Morphological encoding of Mandarin Chinese: evidence from Chinese disyllabic compound words

Wang, J.

Chapter 3 Word and morpheme frequency effects in naming Mandarin Chinese compounds: More than a replication

This article is published as: Wang, J., Schiller, N. O., & Verdonschot, R. G. (2024). Word and morpheme frequency effects in naming Mandarin Chinese compounds: More than a replication. Brain and Language, 259, 105496. https://doi.org/10.1016/j.bandl.2024.105496.

Abstract:

The question whether compound words are stored in our mental lexicon in a decomposed or full-listing way prompted Janssen and colleagues (2008) to investigate the representation of compounds using word and morpheme frequencies manipulations. Our study replicated their study using a new set of stimuli from a spoken corpus and incorporating EEG data for a more detailed investigation. In the current study, despite ERP analyses revealing no word frequency or morpheme frequency effects across conditions, behavioral outcomes indicated that Mandarin compounds are not sensitive to word frequency. Instead, the

response times highlighted a morpheme frequency effect in naming Mandarin compounds, which contrasted with the findings of Janssen and colleagues. These findings challenge the full-listing model and instead support the decompositional model.

Keywords: *Language production; Word frequency; Morpheme frequency; Compound representation; Picture naming; Mandarin Chinese*

3.1 Introduction

Theories of language production (Caramazza, 1997; Dell, 1986; Levelt et al., 1999) have specific assumptions about how words are organized and represented in the mental lexicon and suggest that word production occurs in several stages: conceptual preparation, lexical access, phonological encoding, and articulation. During speech production, such as when naming a picture, the conceptual representation of the intended object becomes active. This activation then extends to the lexical representations associated with this concept. Subsequently, phonological information is retrieved, which involves the encoding of word forms and is ultimately used for articulation by initiating the relevant speech gestures.

Many speech production models include a role for morphology (Koester & Schiller, 2008, 2011; Levelt et al., 1999; Roelofs, 1996; Zwitserlood et al., 2000, 2002). For example, in Levelt and colleagues' word production theory (Levelt et al., 1999), which largely draws on

results from the Dutch language, morphological encoding is the initial stage of word-form encoding. This stage involves the construction of words and defining their internal structures at the word form level. This has raised the question of how morphologically complex words, including derived words (e.g., “happiness”), inflected words (e.g., “running”), and compound words (e.g., “birdhouse”), are represented in our mental lexicon and how they are comprehended and produced. The present study aimed to investigate the representation of Mandarin compound production by reviewing the previous literature on complex words in both language production and comprehension.

3.1.1 The representation of morphologically complex words

3.1.1.1 Production of complex words

For language *production* of complex words typically a distinction is made between decomposition and full-listing models (Caramazza, 1997; Janssen et al., 2008; Janssen et al., 2014) with *hybrid* models (i.e., some complex words have full-listing storage, but other complex words are decomposed or follow a rule-based system) being somewhat scarcer. Levelt et al.’s model (1999; p 25) states that compound words (e.g., “birdhouse”) have a single lemma which then in turn activates two lexemes (<BIRD> + <HOUSE>). Note that Levelt et al.’s model (1999; p 27) may be considered “hybrid” in the sense that it suggests that certain compound words are “degenerate” in production,

meaning they are not decomposed at the form level. For example, the Dutch word *aardappel* "potato" appears semantically composed of *aard* "earth" and *appel* "apple," but it is produced as a single unit: *aar-dap-pel* (not *aard-ap-pel*). In contrast, a non-degenerate opaque compound like *oogappel* which is *oog* "eye" + *appel* "apple," meaning "apple of my eye" (a term of endearment for children), is decomposed at the form level, as evidenced by its pronunciation as *oog-ap-pel*, rather than *oo-gap-pel*. Inflected words (e.g., "escorted") in Levelt et al.'s model (1999) have a marking at the lemma level indicating tense (e.g., escort + PAST will activate the morphemes <ESCORT>+<ED>). Note that other theories, such as the words and rules theory by Pinker and Ullman (2002), similarly state that regular verb forms can be generated by a rule (i.e., a unification operation applied to a specific morpheme), just as how phrases and sentences are formed. When an irregular verb form happens to be stored (e.g., "drank"), it prevents a rule from applying (e.g., blocking *drinked), but anywhere else (by default) the rule applies. In Levelt et al.'s model (1999), the case is somewhat more difficult for complex derivational morphology especially when words would change syntactic class (e.g., the adjective "weak" + NESS forming the noun "weakness") for which Levelt et al. (1999) propose that these most likely are lemmas in their own right (i.e. "weakly" and "weakness" are separate lemmas).

3.1.1.2 Comprehension of complex words

Regarding language comprehension, which is more amply investigated, a similar division between decomposition can be made

(Longtin & Meunier, 2005; Rastle & Davis, 2008; Koester et al., 2004, 2009), full-listing (e.g., Butterworth 1983; Norris & McQueen, 2008) and hybrid models (Caramazza et al., 1988; Frauenfelder & Schreuder, 1991). For example, inflectional and derivational processes have received ample attention in the comprehension literature (Bozic & Marslen-Wilson, 2010; Marslen-Wilson et al., 1994; Penke et al., 1997; Rodriguez-Fornells et al., 2001; Smolka et al., 2015). For inflectional processes, the "morphological violation paradigm" in which a verb takes an incorrect form (e.g., the past participle **getanz-en* meaning "danced" in German, which should be *getanz-t*) has shown that differential EEG patterns occur for regular vs. irregular verbs supporting a hybrid model in which regular/irregular verbs are processed differently (Penke et al., 1997). Koester and colleagues (2009) investigated the time course of semantic integration in auditory compound word processing and found that the lexical-semantic integration of compound constituents occurs incrementally, supporting the decompositional hypothesis.

Others, for instance, Winther Balling and Baayen (2008), have suggested that there are two key moments during the auditory recognition of a complex word where its recognition likelihood changes significantly. First, when unrelated words are ruled out at the (typical) "uniqueness point" in which only one option remains in the competition for word recognition. For instance, for a simple word such as "candle," the uniqueness point might occur at the sound /d/ because up to /kænd/ "cand," it could be confused with words like "candy," but once the /l/

is added, the word becomes clearly distinguishable as "candle." However, for complex words such as derivations, there might be a continuation after the base word is recognized. For example, the word "hope" might receive its first uniqueness point at "p" where it is distinguished from words such as "hole" or "home." However, when acting as a part of a complex word, it could receive another uniqueness point, for example, at the "f" for "hopeful" where it is distinguished from other morphological stem-related words such as "hopeless."

