syntactic condition'. The results showed that the 1 Hz peak was observed in both the semantic and syntactic conditions, but the spectral power in the syntactic condition was significantly stronger than that in the semantic condition, suggesting that syntactic processing can evoke cortical responses more effectively than lexical semantic information of single words (Lu et al., 2022). We wish to note, first, that the semantic correlation between a living and a nonliving noun in the study was probably too weak to drive a strong neural response. For example, Quinn and Kinoshita (2008) argued that words from a broad category had a weaker smantic correlation than words from a narrower category. Federmeier and Kutas (1999) found that a weak semantic correlation could predict a weak priming effect on neural responses. On the basis of these previous findings, we suspect that the weaker 1 Hz neural response in the semantic condition might be attributed to the fact that the nouns in the 'noun + noun' pairs were from two broad categories (i.e., living and nonliving things). Second, the speech sequences in the syntactic condition contained not only syntactic structural information but also semantic information, because the noun and the verb used in the 'noun + verb' pairs were not randomly arranged but rather had some semantic correlations. Consider the 'noun + verb' pair 'rabbits escaped', for example. The noun 'rabbit' and the verb 'escaped' not only form a 'subject + predicate' structure, but also constitute as a semantically sensical pair if we compare them to nonsensical 'noun + verb' pairs like 'rabbit closed'. The two observations led us to wonder whether the observed 1 Hz peak was due to the processing of syntactic information, or semantic information, or the combination of the two. To address these questions, speech sequences that can truly tease apart the respective roles of semantic and syntactic information should be used.

In addition, to verify to what extent different statistical language models can predict the same neural responses in the EEG experiment, Lu et al. (2022) used the same EEG stimuli to simulate the corresponding neural responses using three models, a word2vec-based lexical semantic model, a word occurrence frequency model, and a bigram probability model. However, none of the models can fully capture the neural activities evoked in the corresponding EEG experiment. First, the word2vec-based lexical semantic model is a model that captures the word semantic relations based on vector representation of words in large text datasets, and in Lu et al. (2022) such a model predicted a significant 1 Hz response to the speech sequences in both the semantic and syntactic conditions, but a stronger 1 Hz in the semantic condition than in the syntactic condition, which stands in contrast with the EEG results in Lu et al. (2022), where a stronger 1 Hz power was detected in the syntactic condition than in the semantic condition. Such a contrast implies that the word embedding model that relies purely on semantic information cannot yield the same results as the EEG experiment. It is possible that the neural processing of syntactic information involves more complex mechanisms beyond the semantic information that the word2vec-based lexical semantic model captured. Second, the word frequency model is a statistical model that predicts linguistic patterns via the capture of word frequency information in given texts, and in Lu et al. (2022) such a model predicted significant 1 Hz peaks in both the semantic and syntactic conditions, but no significant spectrum power differences between these two conditions, which was also different from the results of the EEG experiment. Such a difference indicated that language statistical models that rely purely on the word frequency information cannot lead to the same result of the EEG experiments, either. The neural processing of syntactic and semantic information is beyond the word frequency information captured by the word frequency model. Finally, the bigram probability model is a model that estimates the probability of a word on the basis of its immediately preceding word in a given text, and in Lu et al. (2022) it predicted a significant 1 Hz in the syntactic condition but no such 1 Hz in the semantic condition, which was very different from the EEG experiment. Such a difference indicated that the neural processing of semantic information is not driven by simple statistical information. To further explore to what degree the existing language models can predict the neural responses in the EEG experiment, we proposed that deep neural network Large Language Models (LLMs) should be used, because LLMs are built upon the Transformer architecture, get trained on a great amount of data, and take varied factors (e.g. word frequency, statistical probability, syntactic and semantic information, context, etc.) into account. LLMs are therefore more likely to predict the same neural responses as in the EEG experiment.

To overcome the problems in Lu et al. (2022), the current study used speech sequences that can truly tease apart the respective roles of semantic and syntactic information. More specifically, in addition to including the semantic and syntactic conditions in Lu et al. (2022) for replication purposes, we designed a new semantic condition and a new syntactic condition. The new semantic condition contained speech sequences consisting of disyllabic antonymic noun pairs (e.g. '真相谎言失败成功……' 'truth lie failure success…') that had stronger semantic correlations[1] than the living and nonliving noun pairs in Lu et al. (2022). The new syntactic condition contained speech sequences composed of random disyllabic 'noun + verb' pairs (e.g. '窗户沸腾冰雪算帐……' 'window boiling ice accounting…') that had no sensical semantic correlations as compared to the 'noun + verb' combinations in Lu et al. (2022). By creating the new semantic condition, we were interested to see whether the 1 Hz peak could be observed in the new semantic condition and whether a significant spectral power difference between the new semantic condition and the replicated semantic condition could be observed when the semantic correlations between words were stronger. By creating the new syntactic condition, we sought to investigate whether the 1 Hz peak could be observed in the new syntactic condition and whether a significant spectral power difference between the new syntactic condition and the replicated syntactic condition could be observed when the 'subject + predicate' syntactic information was maintained but the semantic correlations were weaker.

In addition, we conducted a model simulation by relying on surprisal values calculated in the Generative Pretrained Transformer-2 (GPT-2) model that could take word frequency, syntactic, semantic, and context into consideration, so as to see to what extent the GPT-2 model can predict the spectral peaks and power differences. To take stock, the present study aims to better understand how neural activity tracks and integrates different levels of linguistic units by using a better experimental design and model simulation.

2. Methods

2.1. Participants

Twenty-five Mandarin-speaking 19- to 30-year-old university students participated in the experiment (M_{age} = 23;4, SD_{age} = 3.10, 16 females). They were all right-handed and had no reported history of speech, hearing or neurological disorders. The study was approved by the Ethics Committee of the School of International Studies, Zhejiang University, SIS2023-04. Written informed consent has been obtained from each participant.

2.2. Stimuli and design

The research design was based on Lu et al. (2022). The stimuli were speech sequences constructed by continuous isochronous Mandarin disyllabic words. Each speech sequence consisted of 24 Mandarin disyllabic words without any pauses between any two adjacent words. All the disyllabic words were synthesized independently using iFLYTEK synthesizer (http://peiyin.xunfei.cn/; Mandarin Chinese, female, Xiaoying) and were adjusted to the same intensity by Praat. Each disyllabic word was cut into an isochronous 500 ms, so each speech sequence lasted for 12 s.

As discussed, four experimental conditions were created (see Table 1), including a new semantic condition, a new syntactic condition, and the original semantic and syntactic conditions in Lu et al. (2022), with each condition containing 30 speech sequences.

In the new semantic condition, the antonym pairs, the disyllabic words in the speech sequences were antonym pairs (e.g. 真话-谎言truth-lie). Each speech sequence contained 12 non-repetitive antonym pairs (N_{word} = 24) that were randomly selected from the total of 120 different antonym pairs (N_{word} = 240).

Table 1
Experimental Conditions.

Condition	Speech Sequence
antonym pairs	$N_1N_2N_1N_2N_1N_2N_1N_2$… 真话谎言赢家败者城市乡村白天夜晚…… Truth lie winner loser city country day night…
NV-nonsensical pairs	NVNVNVNV… 大海丢失细胞讲课果汁扫地厨房生长…… Sea lose cell teach juice sweep kitchen grow…
living-nonliving pairs	LNLNLNLN… 蜘蛛剪刀蝙蝠大门草莓话筒莲藕码头…… Spider scissors bat door strawberry microphone lotus wharf…
NV-sensical pairs	$N_1V_2N_1V_2N_1V_2N_1V_2$… 牡丹盛开情侣散步小孩哭泣客人离去…… Peony bloom spouse walk kid cry guest leave…

Table 2
An Example Filler Sequence.

Condition	Speech Sequence
Filler sequences in the living-nonliving pairs condition	LNLNLNLN…P[1] 老虎学校松鼠堤坝东边黄瓜蜡烛西边 …… Tiger school squirrel dam east cucumber candle west … (24 words in total)

[1] P represents the locational prepositions.

In the new syntactic condition, the NV-nonsensical pairs, the disyllabic words in the speech sequences were a 'noun + verb' combination (e.g. '大楼-打败' 'building-defeat') in which the noun and the verb had no sensical semantic correlations. Each speech sequence consisted of 12 NV-nonsensical pairs that were randomly selected from the total of 80 different pairs (N_{word} = 160).

