Syllables and phonemes as planning units in Mandarin Chinese spoken word production:

Evidence from ERPs

Qingqing Qu[1 2], Chen Feng[1 2], Fengyun Hou[1 2], Markus F. Damian[3]

[1]Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

[2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

[3]School of Psychological Science, University of Bristol, United Kingdom

Word count: 11,272 (incl. abstract and references)

Address for correspondence:

Qingqing Qu
Key Laboratory of Behavioral Science
Institute of Psychology
Chinese Academy of Sciences, Beijing, China
16 Lincui Road, Chaoyang District, Beijing, China
100101
China
Tel: +86-10-64888629
Fax: +86-10-64872010
Email: quqq@psych.ac.cn

Abstract

Speakers of different languages might rely on differential phonological units when planning spoken output. In the present experiment, we investigated the role of phonemes, as well as the relative time course of syllabic vs phonemic encoding, in Mandarin Chinese word production. A form preparation task was combined with encephalography (EEG). In Experiment 1, word-initial phonemic overlap was manipulated; in Experiment 2, overlap was either in terms of phonemes or of syllables. Priming in latencies was found for syllabic but not for phonemic overlap. Phonemic overlap modulated ERPs in a 230-300 ms time window (range across Experiment 1 and 2) whereas syllabic overlap was found in a 200-280 ms time window. These results show that both phonemes and syllables are important planning units for Chinese speakers, and the relatively similar time course of activation provides important constraints on psycholinguistic models of Chinese spoken production. (143 words)

Keywords: Chinese spoken word production; phonological encoding; syllable; phoneme; form preparation task

1. Introduction

Although theories of language production disagree on many details, they concur that lexical access involves two stages of processing: a lexical selection stage, which involves accessing the word's semantically and syntactically specified representations, and a phonological encoding stage, which involves the retrieval of its sound form (see, e.g., Caramazza, 1997; Dell, 1986; Levelt, Roelofs, & Meyer, 1999; Rapp & Goldrick, 2000). Over the last few decades, much work has been carried out to investigate the processes and mechanisms underlying the two stages of lexical access. An important theoretical aim is to identify the functional units that underlie the latter stage, i.e., units that play a fundamental and necessary role in phonological encoding of lexical items. Classic theories of phonological encoding (Dell, 1986; Fromkin, 1971; Garrett, 1975; Levelt, Roelofs, & Meyer, 1999; Shattuck-Hufnagel, 1979) assume the importance of phonemes (or "segments", in the following we use the two terms interchangeably) as planning units of spoken word production. According to this view, sound forms are retrieved from the mental lexicon in terms of abstract phoneme-sized units. At the same time, syllable-sized production units are also assumed to play a role, either as abstract frames (e.g., Costa & Sebastian-Galles, 1998) or as articulatory gestural scores (e.g., Levelt et al., 1999).

The claim that phoneme-sized sounds constitute central phonological units is consistent with the intuition that spoken words are assembled from a small set of single speech sounds (or "phonemes"), but also with the observation that the most common phonological speech errors are phoneme-sized (i.e., they involve the addition, deletion, substitution, or exchange of single phonemes; see, e.g., Fromkin, 1971; Nooteboom, 1969; Shattuck-Hufnagel, 1983; Shattuck-Hufnagel & Klatt, 1979). Surprisingly, relevant studies with laboratory-based psycholinguistic tasks which would provide direct evidence for the notion of phonemes are quite scant. Early empirical evidence originated from studies conducted with the form

preparation task, also known as “implicit priming”, originally developed by Meyer (1990, 1991). In this task, speakers repeatedly produce a small set of spoken responses within short experimental blocks, and phonological overlap within a block is manipulated. A typical finding with the task is that word–initial phonological overlap between responses within a block results in a facilitatory effect on naming latencies (e.g., Meyer, 1990, 1991). This facilitatory effect is commonly interpreted as the online use of partially available information for form preparation (see Roelofs, 1997, for computational modelling of this hypothesis). Interestingly, phonological overlap between response words that involves only a single word-initial phoneme (e.g., moon-map-mouth) is sufficient to produce the form preparation effect (see Alario, Perre, Castel, & Ziegler, 2007 for French; Damian & Bowers, 2003 for English; Meyer, 1991; Roelofs, 1999 for Dutch). At the same time, deviation of word-initial sounds in just a single articulatory feature (e.g., bear-pot-bank) eliminates the preparation effect (Roelofs, 1999). These findings suggest that the phoneme is an effective planning unit for speakers. Evidence from related tasks also supports this inference. For instance, in a colored picture naming task, Damian and Dumay (2007) showed that when participants produced adjective-noun phrases in response to colored objects, a single-phoneme overlap between the color and object name facilitated responses (e.g., red rope), even when the overlapping phoneme occupied different positions within the words (e.g., green flag; Damian & Dumay, 2009). These findings also support the view originally derived from speech errors, i.e., that abstract phonemic representations constitute important planning units in spoken production.

Evidence for the importance of syllables as production-sized units comes from the demonstration of syllable frequency effects in spoken production, reported initially by Levelt and Wheeldon (1994) with Dutch speakers producing words, and subsequently also found in the production of Dutch (Cholin, Levelt, & Schiller, 2006) and Spanish (Carreiras & Perea, 2004) pseudowords, as well as English words and pseudowords (Cholin,

Dell & Levelt, 2011). Somewhat mixed evidence comes from a literature on "masked syllable priming", a technique in which a target picture or word which is to be named is preceded by a briefly presented and masked prime which either matches or mismatches the target in its syllabic structure. Early work conducted with French (e.g., Ferrand, Segui, & Grainger, 1996) and English speakers (Ferrand, Segui & Humphreys, 1997) demonstrated a syllabic priming effect (faster latencies with matching than with mismatching syllabic primes) but this effect was not found in subsequent studies (e.g., Schiller, 1998 in Dutch; Schiller, 1999, 2000, in English; see also Schiller, Costa & Colomé, 2002). Hence it appears that masked syllable priming effects are difficult to reliably obtain. Overall, the empirical pattern suggests that phonemes and syllables play different roles for speakers of Indo-European languages (Dutch, English, French, Spanish, etc.).

Different phonological units for different languages?

The evidence for phonemes and syllables described above is based on studies with speakers of Indo-European languages. However, it cannot be assumed that speakers of all languages use the same phonological planning units. For instance, syllables are particularly prominent representations in spoken Mandarin Chinese, with relatively few syllable types (approximately 1,200 counting and 400 not counting tone), clear syllabic boundaries, and no resyllabification (contrasting with a language such as English which has more than 10,000 distinct syllables, and frequent ambisyllabicity and resyllabification). Cross-linguistic variation of this kind could entail fundamental differences across languages regarding the phonological units underlying speech production. Additionally, languages such as Chinese have a logographic writing script in which orthographic characters map onto spoken syllables, but importantly no orthographic representations specifically map onto individual phonemes. Because languages with alphabetic orthographic systems explicitly code for phonemes but languages with non-alphabetic scripts do not, perhaps phonemes appear more salient for speakers of the former than the latter languages.

Indeed, some evidence suggests that phonemes are less salient for Mandarin Chinese speakers than they are for speakers of other languages, and that initial phonological encoding may be primarily governed by syllable-sized units. For instance, the most frequent phonological speech errors in Mandarin Chinese are syllable-sized, whereas phonemic errors, although occurring are quite rare (Chen, 1993, 2000). A seminal experimental finding comes from the form preparation task: when conducted with Mandarin Chinese speakers, contrary to what is found in alphabetic languages (see above), word-initial overlap in terms of a single phoneme generates a behavioral null finding, however priming is generated from word-initial syllabic overlap (Chen, Chen, & Dell, 2002; O'Seaghdha, Chen, & Chen, 2010; Zhang, 2008). This finding, suggesting that phonemes do not play an identical role for speakers of Mandarin and Western languages, motivated the "proximate unit principle" (O'Séaghdha, 2015; O'Séaghdha, Chen, & Chen, 2010) according to which the first selectable phonological unit below the level of the word (the so-called proximate unit) varies across languages. According to this view, phonemes are proximate units in Indo-European languages whereas in Mandarin Chinese, syllables constitute proximate units (and in Japanese, morae fulfil this role; Kureta, Fushimi, Sakuma & Tatsumi, 2015; Verdonschot, Kiyama, Tamaoka et al., 2011).

Interestingly, some recent evidence suggests that the difference between Indo-European languages and Mandarin Chinese regarding the role of the phoneme originates from the differences in orthography, rather than the characteristics of spoken language. For instance, Li, Wang and Idsardi (2015) replicated the null finding concerning word-initial phonemic overlap with Mandarin Chinese speakers, but critically showed that the effect emerged when stimuli were presented as Pinyin stimuli (phonetic transcription of the characters). They argued that Chinese speakers prepared the word-initial phoneme depending on cues from the specific orthographic context (Pinyin vs characters; see also Kureta et al., 2015, for similar orthographic effects with Japanese speakers). Li and Wang (2017) tested Mandarin Chinese children at a range of ages on a picture

naming version of the form preparation task, and showed that extensive experience with Pinyin cued speakers to prepare the word-initial phoneme, whereas extensive experience with Chinese characters cued them to prepare syllables.