Others working on visual word recognition have posited that there might be an early level of representation where (seemingly) complex words are broken down based on their morpho-orthographic features (Rastle et al., 2004). For example, a semantically transparent morphological relationship of a prime with the target (e.g., cleaner-CLEAN) would give rise to facilitation. However, Rastle et al. (2004) also found that when encountering a word like "corn," a prime such as "corner" would facilitate lexical decision times as well, even though the two words were not morphologically related, but see Baayen et al. (2011) for different views. This suggested a form of morphological decomposition that functioned differently from the semantic-based decomposition involved in the early stages of visual (complex) word recognition.

3.1.2 The representation of compound words

Returning to the issue of the processing of compound words in language production, as stated earlier, the most prominent theoretical

models are either decompositional, full-listing, or hybrid. The decompositional model holds the view that compounds are represented in terms of their constituents unless they are degenerate (Levelt et al., 1999), while the full-listing hypothesis suggests that the whole-word forms are only fully listed in our mental lexicon (Butterworth, 1983; Caramazza, 1997; Dell, 1986). Dual-route models combine elements of both approaches, proposing that transparent compounds are processed by breaking them down into their components while also allowing for parallel access to the whole-form representation. However, for opaque compounds, particularly those with high frequency, the whole-form access is assumed to be the dominant mechanism (MacGregor & Shtyrov, 2013). The present study focused on the decompositional and full listing hypotheses and hence summarized the relevant literature of these two models in the following sections.

3.1.2.1 The decompositional model

The presence of morphologically decomposed entries in the form of lexicon underlying speech production was addressed by manipulating word frequency by Roelofs (1996). He demonstrated the effect of constituent frequency on lexical access during speech production planning by using Dutch compounds. Implicit priming experiments were conducted, including homogeneous blocks where the stimuli shared a common form and heterogeneous conditions where they did not, to investigate whether the speech production system could plan non-initial constituents of a word before the initial ones. These

experiments revealed the task's sensitivity to morphological planning. A more pronounced facilitatory effect was observed when the initial syllable, constituting a morpheme, was shared (e.g., *nasmaak* "after-taste"-*nagalm* "reverberation"-*najaar* "autumn"), compared to producing disyllabic simple words in which the same overlapping part "na" constituted a syllable but not a morpheme (e.g., *nagel* "nail"-*natie* "nation"-*nader* "further"). These results supported the idea that component morphemes of compound words served as planning units in speech production, confirming the decomposition assumption.

Likewise, in a study conducted by Bien and colleagues (2005), four experiments employing a position-response association task were carried out to explore the influence of frequency information on Dutch compound word production. They independently manipulated the frequencies of the first and second constituents and the frequency of the compound itself. Compound production latencies demonstrated notable variability based on factorial contrasts in the frequencies of both constituent morphemes rather than being influenced by a factorial contrast in compound frequency, providing further reinforcement for decompositional models of speech production.

Other evidence came from the studies (Kaczer et al., 2015; Koester & Schiller, 2008, 2011; Lensink et al., 2014; Verdonschot et al., 2012; Zwitserlood et al., 2000, 2002; Wang et al., 2024) that employed the long-lag priming paradigm to investigate morphological processing, showing that morphological priming remained effective even with many intervening trials. These studies above suggested that

priming in those instances occurred at a distinct morphological level rather than at a phonological or semantic level. This finding was corroborated by further investigations. For example, in a study by Koester and Schiller (2008), they found that mere form overlap, as in prime-target pairs like *jasmijn* "jasmine"- *jas* "coat," did not facilitate picture naming, indicating that morphological priming represents a distinct form of priming separate from form-identity priming. While these mentioned findings align with decompositional hypotheses in compound production, it is essential to note that there are other studies that favor a full-listing model.

3.1.2.2 The full-listing model

The evidence of the full-listing model comes from Janssen and colleagues (2008). In their study, they tried to answer the question of the representation of compound words in our mental lexicon by manipulating the compound and constituent frequencies of Mandarin compound words. Native Mandarin speakers were asked to name objects in a picture naming task. There were three conditions: H(h), L(h), and L(l). High (H) or low (L) compound word frequency was denoted by the first upper-case letter, while high (h) or low (l) constituent morpheme frequency was indicated by the second lower-case letter in parenthesis. Their analysis revealed that only the compound frequency influenced naming latencies in Mandarin compound production, providing support for full-listing hypothesis. Another study by the same lab (Janssen et al., 2014) showed that this

pattern (effect of compound but not constituent frequency) extended to naming pictures using English (i.e., “oillamp” is a low-frequency compound, but its constituents are highly frequent; “bobsled” is a low-frequency compound with low-frequency constituents). However, when a lexical decision task (LDT) was used (Experiment 2), constituent frequency effects arose. Janssen et al. stated that when a semantic input representation drove word retrieval, constituent effects were absent, but when the signal in the LDT (i.e., a visually or auditorily presented word) contained the constituents, they would be accessed separately, and in that case constituent frequency had an effect.

Additional support for the full-listing hypothesis comes from a study by Bi and colleagues (2007), which investigated two Chinese aphasic patients with lexical access difficulties in oral and written production in naming disyllabic compound pictures. Their findings indicated that the frequency of the compound word, rather than the constituent, affected the production performance of both patients in Mandarin compound production.

In another investigation by Chen and Chen (2006), Mandarin Chinese speakers participated in an implicit priming task (Meyer, 1991), naming compound words in a response-association task. Their goal was to explore whether morphological encoding played a role in producing Chinese disyllabic transparent compound words. Their results showed that naming latencies were not sensitive to the compound’s constituent frequency.

The studies above provided inconsistent results and the debate on how compound words are organized and produced remains unclear. Additionally, there may be some methodological issues with earlier studies. For instance, Janssen and colleagues' study (2008) based their stimuli on the information found in the Modern Chinese Frequency Dictionary (MCFD) (Language Teaching and Research Institute of Beijing Language Institute, 1986). However, the MCFD was published forty years ago and primarily based on written texts rather than speech production, which could potentially influence the evaluation of stimuli's frequency. For example, 水池 /shui3chi4/ "basin" was deemed a high-frequency compound word in the MCFD, but it was not in the SUBTLEX-CH corpus (Cai & Brysbaert, 2010), which was based on speech (e.g., movie subtitles).

