

Cognition and Behavior

How Do We Segment Text? Two-Stage Chunking Operation in Reading

Jinbiao Yang (杨金骅)^{1,2,3,4,5} Qing Cai (蔡清)^{1,2} and Xing Tian (田兴)^{2,3}<https://doi.org/10.1523/ENEURO.0425-19.2020>

¹Key Laboratory of Brain Functional Genomics– (Ministry of Education & Science and Technology Commission of Shanghai Municipality), Affiliated Mental Health Center, ECNU Shanghai Changning Mental Health Center, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China, ²NYU-ECNU Institute of Brain and Cognitive Science at New York University Shanghai, Shanghai 200062, China, ³Division of Arts and Sciences, New York University Shanghai, Shanghai 200122, China, ⁴Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, The Netherlands, and ⁵Centre for Language Studies, Radboud University, 6500 HD Nijmegen, The Netherlands

Abstract

Chunking in language comprehension is a process that segments continuous linguistic input into smaller chunks that are in the reader's mental lexicon. Effective chunking during reading facilitates disambiguation and enhances efficiency for comprehension. However, the chunking mechanisms remain elusive, especially in reading, given that information arrives simultaneously yet the written systems may not have explicit cues for labeling boundaries such as Chinese. What are the mechanisms of chunking that mediates the reading of the text that contains hierarchical information? We investigated this question by manipulating the lexical status of the chunks at distinct levels in four-character Chinese strings, including the two-character local chunk and four-character global chunk. Male and female human participants were asked to make lexical decisions on these strings in a behavioral experiment, followed by a passive reading task when their electroencephalography (EEG) was recorded. The behavioral results showed that the lexical decision time of lexicalized two-character local chunks was influenced by the lexical status of the four-character global chunk, but not vice versa, which indicated the processing of global chunks possessed priority over the local chunks. The EEG results revealed that familiar lexical chunks were detected simultaneously at both levels and further processed in a different temporal order, the onset of lexical access for the global chunks was earlier than that of local chunks. These consistent results suggest a two-stage operation for chunking in reading, the simultaneous detection of familiar lexical chunks at multiple levels around 100 ms followed by recognition of chunks with global precedence.

Key words: EEG; hierarchy; lexical access; reading; segmentation

Significance Statement

The learners of a new language often read word by word. However, why can proficient readers read multiple words at a time? The current study investigates how we efficiently segment a complicated text into smaller pieces and how we process these pieces. Participants read Chinese strings with different structures while their key-press responses and brain electroencephalography (EEG) signals were recorded. We found that texts were quickly (~100 ms from their occurrences) segmented to varied sizes of pieces, and larger pieces were then processed earlier than small pieces. Our results suggest that readers can use existing knowledge to efficiently segment and process written information.

Received October 14, 2019; accepted May 4, 2020; First published May 11, 2020.

The authors declare no competing financial interests.

Author contributions: J.Y. and Q.C. designed research; J.Y. performed research; J.Y. and X.T. analyzed data; J.Y. and X.T. wrote the paper.

Introduction

Reading is arguably one of the unique human intelligences. However, how we process written texts remains elusive. For instance, how can we comprehend a complex sentence? A sentence consists of many letters/characters that form a hierarchical structure of text chunks (e.g., morphemes, words, and phrases). Readers need to incrementally segment a complex sentence into smaller chunks that map onto their mental lexicon. This process is termed as text chunking (Reali and Christiansen, 2007; Gobet et al., 2016). What are the small chunks during chunking? How do we process the chunks? To answer those questions, this study investigated the cognitive procedure of text chunking.

Words and their sublevel (morphemes) are usually assumed as the basic units in reading models in psycholinguistics and computer science (McClelland and Rumelhart, 1981a; Coltheart et al., 2001; Taft, 2013). However, eye-tracking studies suggested we can perceive the text information longer than a word at one time (Rayner, 1998). Our working memory also allows us to remember familiar multiple words (Miller, 1956). Even more, multiword expressions can be stored in our mental lexicons (Arnon and Snider, 2010; Siyanova-Chanturia et al., 2017). These studies suggest the multiword representations and the beyond-word processing are feasible. Moreover, relying on larger chunks effectively reduces the cognitive load while processing sentences: fewer chunks to be interpreted and integrated (Ellis, 2003; Krishnamurthy, 2003; Blache and Rauzy, 2012). Furthermore, the semantic combination of constituents can be different from holistic meaning (Goldberg, 1995). One extreme example is idioms, as the metaphors of an idiom can be distinct from their literal meanings of smaller constituents. Multiword representation is required in certain contexts to avoid ambiguity. Therefore, multiword chunks, as well as word chunks, could be the units during chunking.

What is the relation between the processes of word chunks and multiword chunks during chunking? The studies of compound words (a single lexical entity but consists of more than one root morphemes, e.g., “flagship”) may offer hints. According to the dual-route models of compound-word processing, both the whole word and its constituents are processed at the same time or are selected to process each level flexibly (Andrews et al., 2004; Koester et al., 2007; MacGregor and Shtyrov, 2013; Semenza and Luzzatti, 2014; Blache, 2015). In a similar

This work was supported by National Natural Science Foundation of China Grants 31871131, 31771210, and 31970987; Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) Grants 17JC1404101, 17JC1404104, and 17JC1404105; the Program of Introducing Talents of Discipline to Universities, Base B16018; and the Joint research institute Seed Grants for Research Collaboration from NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai.

Correspondence should be addressed to Xing Tian at xing.tian@nyu.edu or Qing Cai at qcai@psy.ecnu.edu.cn.

<https://doi.org/10.1523/ENEURO.0425-19.2020>

Copyright © 2020 Yang et al.

This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

vein, we hypothesized that all the familiar lexical chunks, no matter which level it is, could be processed simultaneously. More specifically, the detection of chunks would be the first step in chunking, and the detection of chunks at multiple levels would occur at the same time, as the early lexical familiarity checking assumed in the E-Z reader model (Reichle et al., 2003).

How does the multilevel operation unfold in the chunking process? Which level has the priority after being detected? The word superiority effect indicates that the recognition of letters within words is better than letters in nonwords or stand-alone letters (Reicher, 1969). It suggests that the word has priority over the letter in reading. Similarly, the processing priority of global chunks can reduce the steps of integration and avoid the ambiguity to enhance the efficiency of language processing (Ellis, 2003; Krishnamurthy, 2003; Blache and Rauzy, 2012). Generalizing from the word superiority effect, we hypothesized that global chunks took priority over the parts and would be initiated first in the processing stage after detection.

In this study, we used Chinese four-character strings to investigate the chunking operation in reading. Chinese written system is an ideal model for observing multilevel chunking because the Chinese do not have explicit word boundaries. Each Chinese character is a basic lexical unit with a similar length. Four characters can form two levels of chunks, chunks with two characters (hereafter as the local level chunks) and a chunk with four characters (hereafter as the global level chunk). The lexicality was manipulated at both levels so that four types of stimuli were included (phrase, idiom, random words, and random characters). In the behavioral experiment, we investigated the interaction between the global and local chunks in reading by a lexical decision task at different levels of chunks. Moreover, an electroencephalography (EEG) experiment was conducted to investigate the temporal dynamics of detection and recognition stages in the multilevel chunking operation.

Materials and Methods

Participants

Twenty-one healthy native Chinese speakers (10 males, mean age 21 years, range 18–30 years) with normal or corrected-normal vision participated in both behavior and EEG experiments for financial compensation. Five participants who produced extensive EEG artifacts were excluded from EEG analysis. Hence, a total of 16 participants were included in the EEG study. The experiments were approved by the Research Ethics Committee of East China Normal University and New York University Shanghai. Written informed consents were obtained from all participants before the experiments.

Stimuli

All stimuli are four-Chinese-character strings. Two factors are included when designing these stimuli. The first factor is the chunk size that contains two levels, a global size of four characters and a local size of two characters.

