scientific data

OPEN

DATA DESCRIPTOR

TMNRED, A Chinese Language EEG Dataset for Fuzzy Semantic Target Identification in Natural Reading Environments

Yanru Bai[1,2,3,5] ✉, Qi Tang[1,3,5], Ran Zhao[1,3], Hongxing Liu[1], Shuming Zhang[1], Mingkun Guo[1], Minghan Guo[1], Junjie Wang[1], Changjian Wang[4], Mu Xing[1], Guangjian Ni[1,2,3] ✉ & Dong Ming[1,2,3]

Semantic understanding is central to advanced cognitive functions, and the mechanisms by which the brain processes language information are still being explored. Existing EEG datasets often lack natural reading data specific to Chinese, limiting research on Chinese semantic decoding and natural language processing. This study aims to construct a Chinese natural reading EEG dataset, TMNRED, for semantic target identification in natural reading environments. TMNRED was collected from 30 participants reading sentences sourced from public internet resources and media reports. Each participant underwent 400–450 trials in a single day, resulting in a dataset with over 10 hours of continuous EEG data and more than 4000 trials. This dataset provides valuable physiological data for studying Chinese semantics and developing more accurate Chinese natural language processing models.

Background & Summary

Language is a unique advanced cognitive function of humans, and semantic recognition is one of the important ways to explore the essence of humanity[1]. Moreover, semantic understanding serves as the foundation for human thought and behavior[2]. The human brain can quickly comprehend linguistic information and generate corresponding verbal expressions, showcasing its complex processing capabilities[3,4]. When subjected to linguistic stimuli, the brain encodes semantic information through neural activities, and analyzing these neural activities can reveal the mechanisms of semantic encoding in the brain. Since semantic tasks typically involve the selective retrieval of conceptual knowledge to establish meaningful connections between experimental materials, this retrieval based on cognitive control to meet experimental requirements constitutes an important aspect of semantic understanding tasks[5]. By recording and analyzing electroencephalogram (EEG) data, it is possible to capture the brain's dynamic responses to complex semantics during natural reading, which will aid in uncovering how the brain processes and understands information in uncertain and polysemous contexts[6,7].

Despite the relative abundance of EEG datasets for natural visual stimuli, datasets for natural linguistic stimuli remain scarce. Currently, only a few EEG datasets related to language exist, such as the ZuCo dataset[8] and the BraVL dataset[9]. However, these datasets are predominantly collected using English corpora, significantly limiting the study of neural representations of other languages, such as Chinese. Research indicates that the brain processes different languages through distinct mechanisms. Compared to English, the brain's responses to Chinese exhibit unique characteristics. Chinese differs from English not only in structure—as an ideographic writing system where each character contains information on form, sound, and meaning—but also in semantics, with its extensive presence of polysemous words and homophones making the identification of fuzzy semantic targets in Chinese contexts more complex and challenging[10,11]. Moreover, the syntactic structures and word order rules of Chinese differ markedly from those of English, imposing higher demands on the brain's language processing mechanisms[12–14]. Recent studies have shown that while Chinese and English native speakers

[1]Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, 300072, China. [2]Tianjin Key Laboratory of Brain Science and Neuroengineering, Tianjin, 300072, China. [3]Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration, Tianjin, 300392, China. [4]National University of Defense Technology, Changsha, Hunan, 410000, China. [5]These authors contributed equally: Yanru Bai, Qi Tang. ✉e-mail: yr56_bai@tju.edu.cn; niguangjian@tju.edu.cn

utilize similar brain regions for language processing, they do so in significantly different ways. Ge *et al*.'s research demonstrated that when Chinese native speakers understand speech, Broca's area in the frontal lobe receives information from both anterior parts of the left and right temporal lobes, whereas for English, it receives information from the posterior part of the left temporal cortex[15]. This discovery reveals the information network characteristics of the brain during the processing of different languages, highlighting the special brain network mechanism for processing Chinese. Li *et al*.'s research further elucidates the neural mechanisms underlying the perception of Mandarin Chinese and demonstrates shared local neural encoding mechanisms across languages[16]. EEG datasets based on Chinese corpus stimuli can help uncover commonalities and differences in language processing across languages, providing new perspectives for understanding language processing mechanisms. This not only enriches research in neurolinguistics but also offers a solid data foundation for practical applications, such as developing natural language processing algorithms better suited to Chinese, improving brain-computer interface technologies, or enhancing language learning tools for education[17–19].

During natural reading, information processing is carried out at a natural pace and manner, bringing brain activity closer to the context of everyday language understanding and usage. Studying natural reading allows for a more accurate understanding of how the brain parses and comprehends linguistic information in uncontrolled, complex contexts[20]. Traditional language research often employs simplified experimental designs, such as the presentation of individual words, which, although capable of controlling variables, may overlook the complexity and multi-layered nature of language processing[21]. The importance of natural reading research lies firstly in its ability to provide a comprehensive understanding of processing fuzzy semantics, reflecting the integrated capabilities of language processing. Artificial interventions may introduce experimenter or subject effects, affecting the brain's natural responses[22–24]. For example, instructing subjects to read at a specific speed or to dwell on certain words may alter their reading strategies and brain activity patterns. By allowing subjects to read in a natural state, it is possible to capture brain activity in real-life situations, thereby obtaining more representative and generalizable research results, enabling us to more accurately understand how the brain operates in natural language environments[25–27].

In this article, we present an EEG dataset of 30 subjects while they naturally read sentences, with each participant undergoing between 400 and 450 trials in a single day's recording, resulting in a dataset with over 10 hours of continuous EEG data, encompassing more than 4,000 trials. This dataset provides researchers with rich physiological data to advance the training of natural language processing (NLP) applications. The reading materials in the TMNRED dataset include content from publicly available internet resources such as Baidu Baike and WeChat public account articles, as well as news reports from media sources. These sentences were presented to the subjects in a natural reading scenario, with complete sentences displayed on the screen for the subjects to read at their own pace. Traditionally, EEG studies on reading associations often proceed through word-by-word presentation, which excludes many important aspects of normal reading processes. Preliminary experimental results indicate that these data hold significant potential for applications in natural language processing. For example, researchers can use differences in EEG signals between target and non-target sentences to analyze the neural mechanisms underlying fuzzy semantic processing or explore the integration of visual and language information in attention allocation. The dataset can also be combined with other imaging techniques, such as fMRI, to validate cross-modal consistency of brain activation[28–30]. In summary, the high quality and versatility of this dataset make it a valuable resource for future cognitive neuroscience research, with broad applications in areas such as language processing, visual processing, and brain network analysis.

Methods

Participants. All participants provided written consent to participate in the study. The experimental procedures and paradigms were approved by the Ethics Committee of Tianjin University (No: TJUE-2024-402). Thirty healthy right-handed participants, aged between 18 and 30 years old(averaged 22.07 years old, std = 2.7 years), including eighteen females and twelve males. All participants are right-handed and have normal or corrected-to-normal vision, with no hearing loss, no speech loss, and with no neurological, movement, or psychiatric disorders, joined the experiment and gave their written informed consent. All participants were native Chinese speakers. Each participant completed 8 blocks of 400 trials in total, and participated in approximately one hour of recording. In this work, the participants are identified by aliases "Sub1" through "Sub30". Detailed information about the participants can be found in Table 1.

Materials. The stimulus materials consist of text ranging from 15 to 20 characters, presented in the form of news headlines or short sentences. The stimulus materials are divided into two types: target semantic materials and non-target semantic materials (non-target materials). To ensure a certain level of coverage, target materials were selected from four types of semantics: names, means of transportation, animals, and fruits. Each category is represented by multiple target words; for example, the "means of transportation" category includes "train (火车)", "airplane (飞机)", "ship (轮船)", "bicycle (自行车)" and "high-speed rail (高铁)" as target words. The experimental material categories and specific target words are summarized in Table 2. Names, as a common and significant semantic category, have a high degree of personal relevance and social significance. They not only frequently appear in daily communication but are also closely related to various cognitive processes such as memory and emotion. Using names in the experiment allows for observing the brain's responses when processing information related to individual identity and social relationships, providing valuable data on social cognition and memory mechanisms. Secondly, fruits, as a specific object category, have distinct sensory characteristics and rich semantic associations. The sensory attributes of fruits, such as color, taste, and shape, make them play an important role in visual and gustatory cognition. By using fruits as target semantic materials, it is possible to explore the neural mechanisms involved in processing specific objects and sensory information, which is of great significance for understanding object recognition and sensory integration processes. Animals, as another specific object category,

Participant	Gender	Age
Sub1	Male	30
Sub2	Female	21
Sub3	Female	24
Sub4	Male	26
Sub5	Male	24
Sub6	Male	23
Sub7	Male	23
Sub8	Male	23
Sub9	Female	24
Sub10	Female	23
Sub11	Female	23
Sub12	Female	23
Sub13	Male	23
Sub14	Female	23
Sub15	Male	18
Sub16	Male	18
Sub17	Female	21
Sub18	Male	22
Sub19	Female	21
Sub20	Male	18
Sub21	Male	19
Sub22	Female	26
Sub23	Male	20
Sub24	Female	20
Sub25	Female	19
Sub26	Male	23
Sub27	Female	19
Sub28	Male	20
Sub29	Female	24
Sub30	Female	21

Table 1. Participants information.