3.1.3 The current study

As we deemed the approach to exploring the question of Mandarin compounds representation and production presented in the study of Janssen and colleagues suitable for the purpose of the current study, we opted to replicate their design framework and extend their study by introducing a new set of stimuli based on the SUBTLEX-CH Mandarin speech production corpus (Cai & Brysbaert, 2010). Additionally, we utilized EEG methodology to examine the temporal dynamics of compound production in greater detail. Event-related potentials (ERPs) offer a higher temporal resolution in contrast to reaction times, allowing for more direct observation of cognitive

processes, even before an explicit response is made (Kutas & Federmeier, 2011; Kutas & Van Petten, 1994). Regarding the current issue, ERPs can serve as a direct method to investigate the frequency effect for constituents and the whole word due to words with a higher frequency tend to trigger N400s of reduced amplitude compared to words with a lower frequency when all other factors remain consistent (Kutas & Federmeier, 2009, 2011; Rugg, 1990; Van Petten & Kutas, 1990).

The prediction of the present study is to have similar reaction times when comparing H(h) to L(h) condition and shorter naming latencies for L(h) condition than for L(l) condition. Moreover, a reduced N400 amplitude was predicted in L(h) and L(l) conditions; no reduced N400 amplitude was predicted in H(h) and L(h) conditions.

3.2 Methodology

3.2.1 Participants

Thirty-two native Mandarin Chinese speakers, aged 20 to 32 years (mean age: 26.42, SD: ± 2.71), including six males, were recruited from Leiden University. All participants were from Mandarin-speaking provinces in China and spoke Mandarin as their mother tongue. Participants who had been living in the Netherlands for less than two years were included, while those who had been residing in the Netherlands for longer were excluded due to potentially higher proficiency in English and Dutch. All participants had normal or

corrected-to-normal vision and received monetary compensation for their participation. At the time of testing, none reported color blindness, learning disorders, hearing or visual impairments, or psychological or neurological conditions. Participants read an information sheet and provided informed consent by signing a consent form before the study began.

3.2.2 Materials

In the process of material design for the present replication, our approach followed the methodology outlined by Janssen and colleagues (2008). The design closely mirrored the structure proposed by Jescheniak and Levelt (1994). We curated three sets of images: (a) L(l) pictures featuring names comprised of low-frequency compound words and composed of low-frequency constituents; (b) L(h) pictures, including names formed by low-frequency compound words but consisting of high-frequency constituents; and (c) H(h) pictures, characterized by names with high-frequency compound words and composed of high-frequency constituents. Three conditions of three different frequency distributions were created for the present experiment. Although most of the compound's constituents were noun-noun pairs, there were occasional instances where a verb formed a constituent (e.g., 扫帚 /sao4zhou3/, meaning "broom," where 扫 means "to sweep"). However, these cases were rare and not expected to influence the results.

The average cumulative frequency of each constituent was calculated in the present study. For example, the Chinese compound word 山羊 /shan1yang2/ “goat” was composed of two constituents 山 /shan1/ “mountain” and 羊 /yang2/ “sheep.” The compound word frequency referred to the occurrence of 山羊 /shan1yang2/ “goat.” Cumulative morpheme frequency was the combined frequency of all homophonic constituents, disregarding their written form. This involved summing the frequencies of all homophones for each constituent and then averaging them to obtain the mean cumulative frequency in this study. For instance, when calculating the cumulative frequency of 羊 /yang2/ “sheep” within 山羊 /shan1yang2/, all its homophones like 阳 /yang2/ “sun” and 洋 /yang2/ “foreign,” etc., were considered. The resulting morpheme frequency was derived from averaging the cumulative frequencies of both constituents, namely 山 /shan2/ “mountain” and 羊 /yang2/ “sheep.” Based on the rationale outlined by Janssen and colleagues (2008), the adoption of cumulative frequency was justified by the fact that homophonic morphemes are condensed into a single lexeme node within Levelt’s model (1999) under examination in this study. The decision to utilize average cumulative frequency stemmed from its strong correlation with individual constituent frequencies. This approach offered a reliable estimation of the impact of constituent frequency on compound production.

We calculated the word and morpheme frequencies in the present study based on the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). We selected 28 disyllabic noun compounds for each condition in the stimuli set and incorporated 36 filler pictures, which were also disyllabic noun compounds. Twenty-eight black-and-white line pictures labeled as L(h) were carefully chosen, and each was paired with an L(l) picture, ensuring a match on average whole-word frequency (t (54) = -1.22, p = 0.23). Additionally, each L(h) picture was paired with an H(h) picture, ensuring a match in average morpheme frequency (t (54) < 1). The visual complexity of all target pictures was controlled (t (54) < 1) in the present study. There was a significant difference in the average morpheme frequencies between the L(h) and L(l) pictures (t (54) = -2.02, p = 0.04). The word frequencies of L(h) pictures were found to be lower than the whole-word frequency of H(h) pictures with t (54) = -10.43, p < .0001, and the word frequencies of L(l) pictures were found to be lower than the whole-word frequency of H(h) pictures as well, with t (54) = - 9.30, p < .0001. See Table 3.1 below for the detailed information.

Table 3.1: Mean frequency distribution of the picture names in the experiment.

Condition	Example	English translation	Mean compound frequency (Zipf)	Mean constituent frequency (Zipf)	Mean left constituent frequency (Zipf)	Mean right constituent frequency (Zipf)

H(h)	电 话 (dian4hua4)	telephone (electricity+speech)	3.62	2.52	2.39	2.65
L(h)	长 椅 (chang2yi2)	bench (long+chair)	2.12	2.54	2.56	2.51
L(l)	蜡 烛 (la4zhu2)	candle (wax+candle)	1.93	2.43	2.43	2.42

The fillers were used both as warm-up stimuli at the beginning of each block and as fillers throughout the experiment. We did not control the frequencies of fillers because we did not analyze the reaction time to filler pictures. The complete list of stimuli is listed in *Appendix 3.A*.

3.2.3 Design

A 3 by 3 factorial within-subject design was adopted in this experiment, with frequency distribution and repetition level as fixed factors. A picture-naming task was employed and a pseudo-randomized design per participant was implemented. Pictures with the same category and the same phonological onset would not be subsequently presented to avoid priming effects of the same category and phonological overlap.

Participants were presented with 120 pictures, comprising 84 experimental pictures and 36 fillers. The proper experiment consisted of three blocks and each picture appeared once per block. We

introduced repetition level as an experimental factor (three repetitions), and each participant encountered each picture three times during the experiment. Consequently, there were 360 trials in total for each participant in the entire real experimental session.

3.2.4 Procedure

The experiment was designed and controlled using *E-prime 3.0* (Psychology Software Tools) and was conducted in a soundproof booth. Participants were seated in front of a computer in a dimly lit room. A microphone connected to a Chronos response device containing a voice key was used to record naming latencies.

The experiment comprised three phases. The initial familiarization phase involved familiarizing participants with the pictures for the subsequent experiment. Each trial began with a 500 ms presentation of a fixation cross. Subsequently, the picture appeared for 2,000 ms, followed by the display of its name underneath it after 1,000 ms. Upon seeing its name, participants were instructed to name the picture verbally, and a new trial commenced after a delay of 1,000 ms.