Table 1: Stimuli description

	Local word	Local nonword
Global word	<i>GwLw</i> : lexicalized compound phrase composed of two two-character words. E.g., “希腊神话” (pinyin: xī là shén huà), translation: Greek mythologies. “希腊” and “神话” means “Greek” and “mythologies” in Chinese, respectively	<i>GwLn</i> : lexicalized compound phrases that consist of 4 independent mono-morphemic characters (Chinese idioms “Chengyu”). E.g., “以逸待劳” (pinyin: yì yì dài láo), translation: wait for the exhausted enemy at your ease. “以逸” and “待劳” are not words in Chinese
Global nonword	<i>GnLw</i> : non-lexicalized compound phrase composed of two two-character words. E.g., “存款电脑” (pinyin: cún kuaǎn diàn nǎo), translation: deposit-computer. “存款” and “电脑” means “deposit” and “computer” in Chinese, respectively	<i>GnLn</i> : random character string, nonwords at both levels. E.g., “投其顾此” (pinyin: tóu qí gù cǐ), a nonsense phrase. “投其” and “顾此” are not words in Chinese either

The second factor is lexicality (word or nonword) at each chunk size. These two factors are fully crossed and yield four types of stimuli. We denote chunk size using upper case letters, G for global and L for local, and use lower case letters for lexicality in each chunk size (w for word and n for nonword). For example, *GnLw* stands for the condition of stimuli that are four-character nonwords at the global level made of two two-character words at the local level. Note that the stimuli in *GwLn* are Chinese idioms, “Chengyu.” They are lexicalized compound phrases that consist of four independent mono-morphemic characters. None of the two characters in Chengyu can form a common word, whereas the four characters together form an idiomatic expression. Chengyu generally expresses the gist or moral message of myths, stories, or historical events from which they were derived. Therefore, the meaning of a Chengyu usually surpasses the sum of the meanings from the four characters. The four types of stimuli are listed in Table 1.

We selected and created all stimuli with the following steps. We extracted the *GwLw* and *GwLn* stimuli from a database of Sogou Pinyin (<https://www.sogou.com/labs/resource/w.php>) and a database of Chinese characters (CharDB: data version: 0.98.1; program version: 0.97.2; <https://chardb.cls.ru.nl/>). All the *GwLw* and *GwLn* stimuli satisfied the following criteria at the global level: (1) noun (the part of speech is determined by a record of the lexicon of Jieba v0.36: <https://github.com/fxsjy/jieba>); (2) high-frequency (the frequency was determined by the database of the Sogou Pinyin, and the high-frequency meant the frequency above 3000); and (3) no duplicative characters (e.g., “高高兴兴,” translation: happy). Moreover, the *GwLw* stimuli satisfied the following criteria at the local level: (1) both two-character words were nouns, and (2) high-frequency words. Moreover, the lexicality of *GwLn* stimuli at the local level was verified by checking if the first two or last two characters’ combination did not exist in the Sogou Pinyin database. These selection criteria made the *GwLw* and *GwLn* stimuli consistent in all aspects except the lexical status at the local level.

The *GnLw* stimuli were created by randomly pairing two different two-character words, the Lw in *GnLw* and *GwLw* follow the same criteria. So the only difference between the *GwLw* and *GnLw* was the lexical status at the global level. Finally, the *GnLn* stimuli were created by randomly mixing

four different characters, and none of the first or last two characters’ combinations existed in the Sogou Pinyin database. Characters used in all stimuli have log frequency ranging from 3.011 to 5.344, with stroke counts ranging from 4 to 13. The character’s log frequency was determined by the Subtitle Database (Cai and Brysbaert, 2010).

The distinction between word and nonword was further controlled by familiarity. Twelve participants who were not in the main experiment were asked to rate the familiarities of either the entire four-character or the constituents of two-character strings as being words or not. The rating range was from 1 to 5, where 1 stands for unfamiliar strings/nonwords and 5 for familiar words. The strings that were rated from 2 to 4 were removed and remained the stimuli that were either very familiar words or very unfamiliar nonwords in a pool. Eighty stimuli in each condition were randomly selected from the pool and used in this study.

Procedure behavioral experiment

In each trial, participants were first asked to focus on a cross presented at the center of the screen. After 400 ms, the fixation cross disappeared, and a four-character string was shown until response. A line also appeared either under the entire four-character string or under the two-character string (the first or the last two characters). Participants were asked to make a lexical decision about the underlined string, either the entire string (global task henceforth) or the constitute of first or last two-character string (local task henceforth) by pressing either “F” or “J” on the keyboard as fast as possible. Participants had a maximum of 3 s to respond. Responses and reaction time were collected. Response keys were counterbalanced across participants. The intertrial intervals were randomly selected from a range from 800 to 1000 ms.

Four stimuli types (*GwLw*, *GwLn*, *GnLw*, and *GnLn*) were fully crossed with task types (global task vs local task) and yielded eight conditions; 320 trials were included in this experiment. Half of the trials were randomly selected and used in the global task and the other half in the local task. The order of conditions was randomized. The experimental presentation was programmed on a Python package, Expy, which is a software for presenting and controlling psychological experiments (<https://expy.readthedocs.io/>).

Behavioral data analysis

All participants had response accuracy exceeding 85%, and the average accuracy was 92%. No participant's data were excluded. Trials with error responses were removed before analysis. We applied repeated measures three-way ANOVA (analysis of variance) on the reaction time data with factors of global-level lexicality, local-level lexicality, and task, followed by planned *t* tests for testing specific hypotheses.

Procedure EEG experiment

The same group of subjects participated in the EEG experiment. The EEG experiment shared the same stimuli list with the behavioral experiment, but both the procedure and the task are different. First, the display of each character string lasted for 300 ms. Participants were asked to read the underlined parts of the stimuli (to keep their attention on the stimuli), but they did not perform any lexical decision task. We used all 320 strings with 80 for each stimuli type in the global task and repeated once in the local task. Moreover, 320 four-symbol strings were included as the visual baseline in the EEG experiment. The symbols in a symbol string trial were randomly sampled with replacement from four symbols (\square , \triangle , \diamond , and \circ). Underlines were included in the symbol trials similar to those in the global and local tasks in experimental trials. Half of the trials were randomly selected and used in the global task and the other half in the local task. To guarantee participants' attention on the stimuli, we randomly inserted strings of digits for 100 ms, and participants were asked to report the underlined digits by pressing number buttons on a keyboard. About 48 number-report trials were presented to each participant.

EEG recording

EEG signals were recorded with a 32-electrode active electrodes system (actiChamp system, Brain Products GmbH). FP1 and FP2 were used to monitor vertical eye movements. Electrode impedances were kept below 10 k Ω . Data were continuously recorded in single DC mode. Data were sampled at 500 Hz, online referenced to the Cz.

EEG data analyses

EEG data were preprocessed using EEGLAB (version 13.5.4b; Delorme and Makeig, 2004). Data were band-pass filtered (0.1–30 Hz, Hamming windowed sinc FIR filter), and re-referenced to the average reference. The preprocessed data were epoched between –200 and 800 ms relative to the onset of strings and baseline-corrected using the 200-ms prestimulus period. The trials with eye blinks were rejected if the amplitude within the 1000-ms epoch exceeded $\pm 50 \mu\text{V}$. Remaining trials with apparent noise were rejected manually. Approximately 15% of trials were rejected. Five participants who produced a large number of artifacts or showed continuous α waves were excluded from further analysis. Epochs in each condition were averaged and created an event-related potentials (ERPs). Root mean square (RMS) responses were also calculated as a geometric mean of all channels.

In addition to the univariate analyses on ERPs and RMS responses, our analysis used the topographic patterns or distributions across all sensors rather than the response amplitude in selected groups of sensors, as it can provide more holistic and unbiased information. Such multivariate methods can collectively reflect spatial and temporal information and offer more power to test psychological and neuroscience hypothesis by overcoming problems such as individual differences, sensor selection and reference selection in EEG (Murray et al., 2008; Tian and Huber, 2008; Tian et al., 2011; Yang et al., 2018; Wang et al., 2019). Two multivariate-based methods [clustering and topographic ANOVA (TANOVA)] and one mass-univariate method (analysis of the topographic distribution of amplitude differences) were applied as following.

Clustering. A clustering method on the ERP topographic responses was implemented first. This unsupervised machine learning method groups data across all conditions by forming temporal clusters based on the similarity of their topographical patterns. This clustering method is a data-driven method, in which it explores the pattern similarity in topographies in all conditions. The clustering algorithm organizes data at different time points into distinct clusters, so that we can explore the temporal dynamics of pattern changes. Moreover, if one considers the clusters reflecting different processing stages, this analysis can identify the processing stages in each condition and display the temporal differences of any specific stage among conditions. We used K-means, the most popular algorithm for clustering. The clustering analysis is an omnibus test about neural dynamics, which can detect clusters and set up the time windows of interest for the following analyses.

The procedure of clustering algorithm covers three steps: (1) averaged EEG data across all participants to get ERPs at each time point for each condition; (2) defined ERPs at each time point in each condition as a sample, and the amplitudes of 32 electrodes were used as features in a sample; (3) K-means algorithm was conducted at all samples. The K-means algorithm is data driven. The target cluster number (K) can range from the minimum of one to the number of data points. We assumed that the baseline period to a fixation involved rest or random cognitive processes. The topographies would be consistent random patterns that were different from later sequences of topographies induced by stimuli. If the K-means separated the baseline period into more than one cluster, it was most likely overfitting. Therefore, we set the criterion of getting two clusters at the baseline period as a stopping point for increasing the number of clusters. That is, the K-means algorithm was conducted at all samples. The number of clusters was initially two and increased until the clustering result included more than one cluster at the baseline stage.