Overall, the extant findings suggest that phonemes likely do not to occupy the role of “proximate units” that they do in Western languages. Instead, syllables may be particularly important for Chinese spoken production, a claim which is also supported by recent findings using masked syllable priming. As summarized in the initial section, studies in which briefly presented and masked primes are chosen to match or not with subsequent targets (pictures or words) have rendered mixed evidence with speakers of Indo-European languages. By contrast, with Mandarin Chinese speakers, You, Zhang, and Verdonschot (2012) found a clear masked syllable priming effect; i.e., target words or pictures preceded by masked primes which overlapped in their initial syllable structure were named faster than when preceded by mismatching primes (see also Zhang & Damian’s 2018, behavioral results). At the same time, there is also some evidence for sub-syllabic effects with Chinese speakers. Verdonschot, Lai, Chen et al. (2015) found phonological masked priming with primes and targets which overlapped in the initial consonant (C) + vowel (V) but not in the initial syllable structure. Wong and Chen (2008) employed a picture-word interference task with Cantonese speakers and manipulated form overlap between the picture and word (e.g., onset, rhyme, tone). They found facilitation not only for pictures and distractors which overlapped in terms of syllable, but critically also for rhyme-only overlap. Wong and Chen (2009) additionally showed facilitation not only from overlapping syllable, but also from shared subsyllabic CV, or VC, components. In combination, these findings make it likely that both syllabic and subsyllabic phonological units are employed by Chinese speakers with syllables being of great importance in Chinese (as evidenced by explicit encoding in the orthographic system) and therefore constituting primary or “proximate” phonological planning units, and phonemes taking on a role as "nonproximate" units which are

not the first selectable units.[1]

Experiments using electroencephalography (EEG) might be particularly suited to explore the role of phonemes and syllables in spoken production, indeed this approach has been taken in a few pioneering studies exploring Mandarin Chinese production. EEG is likely to be more sensitive than behavioral measures because mental or cognitive representations are neurocognitive in nature and thus might be reflected in neural activity even if they do not emerge in behavioral measures. Event-related potentials (ERPs) are time-locked to an external stimulus and thus could more directly reveal cognitive processing; critically, their millisecond resolution makes it possible to explore the detailed course of a cognitive process as it unfolds over time. With regard to the hypothesis of "proximate units", a plausible prediction could be a sequential pattern between syllabic and phonemic encoding: first, a syllable frame is retrieved, and subsequently, the segments of the syllable are activated in parallel and linked sequentially to positions in the syllable frame (O'Seaghdha, 2015, p. 12). However, as shown below, direct comparisons concerning the relative time course of syllabic and phonemic encoding in Mandarin Chinese word production are still scarce.

Results from a few recent EEG studies have shown phonemically based effects even with Mandarin Chinese speakers. Qu, Damian, and Kazanina (2012) asked Chinese speakers to name colored objects with adjective-noun phrases and manipulated word-initial phonemic overlap between the color and object name. As expected from previous findings (summarized above), behavioral responses (latencies and errors) showed no effect of phoneme overlap. In contrast, ERP responses were modulated by phonemic overlap from 200 to 300 ms after picture onset, providing positive evidence that phonemes are important processing units for Mandarin Chinese speakers. Similar results were reported by Yu, Mo, and Mo (2014) with a picture-naming

[1] Also note that in a recent neuroimaging study which used the fMRI adaptation paradigm, Yu, Mo, Li and Mo (2015) suggested separate phoneme- and syllable-specific neural representations (phoneme: bilateral basal ganglia; syllable: bilateral superior temporal gyrus) in Mandarin Chinese spoken word production.

priming task, in which they found that overlapping phonemes between response word pairs did not facilitate behavioral measures but modulated ERPs from 180 to 300 ms. These studies support the assumption that phonemic representations also play a role in Mandarin Chinese spoken word production. On the other hand, Wang, Wong, Wang and Chen (2017) used a delayed picture priming naming task with Cantonese speakers, and manipulated phonological overlap between consecutive pictures such that the two picture names shared either the same word-initial syllable, the same multiple word-initial phonemes, or were unrelated. Syllable overlap significant affected ERPs during response planning in a time window of 200-400 ms and 400-600 ms post picture onset. Surprisingly, overlap of multiple word-initial phonemes between prime and target (e.g., xi-xin, bi-bing) did not modulate ERPs in this study. Finally, in a recent study Zhang and Damian (2019) adopted the masked priming task and manipulated the phonological relation between prime words and target object names so that both were matched or not in terms of syllabic structure (match: CV-CV, CVN-CVN; mismatch: CV-CVN, CVN-CV). Zhang and Damian found syllabic priming (i.e., an interaction between prime and target syllable type) on response latencies, and surprisingly an inhibitory effect of phoneme overlap. Syllabic effects were found in ERPs from 300-400 ms after target object onset, whereas phonemic overlap modulated ERPs in a much later time window from 500-600 ms. The latter time interval is considerably later than the time window estimated by Indefrey and Levelt (2004) for phonological encoding, and it also conflicts with the previous ERP findings summarized above (Qu, Kazanina, & Damian, 2012; Yu, Mo, & Mo, 2014). Therefore, the authors acknowledged that the effects of phonemic overlap in their study should be regarded with some caution.

In combination, these studies paint a complex picture: experiments in which syllabic overlap/match was manipulated showed clear evidence of syllable-based effects in ERPs (Wang et al., 2017; Zhang & Damian, 2019) but there is some evidence that phonemic representations are relevant for Chinese speakers as well

(Qu et al., 2012; Yu et al., 2014; but see Wang et al. for a null finding). The time course of the two types of effects is also somewhat inconsistent: syllabic effects were found in a time window between 200-600 ms post target onset (Wang et al; Zhang & Damian); two studies found relatively "early" effects of phonemic overlap (Qu et al; Yu et al.) but one study found a relatively "late" phonemic effect (Zhang & Damian).

The present study

In the present study, we pursued the following two aims: 1) add to the evidence concerning the role of phonemes in Mandarin Chinese word production (Experiment 1); 2) explore the relative time course of syllabic vs phonemic encoding (Experiment 2). We used the form preparation task, which as outlined earlier, is widely popular in the literature on spoken word production. This task is one of the few available tasks in the literature which manipulates form properties in the production of single words; by contrast, the results of Qu et al. (2012) and Yu et al. (2014) involved multiple word production, a context which entails a separate set of complexities (see Bürki & Laganaro, 2014). To reiterate, the behavioral null finding of word-initial phonemic overlap found with Mandarin Chinese speakers in this task (Chen et al, 2002; O'Seaghdha et al., 2010) had originally motivated the "proximate units" claim. Here, we combined this task with the measurement of EEG. In Experiment 1, we manipulated word-initial phonemic overlap between spoken responses. Based on previous findings, we predicted that phonemic overlap would not affect latencies, but it might modulate ERPs. If so, this finding would further highlight the importance of phonemic representations for Mandarin Chinese speakers. In Experiment 2, both syllabic and phonemic overlap between object names were manipulated at the word-initial position: in the syllable overlap condition, object names shared a word-initial syllable, but orthographic overlap was avoided. In the phoneme overlap condition, object names shared a word-initial phoneme. By measuring EEG while participants prepared their spoken responses, we expected to identify separate modulations of ERPs associated with the planning of syllable and phoneme.

Doing so should be informative with regard to the relative time course of access to syllabic and phonemic presentations in spoken word production. Based on the proximate unit principle, our expectation was to find earlier syllabic than phonemic effects.

The form preparation task has recently been scrutinized with regard to attentional effects in the generation of the priming effect (O'Seaghdha & Frazer, 2014). In its classic form (e.g., Meyer, 1990; 1991, as well as all existing studies on Mandarin Chinese spoken production that we are aware of) participants memorize a small set of associated word pairs and subsequently produce response words based on visually presented associated prompt words. O'Seaghdha and Frazer did not recommend this version and suggested that instead, a variant should be used in which responses are elicited via simple picture naming. With speakers of Western languages, a form preparation effect has been reported with this simplified version which is comparable to the one found with the prompt-response version (see Damian, 2003, Experiment 2; O'Seaghdha & Frazer, 2014, Experiment 1; Roelofs, 1999, Experiment 3). For this reason, in our experiment, we used the picture naming version.

Experiment 1

The first experiment further investigated whether phonemes constitute planning units for Mandarin Chinese speakers. In a picture naming version of the form preparation task, we manipulated the presence or absence of word-initial phonemic overlap between picture names within an experimental set. We predicted a null effect of phonemic overlap on latencies and errors, but an effect appearing in ERP. Based on findings such as those by Qu et al. (2012), the effect should emerge in a time window of roughly 200-300 after picture onset.

Method

Participants. Thirty native Mandarin Chinese speakers (11 females; age range: 19–28 years, mean = 22.4

years) from various universities in Beijing participated in the experiment. All were right-handed and had normal or corrected-to-normal vision. The data of six participants were excluded because of excessive artifacts (the rejection rates exceeded 50%).

Materials and Design. Sixteen line drawings of common objects with disyllabic names were selected, from which sets of four items shared a single word-initial phoneme. These objects were arranged in a matrix of 4 × 4 items such that rows corresponded to four homogeneous blocks of four items each (卡车, /**k**a3che1/, truck - 裤子, /**k**u4zi/, pants - 开关, /**k**ai1guan1/, switch - 口哨, /**k**ou3shao4/, whistle) and columns formed the heterogeneous blocks (沙发/sha1fa1/, sofa - 车轮, /che1lun2/, wheel - 裤子, /ku4zi3/, pants - 拇指, /mu3zhi3/, thumb) (Fig. 1A; see Appendix A for the full set of materials). Care was taken to avoid semantic and orthographic overlap between items. Moreover, care was taken to avoid any other phonological overlap except initial phonemes.

The eight experimental blocks (four blocks for each condition) were presented in an alternating sequence of homogeneous and heterogeneous blocks. Half of the participants received homogeneous blocks first and then the heterogeneous blocks; the remaining half received the opposite order. The order in which participants received the four blocks per condition was determined by a Latin square design. Within each block, each object was presented four times, constituting 16 trials in each block, in a pseudo-random order so that there was no repetition on adjacent trials. The eight blocks were repeated three times, and thus, the entire experiment consisted of 384 trials (16 trials by 8 blocks by 3 repetitions).