The living-nonliving pairs and the NV-sensical pairs replicated the semantic and syntactic conditions in Lu et al. (2022) respectively. We renamed them for illustration purposes. The living-nonliving pairs contained speech sequences that were composed of 'living noun + nonliving noun' pairs (e.g. '老虎-茶杯' 'tiger-teacup'). The living nouns included animals (N_{word} = 60) or plants (N_{word} = 60), and the nonliving nouns included small manipulable objects (N_{word} = 60, e.g. '茶杯' 'teacup') or large manipulable objects (N_{word} = 60, e.g. '操场' 'playground'). The living nouns in each sequence were evenly selected either from the category *animals* or from *plants*, and the nonliving nouns were evenly selected either from the category *small manipulable objects* or from *large manipulable objects*. The NV-sensical pairs contained sequences that were composed of 'noun + verb' combinations that had sensical semantic correlations (e.g. '情侣-散步' 'spouse-walk'). Each sequence had 12 'noun + verb' pairs that were randomly selected from the total of 80 pairs (N_{word} = 160).

In addition to the target sequences, seven filler sequences were constructed for each experimental condition. The filler sequences had the same composition as the corresponding target sequences in each condition. The difference between the filler and the target sequences was that each filler sequence contained two non-adjacent disyllabic locational prepositions (e.g. '北边' 'north', '南边' 'south') that were randomly selected from a pool of 42 locational prepositions. An example filler sequence in the living-nonliving pairs condition is provided in Table 2. In addition, we also created eight speech sequences as practice trials, in which four were from the filler sequences and four from the target sequences.

As discussed, we further quantified the strength of the semantic

[1] The semantic correlations between the antonymic noun pairs and between the noun–verb pairs in the syntactic condition was measured by the values calculated through the word2vec-based lexical semantic model. Details are presented in section 2.2.

correlations of word pairs in the four experimental conditions (i.e. the antonym pairs, NV-nonsensical pairs, living-nonliving pairs, and NV-sensical pairs). This process was realized by computing the semantic similarity[2] for each word pair in the four experimental conditions based on BGE-M3 (Chen, et al., 2024) within the word2vec framework. BGE-M3 is a model that uses 1024-demensional dense vectors to represent words. The dataset for training covers 194 languages (Mandarin is also included), which is sufficient for capturing the semantic relations between words.

We used cosine similarity as the metric and obtained values of the semantic similarity for each antonym pair, NV-nonsensical pair, living-nonliving pair, and NV-sensical pair, and then saved all the values in one dataset. Applying the Shapiro-Wilk test to the values (*shapiro.test* function in R (v4.2.3)), we found that the similarity values in some of the conditions did not follow normal distribution (e.g. living-nonliving pairs: $W = 0.9897$, $p = 0.0128$). We therefore adopted the Wilcoxon Rank-Sum Test (*wilcox.test* function in R, one-sided comparison) to compare the differences between the antonym pairs and the living-nonliving pairs, and between the NV-nonsensical pairs and the NV-sensical pairs. The results showed that the values in the antonym pairs were significantly higher than the values in the living-nonliving pairs ($Median_{\text{antonym pairs}} = 0.521$,[3] $Median_{\text{living-nonliving pairs}} = 0.407$, $W = 36451$, $p < 0.0001$), pointing towards stronger semantic correlations in the antonym pairs. In addition, the values in the NV-sensical pairs were also significantly higher than those in the NV-nonsensical pairs ($Median_{\text{NV-sensical pairs}} = 0.542$, $Median_{\text{NV-nonsensical pairs}} = 0.422$, $W = 5768.5$, $p < 0.0001$), indicating stronger semantic correlations in the NV-sensical pairs. Fig. 1 gives the distribution of the cosine similarity values for the four conditions.

2.3. Experimental procedures

The experiment was programmed using PsychoPy. The participants' task was to judge whether or not the auditorily presented speech sequence contained locational prepositions by pressing different keys (i.e., the participants were asked to press the 'up' key if they heard any locational prepositions and to press the 'left' key if they did not hear any locational prepositions). The entire experimental procedure was divided into three sessions. **(1) The word familiarization session.** In this session, the synthesized disyllabic words were presented to the participants both visually and auditorily. When the participants pressed the 'space' key, they heard a word and simultaneously saw this word on the computer screen. They could press the 'up' key to listen to the word again or press the 'down' key to go to the next word. They were asked to go through all the words to get familiarized with these words. **(2) The practice session.** In this session, there were eight practice speech sequences that were evenly divided into two sets. Each set included four speech sequences, two from the filler sequences and two from the target sequences. The participants were asked to go through either set of the sequences, so that they could be familiarized with both the target and filler sequences. More specifically, they listened to speech sequences with eyes closed in order to reduce artifacts, and when a speech sequence ended, they pressed the 'up' or the 'left' key to judge whether or not the sequence they just heard contained locational prepositions. After they made their judgements, the program provided an auditorily presented 'correct' or 'incorrect' feedback. **(3) The experimental session.** After the practice session, the participants proceeded to the experimental session. The experimental session included four blocks (i.e. one block per condition) and each block lasted for about 10 min. The participants were presented with the four blocks in random order and were required to have a rest for at least 3 min between blocks. The task for the participants in the experimental session was to judge whether or not the speech sequence they heard contained locational prepositions, which was exactly the same as what they did in the practice session.

2.4. EEG data recording

EEG data was continuously recorded with a 128-channel system (*NetStation* software, *Electrical Geodesics, Inc.*) at the sampling rate of 1000 Hz (online bandpass filter = 0.3–15 Hz) and referenced to the vertex (Cz). Twenty electrodes (C3, C4, F3, F4, F7, F8, FP1, FP2, FPZ (14), FPZ(21), FPZ(15), Fz, O1, O2, P3, P4, P7, P8, T7, T8) in the 10–10 system were selected for the following data preprocessing and analysis (see Fig. 2). We chose these electrodes based on prior research (Ding et al.,2018; Lu et al., 2022). The electrodes selected in Ding et al. (2018) were distributed across the frontal, central and parietal regions of the brain, and the electrodes chosen in Lu et al. (2022) were primarily distributed in the frontal and central areas of the brain (i.e. Cz, Fz, FCz, FC3, and FC4). We followed the two studies and chose twenty electrodes that were primarily located in the frontal, central, and parietal regions of the brain. We selected these twenty-electrodes, because we used EGI devices and these twenty electrodes had relatively fixed positions in different EGI devices (e.g. 128-channel and 256-channel). The impedance of electrodes was kept below 50kΩ and the pre- and post- experiment impedance of electrodes were checked and saved for each participant.

2.5. EEG data preprocessing

Raw EEG data saved in *NetStation* were exported to matlab (version R2022b) using Fieldtrip (version 20220707) for data analysis. The filler sequences (i.e., those contain locational prepositions) were excluded in the data analysis. The sampling rate was down to 20 Hz, as the current study focused on the 1 Hz and 2 Hz. A linear-phase finite impulse response (FIR) filter was used to bandpass filter the EEG signal between 0.3 and 2.7 Hz, using the same procedure as in Lu et al. (2022) (−6 dB attenuation at the cut-off frequencies, 10-s Hamming window). The least-squares method (Ding et al., 2017, Lu et al., 2022) was used to remove the horizontal and vertical EOG artifacts that included 6 electrodes,[4] the left and right horizontal ones, the left inferior and superior ones, and the right inferior and superior ones. The least-squares method is a mathematical approach used to minimize the difference between observed data and a model. To remove EOG artifacts, we first recorded signals from EOG electrodes to measure artifacts (e.g. eye movements) and then modeled how these artifacts affected the EEG signals. Using the least-squares method, we calculated and subtracted the influence of the EOG signals from the EEG data, leaving cleaner EEG signals for analysis. By using the kurtosis method, we detected bad channels that contained "spikes" that might distort the data and then corrected them. More specifically, we first set a threshold value and identified the channels with kurtosis values higher than the threshold one, and defined these channels as bad ones. We then corrected these bad channels by centering the signal (e.g. subtracting the median) and by capping extreme values to a reasonable range, and we finally updated the data. The data were re-referenced to the average of the left and right mastoid signals.

2.6. EEG data analysis

2.6.1. Frequency-domain analysis

We adopted the frequency-domain analysis as in Lu et al. (2022). The

[2] Note that word semantic similarity here does not simply mean that the words have similar meanings. Instead, it indicates that the two words have semantic correlations in some specific context.