Analysis of the topographic distribution of amplitude differences. We calculated the response amplitudes across all sensors at a given time window. The changes in the amplitude differences can reflect the processing dynamics. Especially, the earliest time point that shows significant differences would indicate the temporal

onset of interest. Moreover, the spatial extent of experimental effects can be estimated by the distribution of sensors in which the amplitude differences are observed. At a window size of 20 ms, we checked the significant electrodes (Yang et al., 2018) to obtain the distribution differences between conditions. Because there were 32 comparisons in each time window, the p values were corrected by false discovery rate (FDR). The primary purpose of this analysis was to identify the possible onset timing of any amplitude differences between conditions. However, the power of detecting the onset using a measure of amplitude could be small. To reduce the Type II error, we did not apply corrections across time so that we can reduce the chance of missing the exact onset time of amplitude differences.

TANOVA. We further investigate the patterns of topographies by considering all sensors at the same time to infer the differences in underlying neural processes across conditions. A single index was calculated to indicate topographic information. Mathematically, each topography can be viewed as an n -dimensional vector, where the n equals the number of sensors. The divergence between the topographies of two experimental conditions can be quantified by the cosine value of the high-dimensional angle between two vectors (Tian and Huber, 2008). The cosine distance has a range from 0 to 2, where 0 stands for identical topographies and 2 exactly opposite patterns. Note that the cosine distance represents the similarity between the response patterns in topographies and is free from the difference of response magnitude because the measure of cosine distance is normalized by the vector length. To statistically test the cosine distance between topographies and to infer the underlying neural processing in different conditions, we applied an algorithm named TANOVA (Murray et al., 2008; Tian and Huber, 2008; Brunet et al., 2011; Tian et al., 2011; Lange et al., 2015). In TANOVA, null hypothesis distribution is generated by shuffling the condition labels, and we here shuffled the condition labels on the subjects' ERPs using the EasyEEG toolbox (Yang et al., 2018; strategy 2, shuffle times: 1000, window size: 10 ms). Furthermore, the temporal clusters in the TANOVA results were identified by a precluster threshold of 0.1 and were tested by the cluster-based permutation with the corrected threshold of 0.05 (Maris and Oostenveld, 2007).

Results

Behavioral experimental results

Reaction time was subject to repeated measures three-way ANOVA with the factors of global-level lexicality, local-level lexicality, and task. The main effect of global-level lexicality was significant ($F_{(1,20)} = 10.83$, $p < 0.01$), suggesting that participants took longer time to identify global-level nonwords than global-level words. The main effect of local-level lexicality is significant ($F_{(1,20)} = 6.30$, $p = 0.02$), suggesting participants took longer time to identify local-level nonwords than local-level words. However, the main effect of task is not significant ($F_{(1,20)} =$

1.07 , $p = 0.31$), suggesting different tasks that require participants to respond to either global or local chunks have a similar level of difficulty. More importantly, all three two-way interactions are significant, global-level lexicality \times local-level lexicality ($F_{(1,20)} = 57.11$, $p < 0.001$), global-level lexicality \times task ($F_{(1,20)} = 16.02$, $p < 0.001$), and local-level lexicality \times task ($F_{(1,20)} = 57.11$, $p < 0.001$).

Planned *post hoc t* tests were further conducted in each factor to specify the observed significant interactions. First, we examined how global information affect processing at the local level (Fig. 1A). In the local task, the reaction time in *GnLw* was significantly longer than that in *GwLw* ($t_{(20)} = 5.7145$, $p < 0.001$, difference = 55 ms), suggesting that the nonwords at the global level significantly slowed down the lexical decision of words at the local level. Moreover, the reaction time of *GwLn* was significantly longer than *GnLn* in the local task ($t_{(20)} = 3.6214$, $p = 0.002$, difference = 54 ms), suggesting that the words at the global level also slowed down the lexical decision of nonwords at the local level. Second, we examined how the local chunk could affect processing at the global level (Fig. 1B). In the global task, we did not find a significant difference between reaction time to *GwLw* and *GwLn* ($t_{(20)} = -1.0091$, $p = 0.32$; Fig. 1B, left), suggesting that the lexical status of local chunks cannot affect the lexical decision of words at the global chunks. These results collaboratively suggest that processing at the global level may take priority.

Another comparison on how the local chunk could affect processing at the global level revealed that the reaction time of *GnLw* was significantly longer than that of *GnLn* in the global task ($t_{(20)} = 6.6874$, $p < 0.001$, difference = 112 ms; Fig. 1B, right). This result suggests that the decisions of global chunks and local chunks could be parallel when the global decision took too long. The lexical information at the local level may leak through to the processing of global chunks and influence the decision of nonwords. The behavioral results demonstrate a unidirectional influence from the global level to the local level when the task targets are words and interactions between levels when the task targets are nonwords that need more time to make decisions. That is, global chunks may take priority in lexical processing. Whereas the processing of global and local chunks could be parallel when global decision took too long, the lexical information at one level may leak through to the process of the other level and influence the decision of nonwords. We further test the processing dynamics in an EEG experiment.

To examine and exclude the possible effects of the underline position, we extracted the trials with two-character underlines and ran a repeated measures three-way ANOVA with the factors of global-level lexicality, local-level lexicality, and underline position. The underline position showed neither main effect ($F_{(1,20)} = 1.61$, $p = 0.22$), nor interaction effect with global-level lexicality ($F_{(1,20)} = 0.01$, $p = 0.92$), nor interaction effect with local-level lexicality ($F_{(1,20)} = 0.38$, $p = 0.54$). These results suggest that positions of stimuli that were relevant to the task did not affect response speed.

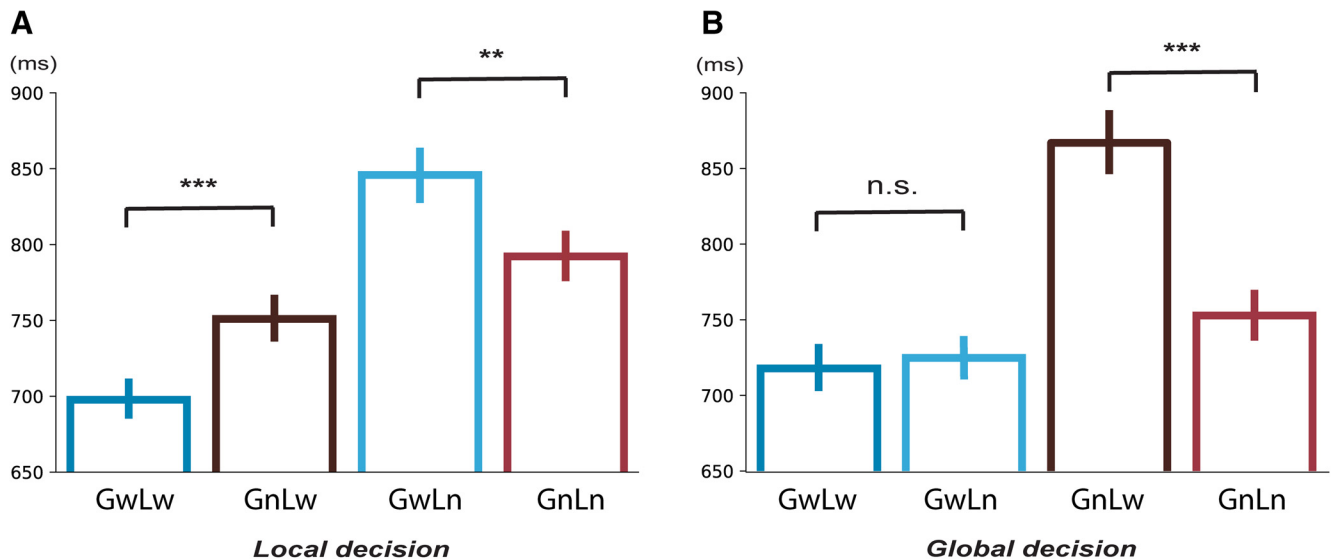


Figure 1. Results of the behavioral experiment. **A, B**, Reaction time results in the global and local tasks, respectively. In each plot, condition labels are provided along the x-axis. Error bars represent ± 1 SEM (standard error of the mean). Each planned paired test was represented by the line linking two bars; n.s., not significant; $**p < 0.01$, $***p < 0.001$.

EEG results

Clustering revealed distinct stages of chunking

We first conducted the clustering analysis to explore the dynamics of ERP responses. The clustering algorithm aimed to separate the continuous ERP responses into distinct stages based on common features observed across time. As shown in Figure 2, the clustering results were reliable as similar clusters were observed continuously in each condition.