Category	Specific Contents	Quantity
Names	Trump, Charles 特朗普、查尔斯	50*2
Means of Transportation	train, airplane, ship, bicycle, high-speed rail 火车、飞机、轮船、自行车、高铁	50*2
Animals	parrot, hedgehog, white crane, steed, kitten 鹦鹉、刺猬、白鹤、骏马、小猫	50*2
Fruits	watermelon, pineapple, tangerine, strawberry, grape 西瓜、菠萝、柑桔、草莓、葡萄	50*2

Table 2. Experimental Material Categories and Specific Contents.

not only play an important role in natural environments but also have a profound impact on human culture and language. Selecting animals as experimental materials helps to study the brain's responses when processing biological information and information related to natural environments. This is of great value for understanding biological cognition and ecological cognition processes. Means of transportation, as a functionally strong and closely related to daily life object category, have unique semantics and functional characteristics. It allows for exploring the neural activities of the brain when processing functional objects and information related to movement and transportation. This is of great significance for understanding functional cognition and spatial cognition processes.

When designing experimental materials, we consider the balance between target strength and response frequency in selecting target words to ensure the reliability and validity of the experimental results. Target strength refers to the stimulus intensity or salience of the target words to the participants. In the experiment, the salience of the target words will affect the brain's response intensity. The target strength of the three categories of words—fruits, means of transportation, and animals—is relatively consistent. Therefore, selecting 5 target words for each category can effectively balance the stimulus intensity and avoid adaptation or fatigue effects caused by the frequent occurrence of a single word. Response frequency involves the number of times the brain responds

Target Word Category	Material Examples
Target-Names	Trump was elected the 45th President in U.S. history in 2016. 特朗普于2016年当选美国历史上第45任总统
	Charles has an almost obsessive passion for charitable causes. 查尔斯对于慈善事业有着近乎偏执的热情
Target-Means of Transportation	The graceful dance of the white crane is so much like that of a "ballet master." 白鹤优雅的舞姿多么像"芭蕾舞大师"
	The feathers shed by parrots, when collected, can be made into pillows. 鹦鹉掉的羽毛收集起来就能做成枕头
Target-Animals	A few leaves hanging on the grape stem are like little parasols. 葡萄梗上挂着几片像遮阳伞一样的叶子
	One can tell if a watermelon is ripe by tapping it with a finger. 西瓜通过手指弹一弹就可以判断是否成熟
Target-Fruits	When the airplane takes off, there is no vibration at all. 飞机起飞的时候，一点儿也不震动
	The ship slowly departs from the harbor, disappearing on the distant horizon. 轮船慢慢地驶离海港，消失在远方的地平线上
Non-target	Accelerate the construction of a new development pattern where domestic and international dual cycles reinforce each other. 加快构建双循环相互促进的新发展格局
	Focus wisdom and strength on achieving the set goals. 把智慧和力量凝聚到实现确定的目标上

Table 3. Examples of Experimental Materials with Target Semantic Words.

Category	Average Character Count (Mean ± SD)	Average Word Frequency (Mean ± SD)	Average Stroke Count (Mean ± SD)
Target-Names	17.0 ± 0.5	13.2 ± 1.1	10.4 ± 0.6
Target-Means of Transportation	16.5 ± 0.6	12.8 ± 1.0	10.7 ± 0.5
Target-Animals	17.0 ± 0.5	13.0 ± 1.2	11.1 ± 0.7
Target-Fruits	16.8 ± 0.4	12.5 ± 1.1	10.6 ± 0.5
Non-Target	17.0 ± 0.6	13.1 ± 1.0	10.9 ± 0.6

Table 4. The statistical analysis of the experimental materials.

to a particular target word. In the experiment, repeating the same target word multiple times can improve the statistical reliability and validity of the data. For the choice of names, selecting two celebrities ensures that there are enough data points to detect and analyze the brain's response to this specific target word. By controlling the number and occurrence of names, we can avoid variable interference caused by differences in the celebrities' popularity, thereby more accurately analyzing the brain's response to the target words. In the categories of fruits, means of transportation, and animals, multiple target words are chosen, while in the category of names, fewer but more frequently appearing target words are selected. This experimental design balances target strength and response frequency, ensuring the reliability and validity of the data. This strategy helps obtain high-quality EEG data, leading to a deeper understanding of the brain's mechanisms in semantic processing, and facilitates the conduct of further in-depth research.

The selection of these categories was made with careful consideration to ensure the coverage and representativeness of the experiment and to effectively elicit semantic processing responses from the subjects. Additionally, to ensure the reliability and validity of the experimental results, the experimental materials were primarily selected from online resources such as Baidu Encyclopedia, "People" magazine, and WeChat public articles. These resources provide rich and diverse semantic information, making the experimental materials highly ecologically valid and representative. The reading materials in the TMNRED dataset include content from publicly available internet resources such as Baidu Baike and WeChat public account articles, as well as news reports from media sources. The choice of neutral materials was mainly to avoid interference from individual subject preferences on their EEG characteristics, thereby ensuring the objectivity and consistency of the experimental results. Examples of experimental materials containing target words and non-target words for each category are shown in Table 3.

To ensure balance at the sentence level, we carefully controlled and statistically validated sentence length (number of characters), overall word frequency, and stroke count. We calculated the average word frequency and stroke count for all words within each sentence. Results from independent-samples t-tests showed no significant differences in character count ($p > 0.05$), word frequency ($p > 0.05$), or stroke count ($p > 0.05$) between target and non-target sentences, confirming that the experimental materials were balanced in terms of visual complexity and semantic distribution. The statistical analysis of the experimental materials is presented in Table 4. No significant differences were found between target and non-target sentences in terms of average character count, word frequency, or stroke count ($p > 0.05$), indicating that the experimental conditions were balanced at the sentence level.