The following practical part involved participants practicing the experimental task for all pictures, while the third part constituted the actual experiment. The trial structure for the second and third parts was identical. Participants were first presented with a 700 ms fixation cross, followed by the picture display for 1,500 ms or until the participant

made a vocal response. Subsequently, there was a blank of 2,000 ms before the start of the subsequent trial (see Figure 3.1).

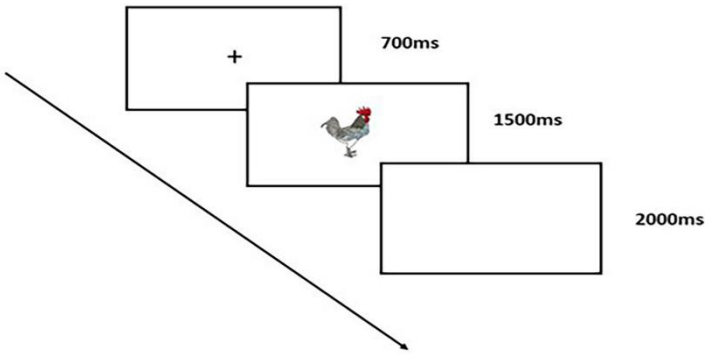


Figure 3.1: A trial sequence for the picture-naming task.

The experimental session consisted of three repetitions with 120 trials in each repetition. There was a break within each repetition. Therefore, the entire experimental session was partitioned into six blocks. The experimenter documented the validity of each trial by noting target language errors, word errors, and voice-key errors. Participants did not receive feedback during the experiment.

3.2.5 Electrophysiological recording and data processing

EEG data were collected using *Brain Vision Recorder* software (Version 1.23.0001) by Brain Products GmbH. An EasyCap electrode cap was employed following the standard 10/20 montage (see Figure 3.B.1 in *Appendix 3.B*). We recorded EEG data from 32 electrodes (*BioSemi Active Two*) placed on the scalp according to the American Electroencephalographic Society standards (1991). The vertical

electrooculogram (*VEOG*) was recorded from two external facial electrodes placed above and below the participant's left eye, and the horizontal electrooculogram (*HEOG*) was recorded from two electrodes at the outer canthus of each eye. Additionally, two flat electrodes were positioned at the mastoids. *CMS* and *DRL* electrodes served as ground reference. The EEG signal was later re-referenced offline using the mean of the two mastoids. Data were sampled at a rate of 512 Hz from DC to 102.4Hz (analogue anti-aliasing filter frequency at 1/5th of the sampling rate). The voltage amplitudes were measured approximately every 1.96 ms. A band-pass filter of 0.01-30 Hz was applied offline, following procedures outlined in previous studies (Koester & Schiller, 2008; Lensink et al., 2014).

3.3 Data analysis and results

3.3.1 Behavioral and EEG data exclusion

The data from three participants were excluded from the analysis due to their high error rates in naming pictures and the high rate of artifacts in their EEG data. Error trials (3.06%) and outlier trials (2.75%) with reaction times that deviated more than 2.5 SDs from the mean per participant per condition were eliminated. Eleven items (12.74%) were removed because participants consistently made mistakes in naming these pictures during the experiment. In total, 27.65% of data trials were excluded from further RT analysis. For EEG data, the data for three participants and eleven items were removed. Error trials

and outliers were also eliminated from the ERP data analysis based on behavioral data. Artifact rejection (22.65%) was administered during the processing stage.

3.3.2 Behavioral data analysis

Behavioral data were analyzed using *RStudio* Version 4.2.2. We first calculated descriptive statistics for naming latencies for each condition (see Table 3.2). Then, we used a single-trial modeling approach applying the *lme4* package. We employed a generalized linear mixed effect model (GLMM) using the *glmer()* function with a gamma distribution to model positively skewed RT data.

Table 3.2: Mean naming latencies (only correct trials included) for each condition and each repetition level (n = 29).

Condition	Naming latencies (ms) per repetition				
	First	Second	Third	Mean	SD
H(h)	660	647	656	655	129
L(h)	667	635	648	650	129
L(l)	686	654	659	666	135

To prevent over-parameterization and strike a balance between Type-I error and power, we employed a strategy for selecting random effects that prioritized simplicity in our model structure given the primary manipulation (Von Grebmer zu Wolfsthurn et al., 2021a, 2021b). In our data analysis, we followed the approach advocated by Matuschek and colleagues (Matuschek et al., 2017), which emphasizes

that model selection should be guided by the underlying data. This approach involves incremental model building, with selection taking place at each step. Model comparisons and likelihood ratio tests were performed using the *anova()* function based on Akaike's Information Criterion (*AIC*) (Akaike, 1974), Bayesian Information Criterion (*BIC*) (Neath & Cavanaugh, 2012), and log-likelihood for model comparisons on each step to see whether the newly added factor had improved the model significantly. Where applicable, Tukey-corrected post-hoc contrasts were executed using the *emmeans()* function. In the model, the fixed effects included *Condition* and *Repetition*. *Subject* and *Item* were introduced as random effects in this single-trial analysis. Additionally, we incorporated an interaction effect between *Condition* and *Repetition*.

For naming latencies, the model with the best fit was *RT ~ Condition + Repetition + (1 + Repetition | Subject) + (1 + Condition | Subject) + (1 | Item)* (see Table 3.C.1 in *Appendix 3.C*). The results showed that the naming latencies of H(h) and L(h) conditions were not significantly different with $\beta = 4.85$, $SE = 4.96$, $z = 0.98$, $p = 0.59$; the conditions of H(h) and L(l) yielded a significant difference in naming latencies with $\beta = -12.38$, $SE = 4.33$, $z = -2{,}86$, $p = 0.01$; L(h) and L(l) conditions elicited significantly different RTs with $\beta = -17.23$, $SE = 5.88$, $z = -2.93$, $p = 0.01$. No interaction effect between *Condition* and *Repetition* was observed ($F = 1.53$, $p = 0.19$), and therefore the interaction was removed from this model.

3.3.3 EEG data analysis

EEG data were preprocessed using *Brain Vision Analyzer* 2.1 (Brain Products GmbH), following the guidelines on its website (https://www.brainproducts.com/downloads/analyzer). The preprocessing pipeline involved several steps: initial visual inspection of the signal, re-referencing, and linear derivation for HEOG and VEOG electrodes. The offline recordings were re-referenced to the average of the left and right mastoid electrodes, followed by filtering with a low-pass filter set at 0.01 Hz and a high-pass filter set at 30 Hz. Subsequently, ocular correction and artifact rejection procedures were applied. Signal segmentations were explicitly applied to correct trials, creating epochs centered around stimulus onsets to investigate voltage amplitudes for the targeted event-related potential (ERP) component. Segments flagged as problematic during artifact rejection were omitted from further analysis. Baseline correction was implemented for each segment by utilizing the average EEG activity in the 200 ms preceding stimulus onset (Von Grebmer zu Wolfsthurn et al., 2021a, 2021b).