More importantly, clear temporal profiles were revealed by the clustering analysis in all conditions. First, the same cluster was observed in the baseline period until around 80 ms after stimulus onset, as well as at the end of epochs (~600 ms after onset) among all types of stimuli. The clustering in these periods was presumably because few cognitive processes that relate to the stimuli or task were available or manifested in the ERP topographies. Second, a novel cluster (the second cluster) appeared after 80 ms across five conditions. The clustering spanned in similar latencies as N1/P2 components, presumably reflecting visual processing. However, different dynamics was observed across conditions after 200 ms. The third cluster appeared earlier in the symbol condition with a much shorter duration than the four experimental string conditions in which the third cluster appeared around 250 ms and lasted ~90 ms. Moreover, in the symbol condition, the third cluster was immediately followed by the sixth cluster that did not appear until 500 ms in the four experimental string conditions (except in *GwLw* condition around 430 ms). The early start and long-lasting sixth cluster in the symbol condition was accompanied by the missing of the first and second clusters that appeared around 320 ms and lasted till 450 ms in other conditions.

More interestingly, a 15-ms period around 130 ms was labeled as the cluster baseline that was grouped in the baseline and the end of epoch periods. This formed a

short gap that broke the early processing into two stages. The clustering results set up the time windows of interest for the following analyses. We focused on the components in early timing to further investigate the underlying processes of chunking operation.

Chunk detection in the earliest stage

To test the hypothesis about the lexical detection in the earliest stage, we conducted analyses to investigate the lexicality effects at the global and local chunk levels. First, we applied repeated measures one-way ANOVA on the ERPs in P3 and P4, as well as RMS waveforms calculated using all channels (Fig. 3A,B). We did not find any significant amplitude differences among the four conditions. Next, compared with the responses to *GnLn* strings that contained no lexical chunks at either level, the topographical pattern of response amplitudes in conditions that include lexical chunks (*GwLw*, *GwLn*, and *GnLw*) did not show any significant differences in any sensors after multiple comparison correction (Fig. 3C). However, the difference topographies showed distinctive patterns of amplitude distribution (90–130 ms, higher on the left frontal area and lower on occipital area). These results suggest that lexical detection could induce changes in the configuration of neural sources, rather than in response amplitude. Therefore, we investigated the topographic patterns to infer the different configurations of neural processing across conditions.

The analysis of TANOVA revealed significant differences between the topographies of *GnLn* and response patterns in *GwLw*, *GwLn*, and *GnLw* conditions (Fig. 3D, highlighted in the red box). The differences were first observed at 90 ms after stimulus onset. The differences were most substantial in the *GwLn-GnLn* comparison as the significant level at $p < 0.01$ for the following 20 ms. The pattern differences were also observed in the *GwLw-*

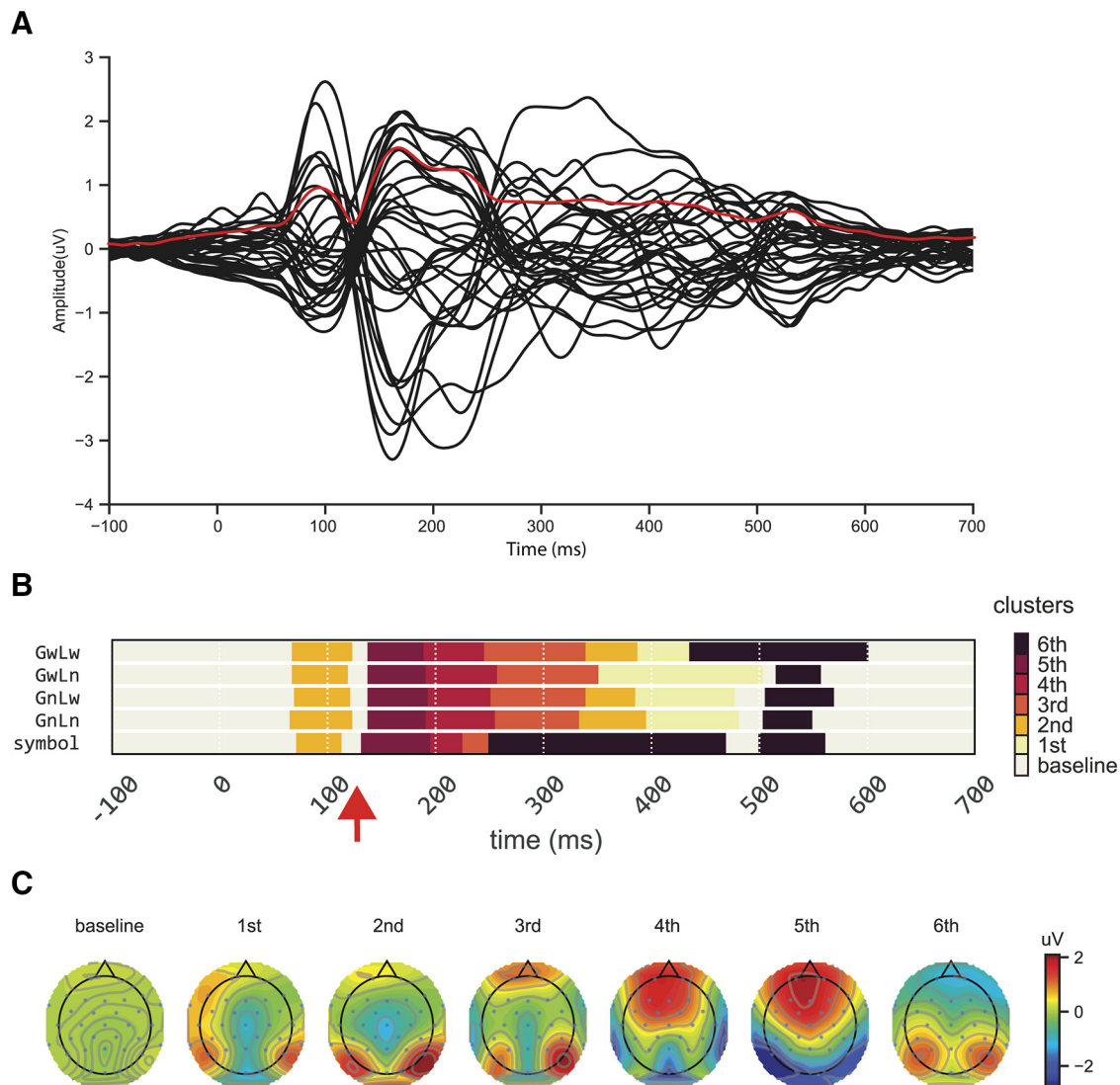


Figure 2. The dynamics of ERP responses and clustering results. **A**, Averaged ERPs waveform responses of all conditions from 32 electrodes (black lines), and RMS waveform response across all electrodes (red line). **B**, Temporal clustering results of topographies for four conditions (*GwLw*, *GwLn*, *GnLw*, and *GnLn*) and a baseline symbol condition (symbol). Different colors represent distinct clusters. Samples in the same color but at different time points indicate that they are grouped into the same cluster, sharing similar features but occurring at different times. The temperature of colors represents the rank of the cluster distance relative to the cluster baseline (cluster defined by the baseline period). Approximately 80 ms after stimulus onset, a novel cluster (the second cluster) appears at the same time across five conditions, followed by another new cluster. However, in the symbol condition, the third cluster (~250 ms) appears earlier with a much shorter duration than four-character string conditions. **C**, The topographies of each cluster.

GnLn comparison, but the significant differences started later at 100 ms and lasted for 30 ms. The p value of the 90- to 100-ms time bin in the *GnLw*-*GnLn* comparison was 0.0539, and the relative Bayes factor (BF10) was 2.3371, which indicates weak evidence in favor of H1 (Held and Ott, 2018). The TANOVA of the comparison between *GnLn* and the average of the other three lexical conditions showed significant differences around 110 ms (Fig. 3E, highlighted in the red box). TANOVA revealed that topographic differences occurred before 130 ms between lexicalized chunks at either level and the non-lexicalized *GnLn* condition. These results suggested that the early-stage process related to the detection of lexicality. The global level information may facilitate the detection

because the effect size ranked from biggest to smallest in the order of *GwLn*, *GwLw*, and *GnLw*.

We also observed significant differences in later responses. The *GwLn* has a significant cluster starting at 150 ms, followed by *GwLw* that has a significant cluster starting around 190 ms. The *GnLw* does not have a significant late cluster until ~230 ms. These latency differences at a later stage suggested that the lexical processes could first occur at the global level. We further investigate these dynamics in the next session.

