Database of word-level statistics for Mandarin Chinese (DoWLS-MAN)

Karl David Neergaard, Hongzhi Xu, James Sneed German, Chu-Ren Huang

Database of Word-Level Statistics for Mandarin Chinese (DoWLS-MAN)

Karl David Neergaard [1], Hongzhi Xu[2], James Sneed German[3], Chu-Ren Huang[4]

[1]Department of English, University of Macau, Macau S.A.R., China
[2]Institute of Corpus Studies and Application, Shanghai International Studies University, Shanghai, China
[3]Laboratoire Parole et Langage, Aix-Marseille University, Aix-en-Provence, France
[4]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong S.A.R., China

Address for correspondence
Karl David Neergaard
Department of English (E21-1060)
University of Macau
Avenida da Universidade, Taipa, Macau S.A.R.
karlneergaard@gmail.com

Abstract
In this article we present the Database of Word-Level Statistics for Mandarin Chinese (DoWLS-MAN). The database addresses the lack of agreement in phonological syllable segmentation specific to Mandarin by offering phonological features for each lexical item according to 16 schematic representations of the syllable (8 with tone and 8 without tone). Those lexical statistics that differ per phonological word and nonword due to changes in syllable segmentation are of the variant category and include subtitle lexical frequency, phonological neighborhood density measures, homophone density, and network science measures. The invariant characteristics consist of each items' lexical tone, phonological transcription, and syllable structure among others. The goal of DoWLS-MAN is to provide researchers both the ability to choose stimuli that are derived from a segmentation schema that supports an existing model of Mandarin speech processing, and the ability to choose stimuli that allow for the testing of hypotheses on phonological segmentation according to multiple schemas. In an exploratory analysis we illustrate how multiple schematic representations of the phonological mental lexicon can aid in hypothesis generation, specifically in terms of phonological processing during reading Chinese orthography. Users of the database can search among over 92,000 words, over 1,600 out-of-vocabulary Chinese characters, and 4,300 phonological nonwords according to either Chinese orthography, pinyin, or ascii phonetic script. Users can also generate a list of phonological words and nonwords according to user defined ranges and categories of lexical characteristics. DoWLS-MAN is available to the public for search or download at https://dowls.site.

Keywords: lexical database, phonological neighborhood density, Mandarin Chinese, syllable segmentation, network phonology

1. Introduction

When it comes to testing hypotheses about the nature of speech processing, researchers from multiple disciplines, such as psychology, linguistics, neuroscience, and education, rely on lexical databases for their selection of experimental stimuli. A defining feature of the existing databases (e.g., Baayen, Piepenbrock, & Gulikers, 1995; Balota et al., 2007; Davis & Perea, 2005; Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013; Marian, Bartolotti, Chabal, & Shook, 2012; New, Pallier, Brysbaert, & Ferrand, 2004; Strand, 2013) is that they offer lexical statistics, for primarily European languages that are a priori built on but one segmentation schema. The ability to rely on a sole schema in languages such as English, Dutch, and French, has depended on evidence of spoken error production (Dell et al. 2000; Shattuck-Hufnagel 1979), on-the-fly resyllabification (Levelt 1989), priming paradigms (e.g., Alario, Perre, Castel, & Ziegler, 2007; Dufour & Peereman, 2003; Meyer & Schriefers, 1991; Schiller, 2000, 2004), and simulations with computational models (e.g., Levelt, Roelofs, & Meyer, 1999; McClelland & Elman, 1986; Norris, 1994) where phonemes that consist of individual segments or near-segmental units, such as diphthongs, linearly construct syllables in metrical (Levelt 1989) or representational frames (Dell 1986). A renewed interest in syllable segmentation over the last decade, particularly with speakers of Mandarin Chinese, has challenged the generalizability of segment-sized phonemes. In the current article, we review the building literature on Mandarin segmentation in speech perception and production and then introduce a lexical database for Mandarin that allows users to source lexical statistics built upon multiple segmentation schemas. Finally, we illustrate how the database can be used to generate hypotheses through an analysis of a megastudy of orthographic lexical decision reaction times.

Mandarin has garnered multiple proposals as to its segmentation schema. In order to encompass them all, we begin with the maximal syllable, CGVX, in which C represents initial consonants, G pre-nuclear glides, V monophthongs, and X post-nuclear glides or final consonants. In line with previous research (Neergaard 2018; Neergaard and Huang 2019; Neergaard, Xu, and Huang 2016), we will rely on underscores between segmental units to denote phonological units. Note that the use of underscores is meant to distinguish a word-level annotation, such as VC for the vowel-consonant monosyllable 昂 *ang2* /aŋ^{35}/ (‘to lift’), from a lexicon-level segmentation schema. In Table 1 we illustrate the use of underscores with the disyllabic word, 小巷 *xiao3xiang4* /ɕiaʊ^{214}ɕiaŋ^{51}/ (‘alley’) and its monosyllabic constituents. The early work on Mandarin segmentation approached the topic through the use of theoretical arguments or language games (for a review and discussion, see Duanmu 2011). Researchers omitted tonal information so as to focus on the unitization of segments. Some proposals argued for complex rimes, including, C_G_VX /ɕ_i_aʊ_ɕ_i_aŋ/ (Cheng 1966), and C_GVX /ɕ_iaʊ_ɕ_iaŋ/ (Xu 1980). Others argued for complex onsets, CG_V_X /ɕi_a_ʊ_ɕi_a_ŋ/ (Ao 1992; Duanmu 2007), or both complex onsets and complex rimes, CG_VX /ɕi_aʊ_ɕi_aŋ/ (Bao 1990). Finally, others still proposed the unitization of vowel information into either diphthongs, C_G_V_C /ɕ_i_aʊ_ɕ_i_a_ŋ/ (Lin 1989; Wan 2006), or triphthongs, C_V_C /ɕ_iaʊ_ɕ_ia_ŋ/ (Sun, 2006).

Table 1. Segmentation schemas, presented in IPA, according to both nontonal and tonal examples of the words, *xiao3* /ɕiaʊ^{214}/, *xiang4* /ɕiaŋ^{51}/, and *xiao3xiang4* /ɕiaʊ^{214}ɕiaŋ^{51}/