[3] As the data did not follow normal distribution, we calculated the median value for the four conditions instead of the mean value.

[4] The EGI device has six EOG electrodes: left horizontal (128), right horizontal (125), left inferior (127), right inferior (126), left superior (25), and right superior (8).

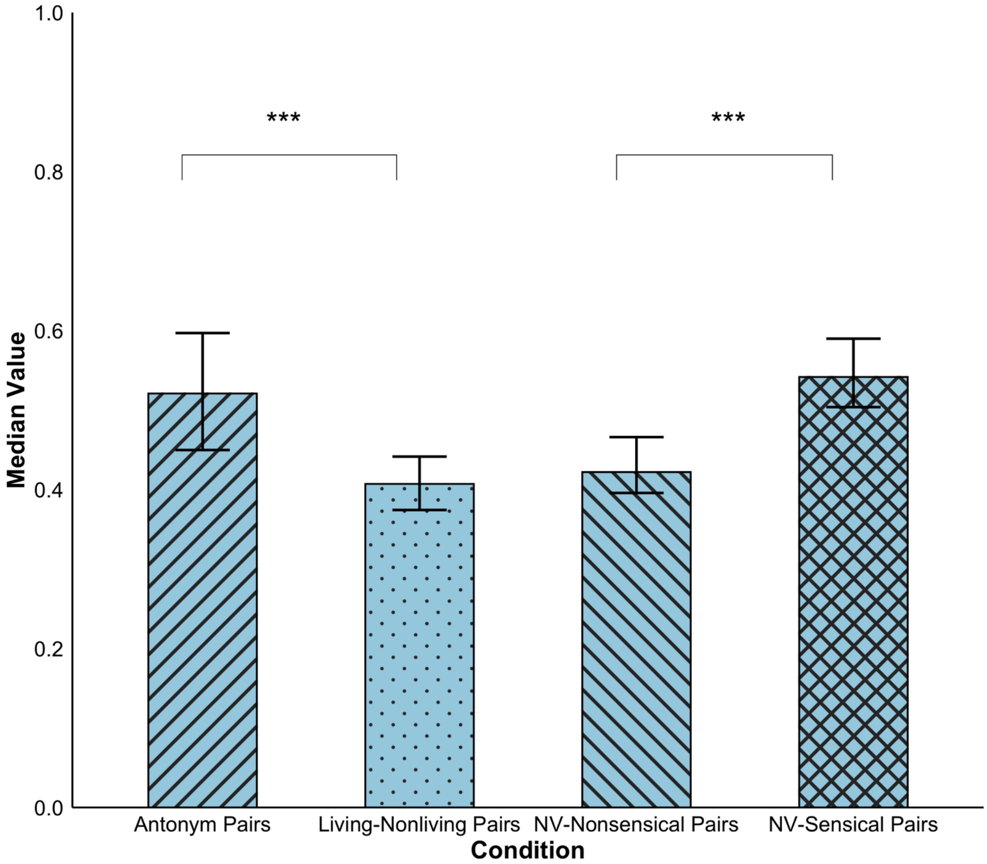


Fig. 1. Distribution of the cosine similarity values for the four conditions.

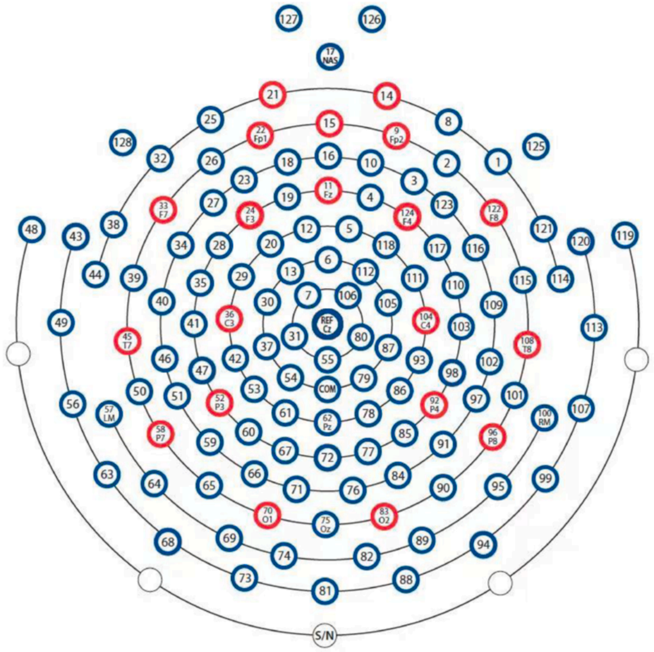


Fig. 2. Distribution of the twenty selected electrodes as marked in red.[1] The picture was adapted based on the EGI 128-Channel Map in the GES Hardware Technical Manual (Electrical Geodesics, 2007)).

first 2 s of each speech sequence were removed to avoid onset response, and thus for each sequence only 10 s were used for analysis. The average of EEG responses of all speech sequences was transformed into the frequency domain by applying the Discrete Fourier Transform (DFT) without the incorporation of any additional smoothing window. The DFT is a method that transforms time-domain signals into frequency components, and the result of the DFT is a spectrum of frequencies. In other words, the DFT decomposes the input signals into different constituent frequencies with each represented by a frequency bin. The DFT was independently applied to each EEG channel.

The process of the DFT analysis was as follows. First, we defined the sampling rate, the DFT length and the frequency range. In the current study, the frequency resolution for the DFT analysis was set at 0.1 Hz ($\Delta f = 0.1$ Hz). As the sampling rate was down to 20 Hz ($f_s = 20$ Hz, see section 2.5), the DFT length N was 200, which was obtained from dividing the sampling rate by the frequency resolution ($N = f_s/\Delta f$). Namely, 200 samples from the input signals were used to compute the DFT, which was also the total number of the frequency bins. In addition, the 20 Hz sampling rate and the frequency resolution together determined the frequency range, which ranged from 0 Hz to 10 Hz (Nyquist frequency that was half of the sampling rate), with steps of 0.1 Hz.

Second, the DFT coefficients were computed using the formula in (1).

$$X[k] = \sum_{n=0}^{N-1} x[n]\cdot e^{-i\frac{2\pi}{N}kn} \tag{1}$$

In this formula, $x[n]$ represents the signal value n sample later; i is a special number defined as the square root of -1; k is an index to identify specific frequency bins; $X[k]$ means that for each frequency bin k (where $k = 0, 1, 2, \ldots, 199$), the coefficient was $X[k]$.

Third, we computed the response power for each frequency bin using the formula in (2).

$$\mathrm{Power}[k] = |X[k]|^2 \tag{2}$$

We averaged the neural responses across all channels when analysing the EEG response power.

2.6.2. Statistical analysis

The statistical analysis process was also consistent with Lu et al. (2022). First, we were interested to see whether there were any significant 1 Hz and 2 Hz spectral peaks in each experimental condition. We pinpointed the frequency of a spectral peak *f* and the two neighboring frequencies around the target frequency *f* (i.e. the frequency bin just below *f* and that just above *f*), and then compared the response power at *f* with the mean of the response power of these two neighboring frequencies. As the frequency resolution was 0.1 Hz, the two neighboring frequency bins around 1 Hz were 0.9 Hz and 1.1 Hz, and the two around 2 Hz were 1.9 Hz and 2.1 Hz. We then compared the response power at 1 Hz with the mean of the response power at 0.9 Hz and 1.1 Hz, and that at 2 Hz with the mean of the response power at 1.9 Hz and 2.1 Hz. The comparison was one-sided using the bias-corrected and accelerated bootstrap method (BCa bootstrap)[5] (Efron & Tibshirani 1994), a method that can create "virtual" samples through repeatedly drawing random samples with replacement from the original data and then access the differences between different datasets.

There were 25 participants, so we obtained 25 response power values for the 1 Hz condition. We also obtained 25 means of response power values under the 0.9 Hz and 1.1 Hz conditions. Next, for the 1 Hz condition, we randomly sampled 25 values with replacement from the 25 original response power values to form a new set of 25 response power values, and we averaged the 25 new response power values, and obtained a mean. We repeated this process 10,000 times and thus obtained 10,000 means and formed a bootstrap distribution. For the [0.9 Hz-1.1 Hz] $_{\text{mean}}$ condition, we conducted the same process and formed another bootstrap distribution. We then conducted a one-sided comparison of the response power values of the two bootstrap distributions.[6] The significance level was set at (100A + 1)/10,001 if the data population in one condition was greater than A% of the data population in the other condition. The same procedure was applied to the 2 Hz condition and the [1.9 Hz-2.1 Hz]$_{\text{mean}}$ conditions.

