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

Wang, J.

Chapter 1 General introduction

1.1 Background

Human communication relies on various representations essential for effectively conveying thoughts. These representations are thought to reside in our mental lexicon - a cognitive “dictionary” that stores the forms of words we know along with their associated meanings. The nature of word representations within the mental lexicon has been the subject of much debate and research. This thesis focuses particularly on the mechanisms of compounding due to the fact that - as an essential method of word formation - compounding is a fundamental mechanism in both language processing and physical contexts.

Compounds combine morphemes (word constituents), enabling the creation of new words from existing words/morphemes by following an internal structure. Compounding plays a significant role in Mandarin Chinese word formation. For example, disyllabic compounds make up approximately 73.6% of word types and 34.3% of word tokens in the large corpus the Modern Chinese Frequency Dictionary (MCFD) is based on (Language Teaching and Research Institute of Beijing Language Institute, 1986), while monomorphemic words represent 12.0% by type and 64.3% by token (Zhou & Marslen-Wilson, 1995). Since compounding is the dominant morphological process in Mandarin - where derivational and inflectional morphology

are relatively rare-it reveals unique properties of morphological processes in Mandarin, distinguishing it from Indo-European languages.

The central theme of the present thesis is the morphological encoding of Mandarin disyllabic noun compounds in language production. Key questions include: How are Mandarin disyllabic noun compounds encoded and processed in language production? What are the underlying behavioral and neural mechanisms involved in producing these compounds? Do speakers first retrieve whole lexical entries from their mental lexicon and then decompose them into their constituent morphemes, or do they process them as complete units? At which level does decomposition occur? These questions will be explored throughout the thesis.

1.2 Language production

Language production is the process of transforming thoughts into spoken words (Schiller, 2020). Speech production models vary in their overall architecture and commonly involve assumptions about how words are organized and represented in the mental lexicon. These language production theories hold the view that language production is a complex process that can be divided into several stages, including conceptual preparation, lexical access, phonological processing, and articulation (Caramazza, 1997; Dell, 1986; Levelt et al., 1999). For instance, when we name an object, the conceptual representation of the object becomes active, activating associated lexical representations. After a lexical item has been selected, the corresponding phonological

information needs to be retrieved, and the lexical item is phonologically encoded. Finally, the abstract phonological information is phonetically encoded to yield motor programs used for articulation through the execution of the corresponding gestural scores.

Models of speech production often propose distinct sequential lexical levels, i.e., the lemma and the word-form level (Dell, 1986; Dell & O'Seaghdha, 1992; Garrett, 1980; Levelt, 1993; Levelt et al., 1999; Roelofs, 1996a, 1996b; Roelofs & Meyer, 1998; Roelofs et al., 1996). Lemmas capture a word's syntactic properties, while word forms convey segmental and metrical or supra-segmental details, including constituent morphemes in polymorphemic words (Roelofs, 1996a, 1996b; Roelofs & Meyer, 1998). In the two-stage model proposed by Levelt et al. (1999), producing compound words begins by activating the relevant semantic concept (i.e., dish washer), which, through spreading activation, co-activates semantically related concepts (e.g., washing machine, dryer, microwave, oven, etc.). This co-activation leads to the simultaneous activation of multiple, semantically related lexical entries at the lemma level, where they compete for selection until a single target lemma is chosen (Abdel Rahman & Aristei, 2010; Abdel Rahman & Melinger, 2009; Damian & Bowers, 2003).

Morphological structure influences speech production, although for a long time, language production models did not assign a distinct role to morphological processing (Schiller & Verdonschot, 2019). Evidence supporting the role of morpheme comes from some speech planning experiments and from speech errors. For instance, Roelofs

(1996) compared the naming latencies of word sets including an overlapping morpheme to a set of words with the same amount of phonological overlap, found a significantly greater facilitation effect for the former group compared to the latter when both were compared to a set of words without phonological overlap. This led him to conclude that morphemes serve as planning units in speech production. Besides, evidence from speech errors, such as “a floor full of holes” becoming “a hole full of floors” or “I carved a pumpkin” turning into “I pumped a carven” (Fromkin, 1973) further supports the crucial role of morphology in language production.

The question of how words and morphemes are structured and represented during Mandarin compound production remains unclear. To answer this question, the current thesis adopts these production models, focusing on the lexical selection stage and the morphological encoding processes of Mandarin compounds in the mental lexicon. In the following sections, we review relevant models of the process of compounds during language production.