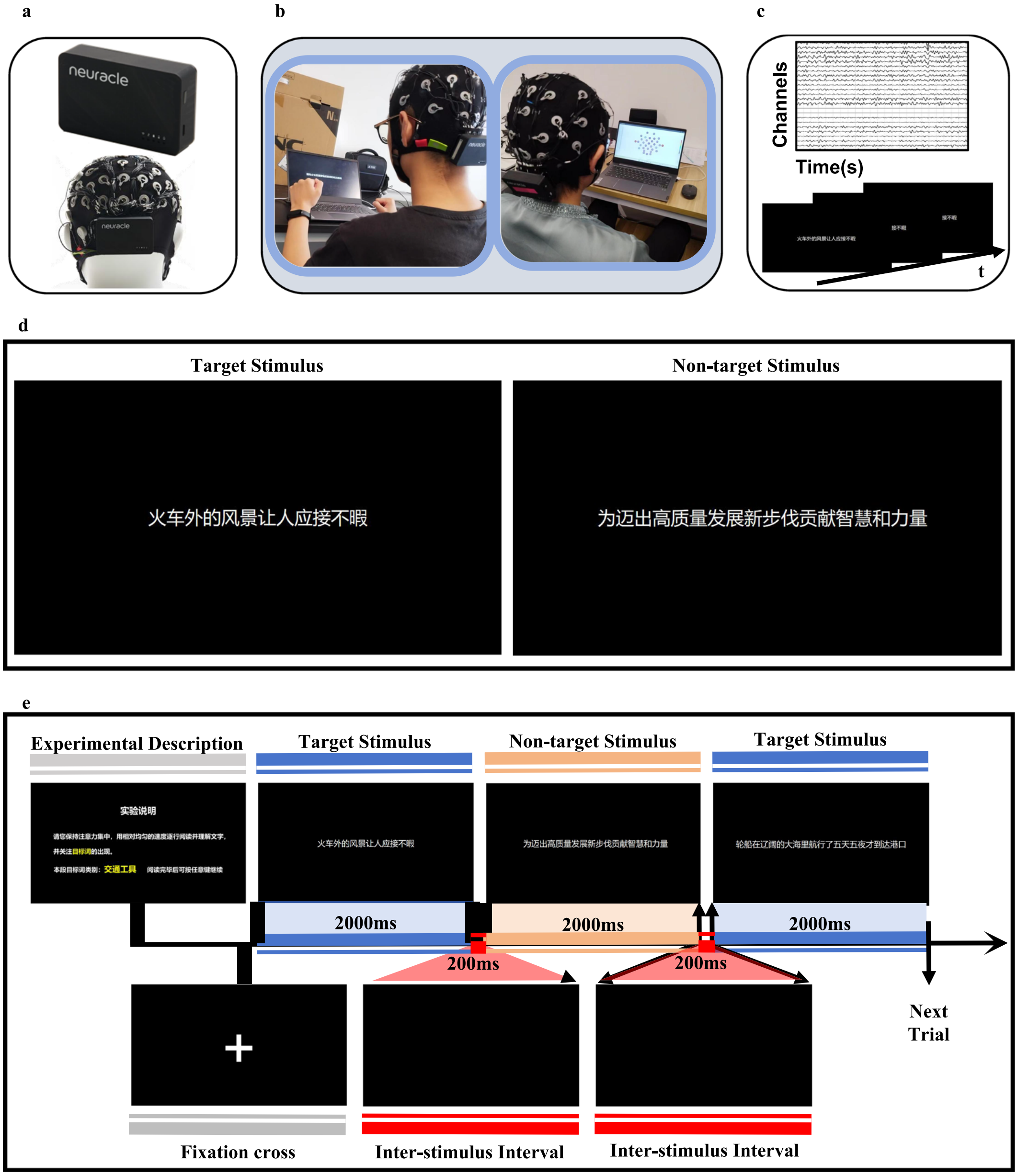


Fig. 1 Overview of the experiment. (**a**) Equipments. The portable brain signal amplifier and EEG cap from neuracle, as well as the NeuSen W series wireless EEG acquisition system were used. (**b**) The experiment setup. Participants were instructed to sit quietly approximately 60 cm from the screen and sequentially read the text. (**c**) The experimental protocol. Participants' 32-channel EEG signals data were recorded while reading the text. (**d**) Target stimulus and non-target stimulu of the experiment.The stimulus protocol was designed using Psychtoolbox-3 running on Matlab and presented on the computer screen, with target word materials and non-target word materials. (**e**) Single block structure of the experiment. Each block could complete 50 trials, with a 0.2-second interval between each trial, meaning that 50 sets of 2.2-second EEG signal data could be collected from a single subject.

EEG data collection. The study was conducted in an electrically shielded room. The participants were seated in a comfortable chair in front of a computer screen where the visual cues were presented. In order to familiarize the participant with the experimental procedure and the room environment, all steps of the experiment were explained, while the EEG headcap and the external electrodes were placed. The setup process took approximately 45 minutes. Figure 1 shows the main experiment setup.

In this experiment, the portable brain signal amplifier and EEG cap from neuracle, as well as the NeuSen W series wireless EEG acquisition system[31], were used, as shown in Fig. 1a. We utilized a portable 32-channel EEG system (30 effective channels), which offers significant advantages in portability and adaptability to experimental environments, especially for large-scale data collection. Despite its lower spatial resolution compared to high-density systems, studies[32,33] have shown that reliable source localization is achievable with low-density EEG systems when combined with advanced signal processing algorithms such as minimum norm estimation (MNE) or sparse source localization methods. These features make it particularly suitable for naturalistic experiments, where higher-density systems might introduce practical constraints.

Participants were asked to sit in front of a computer screen on a chair with adjustable height, backrest, and armrests, with visual cues displayed on the screen. The distance between the eyes and the computer screen was approximately 60 centimeters (resolution: 1,920 × 1,080 pixels, refresh rate: 60 Hz), and adjustments were made during the experiment based on the participant's height and personal comfort. The center of the monitor was aligned with or slightly below the eye level to reduce neck and eye fatigue, as shown in Fig. 1b. To familiarize the subjects with the experimental procedure and indoor environment, all steps of the experiment were explained while placing the EEG cap and external electrodes. The setup process took about 45 minutes. Fig. 1 shows the main experimental setup. These devices provided reliable data acquisition assurance for the experiment with their high precision and portability. During the experiment, the text content was displayed on the screen, with each sentence appearing sequentially, ensuring a consistent and comfortable reading experience, as shown in Fig. 1c. In this study, we designed a reading task in which participants read sentences containing target words within a 2-second window. The target words belonged to four semantic categories: names, means of transportation, animals, and fruits. The position of the target words within the sentences was randomized to avoid systematic positional effects (e.g., always appearing at the beginning or end). Additionally, the target sentences were designed to evoke fuzzy semantic processing, where target words exhibit ambiguity, multiple interpretations, or context dependency, reflecting real-world language comprehension scenarios beyond isolated word recognition. The division of the 2-second window into time intervals aimed to explore the temporal dynamics of fuzzy semantic processing.

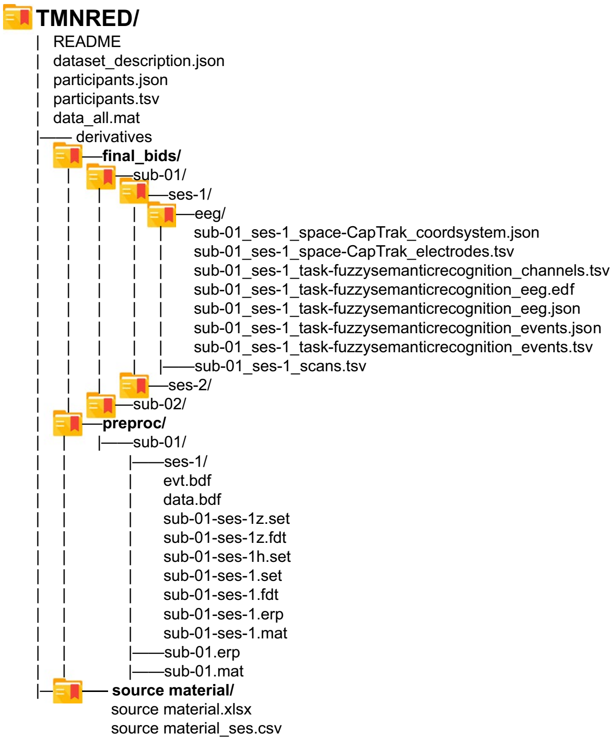
Experimental design. The stimulus protocol was designed using Psychtoolbox-3 running on Matlab and presented on the computer screen, with target word materials and non-target word materials shown in Fig. 1d. Participants were asked to read the sentences and instructed to keep their heads still, focusing their eyes on the Chinese characters on the screen, and read at the speed set by the program.

To prevent dizziness and eye fatigue, the experimental design used a black background with white text, and each trial presented a sentence. When designing the experiment, the number of characters in each stimulus material was controlled between 15 and 20, presented as a short sentence and limited to a single line. Multi-line text could lead to frequent eye movements during reading, increasing the burden on the eyes and cognitive load. Single-line text, however, could keep the gaze relatively fixed, reducing the frequency of eye movements and further lowering the risk of eye fatigue. This design not only improved the participants' reading efficiency but also ensured they could concentrate and accurately understand the content in each trial[34–36]. This range of character counts ensured an appropriate amount of information, allowing participants to easily understand and process the information in a short time. The structure of short sentences is usually simple and grammatically clear, which helps reduce cognitive load, enabling participants to focus more on the content itself. Additionally, controlling the number of characters ensured consistency in all stimulus materials, avoiding variations in reading time and comprehension difficulty due to differences in character count, thus maintaining consistent experimental conditions. Based on the average reading speed of normal individuals and the mechanisms of EEG signal formation, the presentation time for each stimulus material was controlled at 2 seconds. Preliminary experimental results also indicated that participants could typically complete reading and understanding a 15 to 20 character short sentence within 2 seconds in a natural reading scenario. This result validated that a 2-second presentation time was reasonable and effective. EEG signals, when processing linguistic information, usually form significant signal changes within a few hundred milliseconds, and a 2-second presentation time was sufficient to cover the entire process of semantic judgment, ensuring the capture of complete cognitive responses. By controlling the number of characters and presentation time, it was possible to reduce participants' reading fatigue, avoid distraction and increased cognitive load due to fatigue, thereby improving the accuracy and reliability of the data[37].

In the experiment, stimulus materials from four target semantic categories were randomly divided into 2 blocks each, resulting in a total of 8 blocks. Each block contained 50 stimuli, with 15 target word trials and 35 non-target word trials (in a 3:7 ratio), ensuring a balanced distribution. Within each block, the stimuli appeared in a pseudo-random order to prevent consecutive occurrences of the same type, reducing cognitive biases. Each block facilitated 50 trials with a 0.2-second interval between them, allowing for 50 sets of 2.2-second EEG data per subject. Throughout the experiment, all presented stimuli were unique and did not repeat.