	Nontonal				Tonal		
Schema	*xiao*	*xiang*	*xiaoxiang*	Schema	*xiao3*	*xiang4*	*xiao3xiang4*
C_V_C	ɕ_iaʊ	ɕ_ia_ŋ	ɕ_iaʊ_ɕ_ia_ŋ	C_V_C_T	ɕ_iaʊ_^{214}	ɕ_ia_ŋ_^{51}	ɕ_iaʊ_^{214}_ɕ_ia_ŋ_^{51}
C_G_V_C	ɕ_i_aʊ	ɕ_i_a_ŋ	ɕ_i_aʊ_ɕ_i_a_ŋ	C_G_V_C_T	ɕ_i_aʊ_^{214}	ɕ_i_a_ŋ_^{51}	ɕ_i_aʊ_^{214}_ɕ_i_a_ŋ_^{51}
C_G_V_X	ɕ_i_a_ʊ	ɕ_i_a_ŋ	ɕ_i_a_ʊ_ɕ_i_a_ŋ	C_G_V_X_T	ɕ_i_a_ʊ_^{214}	ɕ_i_a_ŋ_^{51}	ɕ_i_a_ʊ_^{214}_ɕ_i_a_ŋ_^{51}
C_G_VX	ɕ_i_aʊ	ɕ_i_aŋ	ɕ_i_aʊ_ɕ_i_aŋ	C_G_VX_T	ɕ_i_aʊ_^{214}	ɕ_i_aŋ_^{51}	ɕ_i_aʊ_^{214}_ɕ_i_aŋ_^{51}
C_GVX	ɕ_iaʊ	ɕ_iaŋ	ɕ_iaʊ_ɕ_iaŋ	C_GVX_T	ɕ_iaʊ_^{214}	ɕ_iaŋ_^{51}	ɕ_iaʊ_^{214}_ɕ_iaŋ_^{51}
CG_V_X	ɕi_a_ʊ	ɕi_a_ŋ	ɕi_a_ʊ_ɕi_a_ŋ	CG_V_X_T	ɕi_a_ʊ_^{214}	ɕi_a_ŋ_^{51}	ɕi_a_ʊ_^{214}_ɕi_a_ŋ_^{51}
CG_VX	ɕi_aʊ	ɕi_aŋ	ɕi_aʊ_ɕi_aŋ	CG_VX_T	ɕi_aʊ_^{214}	ɕi_aŋ_^{51}	ɕi_aʊ_^{214}_ɕi_aŋ_^{51}
CGVX	ɕiaʊ	ɕiaŋ	ɕiaʊ_ɕiaŋ	CGVX_T	ɕiaʊ_^{214}	ɕiaŋ_^{51}	ɕiaʊ_^{214}_ɕiaŋ_^{51}

More recently (for a review see Neergaard and Huang 2021), evidence from speech production paradigms have shown a bias towards the encoding of unsegmented syllables. Using a neighbor generation task, in which participants produced phonological neighbors to auditory stimuli, Mandarin speakers have shown that while they are able to manipulate segmental information at every segment location (Neergaard and Huang 2019), they are faster and more accurate in the manipulation of lexical tone by maintaining the same nontonal syllable (Neergaard and Huang 2019; Wiener and Turnbull 2016). This bias towards the retrieval of nontonal syllables can be explained by evidence from a series of speech production priming studies (Chen and Chen 2013; Chen, Chen, and Dell 2002; Chen, Lin, and Ferrand 2003; O'Séaghdha, Chen, and Chen 2010; Verdonschot et al. 2013; You, Zhang, and Verdonschot 2012) that has led to a proposal wherein nontonal unsegmented syllables are the likely stored representations that are then combined with tone prior to articulation. This evidence suggests a lexicon structured according to nontonal unsegmented syllables, i.e., CGVX /ɕiaʊ_ɕiaŋ/, or its tonal counterpart, CGVX_T /ɕiaʊ_^{214}_ɕiaŋ_^{51}/.

A hypothesis taking shape to account for why Mandarin speakers produce evidence of syllabic processing during speech can be understood in terms of literacy acquisition and Mandarin's orthographic grain size. Mandarin speakers have at their disposal both Chinese characters and an alphabetic script, referred to as *pinyin*. Pinyin is used as both a pronunciation aid to young learners and as one of many keyboard input methods for writing Chinese characters. Meanwhile, with the exception of the phenomena known as *erhua* in which the character 儿 (*er2*) is added to another character yet pronounced as a single syllable (e.g., 哪儿, *na3 er2 = nar3, "where"*), the orthographic grain size for Mandarin speakers is such that syllables map to Chinese characters. The influence of pinyin versus Chinese characters on speech production can be seen in a picture-naming study that featured lists of items whose names either did or did not overlap in target units such as onsets, tones, and nontonal syllables. Children learning by pinyin (Grade 1) showed evidence of segmentation caused by onsets, while older children (Grade 4) who had transitioned to the learning of Chinese characters, were facilitated when homogenous lists overlapped in nontonal syllables, in line with their adult counterparts (Li & Wang, 2017).

The mapping of alphabetic letters to segmental information is likely responsible for restructuring brain areas related to phonological processing. For instance, in an auditory rhyme judgment task, wherein native speakers of English and Mandarin judged the similarity of orthographic/phonologically consistent (e.g., gate /geɪtʰ/, hate /heɪtʰ/; 详 *xiang2* /ɕiaŋ^{35}/, 洋 *yang2* /iaŋ^{35}/) and inconsistent pairs (e.g., pint /pʰaɪntʰ/, mint /mɪntʰ/; 译 *yi4* /i^{51}/, 择 *ze2* /tsɛ^{35}/), English

speakers showed greater activation than Mandarin speakers in the brain area related to phonological processing, likely due to co-activation of phonological and orthographic information. (Brennan et al. 2013). These results, which suggest that the phonological mental lexicon undergoes a restructuring for first language (L1) speakers of languages that use alphabets, help to understand why Mandarin second language (L2) speakers of English would similarly show evidence of segmentation in speech production tasks (Neergaard, Luo, and Huang 2019; Verdonschot et al. 2013). Neergaard, et al. (2019) found evidence to support the influence of orthography on phonology within a verbal fluency task in which participants produced as many phonological neighbors to a given Mandarin monosyllable as possible within 1 minute. Participants low in English proficiency tended to produce syllable neighbors (e.g., target: *wai4*; responses: *ai4, ai3, ai1, etc.*) while those with greater English proficiency exhibited greater segmentation (e.g., target: *an3*; responses: *an2, dan3, gan3, etc.*). They argued that similar to L1 English speakers, native-Mandarin speakers undergo a restructuring of their phonological mental lexicons from a system that is biased towards a syllable grain size to a system sensitive to segmental units. Variation in L2 English proficiency might also explain why in neurological examinations of speech encoding researchers have found evidence of segmental processing (Qu, Damian, and Kazanina 2012; Yu, Mo, and Mo 2014), syllabic processing (Feng, Yue, & Zhang, 2019; Wang et al., 2017; Zhang & Damian, 2019); and both segmental and syllabic processing (Cai, Yin, & Zhang, 2020; Yu, Mo, Li, & Mo, 2015).

In contrast to the production literature, evidence from speech recognition suggests that the mental representations of phonological information are segmented and tonal. Studies implementing a picture-word matching paradigm utilizing ERP (Malins et al. 2014; Malins and Joanisse 2012), eye-tracking (Malins and Joanisse 2010), and computational simulations (Shuai and Malins 2016), have supported the claim that segmental information is processed incrementally, while tone is likely processed separately but parallel in time. The experimental studies to support this proposal critically mismatched stimuli according to onsets, rimes, syllables and lexical tone, but did not with regularity contrast stimuli according to pre-nuclear glides (i.e., the G unit: /i, u, y/) or the post-nuclear glides or final consonants (i.e., the X unit: /ɪ, ʊ, n, ŋ/). Thus, while their proposal entails the tonal fully unsegmented schema, C_G_V_X_T /ɕ_i_a_ʊ_^{214}_ɕ_i_a_ŋ_^{51}/, it best supports a tonal schema with complex onsets and complex rimes, CG_VX_T /ɕi_aʊ_^{214}_ɕi_aŋ_^{51}/. That processing during speech recognition would be segmental follows the literature with nontonal languages (e.g., McClelland & Elman, 1986; Norris, 1994) and matches the reality of recognition being a time-dependent process of hearing phonological information from beginning to end.