1.3 The representations of compounds

Numerous studies emphasize the importance of morphology in speech production (Caramazza, 1997; Dell, 1986; Levelt et al., 1999). For instance, in Levelt et al.’s (1999) framework, morphology is considered the initial stage of word-form encoding. To investigate morphological encoding in compound production, two primary hypotheses are often discussed: the decompositional view, which posits

that compounds are represented at one or more levels based on their constituent morphemes (Levelt et al., 1999), and the full-listing model, which proposes that only whole-word forms are stored in the lexicon (Butterworth, 1983; Caramazza, 1997; Dell, 1986).

A less common hybrid model combines elements of both approaches, suggesting that transparent compounds (those with meanings closely related to their constituents) are processed by decomposing them, allowing for simultaneous access to whole-form and constituent representations. In contrast, opaque compounds (i.e., meaning not related to their parts), especially high-frequency ones, are processed through whole-form access (MacGregor & Shtyrov, 2013). Levelt et al.'s (1999) model may also be seen as somewhat "hybrid," as it suggests that some compound words are "degenerate" in production, meaning they are not decomposed at the form level. The current thesis focuses on the two hypotheses of compound representation, with details on these hypotheses presented in the following two sections.

1.3.1 The full-listing hypothesis

The full-listing hypothesis suggests that only whole-word forms are represented in the lexicon (Butterworth, 1983; Caramazza, 1997; Dell, 1986). There are studies showing support for this account by manipulating the frequency of compound words and their morphemes. For instance, Janssen et al. (2008) conducted a study where participants named pictures by manipulating the frequency of the compound word and its constituents using three conditions: H(h) (high compound and

high morpheme frequency), L(h) (low compound and high morpheme frequency), and L(l) (low compound and low morpheme frequency). For instance, 山羊 /shan1yang2/ “goat” was a low-frequency word, but its constituents 山 /shan1/ “mountain” and 羊 /yang2/ “sheep” were high-frequency morphemes. Their results showed that only the frequency of the whole compound word influenced naming latencies, supporting the full-listing hypothesis.

Further support for the full-listing model comes from Bi et al. (2007), who investigated two aphasic patients with difficulties in lexical access for oral and written production. Their study demonstrated that the frequency of the compound word itself, not the frequency of the constituent morphemes, affected production performance in both patients.

Chen and Chen (2006) used the implicit priming task (Meyer, 1991) to investigate whether morphological encoding played a role in producing disyllabic transparent compound words in Mandarin Chinese. Participants named compound words in a response-association task. Contrary to findings from the Dutch, their results showed that naming latencies were not sensitive to morpheme frequency, which supported a single-stage model of lexical access (Caramazza, 1997). These findings lent support to models suggesting that compound words are stored as unique lexical units, challenging theories that promote morphological decomposition during word production.

1.3.2 The decompositional hypothesis

Several studies provide evidence supporting the decomposition model, which posits that compounds are accessed through their constituent morphemes. Roelofs (1996), for instance, investigated whether the form lexicon used in speech production included morphologically decomposed entries by manipulating word frequency. His findings suggested that morphemes, as components of compound words, served as planning units in speech production, reinforcing the decomposition hypothesis. Both high- and low- frequency morpheme constituents showed a facilitatory effect - an effect that would not be expected if compounds were stored holistically, as a full-listing model would imply that constituent morpheme frequency should not impact production.

Similarly, Bien and colleagues (2005) conducted four experiments to examine the role of frequency in compound production by independently varying the frequencies of each constituent and the compound as a whole. Their results showed that compound production was sensitive to the cumulative frequency of morphemes, further supporting the decomposition model.

Chen and Chen (2015) investigated a word-onset-first planning strategy in Mandarin Chinese monomorphemic and bimorphemic compound words. Using the first character as a cue for the second, they observed a significant preparation effect, indicating that the second morpheme in a compound could be pre-planned in Mandarin. This

finding aligned with the notion of morphological decomposition in compound production.

Most studies have employed paradigms without inherent lag, such as implicit priming or form preparation paradigms, which may involve both semantic and phonological activation. In contrast, the long-lag priming paradigm has proven to be a more robust method for examining morphological representation, consistently producing reliable results across various experimental techniques - behavioral, electrophysiological, and hemodynamic - and across languages (Schiller & Verdonschot, 2018). For example, Zwitserlood and colleagues (2000) used this paradigm to investigate morphological priming, where the prime was morphologically related to the target, finding that morphological priming remained significant even after multiple intervening trials. By contrast, semantic and phonological priming effects diminished at longer lags, suggesting that morphological priming operated at a distinct level from semantic or phonological priming. This conclusion was further supported by additional studies (Zwitserlood, 2000, 2002; Kaczer et al., 2015; Koester & Schiller, 2008, 2011; Verdonschot et al., 2012; Lensink et al., 2014).

The decompositional model for compounds proposes that decomposition occurs at least at one level of lexical selection, potentially at the lemma level or at the lexeme level, or both. Therefore, another central focus in the following section of this thesis is the lemma representation of Mandarin compounds.