In addition, we were also interested to see whether there were any response power differences between the experimental conditions at 1 Hz and 2 Hz. Normalized power was used to conduct the comparisons so as to eliminate potential effects due to individual variations in the experimental conditions. The normalized power was calculated by using the 1 Hz peak response power minus the mean of 0.9 Hz and 1.1 Hz response power and by using the 2 Hz peak response power minus the mean of 1.9 Hz and 2.1 Hz response power. We compared the normalized power at 1 Hz and 2 Hz in the antonym-pair condition with that in the living-nonliving condition, and the normalized power at 1 Hz and 2 Hz in the NV-sensical condition with that in the NV-nonsensical condition. The comparison was two-sided using the bias-corrected and accelerated bootstrap method, too. We resampled all the means 10,000 times with replacement and conducted paired comparisons. In the two-sided comparisons, the significance level was set at (200A + 1)/10,001. In addition, a false discovery rate (FDR) correction was applied in both the one-sided and two-sided comparisons.

2.7. Predictions

Since the speech sequences were composed of continuous isochronous Mandarin disyllabic words that alternated periodically at several levels, neural oscillations were expected to be observed in a specific frequency band if they could track such periodicity. More specifically, a 1 Hz peak was expected to occur in the antonym pairs and living-nonliving pairs if neural activity tracks the semantic properties of the sequences; such a 1 Hz peak might occur in the NV-nonsensical pairs if neural activity tracks the syntactic information of the sequences; a 1 Hz peak might appear in the NV-sensical pairs if neural activity tracks the semantic or/and syntactic information of the sequences. 2 Hz peaks might be observed in all four conditions if neural activity tracks the word-level information of the sequences. In addition, we might also observe a significant power difference between the antonym pairs and living-nonliving pairs as the semantic correlations were stronger in the former condition.

2.8. Model simulation

2.8.1. The pretrained language model GPT-2 and surprisal values

We constructed a model simulation using the surprisal values calculated from the pretrained large language model GPT-2 (Generative Pre-trained Transformer 2). The GPT-2 we chose is gpt2-chinese-cluecorpussmall (Zhao et al., 2019), which was built on the architecture of GPT-2 (12-layer transformer blocks, 1.5B parameters) (Radford et al., 2019) and was trained on a Chinese corpus, CLUE CorpusSmall (Xu et al., 2020) (exceeding 14 GB in size, containing over 5 billion tokens), thereby enabling Chinese sequence generation.

Surprisal is a concept in information theory (Shannon, 1948) that is used to quantify the unexpectedness of an event. In LMMs (e.g. GPT-2), surprisal is adopted to measure how "unexpected" a token (e.g. word) is in a given context. If we represent the sequence of words by $\mathscr{S} = [w_1, w_2, \cdots, w_{i-1}]$, then the GPT-2 model predicts the conditional probability of the next word w_i given the preceding words, represented as $P(w_i|w_1, \cdots, w_{i-1})$, which is determined by the model's weights that are learned during pretraining. The formula for the computation of surprisal values from a language model on a word sequence is given in (3).

$$\text{Surprisal}(w_i) = -\log P(w_i|w_1, w_2, \cdots, w_{i-1}) \tag{3}$$

In order to see whether there were any differences between the distribution of the surprisal values and that of the cosine similarity values for word pairs in the four conditions, we computed the surprisal values for each antonym pair, NV-nonsensical pair, living-nonliving pair, and NV-sensical pair, and conducted the same statistical procedures as we computed the cosine similarities values of word pairs in section 2.2. The results showed that the median of the surprisal values of the antonym pairs were significantly lower than that of the living-nonliving pairs ($Median_{\text{antonym pairs}} = 28.8$, $Median_{\text{living-nonliving pairs}} = 30.9$, $W = 15143$, $p < 0.0001$), and the median of the surprisal values of the NV-sensical pairs were significantly lower than that of the NV-nonsensical pairs ($Median_{\text{NV-sensical pairs}} = 26.9$, $Median_{\text{NV-nonsensical Pairs}} = 33.0$, $W = 389$, $p < 0.0001$). Note that surprisal values represent uncertainty, so lower surprisal values for word pairs indicate lower uncertainty in their connections, which might be associated with higher cosine similarity values. Therefore, these statistical results were consistent with those of the cosine values in the corresponding conditions. But we wish to note that the distribution of the median of the surprisal values for the four conditions was slightly different from that of the median of the 1-cosine similarity values as shown in Fig. 3.[7]

Fig. 3A indicated that the median of the surprisal values of the living-nonliving pairs was lower than that of the NV-nonsensical pairs, whereas Fig. 3B showed that the median of the 1-cosine similarity values of two types of pairs was in the opposite distribution. The median of the surprisal values and that of the 1-cosine similarity values of the other two types of pairs (i.e. the antonym pairs and the NV-sensical pairs)

[5] "Bias-corrected" and "accelerated" mean that the data was adjusted through bias-correction and acceleration so as to make it statistically more precise.

[6] We conducted one-sided comparison in order to see whether the 1 Hz response power was significantly greater than the mean of [0.9Hz-1.1 Hz] power.

[7] When the cosine similarity value of a word pair is low (i.e. semantically less related to the context), the word pair is likely to be more unexpected and thus corresponds to a higher surprisal value. We therefore plotted the 1-cosine similarity values to better capture the similarities and differences in their distribution.

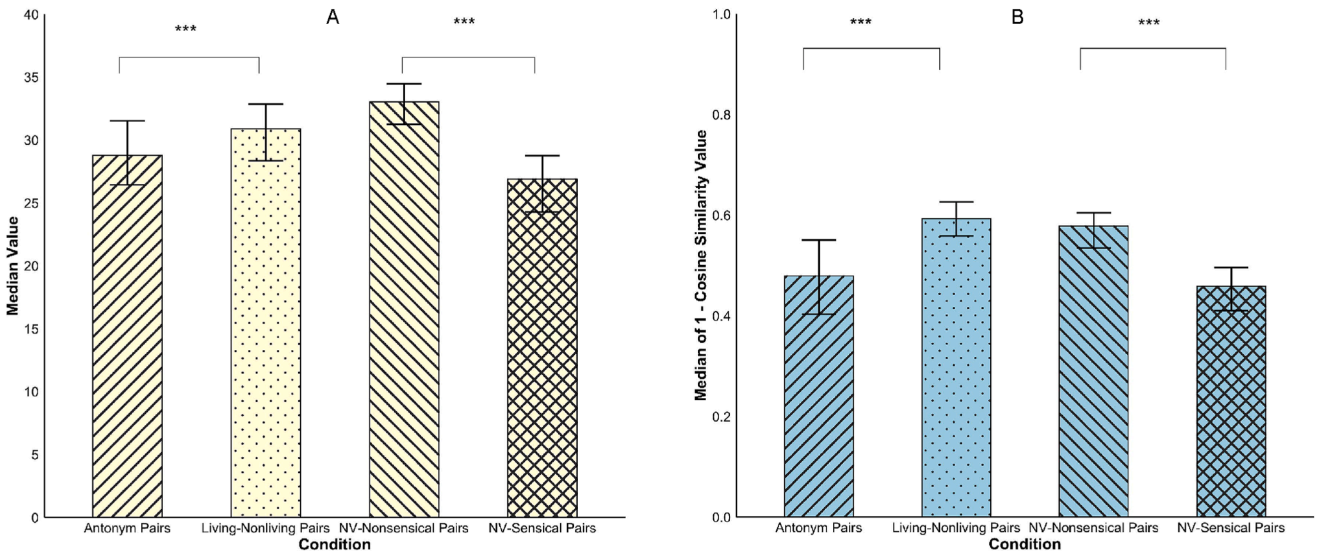


Fig. 3. Distribution of the median of the surprisal values for the four conditions (A) and that of the median of the 1-cosine similarity values for the four conditions (B). *** $p < 0.001$.

exhibited similar distribution.

In addition, as the speech sequences were required to be simulated as neural pulse sequences, we next computed the surprisal values between every two immediately adjacent words, treating the immediate previous one word as the context. For example, the surprisal value of ‘徒弟’ (‘apprentice’) was computed given its immediate preceding word ‘师傅’ (‘master’), which in turn took its immediate preceding word as the context. This process yielded 23 surprisal values for each word sequence.