Procedure. Participants were first instructed that their task was to name objects presented on the screen with the corresponding nouns as fast and accurately as possible. They were asked to refrain from moving or blinking during each trial to minimize potential EEG artifacts. Participants first received a practice list comprising 16 objects. Blocks were then presented one by one, separated by a short break.

In each block, participants were first asked to familiarize themselves with the experimental stimuli by viewing four pictures, with the expected name printed underneath each object. Each trial started with a fixation (500 ms) and then a blank screen (500 ms) followed by a target picture in the center of the screen against a white background. The target picture disappeared once the participants initiated a verbal response or after a time out of 3,000 ms. The interval between two successive trials was 2,000 ms. The experimental task session lasted approximately 30 minutes. The entire experiment lasted about 1.5 hours.

EEG Recordings and Analyses

EEG signals were recorded with 64 electrodes secured in an elastic cap (Electro Cap International) using Neuroscan 4.3 software. The vertical electrooculogram (VEOG) was monitored with electrodes placed above and below the left eye. The horizontal EOG (HEOG) was recorded by a bipolar montage using two electrodes placed on the right and left external cantus. The left mastoid electrode served as a reference. The EEG data were re-referenced offline to the average of both mastoids. All electrode impedances were kept below 5 kΩ. Electrophysiological signals were amplified with a band-pass filter of 0.05 and 70 Hz (sampling rate 1,000 Hz).

The EEGLAB toolbox based on MATLAB was used for the following procedure of preprocessing EEG signals: down-sampled to 500 Hz, filtered with a high-pass cutoff point of 0.1 Hz and a low-pass cutoff point of 30 Hz, ran ICA analysis on segmented data (−800 ms to 1,500 ms relative to the picture onset) to remove artifacts followed by manual inspection, re-referenced the data to bilateral mastoids, segmented into epochs of 600 ms (−100 ms to 500 ms relative to the picture onset, with a 100 ms pre-stimulus baseline), applied baseline correction, and rejected epochs by the criteria of −100 μv to 100 μv. Epochs with response latencies out of range from 200 ms to 2,000 ms or exceeding 3 SDs from participants' mean latency, or with missing/wrong responses, were deleted.

Data Analysis

Trials with missing (0.86%) or incorrect responses (0.39%) or with latencies faster than 200 ms, slower than 2,000 ms (0.47%), or exceeding 3 standard deviations from a participant's mean latency (1.11%) were excluded from the behavioral and ERP analyses. In ERP analysis, epochs with extensive fluctuation or artifacts were also manually rejected (28.1%). The remaining epochs were on average 132 trials per participant in each condition.

Two types of analyses were performed on the ERP data. First, mean amplitudes analyses were conducted. In order to see the distribution of effects on the scalp, 9 regions of interest (ROIs) were selected along the sagittal and coronal axes: the left-anterior (F3, F5, FC3), middle-anterior (Fz), right-anterior (F4, F6, FC4), left-middle (C3, C5), middle-middle (Cz), right-middle (C4, C6), left-posterior (P3, P5, PO3), middle-posterior (Pz), and right-posterior (P4, P6, PO4). The ROIs were in accordance, or had considerable overlap, with those of previous studies (Dell'Acqua et al., 2010; Qu, Damian, & Kazanina, 2012; Yu, Mo, & Mo, 2012; Zhu, Damian, & Zhang, 2015; Zhang & Damian, 2019). Time windows were selected according to the results of an analysis of consecutive 10-ms. At each ROI, paired *t*-test were conducted through the range from -100 ms to 500 ms with a step size of 10 ms. The time windows were selected when at least five consecutive *t*-tests approached significance ($p < 0.05$, two-tailed) or occasionally marginal significance ($p < 0.1$, two-tailed). Thereafter, mean amplitudes of the selected time windows were entered into a 2×3×3 repeated measures ANOVA with factors context (homogeneous vs heterogeneous), anteriority (anterior/middle/posterior), and laterality (left/middle/right). Greenhouse-Geisser correction was applied where appropriate, to control for violations of the sphericity assumption (original degrees of freedom are reported). All main effects and interactions involving the factor context that were significant at $p < .05$ levels were reported.

Onset latency analysis was also performed, with the aim of identifying the latency at which the ERPs of

the two conditions (homogeneous and heterogeneous) started to diverge significantly from each other. To protect against problems associated with multiple comparisons, we performed onset latency analyses using a method developed by Guthrie and Buchwald (1991) (see e.g., Costa, Strijkers, Martin, & Thierry, 2009; Qu, Zhang, & Damian, 2016; Strijkers, Costa, & Thierry, 2010; Thierry, Cardebat, & Demonet, 2003 for the use of this method in recent studies). This method estimates the critical interval for determining statistical significance, rather than happening by chance via computer simulations. If the observed number of consecutive significant time points is larger than the estimated time interval, it would indicate a statistically significant interval; otherwise it suggests it happens by chance. If there is a statistically significant interval, the onset point of a sequence of consecutive significant points is deemed as the "point of divergence".

Results

Response Latency. Behavioral data are summarized in Table 1 and Fig. 1B. A linear mixed-effect model (LMM) was used to estimate fixed effects with the *lme4* package (Bates, Maechler, Bolker & Walker, 2015) in R (R Core Team, 2013). Data were analyzed using an LMM that included context (homogeneous vs heterogeneous) as a fixed-effect factor, and random intercepts and slopes by participant and by item. The maximal random structure was kept to obtain the most reliable results. Results showed no significant effect of context (β = 3.94, *SE* = 5.58, $F_{(1,27)}$ = 0.49, p = 0.486; Bayes Factor BF_{01} = 9.06).[2] Error analysis with the *glmer* (binomial family) also did not find significant difference between the two conditions (β = -0.6, *SE* = 1.02, $z_{(1)}$ = -0.59, p = 0.557).

[2] We conducted additional analyses in which repetition was included as the fixed factor, and we obtained a main effect of repetition and an interaction between repetition and context (ps < .01). But critically, simple effect analysis revealed no significant effect of context under each repetition (all ps > 0.18).

Table 1. Mean latency (in milliseconds) with standard deviations (SD) in parentheses and error rate (%) in each context.

Context	Latency (SD)	Effect	Error	Effect
Homogeneous	582 (126)		1.41	
Heterogeneous	586 (132)	+4	0.89	-0.52

EEG Analyses

Mean Amplitude Analysis. Grand average ERP waveforms are displayed in Figure 1C for homogeneous and heterogeneous conditions and nine regions of interests chosen for the analysis. Based on the 10 ms step analyses, the time window from **240 to 300 ms** post pictures onset reached significance.[3] The results of omnibus ANOVA for the time window revealed a main effect of context ($F_{(1, 23)} = 4.59$, $p < 0.05$). To investigate the distribution of the Context effect, planned pairwise comparisons at each ROI revealed a significant effect of context in the left-anterior region ($t_{(23)} = -1.78$, $p < 0.05$), middle-anterior region ($t_{(23)} = -2.19$, $p < 0.05$), right-anterior region ($t_{(23)} = -2.45$, $p < 0.05$), left-middle region ($t_{(23)} = -1.9$, $p < 0.05$), right-middle region ($t_{(23)} = -1.75$, $p < 0.05$), middle-posterior region ($t_{(23)} = -2.14$, $p < 0.05$), reflecting more negative ERPs in the homogeneous compared to the heterogeneous condition over large electrode sites.

Onset Latency Analysis. ERPs for homogeneous and heterogeneous conditions were compared by running *t*-tests at every sampling point (every 2 ms) starting from the picture presentation. Onset latency analyses were based on the six ROIs where significant context effects emerged in the mean amplitude analysis. Onset latency analyses (Guthrie & Buchwald, 1991) suggested that in four of the six ROIs, the

[3] In Experiment 1, the only time window which approached significance in more than a consecutive 50 ms stretch was 240-300 ms whereas the later time window 350-400 m did not reach significance with five consecutive t-tests but showed a trend. Results for this later time window revealed that none of the effects involving the factor context were significant (Fs < 2.12, ps > .159). Post-hoc analysis for each ROI revealed a significant effect of context in left-middle ($t_{(23)} = -1.75$, $p < .05$) and middle-posterior ($t_{(23)} = -1.71$, $p = .05$) and a marginally significant effect in right-middle ($t_{(23)} = -1.33$, $p < .10$) and right-posterior ($t_{(23)} = -1.64$, $p < .10$) regions.

observed number of consecutive significant time points were longer than their estimated time intervals, i.e., the differences between homogeneous and heterogeneous conditions can be considered statistically reliable (left-anterior: 12-10; middle-anterior: 30-10; right-anterior: 28-12; middle-posterior: 15-11). The averaged splitting point computed from individual ROI of the four ROIs was **248** ms after picture onset.

Insert Fig. 1 here.

Fig 1. (A) Examples of stimuli used in the task. In the homogeneous condition, object names shared the word-initial phoneme in Chinese; in the heterogeneous condition, object names were phonologically unrelated. (B) Behavioral data showed no effect of context on naming latencies (color bars, left axis) or error rates (color dots, right axis). Error bars represent 95% confidence intervals. (C) Grand average ERPs for the homogeneous (red line) and heterogeneous (blue line) conditions at 9 ROIs: left-anterior (F3, F5, FC3), middle-anterior (Fz), right-anterior (F4, F6, FC4), left-middle (C3, C5), middle-middle (Cz), right-middle (C4, C6), left-posterior (P3, P5, PO3), middle-posterior (Pz), and right-posterior (P4, P6, PO4). The onset of an object was represented by 0 ms. Homogeneous ERPs were significantly more negative than heterogeneous ERPs in the 240-300 ms time interval (pink shading). ERPs of the two conditions diverged from each other beginning at 248 ms after picture onset, indicating that word-initial phoneme overlap affects speech preparation of Chinese words.