1.3.3 The lemma representation of compounds

According to prevailing theories of speech production, naming an object (e.g., a bird) involves activating conceptual information, retrieving the relevant lexical entry, and initiating phonological and phonetic encoding prior to articulation (Caramazza, 1997; Dell, 1986; Levelt et al., 1999). While substantial empirical evidence supports the role of morpheme-based representations in speech production (Jacobs & Dell, 2014; Janssen et al., 2008; Lüttmann et al., 2011; Roelofs, 1996b), it remains unclear at which specific level these representations operate - whether at the lemma level (Levelt et al., 1999), the lexical node level (Dell, 1986), or the word-form level. The decompositional model for Mandarin compounds suggests that decomposition occurs at least at one level of lexical selection, indicating that decomposition may occur at both the lemma and lexeme levels.

Therefore, the present thesis included two experiments to explore this question: one employed distractor that was synonymous to the first constituent of compound targets in a picture-word interference task; the other examined the concreteness effect of the constituents of compound targets in a picture-naming task. These two experiments aimed to determine whether compounds were stored solely as holistic units at the lemma level, supporting a single-lemma representation account (Levelt et al., 1999), or if they were stored in a decomposed format alongside a single lemma, supporting hybrid lemma representation accounts (Sprenger et al., 2006; Kuiper et al., 2007).

Below, we discussed two models regarding the lemma representation of compounds.

Levelt and colleagues (1999) proposed that compound words were stored holistically at the lemma level, allowing access to multiple morphemes at the word-form level. After selecting a single lemma, the constituent morphemes were retrieved at the word-form level, followed by morpho-phonological and phonetic encoding before articulation. Evidence from sequential processing suggested that a compound's morphemes were encoded one after another, with the first morpheme being processed before the second (Roelofs, 1996b). Taken together, two-stage models advocated for single, holistic compound lemmas with decomposed (morpheme-based) form representations (Dell, 1986; Levelt et al., 1999).

Alternatively, a model proposed by Sprenger et al. (2006) suggested a hybrid approach, positing that idiom representations existed alongside individual constituent representations at the lemma level. They conceptualized idiomatic forms as "super lemmas" - distinct units that encompassed information about the syntactic constraints linked to the idiom, thus outlining the syntactic properties of the individual lemmas involved. Given the parallels between idioms and compounds in mapping onto a single conceptual representation while consisting of multiple morphemes, this thesis proposed that compounds might share a similar lemma-level representation as idioms, involving a holistic lemma supported by constituent lemmas.

1.4 Electroencephalographical evidence of compounds production

This thesis combined behavioral measures, such as reaction times and naming latencies, with electrophysiological measures using electroencephalography. While most previous studies on compound word production have primarily relied on behavioral data, the precise neural correlates of morphological processing remain unclear. As a non-invasive technique for measuring event-related brain potentials, EEG could provide valuable insights by tracking brain activity that is time-locked to specific events, such as stimulus presentation during a task (Woodman, 2010). Unlike reaction times, ERPs offer high temporal resolution, enabling real-time observation of cognitive processes before an explicit response occurs (Kutas & Van Petten, 1994). For example, ERPs have effectively explored morphological decomposition processes during word comprehension in visual and auditory modalities (Krott et al., 2006; Fiorentino & Poeppel, 2007; Koester et al., 2007).

ERPs include various components that are either negative or positive in polarity, each associated with specific linguistic processing responses. In the present thesis, we focused on the N400 components and the N400 effect to investigate the morphological encoding of Mandarin compounds in language production.

The N400 is a negative voltage peak that reaches its maximum amplitude approximately 400 ms after a stimulus word appeared (Kutas

& Hillyard, 1984). Although every word elicits an N400 component, the difference in N400 amplitude between contextually appropriate and inappropriate words is known as the N400 effect. Initially linked to sensitivity about lexical semantics, the N400 is now understood to reflect the ease of integrating words into context (Schiller & Verdonschot, 2019). In the process of morphological encoding, ERPs can serve as a valuable tool for investigating whether morphological priming occurs at the word form level, offering detailed insights into the temporal dynamics of this process. For instance, one ERPs study on word reading found that the N400 component was sensitive to both lexical status and morphological decomposition (McKinnon et al., 2003).