Meanwhile, speech recognition studies using the variable known as phonological neighborhood density (PND) have similarly supported a segmented and tonal schema. The PND metric entails the summing of phonological neighbors of a target word identified through the addition (*cat* /kʰætʰ/ → *cats* /kʰætʰs/*)*, deletion (*cat* /kʰætʰ/ → at /ætʰ/), or substitution (*cat* /kʰætʰ/ → *kit* /kʰitʰ/) of a single phonological unit. In the case of Mandarin, where segmentation is a theoretical and methodological question, differences in segmentation lead to differences in PND values. As can be seen in Table 1, phonological units can include clusters and/or lexical tone, depending on the segmentation schema used. As such, neighbors from one schema might not overlap with those of another. For instance, while the tonal fully segmented schema (C_G_V_X_T) has 6 neighbors (*qiu2*, *liu2, niu1, niu3, niu4, you2*) for the monosyllabic word 牛 *niu2* /nioʊ^{35}/ (‘cow’), the tonal onset/complex-vowel segmented schema (C_V_C_T) has 5 additional neighbors (*na2, ni2, nuo2, nu2, nao2)* due to its collapsing of vowel information into a single unit. A number

of exploratory studies have exploited the differences between schemas to narrow down the possible candidates to best represent the co-activation of words in the mental lexicon during specific tasks.

An exploratory study that tested an extensive list of monosyllabic lexical items in an auditory lexical decision task showed in a model selection procedure that the model representing the tonal fully segmented schema (C_G_V_X_T) outperformed its nontonal counterpart (C_G_V_X) and the tonal and nontonal onset/rime segmented schema, i.e., C_GVX_T and C_GVX (Yao and Sharma 2017). With similar exploratory methods, both the C_V_C_T and C_G_V_X_T schemas were the top competitors based on evidence from auditory word repetition, also known as shadowing (Neergaard and Huang 2016). Both these results fell in line with previous research wherein English (e.g., Luce & Pisoni, 1998), and French (e.g., Ziegler et al., 2003) speakers showed slower reaction times to greater PND, suggesting that phonological words compete during lexical selection. A third study that unlike the prior exploratory studies controlled for stimuli chosen through the use of the C_G_V_X_T schema, found facilitation to greater PND (Neergaard, Britton, and Huang 2019). While the contradiction in directional effects calls for a need of further investigation, it also reiterates that exploratory and confirmatory methods can result in diverging evidence (Baayen et al. 2017). Meanwhile, the facilitative effect also suggests that Mandarin, like Spanish (Vitevitch and Rodríguez 2004; Vitevitch and Stamer 2006, 2009), and Russian (Arutiunian and Lopukhina 2020) differs from English and French due to the structural characteristics of their respective lexicons, or what has been called psychotypology (Vitevitch and Rodríguez 2004; Vitevitch and Stamer 2006). While research is still ongoing in terms of the psychotypology of lexical access, a benefit of the current database is that it provides a unified resource for researchers seeking to identify how the features of Mandarin influence lexical processing.

One particularly unique contribution of the current database is that it provides resources for researchers seeking to understand the network-like structure of the lexicon. The last decade has shown a heightened interest in the combination of network science methods with psycholinguistic experimentation (Siew et al. 2019). In terms of the use of network science with phonology, researchers have used what has been referred to as phonological networks, in which words act as nodes that are connected to one another, i.e., the network's edges, through the addition, deletion, or substitution of a single phonemic unit (Vitevitch 2008). According to this method of constructing a network, PND, i.e., the sum of a word's phonological neighbors, is equivalent to the network measure known as degree. To expand beyond research on a word's degree, researchers have utilized a word-level network measure, known as clustering coefficient (CC). CC measures the interconnectedness of a target word's neighbors. For instance, among the 4 phonological neighbors (*er4*, *er3*, *ger2*, *e2*) of the monosyllabic word *er2* (儿, 而, 鸸), only 1 pair (*er4*, *er3*) of the 6 possible pairs are phonological neighbors according to the C_G_V_X_T schema; gaining *er2* a low CC value of $1 \div 6 = 0.167$. In research with English speakers, CC has been used to show that the interconnectedness of phonological words affects the structure of the developing lexicon (Carlson, Sonderegger, and Bane 2014), as well as adult spoken word production (Chan and Vitevitch 2010) and recognition (Chan and Vitevitch 2009). With Mandarin speakers, CC has been shown to significantly slow participants' production of phonological neighbors (Neergaard and Huang 2019), suggesting a cognitive cost during mental search for words with greater interconnectedness. Other word-level network measures included in the current database are components size, which is the number of nodes within the component a given word resides in (Siew and Vitevitch 2015; Stella 2018); closeness centrality (Castro, Pelczarski, and Vitevitch 2017; Goldstein and Vitevitch 2017; Iyengar et al. 2012), which measures the average shortest

path length between a given node and all other nodes within its component; betweenness centrality (Stella, Beckage, and Brede 2017), which is the number of times a given node is a bridge between the shortest paths of two other nodes and finally; and an as yet studied measure in the literature of phonological networks: eigenvector centrality, which ranks a nodes influence in lieu of the density of its neighbors.

Analyses of phonological networks that go beyond the word-level, referred to as the network's topology, have been illustrative of the role that segmentation plays in speech processing. Topological features extracted from phonological networks have been used to analyze both participant-level verbal productions (Neergaard, Luo, et al. 2019), and whole-vocabularies (Arbesman, Strogatz, and Vitevitch 2010b, 2010a; Brown et al. 2018; Dautriche et al. 2017; Neergaard and Huang 2019; Shoemark et al. 2016; Siew 2013; Siew and Vitevitch 2019; Stella et al. 2018; Stella and Brede 2015; Turnbull and Peperkamp 2016; Vitevitch 2008). Neergaard and Huang (2019) constructed the sixteen Mandarin phonological networks seen in Table 1. They showed that changes in segmentation and or presence of lexical tone resulted in not only differences in lexical frequency and homophony for a given phonological word (i.e., collapsed homophonous words into one phonological form), but also whether or not words were phonological neighbors. They illustrated that both the number of phonological units within a segmentation schema, and the presence or absence of lexical tone, determined the networks' topological features, such as the size of each network's largest component, mean degree (i.e, mean of every word's PND value), and mean CC (i.e., mean of every word's CC value). This implies that the choice of one or another schema in the creation of a lexical database entails taking a theoretical stance. This is problematic for researchers that are agnostic as to the question of segmentation but still require lexical statistics for the selection of stimuli.