Discussion

As predicted based on earlier results such as those reported by Chen et al. (2002), word-initial phonemic overlap resulted in a behavioral null finding. By contrast, EEG results provided clear evidence for the role of the phoneme in Mandarin Chinese word production, further confirming results by Qu et al. (2012) and Yu et al. (2014) obtained with different tasks. The time window in which the phonemic effect was found (240-300 ms post target onset) is also broadly in line with the earlier studies (but Zhang & Damian's, 2019, results from a masked priming task in which weak phonemic effects emerged at a much later point in time deviate strongly from the other sets of findings). In Experiment 2, we aimed to directly compare the relative time course of phoneme and syllable processing underlying Mandarin Chinese word production. Both syllabic and phonemic overlap between spoken responses were manipulated (orthographic overlap between response words, which by itself may give rise to behavioural effects - see Qu & Damian, 2019a, 2019b – was carefully avoided). Previous studies on Mandarin Chinese spoken production have shown that syllable overlap

accelerates response latencies, relative to a condition in which spoken responses are unrelated (e.g., Chen, Chen & Dell, 2002). We predicted a parallel behavioral result here. By measuring ERPs, we expected to identify separate modulations of ERPs associated with the manipulation of syllable and phoneme. Doing so should be informative with regard to the relative time course of syllable and phoneme encoding in Mandarin Chinese word production.

Experiment 2

Method

Participants. A new group of 24 native Mandarin Chinese speakers (10 females; age range: 18–27 years, mean = 22.1 years) participated in the experiment. All were right-handed and had normal or corrected-to-normal vision.

Materials, Design and Procedure. For the phoneme condition, the same stimuli as in Experiment 1 were used. For the syllable condition, a further sixteen line drawings with disyllabic names were selected so that sets of four items shared the word-initial syllable. They were arranged into a 4*4 matrix that rows corresponded to four homogeneous blocks (橡皮, /**xiang**1pi2/, eraser - 相机, /**xiang**4ji1/, camera - 项链, /**xiang**4lian4/, necklace - 香肠, /**xiang**1chang2/, sausage) and columns to four heterogeneous blocks (橡皮, /xiang4pi2/, eraser - 眼镜, /yan3jing4/, glasses - 喇叭, /la3ba1/, trumpet - 树叶, /shu4ye4/, leaf). Semantic and orthographic relations were carefully avoided between members of each set. Across the phoneme and syllable conditions, pictures were statistically matched on the following variables: image variability, image agreement, concept familiarity, visual complexity, subjective frequency, name agreement, concept agreement, rated age-of-acquisition and objective age-of-acquisition (*p*s ≥ .131; Liu, Hao, Li & Shu, 2011). They were also matched on objectively measured word frequency, and initial and final syllable frequency (*p*s ≥ .618; Chinese Linguistic Data Consortium, 2003). See Appendix B for the full set of materials. Half of the

participants received the phoneme conditions first and the remaining half received the syllable conditions first. The procedure was identical to Experiment 1. The entire experiment consisted of 768 trials (16 trials by 16 blocks by 3 repetitions).

EEG Recordings and Data Analyses. The EEG recordings and analyses were the same as in Experiment 1. Following the same criteria as in Experiment 1, trials with missing (0.46%) or incorrect responses (0.76%) or with latencies faster than 200 ms, slower than 2,000 ms (0.80%), or exceeding 3 standard deviations from a participant's mean latency (0.93%) were excluded from the behavioral and ERP analyses. In the ERP analysis, epochs with extensive fluctuation or artifacts were manually rejected (35.9% in total). The remaining epochs were on average 120 trials per participant in each condition.

Results

Response Latency. Behavioral data are summarized in Table 2 and Fig. 2B. Data were analyzed using a LMM that included type (phoneme vs syllable) and context (homogeneous vs heterogeneous) as fixed-effect factors, random intercepts for participants and items, by-participant random slopes for type and context, and by-item random slopes for context. The results showed a significant main effect of context, $F_{(1, 40)} = 9.34$, $p < .01$, and an interaction between type and context, $F_{(1, 30)} = 13.56$, $p < .001$.[4] *Post-hoc* analysis obtained no significant effect of context for the phoneme condition, $F_{(1, 24)} = 0.06$, $p = .815$; $BF_{01} = 35.66$. For the syllable condition, word-initial syllable overlap facilitated response latencies by 25 ms, $F_{(1, 30)} = 12.86$, $p < .01$. Error analysis revealed no significant main effects or interaction (all ps ≥ .3).

[4] We conducted additional analyses in which the fixed factor repetition was included, and we obtained a main effect of repetition ($p = .039$, response latencies accelerated with repeated naming of the same stimuli). But critically, repetition did not statistically interact with any of the other factors, ps > .116.

Table 2. Mean latency (in milliseconds) with standard deviations (SD) in parentheses and error rate (%) in each context.

Context	Phoneme				Syllable			
	Latency	Effect	Error	Effect	Latency	Effect	Error	Effect
Homogeneous	555 (129)		0.76		553 (124)		0.72	
Heterogeneous	557 (127)	+2	0.63	-0.13	578 (135)	+25	0.95	+0.23

EEG Analyses

Mean Amplitude Analysis. Grand average ERP waveforms for each condition are displayed in Fig. 2C and Fig. 2D. Based on the 10 ms step analyses, the following time windows reached consecutive and robust significance: 200-280 ms for the syllable condition, and 230-290 ms and 350-400 ms for the phoneme condition. Analyses were conducted for the phoneme condition and the syllable condition separately.[5]

For the phoneme condition, in the time window of 230-290 ms, the results of omnibus ANOVA revealed a main effect of context, $F_{(1, 23)} = 4.50$, $p < .05$. All interactions involving context were not significant, all *p*s ≥ .37. *Post-hoc* analyses for each ROI showed a significant context effect on left-middle ($t_{(23)} = -2.42$, $p < .05$), left-posterior ($t_{(23)} = -2.53$, $p < .01$), middle-posterior ($t_{(23)} = -2.87$, $p < .01$), and right-posterior ($t_{(23)} = -2.56$, $p < 0.01$) regions. In the time window of 350-400 ms, ANOVA showed a marginally significant three-way interaction between anteriority, laterality and context, $F_{(4, 92)} = 2.37$, $p = .070$. *Post-hoc* analyses showed significant context effects on left-middle ($t_{(23)} = -1.85$, $p < .05$), left-posterior ($t_{(23)} = -3.510$, $p < .001$), middle-posterior ($t_{(23)} = -2.28$, $p < .05$) and right-posterior ($t_{(23)} = -2.30$, $p < .05$) regions. For the syllable

[5] To compare the magnitude of phonemic and syllabic effects, we conducted an omnibus ANOVA with type, context and brain region as factors in the 230-280 ms time window where both effects emerged. All interactions involving type × context were not significant, *p*s ≥ .135. Therefore, subsequent analyses were conducted for the phoneme condition and the syllable condition separately.

condition in the time course of 200-280 ms, results of omnibus ANOVA showed a main effect of context, $F_{(1, 23)}$ = 6.60, $p < .05$. All interactions involving context were not significant; all ps ≥ 0.38. *Post-hoc* analyses showed significant context effects on left-anterior ($t_{(23)}$ = -2.18, $p < .05$), middle-anterior ($t_{(23)}$ = -1.73, $p < .05$), right-anterior ($t_{(23)}$ = -2.81, $p < .01$), left-middle ($t_{(23)}$ = -2.16, $p < .05$), middle-middle ($t_{(23)}$ = -2.53, $p < .01$), right-middle ($t_{(23)}$ = -2.83, $p < .01$), and middle-posterior ($t_{(23)}$ = -1.98, $p < .05$) regions.[6]

Onset latency Analysis. Onset latency analyses were conducted on ROIs where significant effects emerged in the waveform analysis. For the phoneme condition, four ROIs obtained statistically reliable differences, i.e., consecutive significant time points were longer than estimated time intervals (left middle: 20-10; left posterior: 28-11; middle posterior: 26-12; right posterior: 33-12). The averaged onset latency of the four ROIs was **230** ms after picture onset. For the syllable condition, five ROIs obtained statistically reliable differences (left-anterior: 17-12; right-anterior: 45-12; left-middle: 21-12; middle-middle: 22-9; right-middle: 26-11). The averaged onset latency of the six ROIs was **215** ms after picture onset.

[6] Half of the participants received the phoneme conditions first whereas the other half received the syllable conditions first. We performed an additional analysis to examine whether the sequence of type of overlap (i.e., whether participants received the syllable or phoneme session first) modulated the effect size. Sequence was treated as a between-participants factor. For the phonemic effect, in the time window of 230-290 ms, sequence significantly interacted with context and brain regions [sequence × context × anteriority, $F(2,44) = 8.97$, $p = .001$; sequence × context × laterality, $F(2,44) = 8.03$, $p = .001$], due to the fact that participants showed a larger phonemic effect when they received the phoneme condition first (-0.60 μv) than when the syllable condition first (-0.20 μv). However, separate analyses for each sequence showed robust phonemic effects in both sequences of types. The same pattern was observed in the later time window of 350-400 ms; participants showed a larger phonemic effect in the "phoneme first" sequence. For the syllabic effect, in the 200-280 ms time window, sequence significantly interacted with context and brain regions [sequence × context × anteriority, $F(2,44) = 4.64$, $p = .015$], due to the fact that participants showed a larger syllabic effect in the syllable first (-0.48 μv) than in the phoneme first (-0.22 μv). Again, separate analyses showed robust syllabic effects in both sequences of types. Overall, both phonemic and syllabic effects were more pronounced when the corresponding experimental blocks were presented first, compared to when presented second. However, compared to Experiment 1 (in which there were only phoneme blocks), Experiment 2 showed more pronounced phonemic effect when the phoneme blocks were presented first. If the sequence effect matters here, then the phonemic effect should have been more pronounced in Experiment 1, but it was actually less so than in Experiment 2. Hence it seems that potential sequence effects cannot explain the residual differences in results between the two experiments.