The entire experiment consisted of 50*8 = 400 trials, 400 stimulus materials, with 120 being target semantic materials and 280 being non-target semantic materials. The stimulus materials of a single block were presented randomly until all 50 materials within the block were traversed, with a 0.2-second blank frame connecting the presentation of each pair of materials; the EEG signals (30 channels) and corresponding labels (whether the stimulus material contained a target word, with 0 indicating no appearance and 1 indicating appearance) were recorded during the presentation time of each stimulus material (2 seconds) and the subsequent blank frame time (0.2 seconds); after a short break, the stimulus materials of the next block were played until all 8 blocks were traversed.

Fig. 1e illustrates the experimental flow for a single block for one subject, with each subject required to complete 8 sets of block experiments. Each individual participated in one single recording day comprising eight consecutive sessions. A self-selected break period between sessions, to prevent boredom and fatigue, was given (inter-session break). At the beginning of each session, a fifteen-second baseline was recorded where the participant was instructed to relax and stay as still as possible. Within each session, fifty stimulation runs were presented. Through this design, not only was a sufficient number of trials ensured for statistical analysis, but the ecological validity of the experiment was also enhanced through diverse stimulus materials. The EEG data of



```
TMNRED/
 |  README
 |  dataset_description.json
 |  participants.json
 |  participants.tsv
 |  data_all.mat
 |—— derivatives
 |    —final_bids/
 |    |    —sub-01/
 |    |    |    —ses-1/
 |    |    |    |    —eeg/
 |    |    |    |    |    sub-01_ses-1_space-CapTrak_coordsystem.json
 |    |    |    |    |    sub-01_ses-1_space-CapTrak_electrodes.tsv
 |    |    |    |    |    sub-01_ses-1_task-fuzzysemanticrecognition_channels.tsv
 |    |    |    |    |    sub-01_ses-1_task-fuzzysemanticrecognition_eeg.edf
 |    |    |    |    |    sub-01_ses-1_task-fuzzysemanticrecognition_eeg.json
 |    |    |    |    |    sub-01_ses-1_task-fuzzysemanticrecognition_events.json
 |    |    |    |    |    sub-01_ses-1_task-fuzzysemanticrecognition_events.tsv
 |    |    |    |    |——sub-01_ses-1_scans.tsv
 |    |    |    —ses-2/
 |    |    —sub-02/
 |    —preproc/
 |    |    |——sub-01/
 |    |    |    |——ses-1/
 |    |    |    |    evt.bdf
 |    |    |    |    data.bdf
 |    |    |    |    sub-01-ses-1z.set
 |    |    |    |    sub-01-ses-1z.fdt
 |    |    |    |    sub-01-ses-1h.set
 |    |    |    |    sub-01-ses-1.set
 |    |    |    |    sub-01-ses-1.fdt
 |    |    |    |    sub-01-ses-1.erp
 |    |    |    |    sub-01-ses-1.mat
 |    |    |    |——sub-01.erp
 |    |    |    |——sub-01.mat
 |—— source material/
          source material.xlsx
          source material_ses.csv
```

Fig. 2 Dataset structure, files, and naming.

the subjects during each trial were recorded, and in subsequent analyses, the EEG data were classified and segmented according to the material type labels and stimulus appearance time labels. This detailed recording and classification method could help researchers accurately analyze the impact of different types of stimuli on EEG signals, thereby revealing potential cognitive processes and EEG characteristics[38].

Data Records

The EEG data along with basic personal information (age, gender, native language) are publicly accessible. The full dataset, including the raw and preprocessed EEG data, is publicly accessible via the Openneuro platform (https://doi.org/10.18112/openneuro.ds005383.v1.0.0)[39]. Public data is distributed under the Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/).

Data privacy.

All data are de-identified and participants gave permission for their data to be openly shared as part of the informed consent process.

EEG data organization.

The data are stored in folders by task (https://doi.org/10.18112/openneuro.ds005383.v1.0.0). Combined EEG data can be found in the MATLAB files, one file per subject. The file structure of the dataset is shown in Fig. 2. In addition to the BIDS-formatted EEG data of 30 participants, pre- and post-processed EEG data, and stimulus material records, the main folder of the dataset also contains five additional files: (i) "data-description.json": Provides a description of the dataset and registration details, including location and time. (ii) "participants.tsv": Contains participant information such as gender, age, dominant hand, and native language. (iii) "participants.json": Describes the details of all columns in the "participants.tsv" file. (iv) "README": Provides general information about the dataset, including contact details.

Technical Validation

EEG data preprocessing.

For the raw data collected from the experiment, preprocessing was performed using the EEGLAB toolbox in MATLAB. After loading the raw EEG data, the first step involved the removal of irrelevant channels, such as those referenced to the bilateral mastoids, and the truncation of task-unrelated periods before and after the stimulus sequence to prevent interference with the subsequent independent component analysis (ICA).

To ensure data quality, classic methods for EEG data quality assessment were referenced, and criteria for data quality checks were established at two levels. At the raw data level, for each block, the power spectrum of each channel was calculated before re-referencing. Channels with power spectra greater than twice the standard deviation of the mean power spectrum were marked as bad and supplemented through interpolation from neighboring channels, i.e., replacing bad channel data with the mean of spatially adjacent channels. Of the total 7200 channels (30 subjects * 30 channels * 8 blocks), the percentage of bad channels that required interpolation/removal was 10.10% (727/7200). At the single-trial data segment level, the median variance of each

Level	Indicator	Removal Criteria	Interpolation/Removal Ratio
Raw Data	Channel Power Spectrum	>2 times the standard deviation	10.10% (727/7200)
Single-Trial Data Segment	Median Variance of Each Channel in a Single Trial	>2 times the standard deviation	2.17% (260/12000)
	Median Difference Between Each Channel and Its Mean in a Single Trial	>2 times the standard deviation	

Table 5. Data Quality Inspection Criteria.

channel under a single trial and the median difference between each channel and its mean were taken. Trials with either of these two indicators exceeding twice the standard deviation in all trials were marked and their union was rejected. Of the total 12000 trials (30 subjects * 400 trials), the percentage of rejected trials was 2.17% (260/12000). As shown in Table 5, these are the criteria for inspecting data quality at different levels.

To facilitate the observation and analysis of the time-domain signal waveforms of each channel, the EEG data were re-referenced to the average of all scalp channels. The advantage of average referencing is that it can more easily attenuate noise that is consistently present between electrodes and preserve most of the voltage in specific EEG components. After re-referencing, the EEG data were bandpass filtered between 0.5–80 Hz and notch filtered at 50 Hz to remove noise, artifacts, and irrelevant components, with a finite impulse response digital filter (FIR) being chosen. The lower limit of the bandpass filter was set at 0.5 Hz to remove low-frequency noise, and the upper limit was set at 80 Hz to ensure that EEG signals related to higher brain functions were included within the frequency domain. After filtering, the EEG data were downsampled. Downsampling after filtering can avoid signal distortion caused by aliasing, and by reducing the data length, it saves computation time required for the subsequent ICA process, thereby improving data processing efficiency. Since the physiological information contained in EEG signals is distributed within 100 Hz, downsampling to 200 Hz was performed according to the Shannon sampling theorem. After completing the above processing of the EEG data, independent component analysis was conducted to decompose the EEG data into multiple independent source signals, and then typical noise components were identified and removed based on their characteristics.

EEG signal data are extremely susceptible to interference from various non-neural factors such as environmental conditions, the psychological state of the participants, and additional movements. These noises can severely affect the accuracy of EEG data analysis. We can effectively eliminate induced electrical signals from the experimental environment as well as some bioelectrical signals, such as blinking, eye movements, muscle activity, and skin potentials, through preprocessing to reduce their impact on EEG data analysis. A common approach is to use the EEG data processing toolbox EEGLAB in MATLAB to preprocess the raw data collected from the experiments. This method improves the signal-to-noise ratio of the EEG data, maximizing the removal of low-frequency noise, power line interference, and bioelectric artifacts[40]. To verify the consistency of the signals, we calculated intraclass correlation coefficients (ICCs) across trials, which showed high consistency with an average ICC of 0.87 ($p < 0.001$). For signal quality evaluation, we conducted analyses of SNR, PSD, and test-retest reliability. The SNR for all electrodes exceeded 20 dB, indicating high-quality data. Within the 0.5–40 Hz frequency range, the signals for target and non-target sentences demonstrated stable and consistent power spectral distributions, suggesting that the dataset is well-suited for frequency-domain analysis. The test-retest reliability analysis further confirmed the stability of the data, with an average correlation coefficient of 0.85 ($p < 0.001$).