Processing of chunks at different levels

We applied responses amplitude and pattern analyses between conditions with different lexical status either at

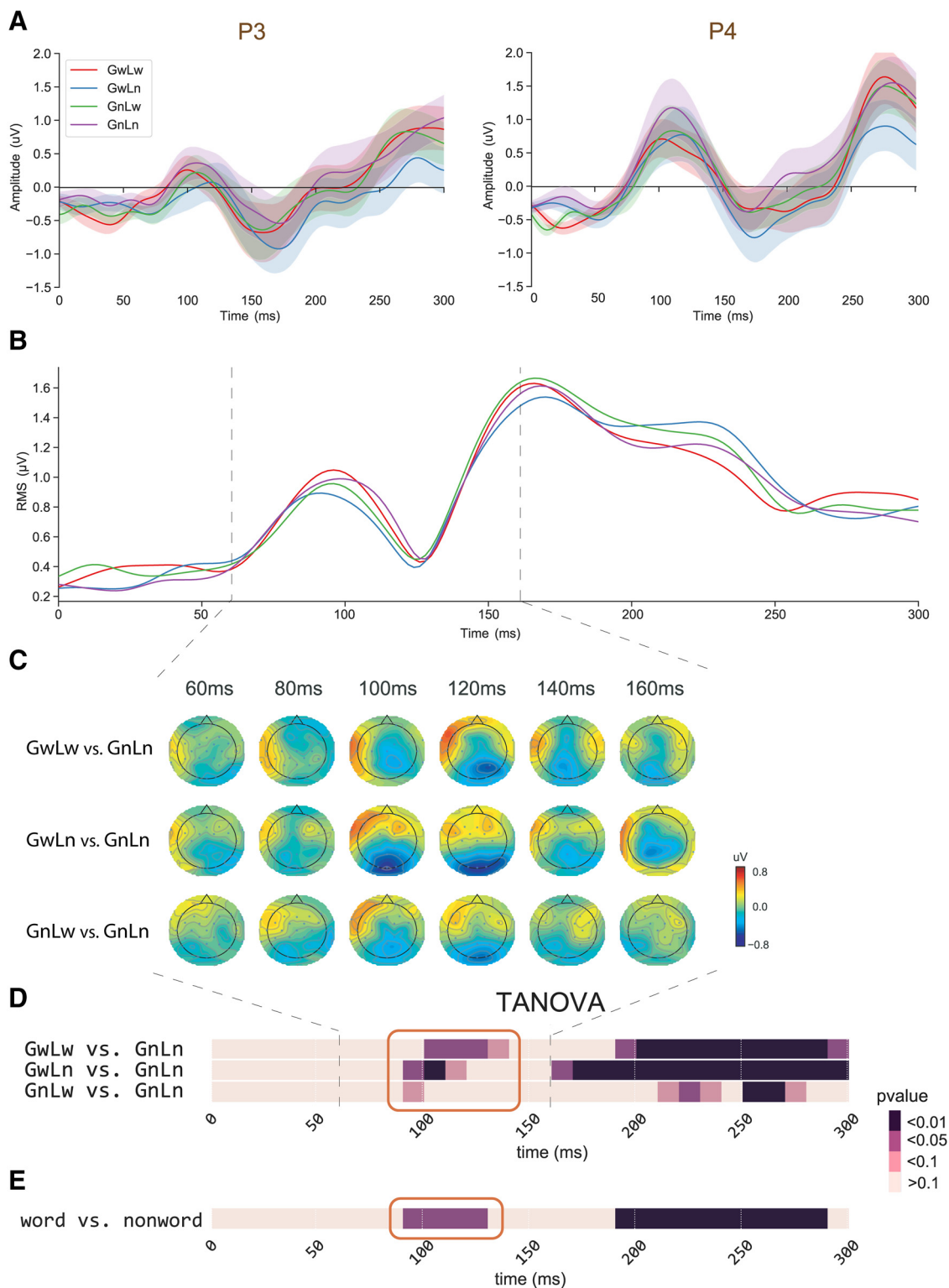


Figure 3. The effects of lexicalized chunks revealed in the paired comparisons between *GnLn* condition and the other three lexical conditions (*GwLw*, *GwLn*, and *GnLw*). **A**, ERP waveform responses in the representative channels P3 and P4. One-way ANOVA did not reveal any response amplitude differences among conditions. **B**, RMS waveform responses of all channels. No amplitude difference was found. **C**, Topographical comparisons of response amplitude. Each row shows a comparison across time. The color scheme depicts the differences in response amplitude between conditions. No significant difference was found on the electrodes after the multiple comparison correction (FDR). **D**, The temporal dynamics of TANOVA on paired comparisons (uncorrected). The red boxes highlight the earliest time when the significant differences were observed. All three conditions show evidence of early lexical detection. **E**, The temporal dynamics of TANOVA on the comparison between *GnLn* and the average of three lexical

continued

conditions, corrected by temporal clustering analysis with a corrected threshold of 0.05 (Maris and Oostenveld, 2007). The lexicality effects emerge around 100 ms.

the global level or at the local level to test the dynamics of chunk processing. We compared ERPs from the channels of P3 and P4 for words and nonwords, separately at the global and local levels (Fig. 4A,B), as well as RMS from all channels (Fig. 4C). Paired *t* tests were applied to these data. We found that the lexicality effects occurred as early as 160 ms at the global level, but much later around 250 ms at the local level. These effects are in selected parietal channels but are absent in the RMS, suggesting the effects are narrowly distributed, which is consistent with the distribution results in Figure 4D. When comparing the strings that contained a global-level word (*GwLw*, *GwLn*) with the strings that contained global-level nonword (*GnLw*, *GnLn*), the significant differences were observed in the electrodes over the middle parietal and left frontal-temporal regions (Fig. 4D, indicated by white points) starting around 170 ms (Fig. 4D, red arrow). When comparing the strings that contained a local-level word (*GwLw*, *GnLw*) with the strings that contained local-level nonword (*GwLn*, *GnLn*), the significant differences in response amplitudes started much later at the latency around 230 ms in the electrodes over frontal and parietal-occipital regions (Fig. 4D, blue arrow).

The TANOVA results in Figure 4E further showed that response patterns were statistically significantly different between processing the distinct lexical status at the global level began around 160 ms after the stimulus onset (Fig. 4E, red arrow), and the significant pattern differences at the local level began around 220 ms (Fig. 4E, blue arrow), consistent with the observation in Figure 4A,D, as well as the observations of late differences in Figure 3D.

Discussion

This study investigated the processing dynamics of written texts that included different levels of chunks, such as word and phrase. With the stimuli of Chinese four-character strings that contain multiple grain-size language chunks, the behavioral results showed that the recognition of lexicalized local chunks was affected by the lexical status of global chunks, but not vice versa. These results suggested that the processing of chunks at the global level was prioritized over the processing of local ones during reading. Moreover, the earliest EEG responses showed distinct patterns between lexicalized and non-lexicalized chunks, and the latency of successive EEG responses was faster when processing chunks at the global level than that for local chunks. These consistent behavioral and electrophysiological results suggested that two distinct stages successively operate in the early stage of reading for the detection of potential chunks and further processes on the detected chunks at multiple levels.

Detection of chunks at 100 ms

In the clustering results (Fig. 2), a “temporal gap” was observed in the early EEG reading responses and

separated the processing from 80 to 200 ms into two distinct clusters, suggesting the different neural bases and possible distinct functions. Furthermore, the response patterns of the earliest cluster around 100 ms were modulated by the lexical status of chunks at both global and local levels (Fig. 3). These findings are consistent with the early lexical familiarity checking mechanism proposed in the E-Z reader model (Reichle et al., 2003). Language chunks and their lexical status should be checked before accessing the semantics. In other words, the familiar lexical chunks are detected before subsequent processes (e.g., semantic retrieval). This is especially important in the language that lacks explicit boundaries for lexicalized chunks/phrases, such as written Chinese. Our results suggest such lexical checking/detection can occur early in the reading process around 100 ms and extend to multiple chunk levels.

What factor enables this early chunk detection in reading? Top-down mechanisms have been proposed to account for perceptual and cognitive functions, such as the prior knowledge or prediction of the global shape information in object recognition (Bar, 2003; Bar et al., 2006; Panichello et al., 2012). The detection of language chunks at multiple levels during reading involved the left frontal regions and occipital regions (Fig. 3A), similar to the top-down modulation by the early feed-forward projection of low spatial frequency information (Bar et al., 2006). In previous research, high-frequency words can be easily detected and recognized (Monsell and Besner, 1991; Ellis, 2002). The transparency (MacGregor and Shtyrov, 2013) and decomposability (Abel, 2003; Vannest et al., 2005) also affect the mental encoding of complex words, phrases, and idioms. However, individual differences in reading may make the perception of these physical attributes vary across individuals. Therefore, the factor that leads to the early chunk detection is likely to be the perceptual consequences, the familiarity of these attributes. The familiarity has been demonstrated in improving language retrieval (Bannard and Matthews, 2008; Zheng et al., 2015). In this study, we controlled the familiarity by only using stimuli that were rated at the extreme degree of familiarity, either very familiar words or strange nonwords. We speculate that the familiarity of lexical-orthographic features (such as frequency and decomposability) is the criterion of chunk detection, and it can apply simultaneously at both global and local levels during early reading processes.