Insert Fig. 2 here.

Fig 2. (A) Examples of stimuli used in the syllable conditions. In the homogeneous condition, object names shared word-initial syllable in Chinese; in the heterogeneous condition, object names were phonologically unrelated. (B) Behavioral data showed no effect of context on naming latencies (color bars, left axis) or error rates (color dots, right axis) in the phoneme condition, but a significant facilitation effect in the syllable condition on latencies (25 ms). Error bars represent 95% confidence intervals. (C) Grand average ERPs for the homogeneous (red line) and heterogeneous (blue line) conditions in the phoneme condition. (D) Grand average ERPs for the homogeneous (red line) and heterogeneous (blue line) conditions in the syllable condition. The onset of an object is represented by 0 ms. Phonemic overlap modulated ERPs in the 230-290 ms and 350-400 ms time windows, and syllabic overlap modulated ERPs in the 200-280 ms time window, indicating that both syllabic and phonemic units affect Mandarin Chinese speech preparation.

Discussion

In Experiment 2, the behavioral results showed the expected facilitation based on syllabic overlap in the absence of a phoneme-based priming effect (e.g., Chen et al., 2002). By contrast, both types of overlap showed reliable effects on EEGs, with a relatively similar onset. Particularly this latter finding – similar onsets for both types of priming – provides important information concerning theories of spoken word production, which will be discussed in more detail below.

General Discussion

The aim of the present study was to investigate the role of the phoneme, and to explore the relative time course of syllabic and phonemic encoding, in Mandarin Chinese spoken word production using an ERP technique. In a picture naming version of the form preparation task, phonological overlap was manipulated so that spoken picture names in an experimental block shared the initial phoneme (in Experiment 1) or the initial phoneme or atonal syllable (in Experiment 2); in both experiments responses were compared to blocks in which picture names were unrelated. Both experiments showed a behavioral null finding of phonemic overlap on spoken response latencies, a result which has been previously reported in this task with speakers of Mandarin Chinese (e.g., Chen et al., 2002). Critically, in both experiments, a phoneme-based effect emerged in ERPs, and with highly similar time windows (240-300 ms after picture onset in Experiment 1, 230-290 ms in Experiment 2). Experiment 2 additionally manipulated syllabic overlap: in the critical condition,

picture names overlapped in the word-initial atonal syllable but were unrelated in the control condition. Again replicating previous findings (Chen et al., 2002), syllabic overlap facilitated naming latencies. Critically, ERPs were modulated in a time window from 200-280 ms. Precise temporal analysis revealed that the phonemic effect emerged with an onset of 230 ms, whereas the syllabic effect had an onset of 215 ms. Our findings demonstrate an asymmetry between response latencies and ERPs: syllabic but not phonemic overlap resulted in behavioral priming, but ERPs reliably indicated priming arising from both types of overlap, and with a relatively similar time course.

On a broad level, the time course of form-based ERP effects in our experiments is compatible with findings from recent ERP studies on the overt speech production of Mandarin Chinese and European languages. In the production of multiple Chinese spoken words, overlap of a single phoneme between successive words modulated ERPs in a time window of 200-300 ms (in a colored picture naming task; Qu, Damian, & Kazanina, 2012) and 180-300 ms (in a picture-picture priming task; Yu, Mo, & Mo, 2014) post picture onset. In a picture naming task carried out by Spanish-Catalan bilingual speakers, Strijkers et al. (2010) observed a “cognate effect” (a facilitation effect which is dependent on phonological similarity across translations) which started at 200 ms after picture presentation. Miozzo, Pulvermüller and Hauk (2015) used a multiple linear regression approach to MEG analysis and found simultaneous effects around 150 ms of variables related to semantic and phonological processing. Similarly, Strijkers, Costa and Pulvermüller (2017) found early activation within 200 ms of phonological word properties in picture naming in a MEG study. These results broadly suggest that speakers access phonological information from pictures quite rapidly. However, it should be noted that EEG results from the picture-word interference task, in which participants are instructed to name pictures while attempting to ignore simultaneously presented distractor words, have suggested a later time course of phonological encoding. For instance, Zhu, Damian, and Zhang (2015) and

Wong, Wang, Ng, and Chen (2016) found phonologically based ERP effects at 450-600 ms and 500-600 ms respectively; Dell'Acqua et al. (2010) observed form-based ERP effects in a slightly earlier time interval of 250-400 ms after picture onset, but this interval is still later than that found in the simple picture naming tasks outlined above. One possible reason for this discrepancy might be that processing of a distractor word itself delays the normally rapid phonological encoding of the spoken utterance. It is worth noting that the interval of approximate 200-300 ms observed here and in relevant studies reviewed above such as Qu et al. (2012) and Yu et al. (2014) is somewhat earlier than the estimated onset time of phonological code retrieval proposed by Indefrey and Levelt (2004; see Indefrey, 2011 for updated version with small adaptations), who postulated that phonological code retrieval starts around 275 ms after picture onset.

In the present study, we interpret the ERP effects as reflecting facilitation from syllable/phoneme repetition arising during phonological encoding proper; syllables/phonemes that are repeatedly retrieved in the planning of a phrase may be easier to retrieve. For both syllabic and phonemic effects, we found consistent directionality of ERP amplitudes, with more negative ERPs in the homogeneous compared to the heterogeneous condition. Because previous ERP results from form preparation tasks with speakers of Indo-European languages are lacking, we are unable to compare this finding against a closely matched counterpart. However, the general pattern is in line with the results from the picture naming task of Strijkers et al. (2010) in which the cognate status of target picture names was manipulated for Spanish–Catalan bilinguals: ERPs were more negative for cognates than for non-cognates, broadly across posterior regions. In another study, Christoffels et al. (2007) also found more negative ERPs for the cognate condition compared with the noncognate condition.

Perhaps the most striking aspect of our findings is how similar the EEG time course of phonemically vs. syllabically based priming is: both types of effects had very similar onset latencies (syllables: 215 ms;

phonemes: 230 ms). How do these results inform current thinking about how phonological encoding takes place in non-Western languages? As reviewed above, behavioral findings from the form preparation task (Chen et al., 2002) had shown that for Mandarin Chinese speakers, word-initial phonemic overlap does not generate a priming effect, but syllabic overlap does. We also found this asymmetry in our own behavioral results. The contrast with findings derived from speakers of Western languages (e.g., Meyer, 1991) motivated the "proximate units" view according to which languages can differ in their primary units of phonological encoding (O'Seaghdha et al., 2010; O'Seaghdha, 2015). Proximate units are defined as the "first selectable phonological units below the level of the word or morpheme" (O'Seaghdha, 2010, p. 285) and it is assumed that for Indo-European languages, phonemes fulfil this role whereas in Mandarin Chinese, syllables are proximate units. The way in which syllables and phonemes (and tones) are selected in Chinese speakers was made more explicit in recent work by Roelofs (2015), and a schema of the proposed processing model is shown in Figure 3.

Insert Fig. 3 here.

Fig 3. Illustration of phonological encoding of the Mandarin Chinese word she2tou0 (舌头, tongue) , adapted from Roelofs (2015). The numbers below the arrows in the tonal frame denote tone numbers (2 = the mid-rising tone; 0 = the neutral tone). The links between atonal syllables and segments specify syllable positions (o = onset; n = nucleus; e = ending).

According to this framework, spoken production of a word proceeds from selection of the corresponding lexeme (word form) in two ways: first, a tonal frame specifies the corresponding tones for each syllable; second, tonally unspecified syllables are accessed. In a subsequent step, segments (phonemes) are accessed, and finally tone and syllable/segment information is merged into tonally specified syllables ("motor programs"). In this scheme, atonal syllables take on the role of "proximate units". In homogeneous experimental blocks in the form preparation task, behavioural priming in the "syllabic overlap" condition arises because the speaker is able in advance (i.e., before the stimulus appears) to select the word-initial

atonal syllable. By contrast, in the “phonemic overlap” condition, the speaker can plan the word-initial phoneme but not the non-initial ones. The processing of these non-shared phonemes sets the pace of responses and hence no behavioral priming effect is predicted for the phoneme overlap condition (as is found empirically).

How do our findings regarding the time course of phoneme- vs. syllable-based priming relate to this theoretical framework? *Prima facie*, this scheme proposes the selection of atonal syllables in a first step, and access to corresponding segments in a subsequent step. From this one may predict a “serial” pattern of syllabic and phoneme-based priming, with syllabic priming emerging earlier than phonemic priming, and this was indeed our expectation at the outset of our study. Contrary to this prediction, our EEG findings showed almost identical onset latencies for the two types of effects.

There are several potential explanations for this pattern. First, the model may be incorrect in assuming that atonal syllables take precedence over access to phonemic information, and instead, Mandarin Chinese speakers may access information about phonemes and syllables in parallel. Although this is possible, the model offers a compelling account of why with Mandarin Chinese speakers, only syllabic but not phonemic overlap results in behavioral priming. Stipulating that both types of codes are accessed in parallel would compromise this account. A second possibility is that although according to the model, syllable selection takes precedence over segmental selection according to the principle of “proximate units”, these two phases of phonological encoding should not be interpreted as successive, “serial” processing stages. On the contrary, one of the fundamental assumptions of the model is that activation “cascades” throughout the network. Hence, it is possible that the “segment” layer receives activation almost as soon as processing begins at the level of atonal syllables. In this way, the logical sequence of syllable followed by phoneme selection could still be compatible with our novel finding that the corresponding priming effects in EEG have a very similar onset.

Further research is clearly needed to disentangle these possibilities.