To date, no lexical database allows for the selection of lexical characteristics from multiple segmentation schemas. Databases providing lexical statistics for Mandarin have offered a number of variables, however, the majority of which are designed for the study of orthographic processing in that they provide lexical frequencies of orthographic words from large-scale corpora (Chen, Huang, Chang, & Hsu, 1996; McEnery & Xiao, 2003; Sun, Huang, Sun, Li, & Xing, 1997; van Esch, 2012), subtitle frequencies for words without phonological transcriptions (Subtlex-CH: Cai & Brysbaert, 2010), multiple orthographic measures and reaction times derived from traditional Chinese characters (Chang et al. 2015), or reaction times for words used in orthographic lexical decision tasks (Sze, Liow, and Yap 2014; Tsang et al. 2018). Liu, Shu, and Li, (2007) provided variables related to speech processing for monosyllabic words, however, their inclusion of homophone density and phonological frequency were limited to values pertaining to unsegmented syllables plus lexical tone (i.e., CGVX_T). Recently, Sun, Hendrix, Ma, and Baayen (2018) made available a large selection of variables for the study of both orthographic and speech processing. The drawback of this database is that users of its phonological variables are forced to rely on values derived from the C_V_C_T segmentation schema wherein all vowel information is collapsed into a single unit. The use of but one schema is not problematic if researchers intend to use the C_V_C_T schema as their model of segmentation, as was done in Wiener and Turnbull, (2016) for the calculation of phonological neighbors. It is however problematic for researchers who depend on a model of Mandarin speech processing that entails greater segmentation, such as the use of the C_G_V_X_T schema in Myers and Tsay (2005), or a model of the Mandarin syllable that lacks lexical tone, such as the use of the C_V_C schema in Tsai (2007).

In the current article we present a database that provides lexical statistics according to multiple segmentation schemas, allowing researchers to 1) choose stimuli that are derived from a

segmentation schema that supports an existing model of Mandarin speech processing, or 2) choose stimuli that allow for the testing of hypotheses on phonological segmentation according to the word-level statistics from multiple schemas.

2. The Database

2.1 The word list

DoWLS-MAN was constructed through the use of the Subtlex-CH wordlist (Cai & Brysbaert, 2010) and its raw corpora. Chinese characters from the wordlist were transcribed into pinyin (Mandarin romanization) using the CKIP Lexicon (Chinese Knowledge Information Processing Group 1995). From the wordlist, roughly 4,300 items were found to be polyphones, meaning that more than one pronunciation was ascribed to identical characters either within a multisyllabic word or for individual monosyllabic words. The majority of the words' pronunciations and their corresponding frequencies were resolved through the help of the PoS assignments made available by Subtlex-CH. For instance, the pronunciation of *le0*, as a modal particle, and *liao3*, as a verb, allowed for automatic assignment from the raw Subtlex-CH subtitle corpus of the orthographic word,了. For more information on PoS annotation see below. The remaining 151 items that did not lend themselves toward PoS disambiguation, such as 分子, which is pronounced as either *fen1zi3* ('molecule', 'numerator') or *fen4zi3* ('part'), were annotated by three native Mandarin speakers as given within sentential context from the raw corpus. Sixty-two token polyphonous items found no annotator agreement and were removed.

The word list was further changed due to corrections made to the original Subtlex-CH word list. We identified 8,415 parsing errors, for instance, entries such as "东尼·本", and "乔治·东尼", which consisted of two or more entries. After parsing these entries, the DoWLS-MAN frequency list garnered new words and adjusted lexical frequencies for numerous existing words. Whereas the original Subtlex-CH word list contains 99,121 entries, DoWLS-MAN is built on an adapted word list and corresponding lexical frequencies for 92,915 orthographic words with corresponding pronunciations.

In order to follow the conventions practiced in similar resources, proper names also needed to be removed. Instead of striking them from the database altogether, we reduced their frequency to 1. This placed them below the threshold (elaborated on below) necessary to contribute to neighborhood calculations.

Pinyin words were then transcribed to an ascii phonological transcription (sampa), according to the syllable inventory of Neergaard and Huang (2019). The Neergaard and Huang (2019) inventory, unlike the existing inventories available (Cheng, 1966; Duanmu, 2011; Lin, 2007; Zhao & Li, 2009), was constructed and validated through the use of two phonological association tasks in which native-Mandarin speaking participants were instructed to verbally produce monosyllabic minimal pairs to auditorily presented stimuli (e.g., stimulus: *ba3*; response: *na3*). The participants' minimal pair responses were measured according to the edit distance rule wherein two syllables are neighbors if they differ by the addition, deletion or substitution of a single segment or tone. The Neergaard and Huang inventory was shown to outperform two existing inventories (Lin, 2007; Zhao & Li, 2009) in terms of aligning with the spoken minimal pair productions.

2.2 Segmentation

DoWLS-MAN offers lexical statistics derived from 16 segmentation schemas. As can be seen in Table 2, we employ the use of underscores between phonological units, and include both tonal and nontonal schemas denoted by the presence or absence of the T unit. The use of underscores is meant to mark a difference between a segmentation schema, in which all words within the lexicon

follow the same pattern of segmentation, and a word level annotation of phoneme categories (described below for the variable "SyStruct"). Note that while schema annotations use underscores to segment units, words transcribed in sampa use blank spaces.

Table 2. Segmentation schemas, presented in sampa, according to both nontonal and tonal examples of the words, *xiao3* /ɕiaʊ^{214}/, *xiang4* /ɕiaŋ^{51}/, and *xiao3xiang4* /ɕiaʊ^{214}ɕiaŋ^{51}/

	Nontonal				Tonal		
Schema	*xiao*	*xiang*	*xiaoxiang*	Schema	*xiao3*	*xiang3*	*xiao3xiang4*
C_V_C	X iaU	X ia N	X iaU X ia N	C_V_C_T	X iaU 3	X ia N 4	X iaU 3 X ia N 4
C_G_V_C	X i aU	X i a N	X i aU X i a N	C_G_V_C_T	X i aU 3	X i a N 4	X i aU 3 X i a N 4
C_G_V_X	X i a U	X i a N	X i a U X i a N	C_G_V_X_T	X i a U 3	X i a N 4	X i a U 3 X i a N 4
C_G_VX	X i aU	X i aN	X i aU X i aN	C_G_VX_T	X i aU 3	X i aN 4	X i aU 3 X i aN 4
C_GVX	X iaU	X iaN	X iaU X iaN	C_GVX_T	X iaU 3	X iaN 4	X iaU 3 X iaN 4
CG_V_X	Xi a U	Xi a N	Xi a U Xi a N	CG_V_X_T	Xi a U 3	Xi a N 4	Xi a U 3 Xi a N 4
CG_VX	Xi aU	Xi aN	Xi aU Xi aN	CG_VX_T	Xi aU 3	Xi aN 4	Xi aU 3 Xi aN 4
CGVX	XiaU	XiaN	XiaU XiaN	CGVX_T	XiaU 3	XiaN 4	XiaU 3 XiaN 4

It is important to note that not all of the annotation units across schemas are equivalent. For instance, the V of the C_V_C and C_V_C_T schemas, entails a complex cluster of all vowel segments, resulting in only three phonological units for *xiao3*: X iaU 3; and four for *xiang4*: X ia N 4. In contrast, the V of the C_G_V_C and C_G_V_C_T schemas, originally proposed for Taiwanese (Lin, 1989), results in four phonological units for *xiao3*: X i aU 3; and five for *xiang4*: X i a N 4. The remaining schemas use the X unit to describe both final consonants and the post-nuclear glide. For these schemas, the V is always in reference to monophthong vowels.