After preprocessing, the data were exported into *RStudio* for statistical analyses. We first selected the Regions of Interest (ROIs) and defined the time windows for analysis (Von Grebmer zu Wolfsthurn et al., 2021a, 2021b). To tentatively explore the locus of the effect of frequency, we conducted a permutation test using the permutes package (Voeten, 2019). This test analyzed voltage amplitudes from all electrodes within a time window between 0 and 1,200 ms across three

conditions and three repetition levels. Larger F-values, indicated by darker colors, suggest a higher likelihood of a statistically significant effect of the manipulations on voltage amplitudes. As shown in Figure 3.2, the test results highlighted potential modulations in frontal-central areas between 300 and 500 ms post-stimulus onset.

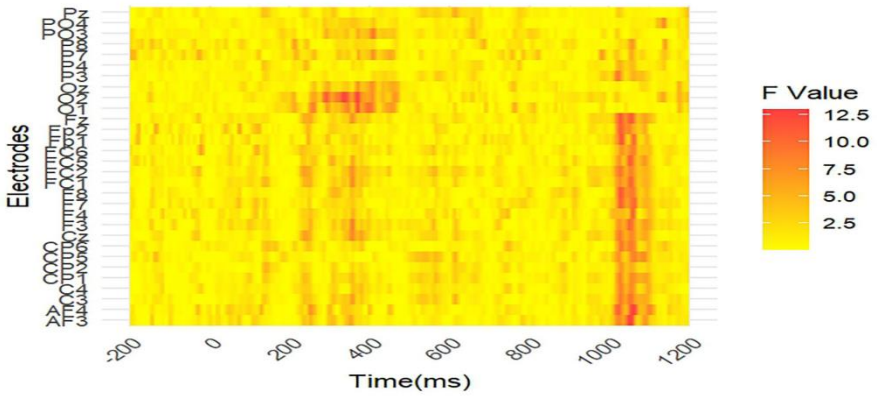


Figure 3.2: Permutation tests for three conditions and three repetition levels across all data electrodes between 0 and 1,200 ms post-stimulus onset. Larger F-values are shown in darker colors and give an increased likelihood of a statistically relevant effect of our manipulations on voltage amplitudes (n = 29).

As for the selection of electrodes in the ROIs, the permutation test highlighted ten relevant electrodes: F3, F4, Fz, FC1, FC2, Cz, C4, C3, P7, and PO3. However, not all the highlighted electrodes were relevant to the N400 component, as signal noise could also have contributed to the observed highlights. Based on previous literature (Brown & Hagoort, 1993; Lau et al., 2008; Šoškić et al., 2022), nine typical N400 electrodes, including F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4 were commonly reported as ROIs. We chose the electrodes that

overlapped the previous literature and the highlighted channels identified in our permutation test. We defined our ROIs as the following frontal and central electrodes: Cz, C4, C3, Fz, F3, and F4. Figure 3.3 illustrates the mean voltage amplitudes for the epoch of 1,200 ms for three conditions at three repetition levels. The mean amplitudes for each channel in the selected ROIs are shown in Figure 3.4.

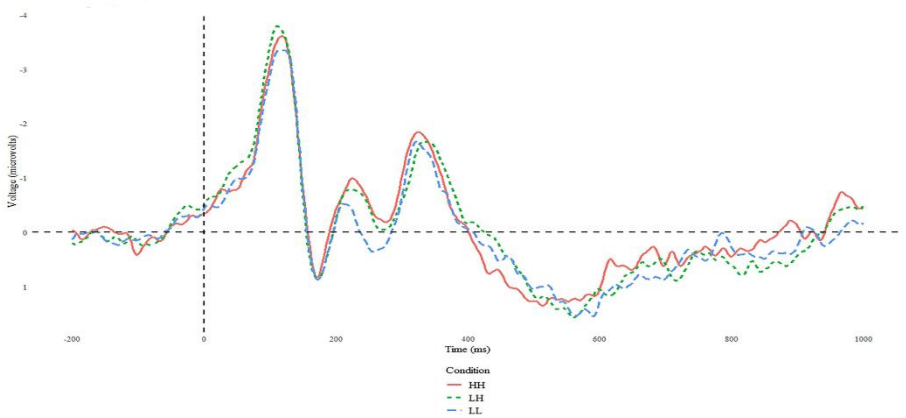


Figure 3.3: Voltage amplitudes for three conditions across three repetition levels over time for channels Fz, F3, F4, Cz, C3, and C4 in the picture-naming task (n = 29). The time window of interest is from 300 to 500 ms. Negativity is plotted up.

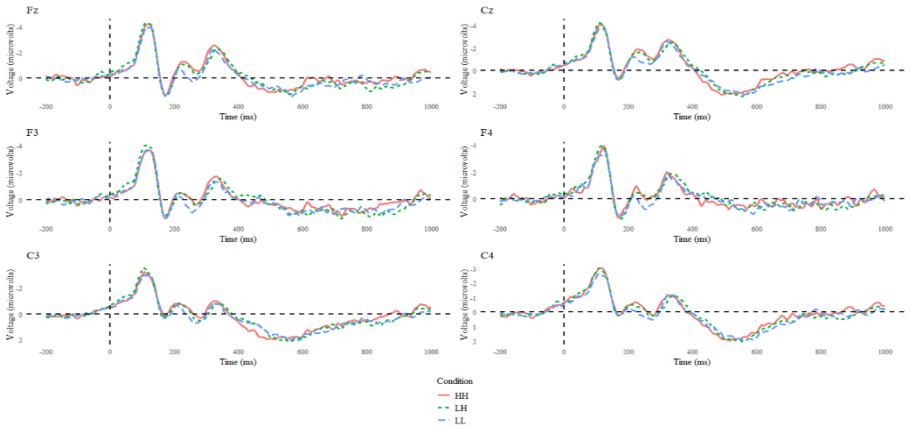


Figure 3.4: Voltage amplitudes for three conditions across three repetition levels over time for each channel of Fz, F3, F4, Cz, C3, and C4 in the picture-naming task (n = 29). The time window of interest is from 300 to 500 ms. Negativity is plotted up.