The results of Experiment 2 showed that the phoneme condition elicited two ERP components, one at the 230-290ms time window and the other at a later 350-400ms time window; the latter component was also present in Experiment 1, although in weaker form and not statistically significant (see footnote 3). What could be an explanation for why phonemic overlap resulted in two separate components? A similar pattern of two stages of phonemically based ERP effects was found in our previous study on phonemic effects in Chinese speaking (Qu, Damian, & Kazanina, 2012). In line with the argument presented previously, later effects might be attributed to internal self-monitoring which monitors abstract phonological codes and is estimated to take place around 355 ms after picture onset (Indefrey, 2011). The later time window of 350-400 ms observed in the present study is broadly consistent with this time estimate. The speculative argument is that when phonemes are repeated among spoken responses, the monitoring system is arguably under higher load to prevent speech errors compared with when there is no such phoneme repetition. This account is rather speculative and further investigations are needed to understand the mechanisms underlying self-monitoring in Chinese speaking.

As far as methodology is concerned, in the present study we employed the form preparation task which has been extensively used to explore word-form encoding underlying spoken production. The critical assumption underlying the paradigm is that the effect is attributable to the advance planning of partial phonological encoding which is possible in a homogeneous context (when response words share a word-initial portion) but not in a heterogeneous context. However, this assumption has been challenged, and as O'Seaghdha and Frazer (2014) recently pointed out, the standard version of the task in which response words are elicited via associated prompt words is particularly susceptible to effects of attention. It could also be that in this version, the priming effect partially reflects memory retrieval processes, with participants

establishing an episodic association between prompt and response and retrieving response words through the episodic association. Relative to heterogeneous blocks, response words in homogeneous blocks may be easier to retrieve from memory due to additional episodic retrieval cues afforded by word form overlap. For these reasons we followed O'Seaghdha and Frazer's suggestions and used a version of the task which is based on simple picture naming, which we agree is preferable to the standard form preparation version.

At least under the constraints of the stimulus selection procedure in our task, a comparison between phoneme and syllable context effects necessitates the use of different picture sets for the two types of context. Despite our best attempts to render the two sets of pictures as comparable as possible, it is acknowledged that a residual difference emerged in naming latencies: in the heterogeneous (baseline) condition, the pictures used for the "phoneme overlap" condition were named 21 ms faster than the pictures for the "syllable overlap" condition[7]; $t = 1.87$, $p = .073$. The variables which contribute to picture naming are still not fully understood; for instance, a recent Bayesian meta-analyses of previous object naming studies (Perret & Bonin, 2019) returned a null finding for visual complexity, a variable which might be of particular importance for the present study. To reiterate, our onset latency analysis suggested onsets of 230 and 215 ms for the "phoneme" and "syllable" overlap conditions. Because naming and onset latencies go in opposite directions, it is possible that our observed difference in the onset latencies might have underestimated the genuine underlying differences by a small fraction. We nonetheless maintain our view that the most striking aspect of the relative onset latencies found in our study is how similar they are across the two conditions.

As alluded to in the Introduction, behavioural evidence suggests that for Japanese speakers, phonemes

[7] Naming latencies in the homogeneous blocks are almost identical for the "phoneme" and "syllable" overlap conditions (555 vs. 553 ms). Given the difference in the heterogeneous condition between the "phoneme" and "syllable" overlap (557 vs. 578 ms), the reviewer proposed a possibility that the behavioural effect for the syllable overlap might be a consequence of the longer naming latencies in the heterogeneous condition. However, we do not think our current findings provide directly relevant evidence concerning this possibility.

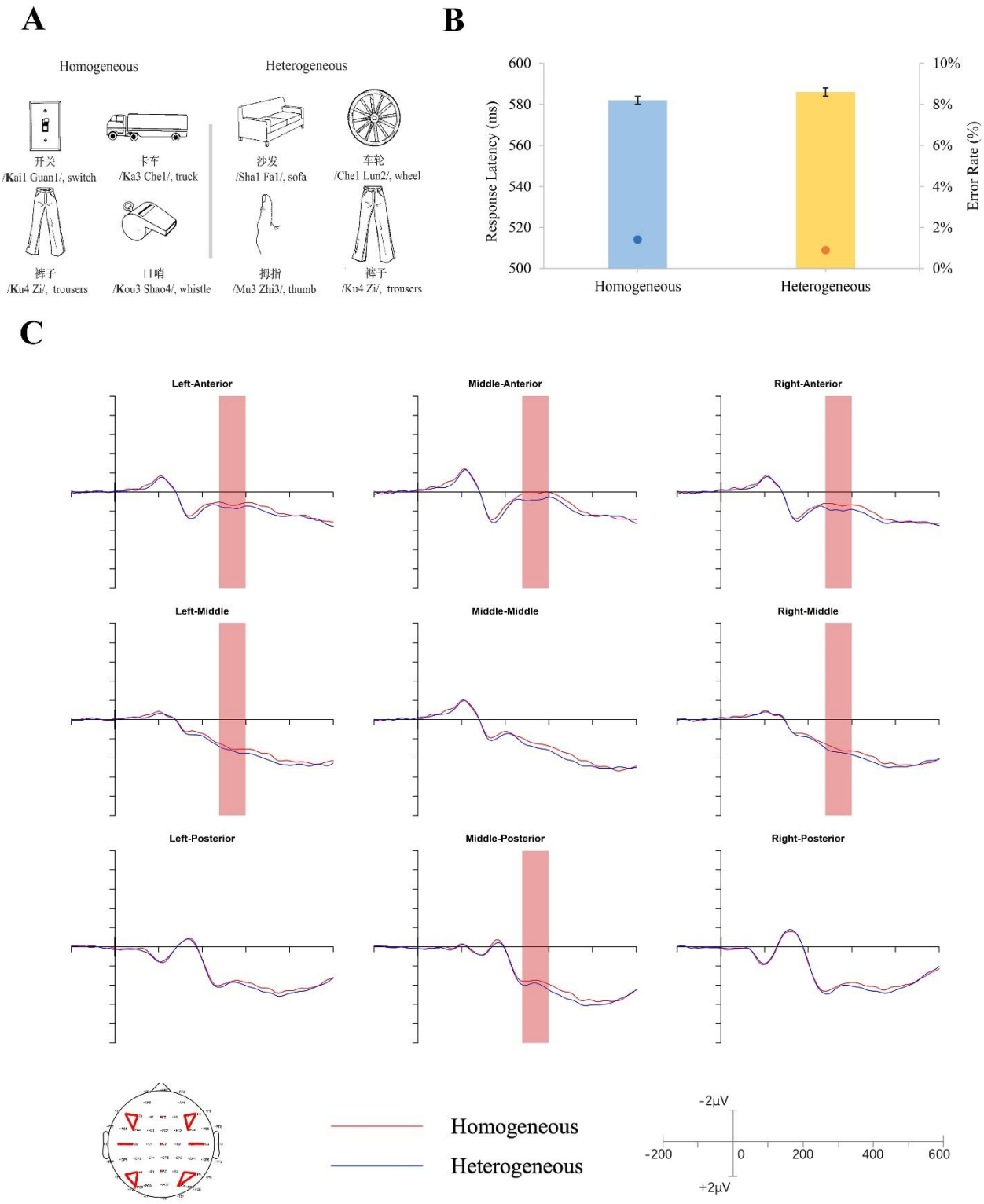
are less salient than they are for speakers of Western languages (as is the case for Chinese Mandarin speakers), and that instead the mora occupies the role of "proximate unit" in Japanese (e.g., Kureta, Fushimi, & Tatsumi, 2006; Verdonschot et al., 2011). An interesting discrepancy arises between our findings and a recent study on Japanese language production. Verdonschot, Tokimoto, and Miyaoka (2019) combined a picture-word interference task with ERP measurement, and observed that mora overlap between picture names and distractor words yielded behavioural and ERP effects. By contrast, overlap in terms of the initial phoneme between picture name and distractor word had no behavioural effects; but more importantly, ERPs were also not affected by phoneme overlap. The latter finding contrasts with our present findings, in which for Chinese speakers, we found that initial phoneme overlap did modulate ERP data. The reason for this discrepancy is at present not fully understood and could have arisen from the different target languages. However, it is also possible that the two experimental paradigms (picture-word interference vs. form preparation task) have different processing characteristics, with the former involving a complex interplay between production (picture naming) and perception (distractor processing), and the latter being relatively simpler with a single dimension (picture naming) but more salient contextual manipulation (repeated naming of few pictures per block, with presence vs. absence of initial phoneme overlap). Future research should compare and contrast the two tasks (or perhaps even combine them, e.g., Roelofs, 2002) regarding the relative role of the phoneme in non-Western language production.

To conclude, the current study provides evidence that Mandarin Chinese speakers utilize both syllables and phonemes as phonological units in word production. The previously reported asymmetry in behavioral findings (syllabic but no phonemic priming with Mandarin Chinese speakers) was also found here, but ERPs showed priming for both types of form overlap, and with similar time course. These findings provide important constraints on psycholinguistic models of Chinese spoken production.

Acknowledgements

This work was supported by the National Natural Science Foundation of China, No. 31771212, Youth Innovation Promotion Association CAS, and the German Research Foundation (DFG) and the NSFC in project Crossmodal Learning, DFG TRR-169/NSFC No. 61621136008 to Qu. We thank Dr Polly Barr for helpful comments on this manuscript.