The process of morphological decomposition affects the N400 component in ERP studies, because it is associated with the ease of integrating words at the lexico-semantic level. In the present thesis, the N400 is particularly responsive to priming and word frequency effects. First, when a target item is primed, the brain is better prepared for the upcoming stimulus, resulting in a smaller N400 amplitude. In the context of language production, Koester and Schiller (2008) examined the temporal aspects of morphological encoding using a picture naming task with Dutch compound words and a long-lag priming paradigm. Their findings highlighted a connection between the N400 component and morphological encoding. Additionally, word frequency influences N400 amplitude, with high-frequency words eliciting smaller N400 responses than low-frequency words (Kutas & Federmeier, 2009, 2011;

Rugg, 1990; Van Petten & Kutas, 1990). High-frequency words, which occur more often, are easier to integrate than low-frequency words, requiring less cognitive effort for retrieval. The present thesis investigates explicitly N400 effects of morphological priming and word frequency to provide a detailed examination of the representation of Mandarin compounds in language production.

1.5 The current study

As previously outlined, the central issue of this thesis concerned the representation of Mandarin compounds, with two dominant hypotheses under debate: the decompositional and full-listing hypotheses. Given the lack of consensus in the existing literature regarding the production of Mandarin compounds, this thesis aimed to explore this question by conducting four studies employing behavioral and electrophysiological approaches. Each study focused on different aspects of Mandarin compound representation to comprehensively investigate the issue.

Chapter 2 of this thesis examined the role of morphology in speech planning in Mandarin Chinese through a long-lag priming experiment. Thirty-two native Mandarin speakers were asked to name target pictures (e.g., 山 /shan1/ “mountain”). The study employed two types of compound primes: morpheme-related (e.g., 山羊 /shan1yang2/ “goat”) and morpheme-unrelated (e.g., 飞机 /fei1ji1/ “airplane”) primes for monomorphemic targets. The target could overlap with the

related prime in either the first constituent (e.g., target 山 /shan1/ “mountain” with prime 山羊 /shan1yang2/ “goat”) or the second constituent (e.g., target 包 /bao1/ “bag” with prime 面包 /mian4bao1/ “bread”). Both behavioral and electrophysiological data were collected using the long-lag priming paradigm. The results supported the decompositional encoding of Mandarin compounds, although the behavioral findings contradicted earlier results from Indo-European languages. Behaviorally, naming latencies for target pictures were not facilitated by morphologically related primes. However, ERP analyses revealed that morpheme-related targets elicited a reduced N400 response compared to morpheme-unrelated targets, but only when the overlap occurred in the first constituent, not the second. The discrepancy between the behavioral and EEG findings led to additional experimental designs and methods to explore this issue further in subsequent chapters.

In **Chapter 3**, we attempted to replicate the study of Janssen et al. (2008) and added the collection of EEG data. 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 frequency manipulations. As we deemed the approach to exploring the question of Mandarin compound representation and production presented in the study of Janssen and colleagues suitable for the current study, we opted to replicate their experimental design but 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. Regarding the current issue, ERPs can serve as a direct method to investigate the frequency effect for constituents and the whole word because words with a higher frequency tend to trigger N400s of reduced amplitude compared to words with a lower frequency. In the current study, despite ERPs 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 challenged the full-listing model and instead supported the decompositional model.

Chapter 4 investigated the lexical representation of Mandarin compounds through a picture-naming task using the picture-word interference (PWI) paradigm. Thirty-nine native Mandarin speakers named 45 pictures of disyllabic noun compounds, such as 公鸡 /gong1ji1/ “rooster” (lit. “male chicken”), with three types of distractors: synonymous distractors of the first morpheme (e.g., 雄 /xiong2/ “male”), distractor words semantically related to the compound (e.g., 鹅 /e2/ “goose”), and unrelated distractors (e.g., 车 /che1/ “car”). The reactions times were recorded in this experiment. Results in the present study showed that synonym distractors produced significantly faster naming times than both semantically related and unrelated distractors,

while semantically related distractors slowed responses compared to unrelated ones. These findings support a decompositional representation and a hybrid lemma representation of Mandarin compounds in the mental lexicon.

In **Chapter 5**, we explored the concreteness effect of morpheme constituents to examine how Mandarin compounds were represented in the mental lexicon, specifically whether abstract and concrete morphemes influenced the retrieval of Mandarin compounds during production. Our study addressed this question using a picture naming task with 41 participants. Two conditions were tested: the "aa" condition, representing compounds with two abstract constituents, and the "cc" condition, representing compounds with two concrete constituents. Behavioral results revealed a concreteness effect in Mandarin compounds, as naming latencies were significantly faster for compounds with two concrete constituents ("cc" condition) than those with two abstract constituents ("aa" condition). These findings supported the decompositional model, highlighting constituents' vital role in producing Mandarin compounds.

Chapter 6 integrated the findings from each of the studies presented in this thesis. In this chapter, we synthesized our research results, discussed their theoretical implications, and proposed future directions. The goal was to answer the central question posed at the beginning of this thesis: How are Mandarin compounds represented and encoded in the mental lexicon during language production?