Classic sensor-level EEG analysis. In this study, to address the potential impact of differences in participants' reading speeds on time series analysis, we applied a standardized time window method for data processing. Specifically, each participant's reading time was normalized to a percentage-based timeline, dividing the entire reading process into four time windows: 0%–25%, 25%–50%, 50%–75%, and 75%–100%. This approach allowed us to analyze the dynamic characteristics of fuzzy semantic processing on a relative time scale, independent of absolute time points. Additionally, we conducted a statistical analysis of participants' average reading times and confirmed through ANOVA that there were no significant differences in reading speeds across participants ($F(1, 29) = 1.25$, $p = 0.32$). This result supports the validity and applicability of the standardized time window method in our study. The core feature of fuzzy semantic processing lies in the dynamic resolution of semantic ambiguity and contextual integration, which exhibits temporal stability on a relative time scale, allowing consistent neural representations across participants with varying reading speeds.

Fig. 3 shows the time-voltage analysis results of four leads: Cz, Oz, C3, and Pz. Our study selected Cz, Oz, C3, and Pz as the primary electrodes for analysis based on the functional characteristics of brain regions involved in semantic processing tasks. The selection of C3 particularly emphasizes the left hemisphere's dominance in language comprehension, validated by comparative analysis with C4 ($t(29) = 3.50$, $p = 0.002$). This result aligns with previous studies, further supporting the critical role of the left Broca's area and Wernicke's area in language processing[41–44]. Additionally, the selection of Cz, Oz, and Pz reflects the importance of attention, visual information processing, and multisensory integration in semantic comprehension. Our findings highlight the critical roles of these electrodes in fuzzy semantic processing and provide methodological insights for future related studies. Fig. 3a shows the brain responses elicited by target and non-target stimuli over time. It can be seen that within 0 to 400 ms, the fluctuations of the two are quite similar, but afterward, the potential of target words is slightly higher. This indicates that the Cz lead shows some differential response in the recognition of target words after initial visual processing. Fig. 3b shows the brain responses elicited by target and non-target stimuli over time. Within the 150 ms to 650 ms time window, significant differences in electrical potentials between

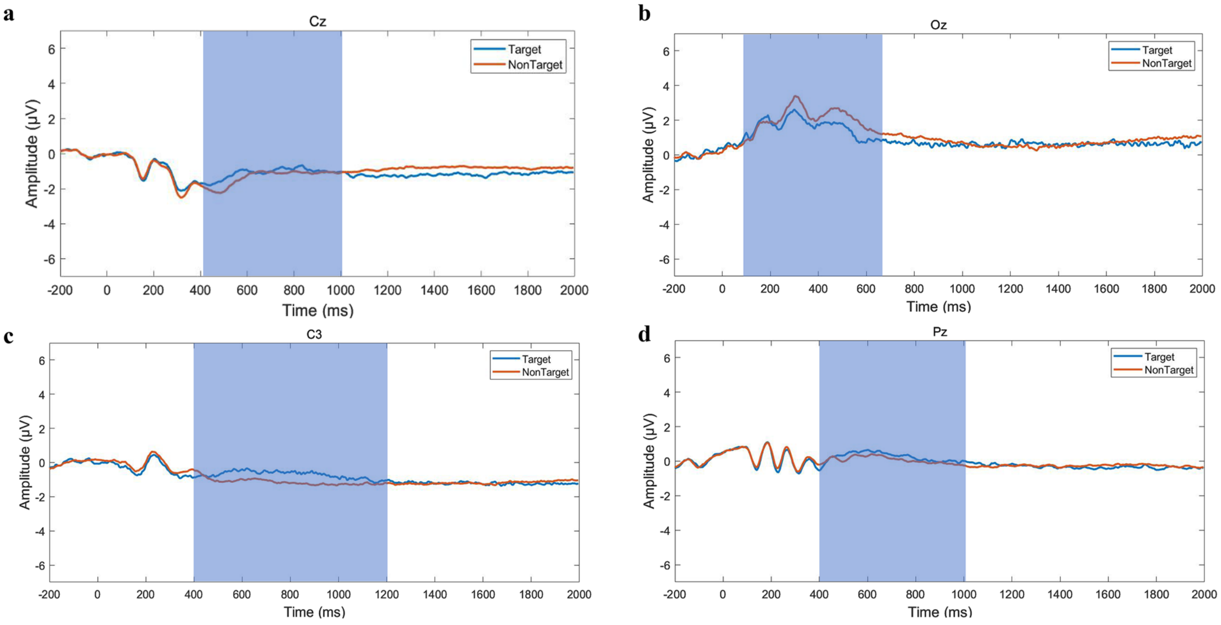


Fig. 3 Time-voltage analysis results of the four leads: Cz, Oz, C3, and Pz. The blue and red curves correspond to the brain responses elicited by target words (Target, hereafter referred to as T) and non-target words (NonTarget, hereafter referred to as nT) stimuli, respectively. The horizontal axis represents time, and the vertical axis represents amplitude. Baseline correction was performed using the pre-stimulus period to avoid voltage shifts caused by factors such as skin hydration or static electricity. (**a**) Cz electrode: EEG responses evoked by stimuli containing target and non-target words. The Cz electrode, located at the central region of the scalp, is typically associated with motor and sensory processing, but it also plays a crucial role in attention and semantic processing tasks. Between 150 ms and 650 ms, the potential for target stimuli was significantly higher than for non-target stimuli (t(29) = 3.85, p = 0.0008, Cohen's d = 0.35). (**b**) Oz electrode: EEG responses evoked by stimuli containing target and non-target words. The Oz electrode, located in the occipital region of the scalp, is highly sensitive to visual stimuli and is primarily responsible for visual information processing. Significant differences in potential were observed between target and non-target stimuli in the 150 ms to 650 ms time window (t(29) = 4.12, p = 0.0003, Cohen's d = 0.45). (**c**) C3 electrode: EEG responses evoked by stimuli containing target and non-target words. The C3 electrode, located in the left central region of the scalp, is commonly associated with motor control and language processing. Within the 150 ms to 650 ms time window, a small but significant difference in potential was observed (t(29) = 2.70, p = 0.011, Cohen's d = 0.28). (**d**) Pz electrode: EEG responses evoked by stimuli containing target and non-target words. The Pz electrode, positioned in the parietal region of the scalp, plays a vital role in attention and information integration. Between 150 ms and 650 ms, the potential for target stimuli was significantly higher than for non-target stimuli (t(29) = 2.95, p = 0.007, Cohen's d = 0.3).

target and non-target words were observed at the Oz electrode. Results of the paired-samples t-test indicated that the average potential for target words (3.20 ± 1.05 μV) was significantly higher than that for non-target words (2.75 ± 0.98 μV) (t(29) = 4.12, $p < 0.001$). The effect size, Cohen's d = 0.45, suggests that the differences between the target and non-target conditions are of moderate strength. These findings highlight the sensitivity of the Oz electrode in capturing visual processing of target stimuli. Additionally, drawing on theoretical support from previous studies[45,46], the Oz electrode is recognized as a key electrode in studies of visual evoked potentials (VEPs). Positioned at the midline occipital region of the scalp, the Oz electrode is instrumental in capturing neural activity associated with visual stimuli. Research demonstrates that the Oz electrode is highly sensitive to visual stimuli, particularly target-related information processing, with its activity primarily associated with responses from the primary visual cortex. This observation further supports our conclusion regarding the critical role of the Oz electrode in visual processing. Fig. 3c shows that while the potential changes of target and non-target words are quite similar for most of the time, between 400 ms and 1200 ms, the potential of target words is slightly higher. This may reflect some specific responses of the C3 lead in language processing or semantic integration of target words. Fig. 3d shows that between 200 ms and 400 ms, the potential changes of target words are similar to those of non-target words, but slightly higher afterward. This indicates that the Pz lead plays a role in higher-level integration and guiding attention to target information. The selection of these leads and their time-voltage analysis results show that different brain regions exhibit different response characteristics in the process of visual semantic text target recognition. The Oz lead is highly sensitive to the recognition of target words during the initial stages of visual processing, while the Cz, C3, and Pz leads show differences in the later stages of semantic processing and integration. These leads complement each other in the visual cognitive task of text, collectively participating in the processing of text information. Through the time-voltage analysis of these leads, we can gain a more comprehensive understanding of how the brain processes target and non-target words

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Target stimuli	×	×	×	×	×	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	×	×	×
Non-target stimuli	×	×	×	×	×	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	×	×	×

Table 6. Results of the statistical analysis of the EEG time-domain features. Note: "√" indicates a significant difference between target and non-target stimuli, while "×" indicates no significant difference between the two.

at different stages, providing important insights into the neural basis of visual semantic processing and target recognition.