Figure 1


A
Homogeneous
Heterogeneous
开关
/Kai1 Guan1/, switch
卡车
/Ka3 Che1/, truck
沙发
/Sha1 Fa1/, sofa
车轮
/Che1 Lun2/, wheel
裤子
/Ku4 Zi/, trousers
口哨
/Kou3 Shao4/, whistle
拇指
/Mu3 Zhi3/, thumb
裤子
/Ku4 Zi/, trousers
B
Response Latency (ms)
Error Rate (%)
600
580
560
540
520
500
10%
8%
6%
4%
2%
0%
Homogeneous
Heterogeneous
C
Left-Anterior
Middle-Anterior
Right-Anterior
Left-Middle
Middle-Middle
Right-Middle
Left-Posterior
Middle-Posterior
Right-Posterior
Homogeneous
Heterogeneous
-2μV
+2μV
-200
0
200
400
600

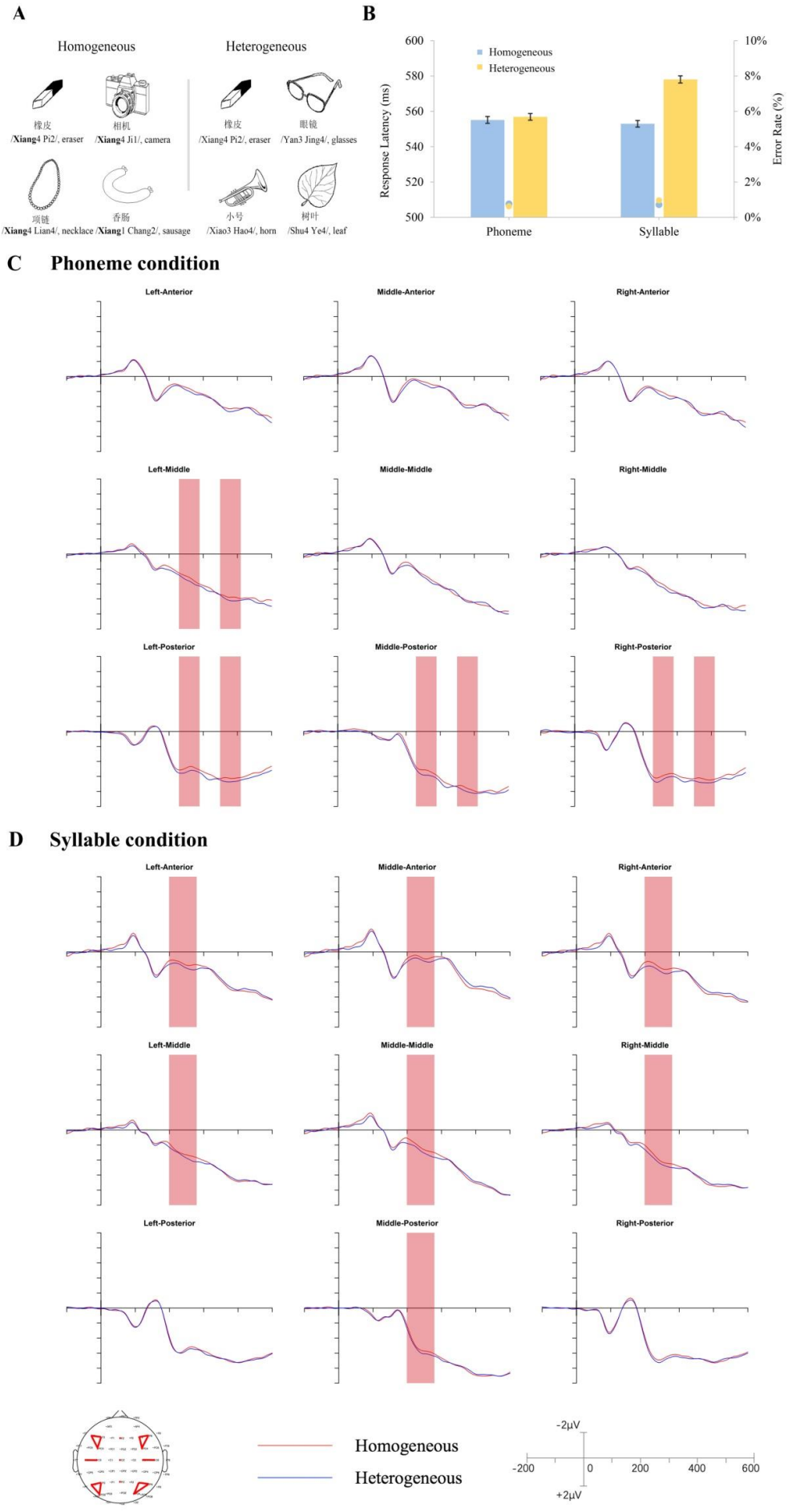
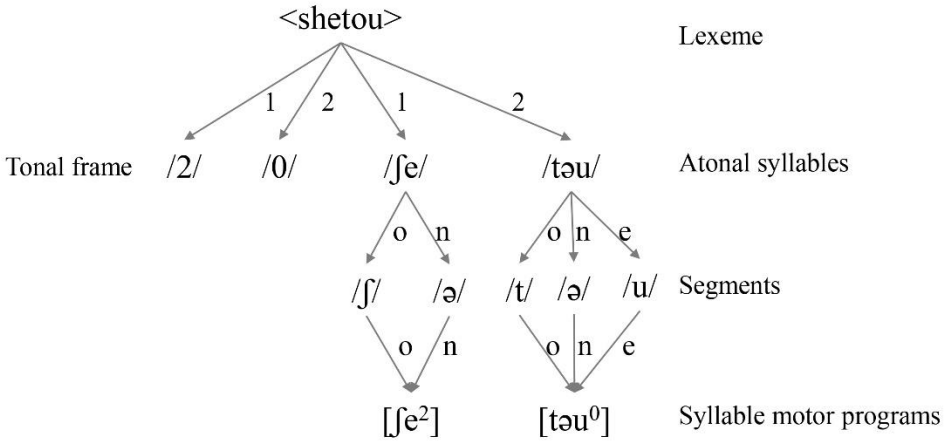

A
Homogeneous
Heterogeneous
橡皮
/Xiang4 Pi2/, eraser
相机
/Xiang4 Ji1/, camera
橡皮
/Xiang4 Pi2/, eraser
眼镜
/Yan3 Jing4/, glasses
项链
/Xiang4 Lian4/, necklace
香肠
/Xiang1 Chang2/, sausage
小号
/Xiao3 Hao4/, horn
树叶
/Shu4 Ye4/, leaf
B
Homogeneous
Heterogeneous
Response Latency (ms)
Error Rate (%)
600
580
560
540
520
500
10%
8%
6%
4%
2%
0%
Phoneme
Syllable
C Phoneme condition
Left-Anterior
Middle-Anterior
Right-Anterior
Left-Middle
Middle-Middle
Right-Middle
Left-Posterior
Middle-Posterior
Right-Posterior
D Syllable condition
Left-Anterior
Middle-Anterior
Right-Anterior
Left-Middle
Middle-Middle
Right-Middle
Left-Posterior
Middle-Posterior
Right-Posterior
Homogeneous
Heterogeneous
-2μV
+2μV
-200
0
200
400
600

Figure 3


<shetou>
Lexeme
1
2
1
2
Tonal frame
/2/
/0/
/ʃe/
/təu/
Atonal syllables
o
n
o
n
e
/ʃ/
/ə/
/t/
/ə/
/u/
Segments
o
n
o
n
e
[ʃe²]
[təu⁰]
Syllable motor programs

References

Alario, F. X., Perre, L., Castel, C., & Ziegler, J. C. (2007). The role of orthography in speech production revisited. *Cognition, 102*, 464-475.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1-48.

Bürki, A., & Laganaro, M. (2014). Tracking the time course of multi-word noun phrase production with ERPs or on when (and why) cat is faster than the big cat. *Frontiers in Psychology, 5,* 586.

Caramazza, A. (1997). How many levels of processing are there in lexical access. *Cognitive Neuropsychology, 14*, 177-208.

Carreiras, M., & Perea, M. (2004). Naming pseudowords in Spanish: effects of syllable frequency. *Brain and Language, 90*, 393–400.

Chen, J. Y., Chen, T. M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language, 46*, 751-781.

Chen, J.-Y. (1993). A small corpus of speech errors in Mandarin Chinese and their classification. *World of Chinese Language, 69*, 26–41

Chen, J.-Y. (2000). Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia, 43*, 15-26.

Chinese Linguistic Data Consortium. (2003). 现代汉语通用词表 [Chinese lexicon] (CLDC-LAC-2003-001). Beijing, China: Tsing hua University, State Key Laboratory of Intelligent Technology and Systems, and Chinese Academy of Sciences, Institute of Automation.

Cholin, J., Dell, G. S., & Levelt, W. J. M. (2011). Planning and articulation in incremental word production: syllable-frequency effects in English. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37,* 109–122.

Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition, 99,* 205–235.

Christoffels, I. K., De Groot, A. M. B., & Waldorp, L. J. (2003). Basic skills in a complex task: a graphical model relating memory and lexical retrieval to simultaneous interpreting. *Bilingualism: Language and Cognition,* 6, 201-211.

Costa, A., & Sebastian-Galles, N. (1998). Abstract Phonological Structure in Language Production: Evidence From Spanish. *Cognition, 24*, 886–903.

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 21442–21446.

Damian, M. F. (2003). Articulatory duration in single-word speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 416-431.

Damian, M. F., & Bowers, J. S. (2003). Effects of orthography on speech production in a form-preparation paradigm. *Journal of Memory and Language, 49*, 119-132.

Damian, M. F., & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language, 57,* 195-209.

Damian, M. F., & Dumay, N. (2009). Exploring phonological encoding through repeated segments. *Language and Cognitive Processes, 24*, 685-712.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.

Dell'Acqua, R., Sessa, P., Peressotti, F., Mulatti, C., Navarrete, E., & Grainger, J. (2010). ERP evidence for ultra-fast semantic processing in the picture–word interference paradigm. *Frontiers in Psychology, 1,*

177.

Ferrand, L., Segui, J., & Grainger, J. (1996). Masked priming of words and picture naming: The role of syllabic units. *Journal of Memory and Language, 35,* 708–723.