The priority of processing global chunks

Our behavioral results demonstrated that the processing of a global chunk significantly affected the lexical decision of lexicalized local chunks. In contrast, the local chunks had no impact on the lexical decision of the lexicalized global chunk. The unidirectional effect suggested that the processing of global level chunks had priority

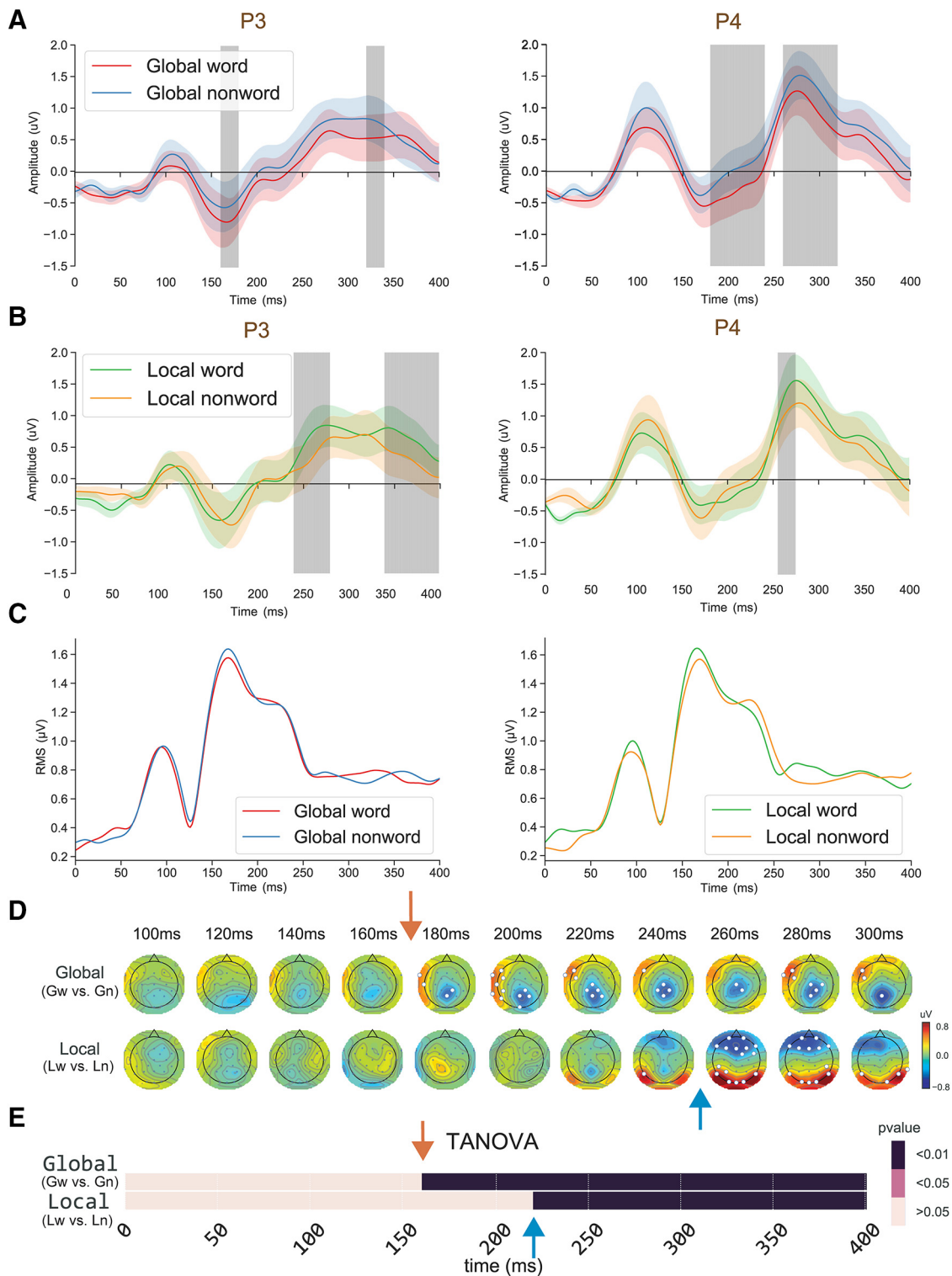


Figure 4. The processing dynamics of chunks at global and local levels. **A**, ERP waveform responses of the global lexicity effect in the representative channels P3 and P4; t tests revealed the effects occurred around 160 ms. The shaded area indicates $p < 0.05$. **B**, ERP waveform responses of local lexicity effect in the representative channels P3 and P4. The effects occurred around 250 ms, later than the global lexicity effect in **A**. The shaded area indicates $p < 0.05$. **C**, RMS waveform responses of all channels on global lexicity and local lexicity comparisons. No amplitude difference was found. **D**, Analysis of response amplitude in topographical comparisons between different lexical status at the global level (upper row) and at the local level (lower row) across time. The color scheme depicts the differences in response amplitude between conditions, and the white points superimposed on the topographies indicate the electrodes that showed significant differences after multiple comparison correction (FDR). **E**, The temporal dynamics of TANOVA results. The results showed a distinct starting time of significant response pattern differences between different lexical

continued

status at the global and local levels. The red arrows in all plots indicate the earliest latency of difference in the global level comparison, and the green arrows indicate the earliest latency of difference in the local level comparison. The results were corrected by a temporal clustering analysis with a corrected threshold of 0.05 (Maris and Oostenveld, 2007).

over the processing of their constituents. EEG further provided evidence supporting the temporal hierarchy in processing global and local chunks. The EEG results showed that the processing of global chunks started around 160 ms, while the onset of local chunk processing was much later (~220 ms). These EEG results, together with our behavioral data, demonstrated that after the simultaneous chunk detection at both levels, the processing of different sizes of lexical chunks began at different times: the processing of global chunks preceded that of local chunks.

The priority of global information has been demonstrated in many cognitive domains. Gestaltism (Heider, 1977; Dewey, 2018) considers the global contains more information than the aggregation of its locals. In vision, the global precedence effect (Navon, 1977/7) suggests that recognizing a scene is hierarchical and global processing has priority over local processing. In contrast, local processing is subject to the top-down reevaluation and integration into global processing. Similarly, the top-down facilitation of visual object recognition also implies that the activation of high-level information will be faster than the lower-level information (Bar, 2003; Bar et al., 2006). In linguistics, the word superiority effect (Reicher, 1969), an advantage of words on recognizing letters, suggests that the processing of a word at the global level interacts with the letter identification (McClelland and Rumelhart, 1981b). This study further demonstrates the influences of phrases on words. Our results expand previous research and suggest that the global-priority mechanism can be applied across multiple levels in a hierarchical manner in the linguistic context. The priority of global chunks is consistent with the information theory (Shannon, 1948): larger chunks contain more context information and less internal entropy, which can prevent ambiguity.

Paralleled processing of chunks at both levels

The behavioral results revealed that the judgment of a non-lexicalized phrase at the global level was more difficult when the task-unrelated chunks were familiar words at the local level. These results indicated that local processing might be initiated before the finish of global processing. The EEG results further supported that processing at both levels temporally overlapped, the response patterns of processing global chunks continued after the start of local processing response patterns (Fig. 4). This observation of partially temporal overlap in the processing part-whole hierarchies is consistent with simultaneous processing mechanisms implemented in the connectionist networks (Hinton, 1990). A scheduler could control the participation of processing at different levels. Should processing a chunk exceed expected duration, the processing of chunks at other levels would occur. Moreover, the topographic patterns showed left lateralization for

processing chunks at the global level, whereas both hemispheres engaged in processing chunks at the local level (Fig. 4), suggesting the possible anatomic differences that mediate the partially temporal paralleled processes at both levels.

Chunking in a broader cognitive perspective

Various cognitive functions can exert a top-down influence on early perceptual responses. For example, attention is one of the most common functions that modulate early perceptual responses, such as increasing the response gain in the visual (Fries et al., 2001), auditory (Poghosyan and Ioannides, 2008), and somatosensory (Steinmetz et al., 2000) domains. The current study offers a new top-down influence in a linguistic context. The lexicality/accessibility of the character combination determines the way of chunking and recombination of characters to form representations at both global and local scales. Such reconstruction of representations may modulate the early visual responses in reading.

The top-down influence provides a common framework that links among cognitive systems. For example, orofacial motion alters speech perception, such as the McGurk effect (McGurk and MacDonald, 1976), and shortens latency of early auditory responses (Van Wassenhove et al., 2005). Speech articulation dampens the auditory responses to speech feedback (Houde et al., 2002) and modulates the sensitivity to auditory stimuli via the motor-to-sensory transformation (Tian and Poeppel, 2010, 2013, 2015; Tian et al., 2016, 2018; Ma and Tian, 2019; Li et al., 2020). The current study provides evidence supporting that the language system can penetrate and influence visual processing.