Time-domain feature analysis of EEG signals was conducted by averaging individual trials of similar stimuli to obtain EEG responses reflecting the text cognitive process. Visual information is transmitted through the visual ascending pathway to the primary visual cortex with a delay of approximately 140 ms, and changes in the EEG response waveform can be observed at the corresponding time in the figure. Further observation and analysis reveal that within the traditional ERP period, i.e., the 0-500 ms range, there are relatively obvious P1-N1-P2 components. The P1-N1-P2 components are exogenous components of the cortical visual evoked potential (CVEP), directly reflecting the brain's perception of visual information. Additionally, after 500 ms, significant differences in the amplitude of EEG responses induced by different types of stimuli can be observed, and the fluctuations in EEG responses during this period may reflect the semantic understanding and text cognitive processing of different types of stimuli. The above time-domain features show similar trends across all scalp channels. In this study, to analyze the statistical differences between Target and non-Target stimuli, we conducted statistical tests on data from all electrodes. Paired-samples t-tests were used to compare differences between the Target and non-Target conditions. Additionally, to control for false positives due to multiple comparisons, we applied the strict Bonferroni correction method. We used a non-overlapping time window of 100 ms to average the amplitude of 24 time segments, numbered 1-24, and compared the target and non-target stimuli using t-tests. The specific results are shown in Table 6, where the peak values and amplitudes within the ERP period (including the P1-N1-P2 components) induced by target and non-target stimuli are basically consistent, indicating that there are no significant differences between the two types of stimuli during the visual information perception stage. However, statistical analysis of each time segment after 500-2400 ms shows that the differences in EEG responses are statistically significant and relatively pronounced, suggesting that changes in EEG responses may reflect the cognitive processing of different stimulus material contents.

Despite individual differences in participants' reading speeds, the standardized time window method employed in this study ensured that the temporal dynamics of fuzzy semantic processing remained stable on a relative time scale. Fuzzy semantic processing is a dynamic process that relies on contextual integration, and its neural mechanisms are not tied to specific time points but rather reflect the continuous resolution of semantic ambiguity. Our analysis demonstrated that neural activity patterns remained consistent on the standardized timeline, even with varying reading speeds across participants, further validating the temporal dynamics of fuzzy semantic processing. Additionally, to more accurately capture the alignment between reading dynamics and neural activity, we propose that future research integrate eye-tracking technology to record the dynamic changes in visual attention during reading. This would help refine the precision of time series analysis and provide deeper insights into the neural mechanisms underlying fuzzy semantic processing.

Signal-to-noise ratio (SNR) analysis.

Signal-to-Noise Ratio (SNR) serves as a crucial metric for evaluating the quality of EEG data, effectively reflecting the contrast between the target signal strength and background noise. In this study, SNR for each trial was calculated using precise methods to ensure high-quality data.

First, EEG signals were segmented within the stimulus presentation window (0–2000 ms), and the Root Mean Square (RMS) values across all channels were computed to represent signal power. Next, EEG segments devoid of task-related activity were extracted during the pre-stimulus baseline period (-200–0 ms), and their RMS values were calculated to serve as noise power. According to the definition,

$$SNR(dB) = 10 * log_{10}\frac{signal\ power}{noise\ power} \tag{1}$$

Using this formula (1), SNRs for all electrodes were calculated.

As shown in Fig. 4, the SNR for all electrodes exceeds 20 dB, with a mean of 24.3 ± 2.1 dB, indicating that the collected data are of excellent quality. Specifically, central electrodes such as Cz, C3, and C4 exhibit the highest SNRs, consistent with the high activation characteristics of the central region during visual-semantic tasks[41]. In contrast, occipital electrodes like Oz, O1, and O2 have slightly lower SNRs, likely due to their proximity to eye muscles, which can introduce electromyographic (EMG) artifacts[42].

Test-retest reliability analysis.

Intraclass correlation coefficients (ICC) and Pearson correlation coefficients (r) were used as statistical metrics to evaluate data consistency and stability. Here we calculated the mean intraclass ICC, which was performed using a two-way mixed-effects absolute agreement model.

$$ICC = \frac{MS_R - MS_E}{MS_R - (k - 1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{2}$$

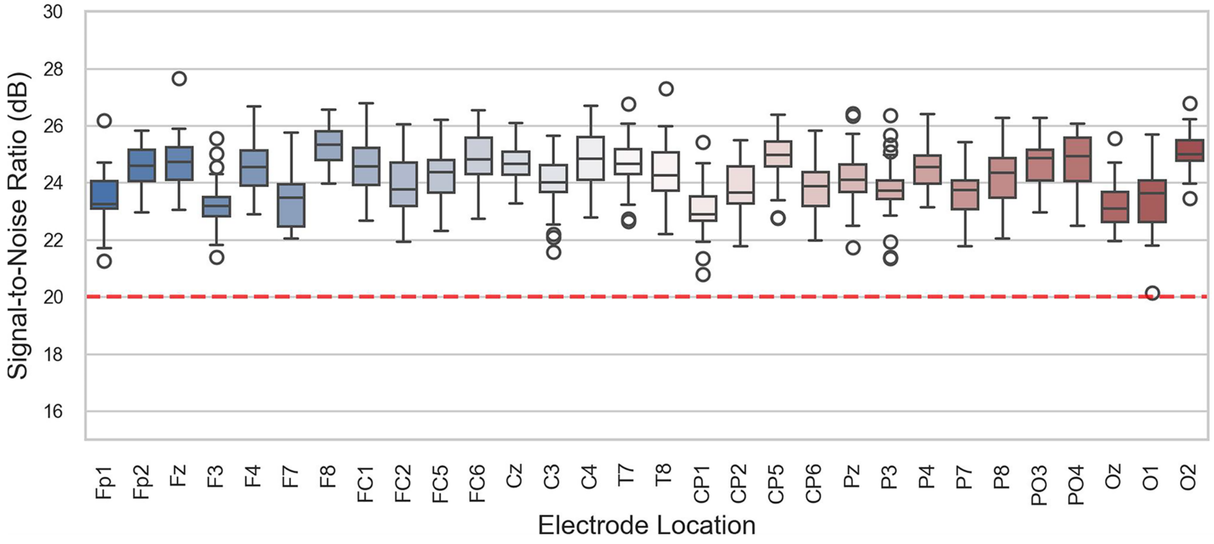


Fig. 4 The distribution of Signal-to-Noise Ratios. The red dashed line indicates the Signal-to-Noise Ratio = 20 dB threshold, with all electrodes having SNRs above this value.

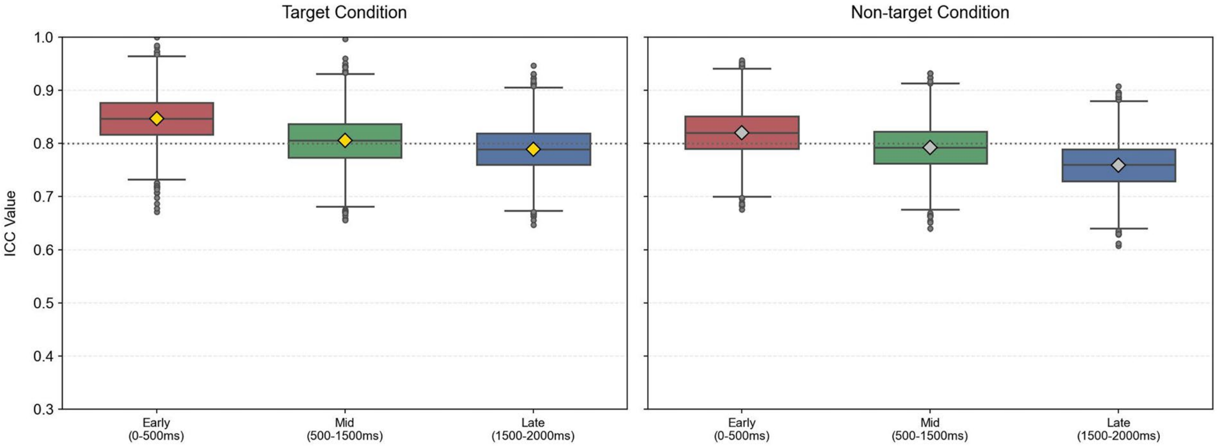


Fig. 5 ICC Distribution in different stages.

This approach quantifies measurement system stability through variance component decomposition, distinguishing between-subject variance (MS_R), between-condition variance (MS_C), and residual error (MS_E). The experimental data reveal a patterned decay in ICC distributions between target and non-target conditions across cognitive stages in Fig. 5, while maintaining high overall stability. During the early stage (0-500 ms), the target condition shows a higher mean ICC compared to non-targets, aligning with early selective attention theory—rapid gamma oscillations in visual cortices prioritize neural resource allocation for target stimuli. In the mid-stage (500-1500 ms), target ICC remains stable through theta-band phase synchronization in prefrontal-parietal networks supporting working memory maintenance, while non-target ICC declines, reflecting gradual decay of task-irrelevant representations. By the late stage (1500-2000ms), target ICC decreases but remains significantly higher than non-targets, driven by stability compensation mechanisms in the anterior cingulate control network. Cross-stage comparisons show shallower ICC decay slope for targets, confirming the temporal stability advantage of goal-directed cognitive processes. Additionally, the Pearson correlation coefficient was r = 0.85 ($p < 0.001$), demonstrating good data stability.