2.8.2. *Pulse sequence and neural response waveform simulation*

The speech sequences were simulated as neural pulse sequences. We first prepared the data for the neural response waveform simulation based on the EEG experimental stimuli. In the EEG experiment, there were four experimental conditions with each condition containing 30 trials (i.e. word sequences), so there were 120 trials in total. We then calculated the surprisal values between every two immediate adjacent words for each trial in each condition and we obtained 23 surprisal values for each trial (see Table 3). We treated the 23 surprisal values for each trial as a pulse sequence that can be used for frequency analysis. The pulse was placed at the onset of each word and its amplitude was assigned a surprisal value. We further convolved the pulse sequences with a 500 ms duration Gaussian window to simulate neural responses. Finally, we conducted a FFT analysis on these simulated responses to see whether or not the specific spectral peaks (i.e. 1 Hz and 2 Hz) were observed. Note that the data of the first two seconds of each pulse sequence were removed from the final analysis, so as to be consistent with the EEG experiment. We put a zero in the beginning of each pulse sequence and regarded it as a marker.

3. Results

3.1. *The spectral peaks in the four experimental conditions*

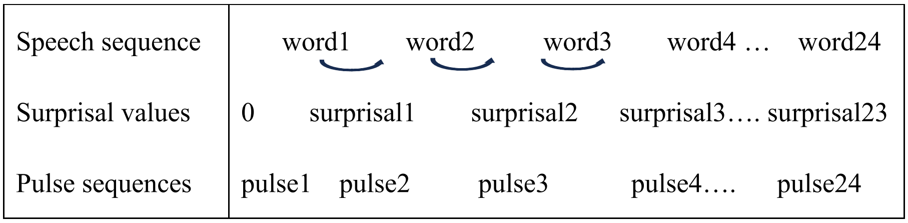
We adopted the paired one-sided bootstrap method with FDR corrected and observed a significant 1 Hz peak in the antonym pairs (Fig. 4A, $p = 0.0064$) and in the living-nonliving pairs (Fig. 4C, $p = 0.0418$). A significant 1 Hz peak was also detected in the NV-sensical pairs (Fig. 4D, $p = 0.0001$), but not in the NV-nonsensical pairs (Fig. 4B, $p = 0.2582$). In addition, a significant 2 Hz peak was observed in all the four conditions (see Fig. 4A-4D, $p = 0.0002$ for the antonym pairs, the NV-nonsensical pairs, and the living-nonliving pairs, $p = 0.0001$ for the NV-sensical pairs), attesting to that low-frequency neural activity can track every single word information.

We then compared the 1 Hz power differences among the three conditions that observed a significant 1 Hz peak. By adopting paired two-sided bootstrap method with FDR corrected, we found significant power differences between the antonym pairs and the NV-sensical pairs (Fig. 5A, $p = 0.0004$), between the living-nonliving pairs and the NV-sensical pairs (Fig. 5A, $p = 0.0002$). No significant power difference was observed between the antonym pairs and the living-nonliving pairs (Fig. 5A, $p = 0.8557$). In addition, there was no significant power difference of 2 Hz peaks between any of the two experimental conditions (see Fig. 5B).

3.2. *The spectral peaks and power of the simulated model*

In addition to the EEG responses, we also constructed a model to simulate how the probability of words contributed to the neural

Table 3
Simulation of Pulse Sequences.



Speech sequence		word1	word2	word3	word4 …	word24
Surprisal values	0	surprisal1	surprisal2		surprisal3….	surprisal23
Pulse sequences	pulse1	pulse2	pulse3		pulse4….	pulse24

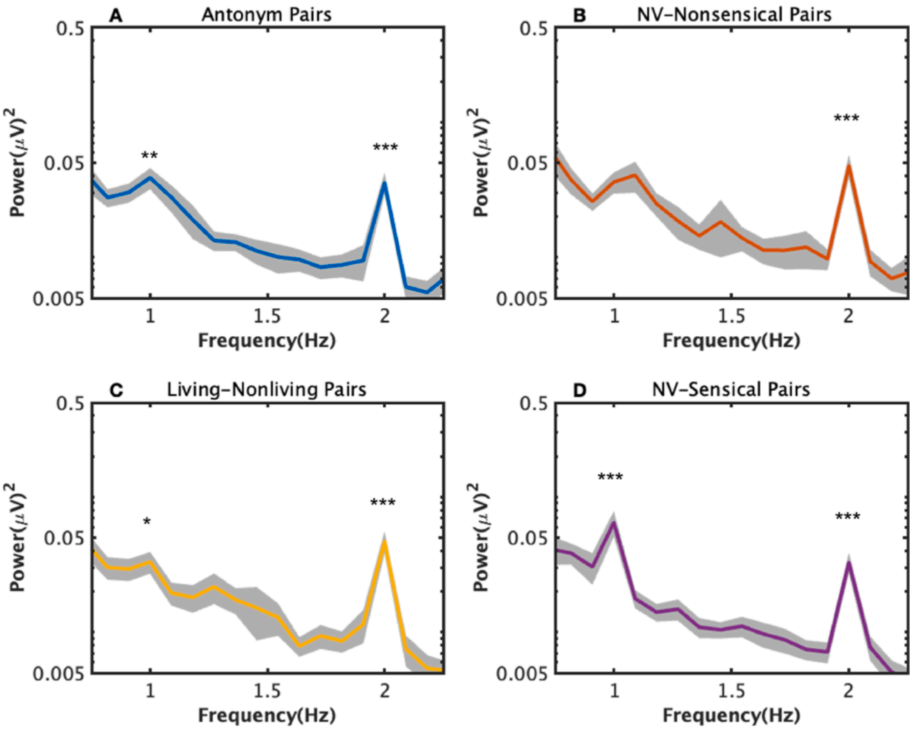


Fig. 4. Spectral peaks in different experimental conditions. A. antonym pairs; B. NV-nonsensical. pairs; C. living-nonliving pairs; D. NV-sentence. *** $p < 0.001$, **$p < 0.01$, *$p < 0.05$.

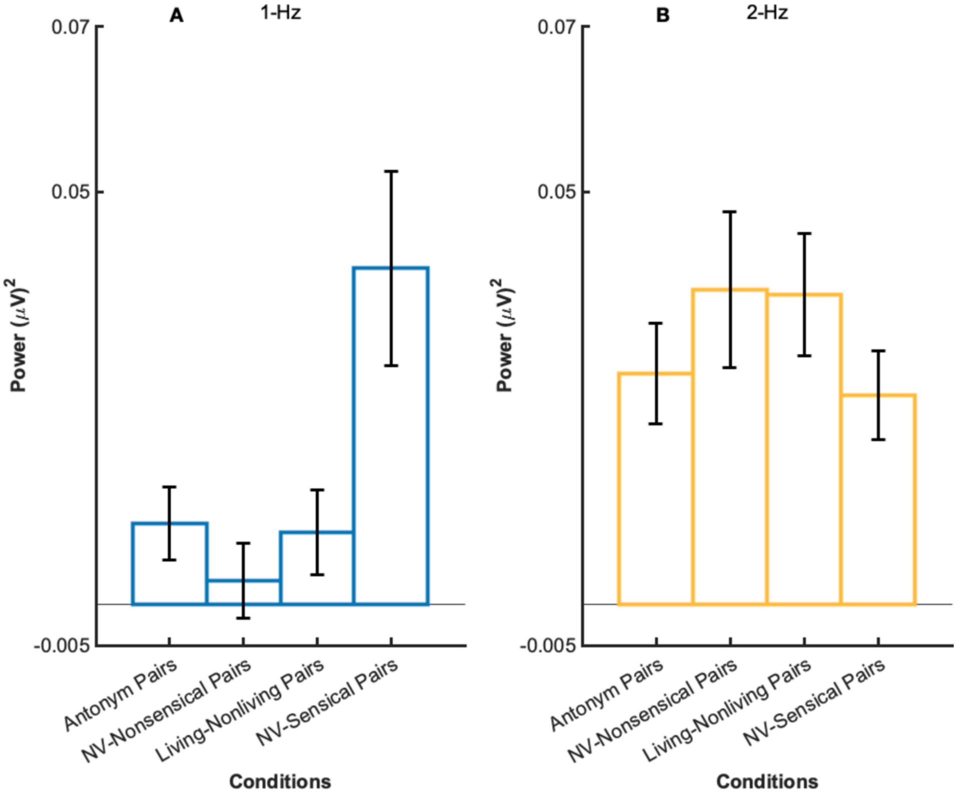


Fig. 5. Peak values of different experimental conditions. A.1 Hz power; B. 2 Hz power.

responses to the sequences. By adopting the paired one-sided bootstrap method with FDR correction, we found a 1 Hz peak in the antonym pairs (Fig. 6A, $p = 0.0002$), in the NV-nonsensical pairs (Fig. 6B, $p = 0.0004$), and in the NV-sensical pairs (Fig. 6D, $p = 0.0002$), and but not in the living-nonliving pairs (Fig. 6C, $p = 0.6326$). In addition, a 2 Hz peak was observed in all the four conditions (Fig. 6A-6D, all $p = 0.0001$).