2.3 Threshold

Neighborhood statistics, such as phonological neighborhood density (PND), neighborhood frequency (NF), and the recently applied network statistics such as clustering coefficient (CC), are calculated from what Vitevitch (2008) referred to as idealized lexicons. Researchers have taken different approaches to creating neighborhood values from such idealized lexicons. The Neighborhood Activation Model (NAM: Luce & Pisoni, 1998), which established a theory of spreading activation among phonological representations in long term memory, used PND values extracted from an electronic version of the 1967 *Webster's Seventh Collegiate Dictionary* consisting of 19,340 words. This lexicon was then matched to frequency counts from the Kučera and Francis (1967) frequency list. Numerous studies have used this lexicon to study phenomena such as word learning (Goldstein and Vitevitch 2014; Storkel, Armbruster, and Hogan 2006), tip of the tongue states (Vitevitch and Sommers 2003) and picture naming (Vitevitch 2002), among numerous other tasks.

Another approach to constructing an idealized lexicon entails the use of targeted media to represent specific subpopulations, such as the use of textbooks for learners (Lété, Sprenger-Charolles, and Colé 2004; Vitevitch, Stamer, and Kieweg 2012) or child corpora for the study of children's speech (Storkel and Hoover 2010). Perhaps the most common method of creating an idealized lexicon entails the use of existing frequency lists (e.g., Davis & Perea, 2005; Holliday, Turnbull, & Eychenne, 2017; Marian et al., 2012). Such databases have used word counts from as low as 7,000 phonological words (Strand 2013) to as high as 129,000 orthographic words (New et al. 2004).

Researchers have either not discussed a reason for choosing a specific word count, or have argued for the validity of their chosen word counts based on measures of receptive vocabulary size,

or the distributional properties of neighborhood values. When discussing the validity of the Webster's dictionary as an idealized lexicon, Vitevitch (2008) argued that the near 20,000 orthographic words were close to the 17,000-lemma receptive vocabulary size estimate of Goulden, Nation, and Read (1990). Vocabulary size has not been used however to motivate most databases, perhaps because estimates range wildly, from as little as 9,800 English lemmas (Treffers-Daller and Milton 2013) to an average of 56,400 lemmas for older English speakers (Brysbaert et al. 2016). In contrast, in a PND database representing five European languages, Marian et al. (2012) used lexical frequency to trim five wordlists as a means of making their frequency distributions comparable. This led to variation in word counts between 27,751 words for English, and 45,027 words for German. Their reasoning for keeping the word counts within a lower range, a contention shared by Davis (2005), was to exclude very low frequency words that would not be regularly perceived or produced in everyday language use, therefore avoiding the inflation of neighborhood values with words that do not likely contribute to lexical selection.

While vocabulary size has been of interest for researchers studying Mandarin-speaking children (Hao et al. 2008, 2015), and Mandarin speakers' L2 English vocabulary (e.g., Wang & Treffers-Daller, 2017; Zhao & Ji, 2018), we have been unable to find any similar estimates of vocabulary size for Mandarin speaking adults. The first version of the current database, published as a conference paper (Neergaard & Huang, 2016), set a threshold at 17,000 phonological words based on the estimate of Goulden et al. (1990). This was later revised in the analysis of network characteristics of the same Subtlex-CH wordlist to a threshold of 30,000 (Neergaard and Huang 2019) so as to increase density amongst tonal segmented schemas, particularly for disyllables.

As with previous iterations of the current database we sought to provide a threshold from which to derive phonological neighborhood and network values. Unique to the current database is the use of a data-driven approach to identify an optimal threshold. In section 2.5.1 we performed a model selection procedure consisting of 48 models (16 schemas * 3 thresholds) that ranked mean marginal r^2 values per three candidate thresholds of 20k, 30k, and 40k. We identified that while no threshold was significantly better in performance, the 30k threshold had the highest performing model.

This number of 30,000 words places our threshold above the size of the original NAM (Luce and Pisoni 1998) word list, and as such, the idealized lexicon of Vitevitch (2008). Its size is closer to the Dutch and English lexicons of the CLEARPOND database (Marian et al. 2012). Importantly, our use of a word count threshold makes DoWLS-MAN novel among databases that report neighborhood, and network statistics. While neighborhood statistics for below-threshold words (i.e., > 30,001) were calculated one at a time in regards to the above-threshold words (i.e., 1-30,000), network statistics were only calculated for above-threshold words. This method resulted in the ability to report neighborhood statistics for all words while simultaneously excluding below-threshold words from contributing to the inflation of neighborhood values with low-frequency words.

2.4 Categories

Due to the presentation of lexical statistics according to 16 segmentations schemas, DoWLS-MAN has both variables that are invariant, and variant. In Table 3 we present the 31 variables offered by the database, divided into nine themes. As can be seen in Table 3, variant characteristics (marked with an asterisk) are those that represent differences due to segmentation, such as segmented sampa (Pho), or log10 lexical frequency (FreqDL). Invariant characteristics are those that describe the items' form, such as pronunciation according to the international phonetic alphabet (IPA_T or IPA_NoT), or the number of syllables within a lexical item (SyLen).

In the database, variant characteristics are labelled according to the format: Variable.SCHEMA. For example, according to the tonal fully segmented schema (C_G_V_X_T), phonological neighborhood density (PND) is labelled as PND.C_G_V_X_T, while according to the nontonal complex vowel segmented schema (C_V_C), homophone density (HD) is labelled HD.C_V_C.

Table 3. Summary of variables by column name and content/description

Column name	Contents/Description
Lexicality	
Lexicality	Item types: words, added, tonegap, syllablegap, systemicgap
Pronunciation	
Key_T, Key_NoT	Sampa transcription with (T) or without (NoT) lexical tone (0-4)
PY_T, PY_NoT	Pinyin transcription with (T) or without (NoT) lexical tone (0-4)
IPA_T, IPA_NoT	IPA transcription with (T) or without (NoT) lexical tone (0-4)
Pho*	Segmented sampa
Length	
SegLen	Number of segmental units within an item
SyLen	Number of syllables within an item
PyLen	Number of letters (pinyin) within an item
PhoLen*	Number of phonological units within an item
Syllable	
Initial	Word initial segment
Tone	Lexical tone (0-4)
SyStruct	Word-level syllable structure
Part of speech (POS)	
Dom_POS	The most frequent POS assignment
Freq_Dom_POS	Token frequency of Dom_POS
Percent_Dom_POS	Percent (0-1) that Dom_POS is the dominant POS
Other_POSes	POS (token frequency) for all POS assignments
Homophony	
Homophones*	Orthographic words that share pronunciation
HD*	Homophone density
Lexical frequency	
FreqDowls	Adjusted lexical frequencies from the Subtlex-CH wordlist
FreqDL	Log10 transformation of FreqDowls
FreqDowls*	FreqDowls collapsed across all phonological words
FreqDL*	Log10 transformation of FreqDowls*
Phonological neighborhood measures	
Neighbors*	Phonological neighbors of target word (in sampa)
PND*	Phonological neighborhood density: total number of Neighbors
Sub_PND*	Number of Neighbors calculated through substitution
Add_PND*	Number of Neighbors calculated through addition
Del_PND*	Number of Neighbors calculated through deletion
NF*	Neighborhood frequency: mean lexical frequency of Neighbors
Network science measures	
CS*	Component Size
CC*	Local clustering coefficient
Close*	Closeness centrality
Btw*	Betweenness centrality
Eigen*	Eigenvector centrality