EEG frequency domain feature analysis based on PSD estimation. In EEG frequency domain feature analysis, changes in EEG signal energy were observed from the perspective of relative power in the above five sub-bands by calculating the Power Spectral Density (PSD), and statistical analysis was conducted on the data. The analysis of frequency bands with significant differences is as follows: the energy distribution in the delta band shows that the energy distribution patterns and trends induced by the two types of stimuli are basically consistent, with the central region being stronger than the bilateral temporal regions, and the activation level in the central region being the most active; the energy distribution in the alpha band shows a pattern of distribution from the central parietal region to the bilateral parietal regions and temporal areas, and the target stimulus

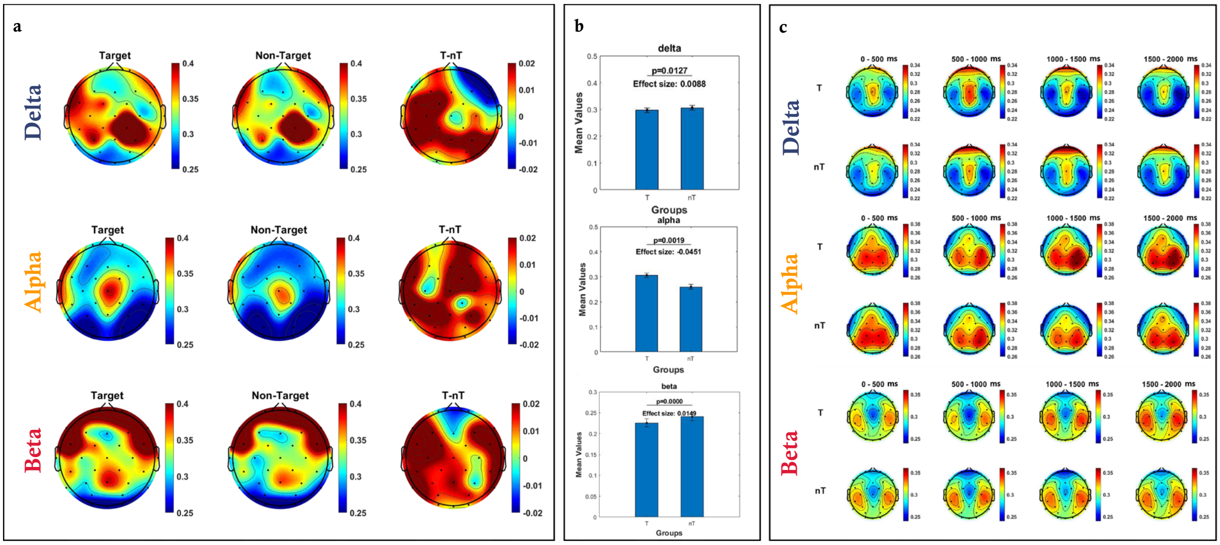


Fig. 6 EEG frequency domain feature analysis and brain region distribution characteristics. (**a**) Energy distribution in the frequency bands. EEG rhythms are believed to participate in higher brain functions, with the delta band (0.5–4 Hz) being associated with cognitive functions during wakefulness; the theta band (4–8 Hz) is thought to be related to memory and learning functions; the alpha band (8–12 Hz) reflects fluctuations in attention levels; the beta band (12–30 Hz) is often associated with brain activities such as stimulus evaluation and decision-making; and the gamma band (30–80 Hz) is related to the activation of working memory functions. (**b**) Relative power ratio. Significant differences were observed in the delta, alpha, and beta frequency bands, suggesting that the relative power in these bands can serve as potential features for distinguishing different stimulus types. Statistical tests were conducted using paired-samples t-tests, with a significance threshold of $p < 0.05$. Specific p-values and effect sizes (Cohen's d) are listed in the figure to quantify the magnitude of the observed differences. To minimize the impact of multiple comparisons, statistical results were corrected using the false discovery rate (FDR) method. (**c**) Relative power distribution by time period. Relative power is defined as the percentage of EEG sub-band energy to the total band energy. Analyzing relative power is chosen to avoid the influence of inter-subject energy level differences on frequency domain analysis results. At the same time, compared to absolute power, which reflects the intensity of brain region activation, relative power reflects the energy distribution of EEG signals across different frequency bands and functions.

causes a more significant activation level across the entire brain region. This result may reflect the cognitive process of the subject making evaluations and judgments on the target stimulus; the energy distribution in the beta band has a distribution pattern opposite to that of the delta and alpha bands, with the left hemisphere being more active and the central region being less active. This distribution pattern may reflect the coupling between frequency bands.

As shown in Fig. 6a, the different frequency domain features and the statistical analysis results in Fig. 6b indicate that the relative power of the delta, alpha, and beta bands induced by the two types of stimuli can be used as candidate features to distinguish between different stimulus types. Additionally, based on different frequency bands divided by 500 ms, we analyzed the corresponding relative power spectral density changes for target and non-target stimuli, further exploring precise time periods and rhythms to distinguish between target and non-target stimuli. From Fig. 6c, it can be seen that for target and non-target stimuli in the delta, alpha, and beta bands, from the initial stage of stimulation (0–500 ms) to the middle stage of stimulation (500–2000 ms), the power spectral density shows a trend of first increasing, then decreasing, and then increasing again, with different brain region activations. In the final stage of stimulation (2000–2400 ms), the PSD value decreases again, basically returning to the PSD value at the initial stage of stimulation. Based on these results, time-voltage analysis of the data can be conducted to further explore the specific differences between the two types of stimuli. The analysis of EEG signals across time intervals revealed that, within the early time window of 0–500 ms, brain responses to target and non-target sentences were primarily associated with visual processing, reflecting participants' initial attention allocation to semantic ambiguity. In the intermediate time window of 500–1500 ms, target sentences significantly activated brain regions associated with semantic integration (e.g., parietal and temporal regions, $p < 0.01$). In the late time window of 1500–2000 ms, EEG signals associated with target sentences exhibited higher amplitudes ($p < 0.05$), consistent with higher-order semantic decision-making and contextual confirmation processes. We acknowledge that the dynamic characteristics of semantic processing during natural reading may extend beyond the operational time window segmentation used in this study. Natural reading is a non-linear process that often involves eye regressions and cross-interval semantic integration. However, the time window division used in this study was based on an operational framework informed by prior literature[44–47] and cognitive processing characteristics. The early visual processing stage (0–500 ms) primarily reflects initial detection of semantic ambiguity, while the intermediate (500–1500 ms) and late (1500–2000 ms) time windows correspond to semantic integration and contextual confirmation processes, respectively. Future studies could incorporate eye-tracking technology to link gaze points with EEG signals dynamically, further enhancing the precision of fuzzy semantic processing analysis.

Usage Notes

The code for all modules is openly available on GitHub (https://github.com/tym182319/TMNRED). All scripts were developed in Python 3.12 and MatLab2022. MNE v1.7.1, pybv v0.7.5, pyprep v0.4.3, mne-iclabel v0.6.0 were used to implement the pre-processing pipeline, while mne-bids v0.15.0 was used to organize the data into BIDS format. For more details about code usage, please refer to the GitHub repository. The code for EEG data pre-processing is highly configurable, permitting flexible adjustments of various pre-processing parameters, such as data segmentation range, downsampling rate, filtering range, and choice of ICA algorithm, thereby ensuring convenience and efficiency. Researchers can modify and optimize this code according to their specific requirements. The stimulation protocols were developed using Psychtoolbox-3 in MatLab R2022.

Code availability

In line with reproducible research philosophy, all codes used in this paper are publicly available and can be accessed at https://github.com/tym182319/TMNRED. The stimulation protocol and the auxiliary MatLab functions are also available. The processing Python scripts are also available. The repository contains all the auxiliary functions to facilitate the load, use and processing of the data, as described above. By changing a few parameters in the main processing script, a completely different process can be obtained, allowing any interested user to easily build his/her own processing code. Additionally, all scripts for generating the time-voltage Representations and the plots here presented, are also available.