We then compared the power differences between conditions. Since there was no significant 1 Hz peak in the living-nonliving pairs

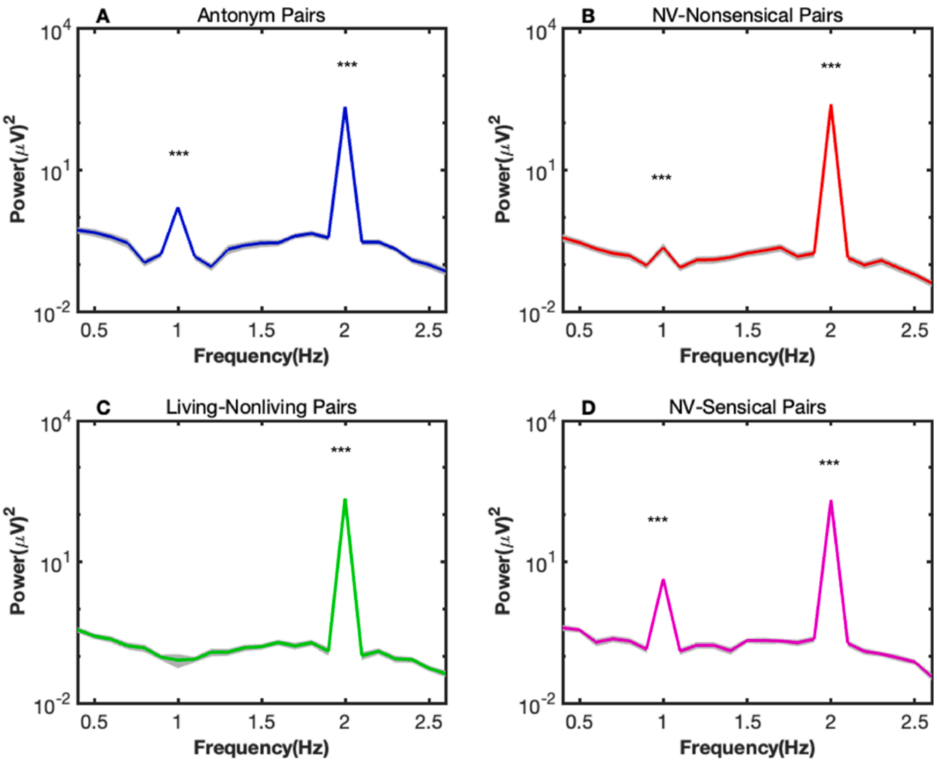


Fig. 6. Spectral peaks in the model. A. antonym pairs; B. NV-nonsensical pairs; C. living-nonliving pairs; D. NV-sentence. *** $p < 0.001$.

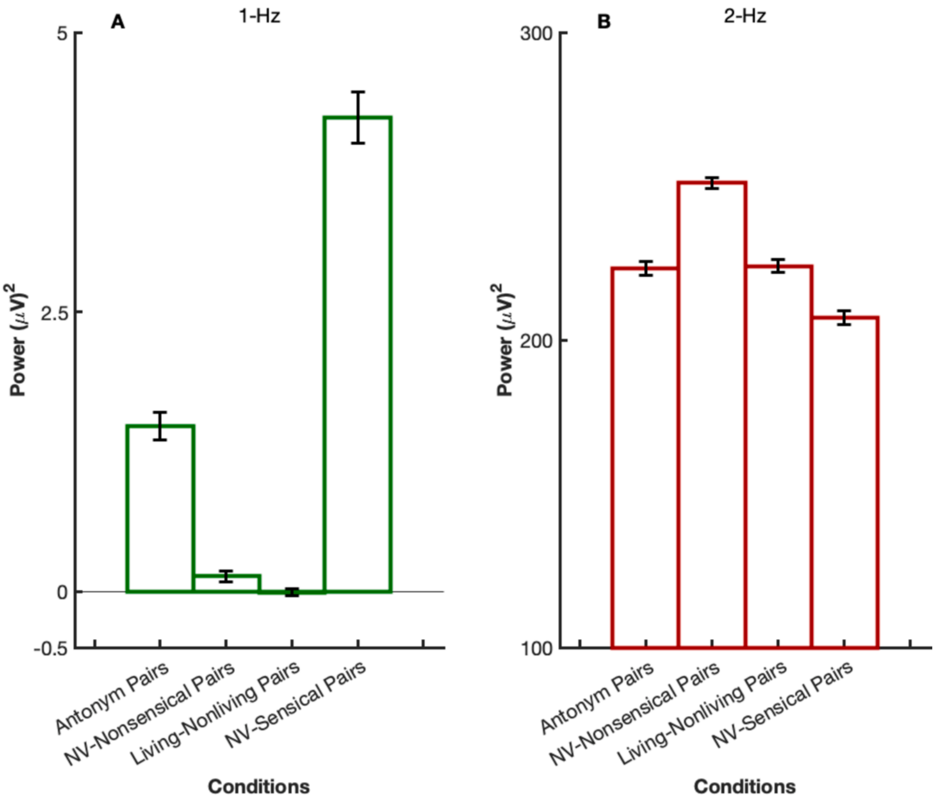


Fig. 7. Peak values of different experimental conditions. A. 1 Hz; B. 2 Hz. [1] We adopted different y scales for Fig. 7A and Fig. 7B because the difference between the 1 Hz response power and the 2 Hz response power was too large, and if we used the same y scale for them, the significant response power differences would be masked in the 2 Hz condition.

condition, we only conducted a paired-wise comparison between the other three conditions. The results showed a significant 1 Hz power difference between each pair of conditions among the three (Fig. 7A, all $p = 0.0002$). In addition, given that significant 2 Hz peaks were observed in all the four conditions, we then conducted a paired-wise comparison between the four conditions. The results showed that the 2 Hz power difference was not significant between the antonym pairs and the living-nonliving pairs (Fig. 7B, $p = 0.7559$), but it was significant between any other two conditions (Fig. 7B, all $p = 0.0002$).

4. Discussion

4.1. *Neural tracking of semantic properties of words in word sequences with different levels of semantic correlations*

The current study explored whether cortical activity can track the semantic properties of words in word sequences with different levels of semantic correlations in Mandarin. Previous studies on this topic mainly focused on semantic features in general, with few looking at how cortical activities track word sequences with different levels of semantic correlations. So, the present study is the first of this kind. More specifically, this study investigated whether changing the strength of the lexical semantic correlations of word sequences would lead to different spectral peaks and response power differences. Consistent with Lu et al. (2022), we found a significant 1 Hz peak in the living-nonliving pairs. In addition, we observed a 1 Hz peak in the antonym pairs. As they were both semantic conditions, a significant 1 Hz peak indicates that neural activity tracks the semantic properties of words in Mandarin. In the antonym pairs, words related to each other in contrast meanings, and in the living-nonliving pairs, words were co-hyponyms of the same superordinate word, both of which were associative relations that exhibited the effects of priming in word recognition tasks (Niemi et al., 1980; Mandler et al., 1990; Perea and Rosa, 2002; Jakić et al., 2011; Was et al., 2019; Chen et al., 2022). According to Niemi et al. (1980), the representation of semantic concepts is activated by priming due to previously attended related materials, and thus increases the availability of the subsequent item processing, leading cortical activity to track the semantic correlations between words. In addition, neuroimaging evidence also showed that the processing of antonyms and superordinate relations evokes stronger cortical responses in specific brain areas than the processing of unrelated semantic concepts (Jeon, et al., 2009; Raposo, et al., 2012; Zhuang et al., 2023).