"*" = Variant categories

2.4.1 Lexicality

DoWLS-MAN allows for the search of lexical items belonging to five types: “words”, “added”, “tonegap”, “syllablegap”, and “systemicgap”. The database adopted the method used in Neergaard & Huang (2019) to define which items belonged to each group. First, granting items as belonging to the “words” status involved using www.zdic.net to verify whether items other than those within the Subtlex-CH orthographic word list corresponded to an existing orthographic word. Zdic.net is an online resource including definitions and pronunciations for 75,983 characters and has been used in the disambiguation of out-of-vocabulary words (Li, 2011; Li, Zong, & Su, 2015; Ma, Kit, & Gerdemann, 2012; Zhang, Niehues, & Waibel, 2016). The database expanded on the wordlist through the inclusion of 5,973 lexical items that we classified as either added (1,669 orthographic/226 phonological items), tone gap nonwords (740 phonological items), syllable gap nonwords (609 phonological items), or systemic gap nonwords (2,955 phonological items).

Added items are those that correspond to Chinese characters but were not considered monosyllabic words in the Subtlex-CH wordlist. We first identified added items by looking within the wordlist for syllables that were only featured within multisyllabic words. They included items such as 火 *huo3* /xuo^{214}/, which occurs in multisyllabic words such as 柴火 *chai2huo3* “fire wood”, and 烽火 *feng1huo3* “fire beacon”. Next, the syllable inventory of Neergaard and Huang (2019) was used to identify missing syllables. These syllables were then verified as to their correspondence to one or more Chinese characters through the use of Zdic.net. This process identified syllables such as *an2* (俺, 唵, 玵, 雸). The featured items were added to aid researchers needing lexical characteristics for monosyllabic items, while also providing a basis of all extant monosyllables in Mandarin.

DoWLS-MAN offers three classes of phonological nonwords. Each class of nonword was constructed based on the phonotactics of the syllable inventory of Neergaard and Huang (2019). Tone gap nonwords are lexical items that correspond to an existing syllable in the Mandarin syllable inventory combined with one of the five lexical tones (tones 0-4) that do not correspond to an existing Chinese character. For instance, the nontonal syllable, *mei*, can be ascribed to tone 2: *mei2* (ex: 没); tone 3: *mei3* (ex: 美); and tone 4: *mei4* (ex: 妹); but not to tone 0: *mei0*; or tone 1: *mei1*. Syllable gap nonwords are those items that contained biphone combinations that exist in the syllable inventory. For instance, the nontonal syllable gap nonword, *fao*, was constructed based on existing instances of /fa/ and /aʊ/. Tonal versions of *fao* were then made through the addition of lexical tones 0-4. Systemic gap nonwords are those items built from biphone combinations that do not exist in the syllable inventory, and include syllables such as /fyn/ and /xon/.

Note that the current list of nonwords should be used with knowledge of the local dialect/topolect of the population being studied. For instance, we included items that did not match our listed resources despite the known use of certain items within some Mandarin spoken dialects, such as *gin2* /kin^{35}/ (琴), which is in use by speakers of Taiwan and likely understood by many mainland Mandarin speakers.

Because added, syllablegap, systemicgap, and tonegap categories were not in the Subtlex-CH wordlist, they were given a frequency of zero. The lexical statistics featured for the item types were calculated in the same way as below-threshold items, i.e., one at a time in regards to the top 30,000 most frequent phonological words.

2.4.2 Pronunciation

The database provides four forms of pronunciation transcription, three of which are invariant and one variant. The database was constructed on the sampa (ascii version of IPA) transcriptions of each item. Based on the key role that sampa played in organizing the database, it is labeled as

Key in both its tonal (Key_T) and nontonal (Key_NoT) versions. Note that neither Key_T or Key_NoT feature spaces to note segmentation, resulting in the transcription of *xiao3xiang4* /ɕiaʊ^{214}ɕiaŋ^{51}/ according to Key_T as XiaU3XiaN4, and Key_NoT as XiaUXiaN. Next, each lexical item is accompanied by its corresponding transcription in IPA, with either the inclusion (IPA_T) or exclusion (IPA_NoT) of tonal information. The final invariant pronunciation transcription is that of pinyin, (i.e. Romanized Mandarin pronunciation) in both its tonal (PY_T) and nontonal (PY_NoT) versions.

Our final transcription varies according to the schema in which it belongs. Sampa transcriptions for each schema, such as those in Table 2, can be found as Pho.SCHEMA. for example, if a user desires the transcriptions of *xiao3xiang4* according to both the tonal and nontonal versions of the fully segmented schema they would need to refer to Pho.C_G_V_X_T (X i a U 3 X i a N 4) for the tonal version, and Pho.C_G_V_X (X i a U X i a N) for the nontonal version. Note that, as illustrated in Table 2, the Pho.SCHEMA column is segmented according to blank spaces between units rather than through the use of underscores.

For assistance on identifying the matches between sampa, IPA and pinyin we have included a chart at https://dowls.site, where this database is freely available. This chart can be accessed on the home page by selecting the tab labeled ‘Pronunciation Chart’.

2.4.3 Length

The combinatorial nature of linguistic units, such as segments, phonemes, syllables, and orthographic units entails the need to measure multiple levels of word lengths. DoWLS_MAN offers three invariant length measures. The first of which, segment length (SegLen), has been used in the literature to investigate the syllable’s internal phonological structure (Wu and Kenstowicz 2015). Meanwhile, syllable length (SyLen) has played a role in the study of Mandarin, particularly as to whether differential processing costs exist between monosyllabic and disyllabic words (Ma, Wang, & Li, 2016). Finally, we offer Pinyin length (PyLen). Pinyin is an orthographically transparent alphabetic representation of Mandarin phonology, meaning that there is a high correspondence between pinyin letters and phonological segments. For a recent review of orthographic transparency see Borleffs, Maassen, Lyytinen, and Zwarts (2019). PyLen, while not explicitly investigated in the literature, was included due to the role that pinyin awareness has taken in the study of second language acquisition (Ding et al. 2018; Qi et al. 2015), and due to the contention that the reliance on pinyin as an input writing method, and educational aid to young learners has a negative influence on Chinese character reading proficiency (Li et al., 2017; Tan, Xu, Chang, & Siok, 2013; Zhou, Kwok, Su, Luo, & Tan, 2020).

In the database, SegLen refers to the number of segments (excluding tone) within each lexical item, regardless of segmentation. For instance, regardless of the segmentation schema, *xiao3xiang4* contains 8 segments: XiaUXiaN. DoWLS-MAN consists of 33 possible segments (listed below in 2.4.4), while the average phonological word is 7 segments in length (M: 7.05; SD: 2.38).