To surpass the limitations of conventional average-type analysis, we adopted a single-trial Linear Mixed Models (LMM) approach (Frömer et al., 2018; Spinnato et al., 2015; Von Grebmer zu Wolfsthurn et al., 2021a, 2021b). The traditional average-type analysis assumes equal weight for observations across conditions and participants. It assumes independence of factor levels, which often become problematic due to the intricacies of experimental designs. During EEG data preprocessing stages (Von Grebmer zu Wolfsthurn et al., 2021a, 2021b), the LMM approach addresses these limitations by adding random effects and is suitable for datasets with varying effect sizes and unbalanced designs (Baayen et al., 2008; Fröber et al., 2017). In the single-trial Linear Mixed Model (LMM) approach, we included all individual voltage values for each epoch within the selected time window of interest (300-500 ms). Rather than averaging the voltage

values across segments from the same condition, we preserved the distinct voltage values to maintain by-subject and by-item variance. The model fitting procedure followed the same steps as in the behavioral data analysis. In this model, the fixed effects included *Condition* and *Repetition*, with *Region* as a covariate to investigate differences among various brain regions. Based on previous studies (Lensink et al., 2014; Von Grebmer zu Wolfsthurn et al., 2021a, 2021b), we divided the brain into seven regions: left-anterior, right-anterior, left-medial, right-medial, left-posterior, right-posterior, and midline. *Subject* and *Item* were included as random effects in this single-trial analysis. We incorporated an interaction effect between *Condition* and *Repetition*. Model comparisons were conducted using the *anova()* function, and factors that did not significantly improve the model fit were excluded during the final model selection procedure.

The model with the best fit was *Amplitude ~ Condition + Repetition + Region+ (*1 | *Subject) + (*1 | *Item)* (see Table 3.D.1 in *Appendix 3.D*). The differences in amplitude between the H(h) and the L(h) conditions were not significant (β = 0.30, *SE* = 0.26, *t* = 1.13, *p* = 0.26). The differences in amplitude between the conditions L(h) and L(l) were not significant (β = 0.16, *SE* = 0.25, *t* = 0.67, *p* = 0.51). The effect of *Region* was significant in the right medial region (β = 0.61, *SE* = 0.02, *t* = 38.22, *p* < 0.001), right anterior region (β = - 0.18, *SE* = 0.02, *t* = - 11.24, *p* < 0.001), left medial region (β = 0.81, *SE* = 0.02, *t* = 50.81, *p* < 0.001) and mid-line channels (β = - 0.05, *SE* = 0.01, *t* = - 3.51, *p* < 0.001). These effects showed that there were significant differences

among different brain regions, indicating different processes related to word frequency. Different repetitions yielded significant differences between the first and second repetition levels (β = - 0.12, *SE* = 0.01, *t* = - 11.06, $p < 0.001$) and between the first and third repetition levels (β = - 0.35, *SE* = 0.01, *t* = - 31.07, $p < 0.001$). No interaction effect between *Condition* and *Repetition* was observed because after the model comparison of the ANOVA test, no significant improvement was found. Therefore, the interaction was not to be retained in this model.

Furthermore, based on the average amplitude of the selected channels, an N2 component was observed in Figure 3.3. To further investigate this, we conducted an exploratory analysis to test for the presence of an N2 effect within a time window between 200ms and 300ms. The results were β = - 0.07, *SE* = 0.33, *t* = - 0.24, $p = 0.81$ for H(h) and L(h) conditions and $\beta = 0.04$, *SE* = 0.31, *t* = 1.20, $p = 0.23$ for L(h) and L(l) conditions. Although no significant differences were found within the initially selected time window of 300-500 ms, we conducted an exploratory analysis using a narrower time window of 400-500 ms to assess the N400 effect. In this analysis, no significant differences were found between the H(h) and L(h) conditions with β = 0.53, *SE* = 0.35, *t* = 1.51, $p = 0.14$ or between the L(h) and L(l) conditions with $\beta = 0.14$, *SE* = 0.33, *t* = 0.43, $p = 0.67$.

3.4 Discussion

The ongoing discussion about the representation of compound words in our mental lexicon has prompted numerous studies to examine their representation. Some studies (Bien et al., 2005; Koester & Schiller, 2008, 2011; Levelt et al., 1999; Verdonschot et al., 2012) advocate the decomposed representation of compounds, while others (Chen & Chen, 2006; Janssen et al., 2008) support the full-listing approach. To further explore this topic, we replicated the design framework established by Janssen and colleagues (2008), using a new set of stimuli and integrating EEG into the design in order to investigate the research question whether the production of Mandarin compound words is influenced by morpheme or compound frequency in the present study.

Interestingly, the reaction times across conditions in our study did not align with the earlier findings of Janssen and colleagues (2008). Specifically, in our study, the L(h) condition, where word frequency was low but morpheme frequency was high, exhibited similar reaction times to the H(h) condition, where both word and morpheme frequencies were high, in picture naming tasks. In contrast, the L(h) condition was notably faster than the L(l) condition, where both word and morpheme frequencies were low. Due to the reported morpheme frequency effect, our current results aligned with other studies (Bien et al., 2005; Chen & Chen, 2015; Koester & Schiller, 2008; Roelofs, 1996) and lent support to decompositional model (Levelt et al., 1999).

The differences between two sets of stimuli may help explain the discrepancies observed between the two studies. Different corpora used could lead to different frequency distributions. Janssen and colleagues' study (2008) based their stimuli on the information found in the Modern Chinese Frequency Dictionary (Language Teaching and Research Institute of Beijing Language Institute, 1986). Our present study was based on SUBTLEX-CH corpus (Cai & Brysbaert, 2010) from movie subtitles. The differences between written and spoken corpora could potentially influence the evaluation of stimuli's frequency distributions.

The ERP results in the present study did not align with the behavioral data observed in our study. The current ERP findings revealed no significant differences in word frequency between the H(h) and L(h) conditions, nor were there morpheme frequency effects between the L(h) and L(l) conditions. While the early N2 component was noted for the L(h) and L(l) conditions and showed consistent tendency in the frontal and central regions, statistical analysis did not reveal significant differences. The time course of N2 component observed in this study was consistent with previous research (Strijkers et al., 2010), which identified an early component with approximately 180 ms after picture presentation and provided electrophysiological evidence for an early influence of frequency on speech production. Additionally, while regional effects were observed in our study, the main effect of frequency was not significant, so we did not explore this

aspect further. Future research could aim to refine these areas to gain a deeper understanding of the underlying mechanisms.