Note that the words in the antonym pairs had stronger lexical semantic correlations than those in the living-nonliving pairs, because the antonym pairs were conventionalized expressions that had high correlations and “are entrenched in memory and perceived as strongly coupled pairings by speakers” (Paradis et al., 2009: 386). By contrast, the words in the living-nonliving pairs covered four different subcategories: animals, plants, small manipulatable objects, and large manipulatable objects. Thus, the combination of these words had much weaker lexical semantic correlations. The difference in the lexical semantic correlations of words in the two conditions, however, did not yield any significant power difference. This finding might suggest that changing the strength of the lexical semantic correlation between words in the ‘noun + noun’ pairs might not necessarily cause significant power difference.

4.2. *Neural tracking of syntactic information through semantic mediation*

The current study also explored whether cortical activity can track the syntactic structural information of word sequences with different levels of semantic correlations. In addition, the semantic factors were teased apart from the syntactic factors in the word sequences. Most of the prior research did not tease apart the semantic information from the syntactic information, and they treated the NV-sensical word pairs condition as a purely syntactic one. A purely syntactic condition, however, should be a jabberwocky one where the word sequence is nonsensical at all, yet perfectly conforms to syntactic rules.[8] For example, Rafferty et al. (2023) designed a jabberwocky condition that was composed of meaningless two-word phrases (e.g. *the moop*)[12] to see whether the cortical activity can track the syntactic structure of the phrase. The results exhibited a significant spectral peak, suggesting that cortical activity can track syntactic structure information. In the current study, we sought to investigate the role of semantic properties of word pairs in mediating the syntactic information of the pairs. In both the NV-nonsensical pairs and NV-sensical pairs, the nouns and verbs in the sequences formed a ‘subject + predicate’ syntactic structure where the nouns and verbs were either strongly or weakly semantically correlated. The findings showed no significant 1 Hz peak in the semantically weakly correlated ‘subject + predicate’ syntactic structure condition (i.e. NV-nonsensical pairs), indicating that cortical activity in response to that syntactic structure was reduced so that no 1 Hz peak was detected in that condition. A significant 1 Hz peak was observed in the semantically strongly correlated ‘subject + predicate’ syntactic structure condition only (i.e. NV-sensical pairs), suggesting that cortical activity in response to that syntactic structure was enhanced so that a significant 1 Hz peak was observed in that condition. These findings seem to suggest that syntactic information was tracked through the semantic mediation.

4.3. *Model prediction*

Different from previous simulations that relied on models based on simple statistical methods, the current study adopted the LLM GPT-2 to explore to what extent such a model can predict the same neural responses that were obtained from the EEG experiment.

4.3.1. *Limitations of the model in tracking semantic properties of words in word sequences with weak semantic correlations*

Both the simulated model and the EEG results showed a significant 1 Hz peak in the antonym pairs, indicating that both the simulated model and our brain’s neural activity can track the semantic properties of words in word sequences with strong semantic correlations. However, the simulated model did not show a significant 1 Hz peak in the living-nonliving pairs, whereas the EEG result showed a significant 1 Hz peak in that condition, suggesting that our brain can still track the semantic properties of words though the words in the word sequences were semantically weakly correlated. This further suggested that our brain might also utilize memory or other cognitive resources except for the direct semantic relations between words in order to capture the semantic correlations of words in word sequences (e.g. Frisby et al., 2023; Kowialiewski et al., 2023). The model, however, had difficulties in tracking the semantic properties of words in word sequences with weak semantic correlations. Such a difference indicated that the model was not efficient as human brain in the integration of weak semantic relations of words in word sequences.

4.3.2. *Limitations of the model in relying too heavily on structural information*

Both the simulated model and the EEG results showed a significant 1 Hz peak in the NV-sensical pairs, indicating that both the simulated model and our brain can track the syntactic structural information of word sequences with strong semantic correlations. However, only the simulated model further showed a significant 1 Hz peak in the NV-nonsensical pairs, suggesting that the simulated model might rely too heavily on structural information so that it still predicted a significant 1 Hz peak even when the semantic correlations of the word sequences were rather weak. In other words, the model would predict a significant 1 Hz peak as long as the structure is ‘subject + predicate’, indicating that

[8] We thank an anonymous reviewer for this idea of ‘jabberwocky’ as a purely syntactic condition and for pointing out the reference by Rafferty et al. (2023).

the model was not as flexible as the brain in processing structural information, especially in integrating syntactic structural information of word sequences that had weak semantic correlations (e.g. Cai, et al., 2024).

The simulated model also predicted that the response power of the 1 Hz peak in the NV-sensical pairs was significantly higher than that in the NV-nonsensical pairs, suggesting that the response power evoked in tracking such syntactic structure was also mediated by the strength of the semantic correlations of the word sequences. In addition, the significant 2 Hz peaks in all the four conditions in the EEG experiment as well as in the model prediction indicated that both our brain and the model were sensitive to the word boundary information.

5. Conclusion

The current study, through a fine-grained design on the linguistic stimuli, explored how our brain integrates single word information into larger linguistic units by conducting an EEG experiment using the frequency-tagging paradigm as well as by constructing a model simulation based on the surprisal values of the pretrained language model GPT-2. In terms of the neural tracking of semantic information of words, the EEG results showed that low-frequency neural activity can track the semantic properties of words, and changing the strength of the semantic correlations of word sequences did not lead to significant response power differences. In addition, the EEG results showed that the cortical activity can also track syntactic structural information, but such a tracking process was mediated by the strength of the semantic correlations of the word sequences. Compared with the EEG results, the simulated model based on GPT-2 seems to have limitations in tracking word sequences that have weak semantic correlations and in relying too heavily on structural information. Taken together, both the EEG experiment and the model simulation results suggest that low-frequency neural activity tracks syntactic information through semantic mediation, supporting an integral point of view that cortical activity tracks syntactic information, but the response power evoked in tracking syntactic information is mediated by semantic information.

Data availability.

All the test materials, data and analysis codes for the study are available in Open Science Framework via the link:

https://osf.io/xhqkd/?view_only

CRediT authorship contribution statement

Yuan Xie: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Peng Zhou:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Likan Zhan:** Methodology, Formal analysis, Conceptualization. **Yanan Xue:** Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to sincerely thank Yuhan Lu for his helpful suggestions and valuable advice on statistical analyses.

Data availability

I have shared the link to my data/code at the Attach File step.

References

Bai, F., Meyer, A. S., & Martin, A. E. (2022). Neural dynamics differentially encode phrases and sentences during spoken language comprehension. *PLoS Biology, 20*(7), Article e3001713. https://doi.org/10.1371/journal.pbio.3001713

Barber, H., & Carreiras, M. (2003). Integrating gender and number information in Spanish word pairs: An ERP study. *Cortex, 39*(3), 465–482. https://doi.org/10.1016/S0010-9452(08)70259-4

Barber, H., & Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of cognitive neuroscience, 17*(1), 137–153. https://doi.org/10.1162/0898929052880101

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience, 31*(8), 2801–2814. https://doi.org/10.1523/JNEUROSCI.5003-10.2011

Burton, M. W., Krebs-Noble, D., Gullapalli, R. P., & Berndt, R. S. (2009). Functional neuroimaging of grammatical class: Ambiguous and unambiguous nouns and verbs. *Cognitive Neuropsychology, 26*(2), 148–171. https://doi.org/10.1080/02643290802536090

Cai, Z., Duan, X., Haslett, D., Wang, S., & Pickering, M. (2024). Do large language models resemble humans in language use?. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 37-56). Bangkok, Thailand. https://aclanthology.org/2024.cmcl-1.4.

Carreiras, M., Carr, L., Barber, H. A., & Hernandez, A. (2010). Where syntax meets math: Right intraparietal sulcus activation in response to grammatical number agreement violations. *NeuroImage, 49*(2), 1741–1749. https://doi.org/10.1016/j.neuroimage.2009.09.058

Chen, L., Li, W., Shi, X., & Han, M. (2022). Cognitive processing differences between stereotype activation and semantic activation. *Applied Neuropsychology: Adult, 1–11*. https://doi.org/10.1080/23279095.2022.2145199

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*. https://doi.org/10.48550/arXiv.2402.03216.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience, 11*, 481. 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. *Journal of Neuroscience, 38*(5), 1178–1188. https://doi.org/10.1523/JNEUROSCI.2606-17.2017

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. *CRC Press*. https://doi.org/10.1201/9780429246593

Electrical Geodesics, Inc. (2007). *Geodesic Sensor Net Technical Manual*. https://www.documents.philips.com/assets/20180705/6f388e7ade4d41e38ad5a91401755b6f.pdf.