Ferrand, L., Segui, J., & Humphreys, G. W. (1997). The syllable's role in word naming. *Memory & Cognition, 25*, 458–470.

Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.

Garrett, M. F. (1975). The analysis of sentence production. In G. Bower (ed.), *The Psychology of Learning and Motivation,* Vol. 9. New York: Academic Press.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology, 28,* 240-244.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2,* 255. doi:10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*, 101–144.

Kureta, Y., Fushimi, T., Sakuma, N., & Tatsumi, I. F. (2015). Orthographic influences on the word-onset phoneme preparation effect in native Japanese speakers: Evidence from the word-form preparation paradigm. *Japanese Psychological Research, 57,* 50-60.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50,* 239-269.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-38.

Li, C., & Wang, M. (2017). The influence of orthographic experience on the development of phonological

preparation in spoken word production. *Memory & Cognition, 45*, 956-973.

Li, C., Wang, M., & Idsardi, W. (2015). The effect of orthographic form-cuing on the phonological preparation unit in spoken word production. *Memory & Cognition, 43,* 563-578.

Liu, Y., Hao, M., Li, P., & Shu, H. (2011). Timed picture naming norms for Mandarin Chinese. *PLoS ONE, 6,* e16505.

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language, 29,* 524-545.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language, 30,* 69-89.

Miozzo, M., Pulvermüller, F., & Hauk, O. (2015). Early parallel activation of semantics and phonology in picture naming: Evidence from a multiple linear regression MEG Study. *Cerebral Cortex, 25,* 3343–3355.

Nooteboom, S.G. (1969). The tongue slips into patterns. In A.G. Sciarone, A.J. van Essen, & A.A. van Raad (Eds.), *Nomen: Leyden studies in linguistics and phonetics* (pp. 114-132). The Hague: Mouton

O'Séaghdha, P. G. (2015). Across the great divide: Proximate units at the lexical-phonological interface. *Japanese Psychological Research, 57*, 4-21.

O'Séaghdha, P. G., & Frazer, A. K. (2014). The exception does not rule: Attention constrains form preparation in word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 797-810.

O'Séaghdha, P. G., Chen, J. Y., & Chen, T. M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition, 115*, 282-302.

Perret, C., & Bonin, P. (2019). Which variables should be controlled for to investigate picture naming in adults? A Bayesian meta-analysis. *Behavior Research Methods, 51,* 2533–2545.

Qu, Q. Q., & Damian, M. F. (2019a). Orthographic effects in Mandarin spoken language production. *Memory & Cognition, 47*, 326-334.

Qu, Q. Q., & Damian, M. F. (2019b). The role of orthography in second-language spoken word production: Evidence from Tibetan–Chinese bilinguals. *Quarterly Journal of Experimental Psychology, 72*, 2597-2604.

Qu, Q. Q., Damian, M. F., & Kazanina, N. (2012). Sound-sized segments are significant for Mandarin speakers. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 14265-14270.

Qu, Q. Q., Zhang, Q., & Damian, M. F. (2016). Tracking the time course of lexical access in orthographic production: An event-related potential study of word frequency effects in written picture naming. *Brain and Language, 159,* 118–126.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 460-499.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition, 64,* 249-284.

Roelofs, A. (1999). Phonological segments and features as planning Units in Speech Production. *Language and Cognitive Processes, 14*, 173-200.

Roelofs, A. (2002). Spoken language planning and the initiation of articulation. *The Quarterly Journal of Experimental Psychology Section A, 55*, 465-483.

Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: From Germanic languages to Mandarin Chinese and Japanese. *Japanese Psychological Research, 57*, 22-37.

Schiller, N. O. (1998). The effect of visually masked syllable primes on the naming latencies of words and

pictures. *Journal of Memory and Language, 39,* 484-507.

Schiller, N. O. (1999). Masked syllable priming of English nouns. *Brain and Language, 68,* 300-305.

Schiller, N. O. (2000). Single Word Production in English : The Role of Subsyllabic Units During Phonological Encoding. *Cognition, 26,* 512–528.

Schiller, N., Costa, A., & Colome, A. (2002). Phonological encoding of single words: In search of the lost syllable. In C. Gussenhoven & N. Warner (Eds.), *Papers in laboratory phonology VII* (pp. 35–59). Cambridge, England: Cambridge University Press.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In: Sentence processing: Psycholinguistic studies presented to Merrill Garrett, W. E. Cooper & E, C. T. Walker (eds.). Lawrence Erlbaum.

Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. E. MacNeilage (Ed.), *The production of speech* (pp. 109-136). New York: Springer.

Shattuck-Hufnagel, S. & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior, 18,* 41-55.

Strijkers, K., Costa, A., & Pulvermüller, F. (2017). The cortical dynamics of speaking : Lexical and phonological knowledge simultaneously recruit the frontal and temporal cortex within 200 ms. *NeuroImage, 163,* 206–219.

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking lexical access in speech production: electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex, 20,* 912-928.

Thierry, G., Cardebat, D., & Démonet, J.-F. (2003). Electrophysiological comparison of grammatical processing and semantic processing of single spoken nouns. *Cognitive Brain Research, 17*, 535-547.

Verdonschot, R. G., Kiyama, S., Tamaoka, K., Kinoshita, S., La Heij, W., & Schiller, N. O. (2011). The functional

unit of Japanese word naming: Evidence from masked priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1458.

Verdonschot, R. G., Lai, J., Chen, F., Tamaoka, K., & Schiller, N. O. (2015). Constructing initial phonology in Mandarin Chinese: Syllabic or subsyllabic? A masked priming investigation. *Japanese Psychological Research,* 57, 61-68.

Verdonschot, R. G., Tokimoto, S., & Miyaoka, Y. (2019). The fundamental phonological unit of Japanese word production: An EEG study using the picture-word interference paradigm. *Journal of Neurolinguistics, 51,* 184-193.

Wang, J., Wong, A. W. K., Wang, S., & Chen, H. C. (2017). Primary phonological planning units in spoken word production are language-specific: Evidence from an ERP study. *Scientific Reports, 7,* [5815].

Wong, A. W. K., & Chen, H. C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34,* 1172-1190.

Wong, A. W. K., & Chen, H. C. (2009). What are effective phonological units in Cantonese spoken word planning? *Psychonomic Bulletin & Review, 16*, 888-892.

Wong, A. W. K., Wang, J., Ng, T. Y. & Chen, H. C. (2016). Syllabic encoding during overt speech production in Cantonese: Evidence from temporal brain responses. *Brain Research, 1648,* 101-109.

You, W., Zhang, Q., & Verdonschot, R. G. (2012). Masked syllable priming effects in word and picture naming in Chinese. *PLOS ONE, 7*(10), e46595

Yu, M., Mo, C., & Mo, L. (2014). The role of phoneme in mandarin Chinese production: Evidence from ERPs. *PLOS ONE, 9*(9), e106486.

Yu, M., Mo, C., Li, Y., & Mo, L. (2015). Distinct representations of syllables and phonemes in Chinese production: Evidence from fMRI adaptation. *Neuropsychologia, 77*, 253-259.

Zhang, Q. (2008). Phonological encoding in monosyllabic and bisyllabic Mandarin word production: Implicit priming paradigm study (in Chinese). *Acta Psychologica Sinica, 40*, 253-262.

Zhang, Q., & Damian, M. F. (2019). Syllables constitute proximate units for Mandarin speakers: Electrophysiological evidence from a masked priming task. *Psychophysiology, 56,* 1–15.

Zhu, X., Damian, M. F., & Zhang, Q. (2015). Seriality of semantic and phonological processes during overt speech in Mandarin as revealed by event-related brain potentials. *Brain and Language, 144,* 16-25.

Appendix A

Materials used in Experiment 1 and the phoneme condition of Experiment 2

	Heterogeneous Blocks			
Homogeneous Blocks	舌头(/**sh**e2tou0/, tongue)	沙发(/**sh**a1fa1/, sofa)	书包(/**sh**u1bao1/, schoolbag)	试管(/**sh**i4guan3/, test tube)
	尺子(/**ch**i3zi0/, ruler)	车轮(/**ch**e1lun2/, wheel)	插头(/**ch**a1tou2/, plug)	厨房(/**ch**u2fang2/, kitchen)
	卡车(/**k**a3che1/, truck)	裤子(/**k**u4zi0/, pants)	开关(/**k**ai1guan1/, switch)	口哨(/**k**ou3shao4/, whistle)
	蘑菇(/**m**o2gu1/, mushroom)	拇指(/**m**u3zhi3/, thumb)	蜜蜂(/**m**i4feng1/, bee)	马桶(/**m**a3tong3/, toilet)

Note: The number denotes the tone for the preceding syllable. There are four tones in Mandarin. The number 0 represents a neutral tone.

Appendix B

Materials used in the syllable condition of Experiment 2

	Heterogeneous Blocks			
Homogeneous Blocks	燕子(/**yan**4zi0/, swallow)	颜料(/**yan**2liao4/, paint)	烟囱(/**yan**1cong1/, chimney)	眼镜(/**yan**3jing4/, glasses)
	拉链(/**la**1lian4/, zipper)	蜡烛(/**la**4zhu2/, candle)	辣椒(/**la**4jiao1/, pepper)	喇叭(/**la**3ba0/, trumpet)
	书架(/**shu**1jia4/, bookshelf)	竖琴(/**shu**4qin2/, harp)	鼠标(/**shu**3biao1/, mouse)	树叶(/**shu**4ye4/, leaf)
	香肠(/**xiang**1chang2/, sausage)	项链(/**xiang**4liang4/, necklace)	相机(/**xiang**4ji1/, camera)	橡皮(/**xiang**4pi2/, eraser)

Note: The number denotes the tone for the preceding syllable. There are four tones in Mandarin. The number 0 represents a neutral tone.