Additional factors might have accounted for the current results. For instance, this study did not control for semantic transparency, which could impact the mechanism of compound word composition. Transparent compound words are easily understood because their meaning can be inferred from the meanings of their constituents. In contrast, opaque compound words do not allow for such straightforward inference, making it necessary to store their meanings as whole units in the lexicon (Schiller, 2020; Schiller & Verdonschot, 2019). Though some studies (MacGregor & Shtyrov, 2013; Tsang et al., 2022) have reported the effects of semantic transparency, other studies also showed that opaque and transparent compound words often do not differ significantly regarding morphological priming (Koester & Schiller, 2008; Verdonschot et al., 2012; Zwitserlood et al., 2000, 2002). For example, Koester and Schiller's paper (2008) found that the production of morphologically related and complex words facilitated subsequent picture naming and resulted in a reduced N400 compared to unrelated prime words, with no significant difference observed between transparent and opaque relations. Future research could account for the semantic transparency of stimuli to better understand its influence on morphological processing.

In addition, although we recruited native Mandarin speakers and controlled the duration of their stay in other countries to mitigate the effects of their multilingual background, this factor could still

potentially influence the results. Furthermore, we did not account for the age of acquisition during the stimulus design. Despite a post-hoc analysis showing no significant difference ($t < 1$) in the effect of age of acquisition among three conditions in the present study, these speaker-related variables should be considered in future research.

3.5 Conclusion

In summary, the present study revealed that the picture naming latencies of Mandarin compound words at the behavioral level were influenced by the frequency of constituent morphemes rather than the frequency of the whole word. The behavioral data supported the predictions of a two-stage model of lexical access (Levelt et al., 1999) over a single-stage model (Caramazza, 1997). However, the ERP results did not show frequency effects at either the whole-word or morpheme-constituent levels. This discrepancy between the behavioral and ERP findings underscored the need for further research to explore and understand these deviations. The present results emphasized the importance of additional studies on morphological representations in Mandarin Chinese, which could contribute to our understanding of linguistic processes, even though they may differ from those observed in languages with more complex morphology, such as Dutch.

CRediT author contribution statement

Jiaqi Wang: Conceptualization, Methodology, Validation, Investigation, Formal analysis, Data curation, Funding acquisition, Writing-original Draft, Writing-review and editing, Visualization. **Niels O. Schiller:** Conceptualization, Writing-review and editing, Funding acquisition, Supervision. **Rinus G. Verdonschot:** Conceptualization, Methodology, Writing-review and editing, Supervision.

Acknowledgments

We sincerely thank Sarah von Grebmer zu Wolfsthurn and other group members for their invaluable support throughout the stimuli selection, data collection, and analysis phases of this study. We also extend our gratitude to the lab technicians for their assistance. We extend our gratitude to Claartje Levelt for her valuable feedback on this manuscript. Finally, we are grateful to all our participants for their participation and our anonymous reviewers for their comments and feedback.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by the China Scholarship Council (No. 202008130142).

Data availability

Data will be made available on request.

Appendix

3.A Experimental stimuli

Chinese	English	Pinyin	English for constituent	Condition
电话	telephone	diànhuà	electricity + speech	HH
警察	police	jǐngchá	police + check	HH
医生	doctor	yīshēng	medical + scholar	HH
照片*	photo	zhàopiàn	photo + slice	HH
衣服	clothing	yīfú	clothes + clothes	HH
学校*	school	xuéxiào	learn + school	HH
眼睛	eye	yǎnjīng	eye + eye	HH
医院*	hospital	yīyuàn	medical + yard	HH
飞机	airplane	fēijī	fly + machine	HH
电视	television	diànshì	electricity +vision	HH
监狱	prison	jiānyù	supervise + prison	HH
礼物	gift	lǐwù	gift + object	HH
头发	hair	tóufǎ	head + hair	HH
老师*	teacher	lǎoshī	old + master	HH
钥匙	key	yàoshi	key + key	HH

手机	cell phone	shǒujī	hand + machine	HH
城市*	city	chéngshì	city + city	HH
电脑	computer	diànnǎo	electricity + brain	HH
酒吧*	bar	jiǔbā	wine + bar	HH
汽车	car	qìchē	gasoline + car	HH
啤酒	beer	píjiǔ	beer + wine	HH
地球	earth	dìqiú	earth + ball	HH
炸弹	bomb	zhàdàn	bomb + bullet	HH
音乐	music	yīnyuè	sound + music	HH
蛋糕	cake	dàngāo	egg + cake	HH
乐队*	band	yuèduì	music + team	HH
法官	judge	fǎguān	law + officer	HH
教堂*	church	jiàotáng	teach + hall	HH
长椅	bench	chángyǐ	long + chair	LH
海马	seahorse	hǎimǎ	sea + horse	LH
水井*	well	shuǐjǐng	water + well	LH
弹弓	slingshot	dàngōng	bullet + bow	LH
手电	flashlight	shǒudiàn	hand + electricity	LH

风车	windmill	fēngchē	wind + car	LH
树枝	branches	shùzhī	tree + branch	LH
日历	calendar	rìlì	day + history	LH
钱包	wallet	qiánbāo	money + bag	LH
电车	tram	diànchē	electricity + car	LH
书包	bag	shūbāo	book + bag	LH
背心	vest	bèixīn	back + heart	LH
旗帜	banner	qízhì	flag + banner	LH
别针	pin	biézhēn	pin + needle	LH
鹿角	antlers	lùjiǎo	deer + horn	LH
蚂蚁	ant	mǎyǐ	ant + ant	LH
长城	great wall	chángchéng	long + wall	LH
轮椅	wheelchair	lúnyǐ	wheel + chair	LH
水杯	water cup	shuǐbēi	water + cup	LH
试管	test tube	shìguǎn	test + tube	LH
海盗	pirate	hǎidào	sea + robber	LH
奖杯	trophy	jiǎngbēi	prize + cup	LH
毛巾	towel	máojīn	hair + towel	LH