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *PloS One, 13*(5), Article e0197304. https://doi.org/10.1371/journal.pone.0197304

Frisby, S. L., Halai, A. D., Cox, C. R., Ralph, M. A. L., & Rogers, T. T. (2023). Decoding semantic representations in mind and brain. *Trends in cognitive sciences, 27*(3), 258–281. https://doi.org/10.1016/j.tics.2022.12.006

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Hasting, A. S., & Kotz, S. A. (2008). Speeding up syntax: On the relative timing and automaticity of local phrase structure and morphosyntactic processing as reflected in event-related brain potentials. *Journal of Cognitive Neuroscience, 20*(7), 1207–1219. https://doi.org/10.1162/jocn.2008.20083

Herrmann, B., Obleser, J., Kalberlah, C., Haynes, J. D., & Friederici, A. D. (2012). Dissociable neural imprints of perception and grammar in auditory functional imaging. *Human Brain Mapping, 33*(3), 584–595. https://doi.org/10.1002/hbm.21235

Iwabuchi, T., Nakajima, Y., & Makuuchi, M. (2019). Neural architecture of human language: Hierarchical structure building is independent from working memory. *Neuropsychologia, 132*, Article 107137. https://doi.org/10.1016/j.neuropsychologia.2019.107137

Jakić, M., Filipović-Đurđević, D., & Kostić, A. (2011). The facilitation effect of associative and semantic relatedness in word recognition. *Psihologija, 44*(4), 367–385. https://doi.org/10.2298/PSI1104367J

Jeon, H. A., Lee, K. M., Kim, Y. B., & Cho, Z. H. (2009). Neural substrates of semantic relationships: Common and distinct left-frontal activities for generation of synonyms vs. antonyms. *Neuroimage, 48*(2), 449–457. https://doi.org/10.1016/j.neuroimage.2009.06.049

Jin, P., Lu, Y., & Ding, N. (2020). Low-frequency neural activity reflects rule-based chunking during speech listening. *Elife, 9*, Article e55613. https://doi.org/10.7554/eLife.55613

Kang, A. M., Constable, R. T., Gore, J. C., & Avrutin, S. (1999). An event-related fMRI study of implicit phrase-level syntactic and semantic processing. *Neuroimage, 10*(5), 555–561. https://doi.org/10.1006/nimg.1999.0493

Kowialiewski, B., Krasnoff, J., Mizrak, E., & Oberauer, K. (2023). Verbal working memory encodes phonological and semantic information differently. *Cognition, 233*, Article 105364. https://doi.org/10.1016/j.cognition.2022.105364

Li, J., Lai, M., & Pylkkänen, L. (2024). Semantic composition in experimental and naturalistic paradigms. *Imaging Neuroscience*. https://doi.org/10.1162/imag_a_00072

Lo, C. W., Tung, T. Y., Ke, A. H., & Brennan, J. R. (2022). Hierarchy, not lexical regularity, modulates low-frequency neural synchrony during language comprehension. *Neurobiology of Language, 3*(4), 538–555. https://doi.org/10.1162/nol_a_00077

Lu, Y., Jin, P., Pan, X., & Ding, N. (2022). Delta-band neural activity primarily tracks sentences instead of semantic properties of words. *NeuroImage, 251*, Article 118979. https://doi.org/10.1016/j.neuroimage.2022.118979

Lu, Y., Jin, P., Ding, N., & Tian, X. (2023). Delta-band neural tracking primarily reflects rule-based chunking instead of semantic relatedness between words. *Cerebral Cortex, 33*(8), 4448–4458. https://doi.org/10.1093/cercor/bhac354

Mandler, G., Hamson, C. O., & Dorfman, J. (1990). Tests of dual process theory: Word priming and recognition. *The Quarterly Journal of Experimental Psychology Section A, 42*(4), 713–739. https://doi.org/10.1080/14640749008401246

Maran, M., Friederici, A. D., & Zaccarella, E. (2022a). Syntax through the looking glass: A review on two-word linguistic processing across behavioral, neuroimaging and neurostimulation studies. *Neuroscience & Biobehavioral Reviews, 104881*. https://doi.org/10.1016/j.neubiorev.2022.104881

Maran, M., Numssen, O., Hartwigsen, G., & Zaccarella, E. (2022b). Online neurostimulation of Broca's area does not interfere with syntactic predictions: A combined TMS-EEG approach to basic linguistic combination. *Frontiers in psychology, 13*, Article 968836. https://doi.org/10.3389/fpsyg.2022.968836

Martin, A. E., & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology, 15*(3), Article e2000663. https://doi.org/10.1371/journal.pbio.2000663

Münte, T. F., Heinze, H. J., & Mangun, G. R. (1993). Dissociation of brain activity related to syntactic and semantic aspects of language. *Journal of cognitive neuroscience, 5*(3), 335–344. https://doi.org/10.1162/jocn.1993.5.3.335

Niemi, P., Vauras, M., & von Wright, J. (1980). Semantic activation due to synonym, antonym, and rhyme production. *Scandinavian Journal of Psychology, 21*(1), 103–107. https://doi.org/10.1111/j.1467-9450.1980.tb00347.x

Paradis, C., Willners, C., & Jones, S. (2009). Good and bad opposites: Using textual and experimental techniques to measure antonym canonicity. *The Mental Lexicon, 4*(3), 380–429. https://doi.org/10.1075/ml.4.3.04par

Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychological research, 66*(3), 180–194. https://doi.org/10.1007/s00426-002-0086-5

Quinn, W. M., & Kinoshita, S. (2008). Congruence effect in semantic categorization with masked primes with narrow and broad categories. *Journal of Memory and Language, 58*(2), 286–306. https://doi.org/10.1016/j.jml.2007.03.004

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9. 19. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe.

Rafferty, M. B., Saltuklaroglu, T., Reilly, K., Paek, E. J., & Casenhiser, D. M. (2023). Neural synchrony reflects closure of jabberwocky noun phrases but not predictable pseudoword sequences. *European Journal of Neuroscience, 57*(11), 1834–1847. https://doi.org/10.1111/ejn.15982

Raposo, A., Mendes, M., & Marques, J. F. (2012). The hierarchical organization of semantic memory: Executive function in the processing of superordinate concepts. *Neuroimage, 59*(2), 1870–1878. https://doi.org/10.1016/j.neuroimage.2011.08.072

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex, 96*, 105–120. https://doi.org/10.1016/j.cortex.2017.09.002

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sinai, A., & Pratt, H. (2002). Electrophysiological evidence for priming in response to words and pseudowords in first and second language. *Brain and Language, 80*(2), 240–252. https://doi.org/10.1006/brln.2001.2597

Tsigka, S., Papadelis, C., Braun, C., & Miceli, G. (2014). Distinguishable neural correlates of verbs and nouns: A MEG study on homonyms. *Neuropsychologia, 54*, 87–97. https://doi.org/10.1016/j.neuropsychologia.2013.12.018

Was, C., Woltz, D., & Hirsch, D. (2019). Memory processes underlying long-term semantic priming. *Memory & cognition, 47*, 313–325. https://doi.org/10.3758/s13421-018-0867-8

Xu, L., Zhang, X., & Dong, Q. (2020). CLUECorpus2020: A large-scale Chinese corpus for pre-training language model. *arXiv preprint arXiv:2003.01355*. https://doi.org/10.48550/arXiv.2003.01355.

Zaccarella, E., & Friederici, A. D. (2015). Merge in the human brain: A sub-region based functional investigation in the left pars opercularis. *Frontiers in psychology, 6*, 1818. https://doi.org/10.3389/fpsyg.2015.01818

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., ... & Du, X. (2019). UER: An open-source toolkit for pre-training models. *arXiv preprint arXiv:1909.05658*. https://doi.org/10.48550/arXiv.1909.05658.

Zhuang, T., Kabulska, Z., & Lingnau, A. (2023). The representation of observed actions at the subordinate, basic, and superordinate level. *Journal of Neuroscience, 43*(48), 8219–8230. https://doi.org/10.1523/JNEUROSCI.0700-22.2023