Chunking, which deducts combinatory representations into more basic linguistic units for processing, plays a crucial role in language comprehension. Previous studies suggest that linguistic chunking arguably occurs in complex morphology such as decomposing compounds into morphemes – the smallest linguistic unit that carries meaning (Stockall and Marantz, 2006; Fiorentino and Poeppel, 2007). The current study further demonstrates that phrases can be segmented into smaller linguistic units based on lexicality at both global and local levels. Our results bridge chunking in morphology with chunking in sentences based on semantics and syntax (Ding et al., 2016), and even higher linguistic levels such as paragraphs or an entire text based on formal structures and conceptual flow (Teng et al., 2020). A complete picture of chunking operation across all levels of linguistic hierarchy emerges.

The two-stage processing suggested by our results may contribute to the debate regarding the accessible units in complex words (Giraudo and Dal Maso, 2016). Some studies suggest that the morphologic decomposition occurs

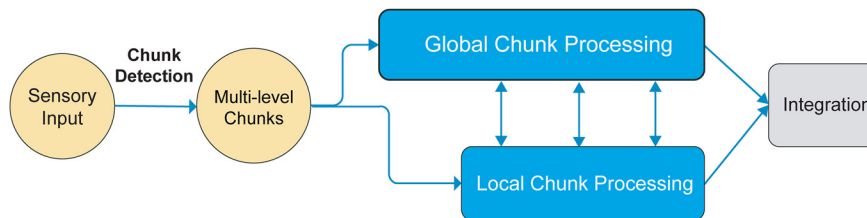


Figure 5. Schematic diagram of proposed two-stage chunking operation in reading.

only on the semantically transparent morphologic pairs (e.g., hunter–HUNT; Meunier and Longtin, 2007). In contrast, other studies found the semantics-opaque but morphology-complex words (e.g., corner = corn + er) also showed decomposition (Davis, 2004; Devlin et al., 2004; Gold and Rastle, 2007). That is, the surface morpheme-like unit that could be an interface between form and meaning is accessible regardless of the semantic relation between the global level and its constituents (Devlin et al., 2004; Gold and Rastle, 2007). Our results are more consistent with the latter view and suggest that this surface morpheme-like unit could be detected automatically as long as it is available. Specifically, these results show that bi-character words, which are bi-morpheme units, are also automatically decomposed from phrases, suggesting that the surface morpheme-like unit in the decomposition is not limited to the basic linguistic morphemes. Furthermore, the access of the surface morpheme-like unit has been localized over the occipito-temporal and left inferior frontal regions (Devlin et al., 2004; Gold and Rastle, 2007; Meinzer et al., 2009; Pliatsikas et al., 2014), which is consistent with our EEG topographic pattern (Fig. 3C). Orthographic typicality and lexicality modulate reading responses around 100 ms (Hauk et al., 2006; Faisca et al., 2019), which is also consistent with the detection timing in our observations.

The “global first” principle in different levels of accessible units has been observed in morphology (Bybee, 1995), letter detection (Han et al., 2003), and general vision (Chen, 1982; Wang et al., 2007). All these results are consistent with our findings that processing global level information possesses priority (Fig. 4). Last, the discrepancy between global and local information affects ERP responses as early as 250 ms (Han et al., 1999), suggesting a possible initiation time of parallel processing that is consistent with our results (Fig. 4).

The chunking operation is universal among sensory modalities for processing information that is beyond cognitive capacity. However, the nature of stimuli among sensory modalities may differentiate the possible neural mechanisms that mediate chunking. For example, linguistic information unfolds over time in speech, whereas the information can be available at the same time in reading (e.g., visual field and reading span). Therefore, temporal processing such as neural oscillations might be a potential dominant mechanism for chunking in the auditory domain with neural entrainment to acoustic features (such as prosodic cues and speech envelope; Luo and Poeppel, 2007), top-down rhythmic and melodic template (Nozaradan et al., 2011; Di Liberto et al., 2020),

semantic and syntactic cues (Ding et al., 2016, 2017), as well as structures and formats of language (Teng et al., 2020). However, in the visual domain, additional spatial information can be available at a time. Chunking is more likely based on the template from higher hierarchy, such as orthographic template in global/local letters (Kimchi, 1992) and mental representation of lexicality in the current study.

Based on all results, we tentatively put forward a workflow of processing multiple-level information in reading (Fig. 5). The segmentation occurs in an early short time window, and possible chunks at all levels are detected based on the familiarity of lexical-orthographic features (detection stage). The chunks at each level are further processed with distinct temporal characteristics (processing stage). Specifically, the processing of global chunks possesses priority over the local chunks, while the processing of local chunks can launch before the finish of global chunk processing. Hence, the processes of chunks at two levels have a partially temporal overlap that enables interaction across levels before final integration.

Because our primary goal was to test the relation between lexicality and chunking at different levels, we controlled the lexical-orthographic features such as the number of strokes, and frequencies. Theoretically, lexical access arguably occurs earlier than the semantic process. It is more likely that lexical factors are the primary factors mediating the effects that we observed. Semantic attributes could be another factor influencing the late process of chunking. It would be of interest to study semantics in chunking and obtain a complete understanding. Moreover, we investigated the computational dynamics of chunking in reading by testing the response latencies. EEG is one of the optimal tools to test the dynamics and latency, but not an optimal tool for inferring the spatial location of sources. The spatial distribution in topography is a distorted and incomplete representation of underlying neural sources because the topography is most likely a manifestation of a mixture from multiple neural sources. To avoid confusion, we only take advantage of changes in topographies across time or across conditions to infer the neural dynamics (Tian and Huber, 2008; Tian et al., 2011; Yang et al., 2018; Wang et al., 2019). Nevertheless, the location of the chunking operation is another aspect of interest. We planned to use fMRI for further investigation.

Conclusion

The current study investigated the chunking mechanism in reading. Consistent behavioral and EEG results

suggested that multiple levels of chunks were realized via two distinct stages of chunking in the early time course of reading. The first stage detected lexicalized chunks at all levels of grain-size. In the second stage, in contrast, the processing at the global level led the local level and resulted later in a parallel and interactive process. This study revealed the rich dynamics of chunking operation during reading, which provides the starting computation for comprehension of hierarchical language systems.