Received: 31 July 2024; Accepted: 22 April 2025;
Published online: 25 April 2025

References

1. Avitan, L., Teicher, M. & Abeles, M. EEG generator–a model of potentials in a volume conductor[J]. *Journal of Neurophysiology* **102**(5), 3046–3059 (2009).
2. da Silva, F. L. EEG: Origin and Measurement[M]//MULERT C, LEMIEUX L. EEG - fMRI: Physiological Basis, Technique, and Applications. Cham: Springer International Publishing, 23–48 (2022).
3. Dale, A. M. & Halgren, E. Spatiotemporal mapping of brain activity by integration of multiple imaging modalities[J]. *Current Opinion in Neurobiology* **11**(2), 202–208 (2001).
4. Shih, J. J., Krusienski, D. J. & Wolpaw, J. R. Brain-computer interfaces in medicine[J]. *Mayo Clinic Proceedings* **87**(3), 268–279 (2012).
5. Khan, M. A. *et al*. Review on motor imagery based BCI systems for upper limb post-stroke neurorehabilitation: From designing to application[J]. *Computers in Biology and Medicine* **123**, 103843 (2020).
6. Jeong, J. H. *et al*. Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions[J]. *GigaScience* **9**(10), giaa098 (2020).
7. Cheng, M. *et al*. Design and implementation of a brain-computer interface with high transfer rates[J]. *IEEE Transactions on Biomedical Engineering* **49**(10), 1181–1186 (2002).
8. Hollenstein, N. *et al*. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading[J]. *Scientific data* **5**(1), 1–13 (2018).
9. Du, C. *et al*. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2023).
10. Cheng, C. M. & Lin, S. Y. Chinese orthographic decomposition and logographic structure[J]. *Reading and Writing* **26**, 1111–1131 (2013).
11. Wang, M., Cheng, C. & Chen, S. W. Contribution of morphological awareness to Chinese-English biliteracy acquisition[J]. *Journal of Educational Psychology* **98**(3), 542 (2006).
12. Shu, H. *et al*. Understanding Chinese developmental dyslexia: morphological awareness as a core cognitive construct[J]. *Journal of educational psychology* **98**(1), 122 (2006).
13. Zhao, J. *et al*. Neural basis of phonological processing in second language reading: An fMRI study of Chinese regularity effect[J]. *NeuroImage* **60**(1), 419–425 (2012).
14. Perani, D. & Abutalebi, J. The neural basis of first and second language processing[J]. *Current opinion in neurobiology* **15**(2), 202–206 (2005).
15. Ge, J. *et al*. Cross-language differences in the brain network subserving intelligible speech[J]. *Proceedings of the National Academy of Sciences* **112**(10), 2972–2977 (2015).
16. Li, Y. *et al*. Human cortical encoding of pitch in tonal and non-tonal languages[J]. *Nature communications* **12**(1), 1161 (2021).
17. Mouli, S. & Palaniappan, R. DIY hybrid SSVEP-P300 LED stimuli for BCI platform using EMOTIV EEG headset[J]. *HardwareX* **8**, e00113 (2020).
18. Chen, Y., Fazli, S. & Wallraven, C. An EEG Dataset of Neural Signatures in a Competitive Two-Player Game Encouraging Deceptive Behavior[J]. *Scientific data* **11**(1), 389 (2024).
19. Melamud, A., Hagstrom, S. & Traboulsi, E. Color vision testing[J]. *Ophthalmic Genetics* **25**(3), 159–187 (2004).
20. Lee, E. H. Review of the psychometric evidence of the perceived stress scale[J]. *Asian nursing research* **6**(4), 121–127 (2012).
21. Arnott, S. R. & Alain, C. Effects of perceptual context on event-related brain potentials during auditory spatial attention[J]. *Psychophysiology* **39**(5), 625–632 (2002).
22. Jaeggi, S. M. *et al*. The concurrent validity of the N-back task as a working memory measure. *Memory* **18**(4), 394–412 (2010).
23. Delplanque, S. *et al*. Modulation of cognitive processing by emotional valence studied through event-related potentials in humans[J]. *Neuroscience letters* **356**(1), 1–4 (2004).
24. Rudkin, S. J., Pearson, D. G. & Logie, R. H. Executive processes in visual and spatial working memory tasks[J]. *Quarterly Journal of Experimental Psychology* **60**(1), 79–100 (2007).
25. Whitney, C. *et al*. Executive semantic processing is underpinned by a large-scale neural network: revealing the contribution of left prefrontal, posterior temporal, and parietal cortex to controlled retrieval and selection using TMS[J]. *Journal of cognitive neuroscience* **24**(1), 133–147 (2012).
26. Davey, J. *et al*. Automatic and controlled semantic retrieval: TMS reveals distinct contributions of posterior middle temporal gyrus and angular gyrus[J]. *Journal of Neuroscience* **35**(46), 15230–15239 (2015).
27. Ralph, M. A. L. *et al*. The neural and computational bases of semantic cognition[J]. *Nature Reviews Neuroscience* **18**(1), 42–55 (2017).

28. He, B. *et al*. Electrophysiological source imaging: a noninvasive window to brain dynamics[J]. *Annual review of biomedical engineering* **20**(1), 171–196 (2018).
29. Seeber, M. *et al*. Subcortical electrophysiological activity is detectable with high-density EEG source imaging[J]. *Nature communications* **10**(1), 753 (2019).
30. He, B., Ding, L. Electrophysiological mapping and neuroimaging[M]//Neural engineering. Boston, MA: Springer US, 499-543 (2012).
31. Hallam, G. P. *et al*. Charting the effects of TMS with fMRI: Modulation of cortical recruitment within the distributed network supporting semantic control[J]. *Neuropsychologia* **93**, 40–52 (2016).
32. Asadzadeh, S. *et al*. A systematic review of EEG source localization techniques and their applications on diagnosis of brain abnormalities. *Journal of neuroscience methods* **339**, 108740 (2020).
33. Liu, Q. *et al*. Open access EEG dataset of repeated measurements from a single subject for microstate analysis[J]. *Scientific Data* **11**(1), 379 (2024).
34. Teige, C. *et al*. Dynamic semantic cognition: Characterising coherent and controlled conceptual retrieval through time using magnetoencephalography and chronometric transcranial magnetic stimulation[J]. *Cortex* **103**, 329–349 (2018).
35. Whitney, C. *et al*. The neural organization of semantic control: TMS evidence for a distributed network in left inferior frontal and posterior middle temporal gyrus[J]. *Cerebral cortex* **21**(5), 1066–1075 (2011).
36. Tan, L. H. *et al*. Neuroanatomical correlates of phonological processing of Chinese characters and alphabetic words: A meta-analysis[J]. *Human brain mapping* **25**(1), 83–91 (2005).
37. Bolger, D. J., Perfetti, C. A. & Schneider, W. Cross-cultural effect on the brain revisited: Universal structures plus writing system variation[J]. *Human brain mapping* **25**(1), 92–104 (2005).
38. Sun, G., Hu, J., Wu, G. A novel frequency band selection method for common spatial pattern in motor imagery based brain computer interface[C]//The 2010 International Joint Conference on Neural Networks (IJCNN). *IEEE*, 1–6 (2010).
39. Bai, Y. *et al*. TMNRED, A Chinese Language EEG Dataset for Fuzzy Semantic Target Identification in Natural Reading Environments. OpenNeuro. [Dataset] https://doi.org/10.18112/openneuro.ds005383.v1.0.0 (2024).
40. Sadiq, M. T. *et al*. Exploiting pretrained CNN models for the development of an EEG-based robust BCI framework[J]. *Computers in Biology and Medicine* **143**, 105242 (2022).
41. Rossion, B. *et al*. Functional imaging of visual semantic processing in the human brain[J]. *Cortex* **36**(4), 579–591 (2000).
42. Park, S. *et al*. Systematic Investigation of Optimal Electrode Positions and Re-Referencing Strategies on Ear Biosignals[J]. *International Journal of Human–Computer Interaction* **41**(2), 1323–1342 (2025).
43. Spyropoulos, G., Bosman, C. A. & Fries, P. A theta rhythm in macaque visual cortex and its attentional modulation[J]. *Proceedings of the National Academy of Sciences* **115**(24), E5614–E5623 (2018).
44. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system[J]. *nature* **381**(6582), 520–522 (1996).
45. Bašnáková, J. *et al*. Beyond the language given: The neural correlates of inferring speaker meaning[J]. *Cerebral Cortex* **24**(10), 2572–2578 (2014).
46. Friederici, A. D. Towards a neural basis of auditory sentence processing[J]. *Trends in cognitive sciences* **6**(2), 78–84 (2002).
47. Kutas, M. & Federmeier, K. D. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)[J]. *Annual review of psychology* **62**(1), 621–647 (2011).

Acknowledgements

This work was mainly supported by the National Key R&D Program of China (2023YFF1203503), Haihe Laboratory of Brain-computer Interaction and Human-machine Integration Independent Project (24HHNJSS00015), and the National Natural Science Foundation of China (82202290). We also thank all research assistants who provided general support in participant recruiting and data collection.

Author contributions

Qi Tang, Ran Zhao, Shuming Zhang and Mu Xing participated in the EEG data collection. Qi Tang and Ran Zhao participated in the manuscript revision and participated in the data collection and analysis. Yanru Bai, Changjian Wang, Guangjian Ni, and Dong Ming designed the study. All the authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.B. or G.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.