SyLen counts the number of syllables within an item based on its pronunciation. Using our disyllabic example, *xiao3xiang4*, we find SyLen = 2. However, not all Mandarin words have a one-to-one character-to-syllable correspondence. An example of this distinction can be found with words that utilize the erhua feature, in which syllable final rhoticization is denoted by the addition of the character 儿, e.g., 船儿 *chuanr2* /tʂʰuaɹ^{35}/ “boat”. Distributional features of SyLen include, 438 nontonal monosyllables (40 of which are erhua monosyllables), 1,207 tonal monosyllables (47 of which are erhua), while tonal disyllables (41,788) and nontonal disyllables (30,893) account for 50.2% and 43.9% of syllable lengths respectively.

Finally, PyLen is the number of letters, excluding tone numbers, used to construct its pinyin spelling. Using our example word, *xiao3xiang4*, we see that PyLen = 9 due to it consisting of 9 letters. Due to the high transparency between pinyin and Mandarin phonology, the average number of letters per word (M: 7.70; SD: 2.62) is very similar to that of segments per word (M: 7.05; SD: 2.38).

The final length measure is of the variant category. Phoneme length (PhoLen.SCHEMA) counts the number of phonological units within an item based on the units within Pho.SCHEMA. The disyllabic example, *xiao3xiang4*, has a different number of phonological units depending on whether the schema is tonal or nontonal, and the extent to which segments are clustered. For instance, according to the tonal fully segmented schema (C_G_V_X_T), *xiao3xiang4* (sampa: X i a U 3 X i a N 4) has 10 units, yet with the removal of lexical tone, as seen in the nontonal fully segmented schema (C_G_V_X), *xiao3xiang4* (sampa: X i a U X i a N) has 8 units. If we then examine the same word according to the complex onset/rime schemas (CG_VX_T and CG_VX) *xiao3xiang4* has just 6 units when tonal (sampa: Xi aU 3 Xi aN 4), and 4 units when nontonal (sampa: Xi aU Xi aN).

2.4.4 Syllable

The study of Mandarin's syllable constituents is an active area in research dedicated to behavioral (Sereno and Lee 2015; Wiener and Turnbull 2016), neuropsychological (Wang et al., 2017; Yu et al., 2015), clinical (Peng et al. 2017), developmental (Yeh et al. 2015), and second-language learning, (Li, Wang, & Davis, 2015). DoWLS-MAN makes available for this community three invariant categories. Initial, and Tone, respectively present the items' initial segment in sampa pronunciation, and lexical tone according to tones 0-4. Under the feature titled, SyStruct, we implement a word-level annotation, in which C refers to consonants at initial position, /f, k, k^h, l, m, n, p, p^h, ɹ, s, t, t^h, ʂ, ɕ, tɕ, $tɕ^h$, ts^h, $tʂ^h$, ts, tʂ, x/; G signifies medial glides, /i, u, y/, either at initial position or following an initial consonant; V denotes both monophthongs, /a, e, ɛ, ə, i, ɨ, u, y/ and the post-nuclear glides /ʊ, ɪ/; R indicates the final rhotic consonant /ɻ/; and N the final nasal consonants, /n, ŋ/. For instance, our example word 小巷 *xiao3xiang4* /ɕiaʊ^{214}ɕiaŋ^{51}/ is annotated in SyStruct as CGVV CGVN, while the entry, 一丁点儿 *yi1 ding1 dianr3* /i^{55}tiŋ^{55}tiɛɻ^{214}/ ("a tiny bit"), is annotated in SyStruct as: V CVN CGVR.

2.4.5 Part of speech

Parts of speech, particularly nouns and verbs (Li, Jin, & Tan, 2004; Xia, Wang, & Peng, 2016), lead to differential processing. DoWLS-MAN presents four invariant POS characteristics adopted from the Subtlex-CH wordlist: Dom_POS refers to the dominant POS assignment for a given phonological word; Freq_Dom_POS entails the lexical frequency of usage noted in Dom_POS; Percent_Dom_POS is the percent to which that usage is dominant; and Other_POSes lists the non-dominant POS assignments associated with the same phonological word and their respective lexical frequencies.

2.4.6 Homophony

Mandarin is a highly homophonous language. It has been reported that whereas 3.2% of English consists of homonyms, 11.6% of Mandarin is homophonous (Wen 1980). An example of high homophony is the monosyllable *yi4* /i^{51}/, which has been reported to have 48 homophone neighbors (Wang, Li, Ning, & Zhang, 2012). Homophones have a cost on processing in Mandarin. Phonological words with a high rate of associated homophones, i.e., words high in homophone density, have been shown to lead to lexical competition in spoken word recognition, as seen by slower reaction times, and lower accuracy (Chen et al. 2016; Wang et al. 2012).

DoWLS-MAN includes two variant categories based on Mandarin homophony: Homophones (Homophones.SCHEMA) and homophone density (HD.SCHEMA). The Homophones category is a list of orthographic words associated to the same phonological word under the threshold of the top 30,000 most frequent phonological words. HD entails the number of items in the Homophones column, such that a value of 1 implies that a given item has no homophone neighbors. Values for HD and Homophones do not vary across tonal schemas and only rarely do across nontonal schemas. For instance, the highest HD item, *yi4* /i^{51}/, has 37 homophone neighbors across all tonal schemas, and 71 across the nontonal schemas. The reason for this lies in the fact that lexical tone, belonging to each syllable, creates a barrier between syllables. In cases where nontonal schemas differ in HD, it is due to the collapsing of two syllables. For instance, when tone is removed, so is the barrier between a monosyllable and disyllables sharing the same segments. For instance, the tonal monosyllable, *liang4* /liaŋ^{51}/, is tied to five Chinese characters (亮，晾，谅，辆，量), yet its nontonal counterpart is tied to an additional 6 monosyllabic words (凉，梁，粮，良，两，俩) and three disyllabic words (李昂，里昂，利昂) due to the collapsing of *li* and *ang* into *liang* /liaŋ/.

2.4.7 Lexical frequency

Lexical frequency is known to affect almost all aspects of lexical processing. We chose to use Subtlex-CH due to it being constructed on subtitle frequency, a genre of frequency shown to better predict lexical processing than counts generated from written sources such as books (Brysbaert and New 2009; Cai and Brysbaert 2010; Keuleers, Brysbaert, and New 2010; Mandera et al. 2015; New et al. 2004), possibly due to its inclusion of words associated with greater emotional content (Baayen, Milin, and Ramscar 2016).

As described in section 2.1, the Subtlex-CH word list was altered in the making of DoWLS-MAN. After accounting for the reallocation of lexical entries and lexical frequencies due to parsing errors, and the disambiguation of pronunciation for polyphonous words, the original 99,121 entries from the Subtlex-CH word list was reduced to 92,915 words for the DoWLS-MAN word list. Upon merging the two lists we found 90,607 words in common that showed a lower correlation than would be expected: 0.838. The gap between the two word lists is due to the presence of polyphonous words. Upon removing words with multiple pronunciations, accounting for 402 unique orthographic words and 876 words enriched with specific pronunciations, we found a correlation of 0.999 between the remaining portions of the two word lists. As a result, the two word lists are identical for 89,818 words.