References

- Abel B (2003) English idioms in the first language and second language lexicon: a dual representation approach. *Second Lang Res* 19:329–358.
- Andrews S, Miller B, Rayner K (2004) Eye movements and morphological segmentation of compound words: there is a mouse in mousetrap. *Eur J Cogn Psychol* 16:285–311.
- Arnon I, Snider N (2010) More than words: frequency effects for multi-word phrases. *J Mem Lang* 62:67–82.
- Bannard C, Matthews D (2008) Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychol Sci* 19:241–248.
- Bar M (2003) A cortical mechanism for triggering top-down facilitation in visual object recognition. *J Cogn Neurosci* 15:600–609.
- Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Schmidt AM, Dale AM, Hämäläinen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E (2006) Top-down facilitation of visual recognition. *Proc Natl Acad Sci USA* 103:449–454.
- Blache P (2015) Hybrid parsing for human language processing. In: *Natural language processing and cognitive science*, pp 9–20. Venice, Italy: Libreria Editrice Cafoscarina.
- Blache P, Rauzy S (2012) Robustness and processing difficulty models. a pilot study for eye-tracking data on the french treebank. 24th International Conference on Computational Linguistics, p 21, December 2012, Mumbai, India.
- Brunet D, Murray MM, Michel CM (2011) Spatiotemporal analysis of multichannel EEG: CARTOOL. *Comput Intell Neurosci* 2011: 813870.
- Bybee J (1995) Regular morphology and the lexicon. *Lang Cogn Process* 10:425–455.
- Cai Q, Brysbaert M (2010) SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One* 5:e10729.
- Chen L (1982) Topological structure in visual perception. *Science* 218:699–700.
- Coltheart M, Rastle K, Perry C, Langdon R, Ziegler J (2001) DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol Rev* 108:204–256.
- Davis MH (2004) Units of representation in visual word recognition. *Proc Natl Acad Sci USA* 101:14687–14688.
- Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134:9–21.
- Devlin JT, Jamison HL, Matthews PM, Gonnerman LM (2004) Morphology and the internal structure of words. *Proc Natl Acad Sci USA* 101:14984–14988.
- Dewey RA (2018) Senses. In: *Gestalt psychology*. Available at <http://www.psywww.com/intropsych/ch04-senses/gestalt-psychology.html>.
- Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, de Cheveigné A, Shamma S, Mesgarani N (2020) Cortical encoding of melodic expectations in human temporal cortex. *Elife* 9:e51784.
- Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19:158–164.
- Ding N, Melloni L, Yang A, Wang Y, Zhang W, Poeppel D (2017) Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Front Hum Neurosci* 11:481.
- Ellis NC (2002) Frequency effects in language processing. *Stud Second Lang Acquis* 24:143–188.
- Ellis NC (2003) Constructions, chunking, and connectionism: the emergence of second language structure. In: *The handbook of second language acquisition*, pp 63–103. Oxford, UK: Blackwell.
- Faisca L, Reis A, Araújo S (2019) Early brain sensitivity to word frequency and lexicality during reading aloud and implicit reading. *Front Psychol* 10:830.
- Fiorentino R, Poeppel D (2007) Compound words and structure in the lexicon. *Lang Cogn Process* 22:953–1000.
- Fries P, Reynolds JH, Rorie AE, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291:1560–1563.
- Giraudo H, Dal Maso S (2016) The salience of complex words and their parts: which comes first? *Front Psychol* 7:1778.
- Gobet F, Lloyd-Kelly M, Lane PCR (2016) What's in a Name? The Multiple Meanings of “Chunk” and “Chunking.” *Front Psychol* 7:102.
- Goldberg AE (1995) *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Gold BT, Rastle K (2007) Neural correlates of morphological decomposition during visual word recognition. *J Cogn Neurosci* 19:1983–1993.
- Han S, Fan S, Chen L, Zhuo Y (1999) Modulation of brain activities by hierarchical processing: a high-density ERP study. *Brain Topogr* 11:171–183.
- Han S, Yund EW, Woods DL (2003) An ERP study of the global precedence effect: the role of spatial frequency. *Clin Neurophysiol* 114:1850–1865.
- Hauk O, Patterson K, Woollams A, Watling L, Pulvermüller F, Rogers TT (2006) [Q:] When would you prefer a SASSAGE to a SAUSAGE? [A:] At about 100 msec. ERP correlates of orthographic typicality and lexicality in written word recognition. *J Cogn Neurosci* 18:818–832.
- Heider GM (1977) More about Hull and Koffka. *Am Psychol* 32:383.
- Held L, Ott M (2018) On p-values and Bayes factors. *Annu Rev Stat Appl* 5:393–419.
- Hinton GE (1990) Mapping part-whole hierarchies into connectionist networks. *Artif Intell* 46:47–75.
- Houde JF, Nagarajan SS, Sekihara K, Merzenich MM (2002) Modulation of the auditory cortex during speech: an MEG study. *J Cogn Neurosci* 14:1125–1138.
- Kimchi R (1992) Primacy of wholistic processing and global/local paradigm: a critical review. *Psychol Bull* 112:24–38.
- Koester D, Gunter TC, Wagner S (2007) The morphosyntactic decomposition and semantic composition of German compound words investigated by ERPs. *Brain Lang* 102:64–79.
- Krishnamurthy R (2003) Language as chunks, not words. In: *JALT2002 conference proceedings: waves of the future* (Swanson M, Hill K, ed), pp 288–294. Tokyo, Japan: The Japan Association for Language Teaching.
- Lange VM, Perret C, Laganaro M (2015) Comparison of single-word and adjective-noun phrase production using event-related brain potentials. *Cortex* 67:15–29.
- Li S, Zhu H, Tian X (2020) Distinct neural signals in speech preparation differentially modulate auditory responses. *bioRxiv*, 905620. doi:10.1101/2020.01.14.905620.
- Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54:1001–1010.
- MacGregor LJ, Shtyrov Y (2013) Multiple routes for compound word processing in the brain: evidence from EEG. *Brain Lang* 126:217–229.
- Ma O, Tian X (2019) Distinct mechanisms of imagery differentially influence speech perception. *eNeuro* 6:ENEURO.0261-19.2019.
- Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* 164:177–190.

- McClelland JL, Rumelhart DE (1981a) An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol Rev* 88:375–407.
- McClelland JL, Rumelhart DE (1981b) An interactive activation model of context effects in letter perception. *Psychol Rev* 88:375–407.
- McGurk H, MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264:746–748.
- Meinzer M, Lahiri A, Flaisch T, Hannemann R, Eulitz C (2009) Opaque for the reader but transparent for the brain: neural signatures of morphological complexity. *Neuropsychologia* 47:1964–1971.
- Meunier F, Longtin C-M (2007) Morphological decomposition and semantic integration in word processing. *J Mem Lang* 56:457–471.
- Miller GA (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 63:81–97.
- Monsell S, Besner D (1991) The nature and locus of word frequency effects in reading. *Basic processes in reading: visual word recognition*, pp 148–197. Hillsdale, NJ:Lawrence Erlbaum Associates, Inc.
- Murray MM, Brunet D, Michel CM (2008) Topographic ERP analyses: a step-by-step tutorial review. *Brain Topogr* 20:249–264.
- Navon D (1977) Forest before trees: the precedence of global features in visual perception. *Cogn Psychol* 9:353–383.
- Nozaradan S, Peretz I, Missal M, Mouraux A (2011) Tagging the neuronal entrainment to beat and meter. *J Neurosci* 31:10234–10240.
- Panichello MF, Cheung OS, Bar M (2012) Predictive feedback and conscious visual experience. *Front Psychol* 3:620.
- Pliatsikas C, Wheeldon L, Lahiri A, Hansen PC (2014) Processing of zero-derived words in English: an fMRI investigation. *Neuropsychologia* 53:47–53.
- Poghosyan V, Ioannides AA (2008) Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron* 58:802–813.
- Rayner K (1998) Eye movements in reading and information processing: 20 years of research. *Psychol Bull* 124:372–422.
- Realí F, Christiansen MH (2007) Word chunk frequencies affect the processing of pronominal object-relative clauses. *Q J Exp Psychol* 60:161–170.
- Reicher GM (1969) Perceptual recognition as a function of meaningfulness of stimulus material. *J Exp Psychol* 81:275–280.
- Reichle ED, Rayner K, Pollatsek A (2003) The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behav Brain Sci* 26:445–476; discussion 477–526.
- Semenza C, Luzzatti C (2014) Combining words in the brain: the processing of compound words. Introduction to the special issue. *Cogn Neuropsychol* 31:1–7.
- Siyanova-Chanturia A, Conklin K, Caffarra S, Kaan E, van Heuven WJB (2017) Representation and processing of multi-word expressions in the brain. *Brain Lang* 175:111–122.
- Steinmetz PN, Roy A, Fitzgerald PJ, Hsiao SS, Johnson KO, Niebur E (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature* 404:187–190.
- Stockall L, Marantz A (2006) A single route, full decomposition model of morphological complexity: MEG evidence. *ML* 1:85–123.
- Taft M (2013) *Reading and the mental lexicon*. Hove: Psychology Press.
- Teng X, Ma M, Yang J, Blohm S, Cai Q, Tian X (2020) Constrained structure of ancient Chinese poetry facilitates speech content grouping. *Curr Biol* 30:1299–1305.e7.
- Tian X, Huber DE (2008) Measures of spatial similarity and response magnitude in MEG and scalp EEG. *Brain Topogr* 20:131–141.
- Tian X, Poeppel D (2010) Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front Psychol* 1:166.
- Tian X, Poeppel D (2013) The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. *J Cogn Neurosci* 25:1020–1036.
- Tian X, Poeppel D (2015) Dynamics of self-monitoring and error detection in speech production: evidence from mental imagery and MEG. *J Cogn Neurosci* 27:352–364.
- Tian X, Ding N, Teng X, Bai F, Poeppel D (2018) Imagined speech influences perceived loudness of sound. *Nat Hum Behav* 2:225–234.
- Tian X, Poeppel D, Huber DE (2011) TopoToolbox: using sensor topography to calculate psychologically meaningful measures from event-related EEG/MEG. *Comput Intell Neurosci* 2011:674605.
- Tian X, Zarate JM, Poeppel D (2016) Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex* 77:1–12.
- Vannest J, Polk TA, Lewis RL (2005) Dual-route processing of complex words: new fMRI evidence from derivational suffixation. *Cogn Affect Behav Neurosci* 5:67–76.
- Van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci USA* 102:1181–1186.
- Wang B, Zhou TG, Zhuo Y, Chen L (2007) Global topological dominance in the left hemisphere. *Proc Natl Acad Sci USA* 104:21014–21019.
- Wang X, Zhu H, Tian X (2019) Revealing the temporal dynamics in non-invasive electrophysiological recordings with topography-based analyses. *bioRxiv*, 779546. doi:10.1101/779546.
- Yang J, Zhu H, Tian X (2018) Group-level multivariate analysis in EasyEEG toolbox: examining the temporal dynamics using topographic responses. *Front Neurosci* 12:468.
- Zheng Z, Li J, Xiao F (2015) Familiarity contributes to associative memory: the role of unitization. *Adv Psychol Sci* 23:202.