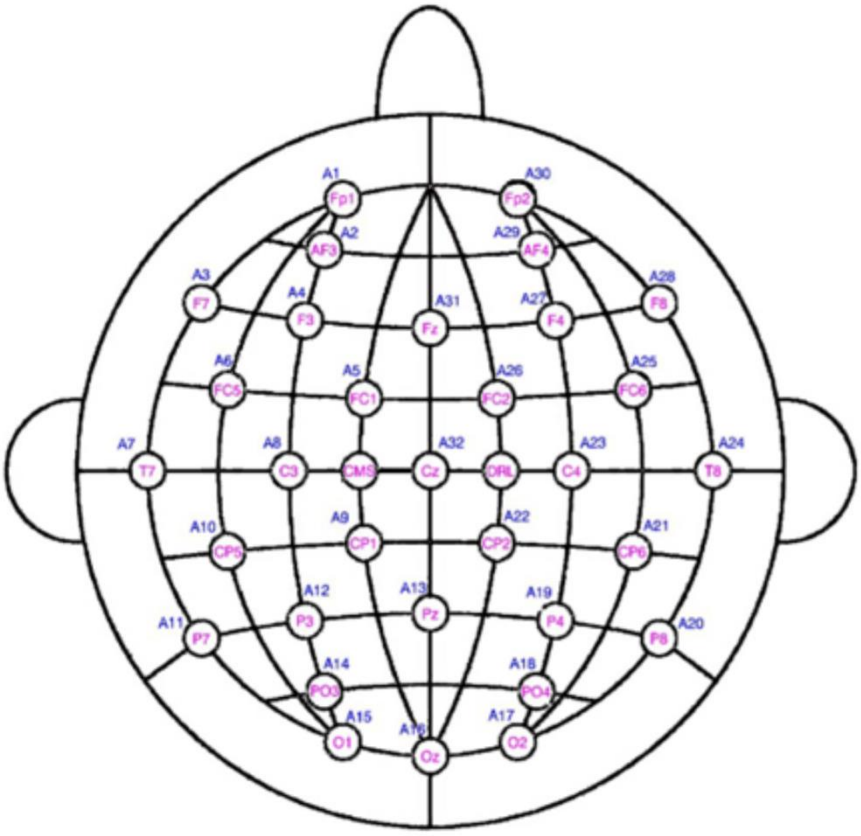
箭头	arrow	jiàntóu	arrow + head	LH
河马	hippo	hémǎ	river + horse	LH
火箭	rocket	huǒjiàn	fire + arrow	LH
面具	mask	miànjù	face + tool	LH
手铐	handcuffs	shǒukào	hand + shackle	LH
钮扣	buttons	niǔkòu	button + buckle	LL
插头	plug	chātóu	insert + head	LL
卷尺	tape measure	juǎnchǐ	roll + ruler	LL
衬衫	shirt	chènshān	lining + shirt	LL
菜板	chopping board	càibǎn	vegetable +board	LL
肩膀	shoulder	jiānbǎng	shoulder + arm	LL
仓库	storehouse	cāngkù	warehouse + storage	LL
翅膀	wing	chìbǎng	wing + arm	LL
豌豆	pea	wāndòu	pea + bean	LL
盾牌	shield	dùnpái	shield + card	LL
拱桥	arch bridge	gǒngqiáo	arch + bridge	LL
黑板	blackboard	hēibǎn	black + board	LL
胶囊*	capsule	jiāonáng	glue + bag	LL

孔雀	peacock	kǒngquè	hole + sparrow	LL
恐龙	dinosaur	kǒnglóng	fear + dragon	LL
蜻蜓	dragonfly	qīngtíng	dragonfly + dragonfly	LL
奶酪	cheese	nǎilào	milk + cheese	LL
鞭炮	firecracker	biānpào	whip + cannon	LL
竹简*	bamboo slips	zhújiǎn	bamboo + slips	LL
琵琶	lute	pípá	lute + lute	LL
漏斗	funnel	lòudǒu	leak + cane	LL
墓碑	tombstone	mùbēi	grave + monument	LL
葡萄	grape	pútáo	grape + grape	LL
拐杖	crutch	guǎizhàng	turn + cane	LL
扫帚	broom	sàozhǒu	sweep + broom	LL
雕塑	sculpture	diāosù	carve + sculpture	LL
熨斗	iron	yùndǒu	iron + cane	LL
萝卜	radish	luóbo	radish + radish	LL

*: deleted items

3.B EEG electrode montage

Figure 3.B.1: 10/20 32-channel montage from BioSemi including CMS and DRL but excluding external channels.



3.C Model parameters: response times

Table 3.C.1: Specification of best-fit model for response times (RTs) for picture naming task (n = 29). Note that estimates are reported in milliseconds.

Formula: RT ~ Condition + Repetition + (1 + Repetition | Subject) + (1 + Condition | Subject) + (1 | Item)				
Term	Estimate	[95%CI]	t-value	p-value
(Intercept)	679.03	[665.55, 685.64]	144.1	<0.001***
Condition: LH	-4.85	[-13.27, 2.93]	-0.98	0.328
Condition: LL	12.38	[2.77, 18.08]	2.86	0.004**
Repetition: 2	-28.34	[-30.02, -18.32]	-7.38	<0.001***
Repetition: 3	-17.96	[-21.96, -10.29]	-3.36	<0.001***
Random effects				
σ^2	0.03			
τ00Item	144.63			
τ00Subject [Repetition]	562.46			
τ00Subject [Condition]	509.57			
τ11Subject [Repetition2]	331.71			
τ11Subject [Repetition3]	810.72			
τ11Subject [ConditionLH]	219.55			
τ11Subject [ConditionLL]	441.29			
ICC	0.84			
NSubject	29			
NItem	73			
Observations	6033			
Marginal R^2	0.16			
Conditional R^2	1.00			

Note: σ^2: Residual variance, representing the unexplained variability in the model. τ_{00}: Variance of random intercepts, indicating how much baseline levels vary across groups (e.g., subjects or items). τ_{11}: Variance of random slopes, reflecting how much the effect of a predictor varies across groups. R^2: Proportion of variance explained by the model, indicating overall model fit.

3.D Model parameters: N400 component

Table 3.D.1: Specification of the model of best fit for Voltage Amplitudes (microvolts) (n = 29).

Formula : Amplitude ~ Condition + Repetition + Region + (1 | Subject) + (1 | Item)				
Fixed effects	Estimate	95%CI	t-value	p-value
(Intercept)	0.45	[-1.23, 0.32]	-1.15	0.26
Condition: HH	0.30	[-0.22, 0.81]	1.13	0.26
Condition: LL	0.16	[-0.32, 0.65]	0.67	0.51
Repetition: 2	-0.12	[-0.15, -0.10]	-11.06	<0.001***
Repetition: 3	-0.36	[-0.37, -0.33]	-31.07	<0.001***
Region: anterior right	-0.18	[-0.21, -0.15]	-11.24	<0.001***
Region: central left	0.81	[0.77, 0.84]	50.81	<0.001***
Region: central right	0.61	[0.58, 0.64]	38.22	<0.001***
Region: midline	-0.05	[-0.08, -0,02]	-3.51	<0.001***
Random effects				
σ^2	71.09			
τ00Item	0.80			
ICC	0.03			
NSubject	29			
NItem	73			
Observations	3389256			
Marginal R^2	0.002			
Conditional R^2	0.06			

Note: σ^2: Residual variance, representing the unexplained variability in the model. τ_{00}: Variance of random intercepts, indicating how much baseline levels vary across groups (e.g., subjects or items). τ_{11}: Variance of random slopes, reflecting how much the effect of a predictor varies across groups. R^2: Proportion of variance explained by the model, indicating overall model fit.