With the DoWLS-MAN word list, we offer two invariant categories and two variant categories. The first invariant category, FreqDowls, consists of the raw lexical frequency counts adapted from Subtlex-CH. For convenience sake, we also provide this variable in its log10 transformation: FreqDL. The first of the variant categories, FreqDowls.SCHEMA, consists of the raw FreqDowls reallocated per schema based on the collapsing of homophonous words into single word forms, i.e., phonological words. As with homophone density (HD) above, FreqDowls.SCHEMA does not vary across tonal schemas but does vary across nontonal schemas due to the collapsing of multisyllabic words into monosyllabic words. The second variant category is again a log10 transformation: FreqDL.SCHEMA.

2.4.8 Phonological neighborhood measures

Phonological neighborhood measures have long been shown to influence lexical processing. While research on the effects of phonological neighbors has primarily taken place with English and Spanish speakers (For a review, see Vitevitch & Luce, 2016), a recent focus has looked to

Mandarin (Neergaard, Britton, et al. 2019; Neergaard and Huang 2019; Neergaard, Luo, et al. 2019; Wiener and Turnbull 2016).

The current database provides six variant categories. The first variant category, Neighbors.SCHEMA, entails all phonological neighbors of a given word presented in sampa transcription. For instance, in Figure 1D, the example word *niang2* /niaŋ^{35}/ 娘 'effeminate', has seven neighbors that in sampa are: XiaN2 (*xiang2*), liaN2 (*liang2*), QiaN2 (*qiang2*), naN2 (*nang2*), niN2 (*ning2*), iaN2 (*yang2*), and niaN4 (*niang4*).

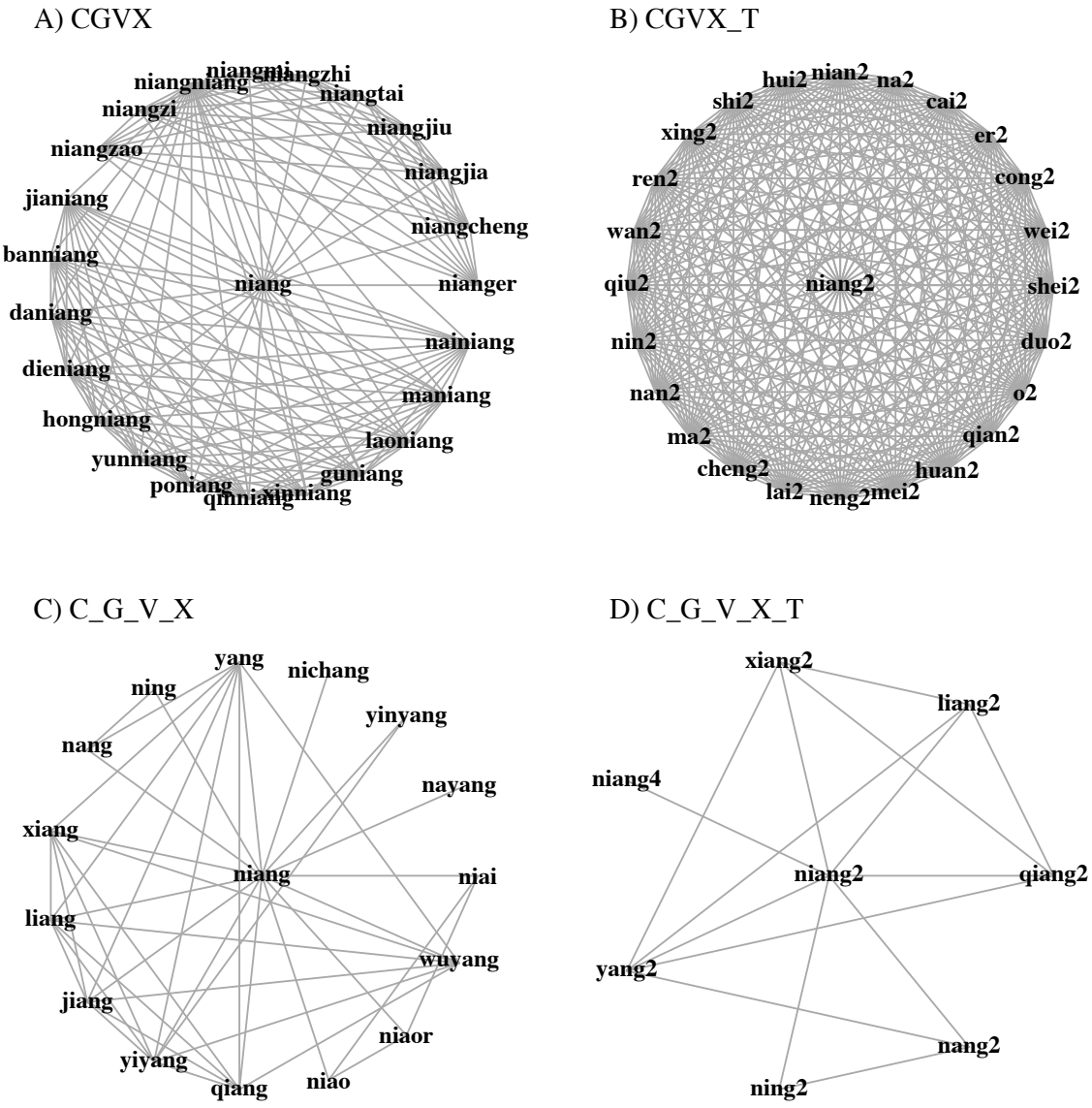

Figure 1. Word-level phonological networks for the monosyllabic word *niang2* /niaŋ^{35}/ 娘 'effeminate', according to **A**) the nontonal unsegmented schema (CGVX), **B**) the tonal

unsegmented schema (CGVX_T), **C**) the nontonal fully segmented schema (C_G_V_X), and **D**) the tonal fully segmented schema (C_G_V_X_T). Note that the number of visual neighbors was truncated in **B**) for display purposes.

The next variant category is PND.SCHEMA, which includes the number of neighbors presented in Neighbors.SCHEMA. As illustrated with Figure 1D, representing the tonal fully segmented schema (C_G_V_X_T), *niang2* has a PND count of seven. Neergaard and Huang (2019) illustrated that syllable segmentation and the existence of lexical tone has an effect on which words are considered neighbors and as such, how high a PND count will likely be per word. The more units within a segmented schema, the less likely a word will have many neighbors. C_G_V_X_T, shown in Figure 1D, has lower PND than its nontonal counterpart (C_G_V_X) in Figure 1C. One reason for the increase in neighbors for nontonal segmented schemas when compared to their tonal counterparts is that monosyllables, like the example nontonal monosyllable *niang*, pick up nontonal disyllabic neighbors due to the absence of lexical tone, like *nayang* /naiaŋ/ 哪样 'which'. This tendency holds across all segmented schemas, a fact that can be seen by taking a mean of PND ($\overline{PND}$) per each schema. Figure 2 shows how tonal schemas have lower $\overline{PND}$ than nontonal schemas, and that those schemas with more units (e.g., 5-unit schemas: C_G_V_X_T & C_G_V_C_T) have lower $\overline{PND}$ than those consisting of fewer units (e.g., 3-unit schemas: C_V_C, C_GVX_T, CG_VX_T & CG_V_X). This pattern is the same for both monosyllables (Figure 2A) and disyllables (Figure 2B). Note, we labelled the two unsegmented schemas (CGVX and CGVX_T) to illustrate their unique behavior.

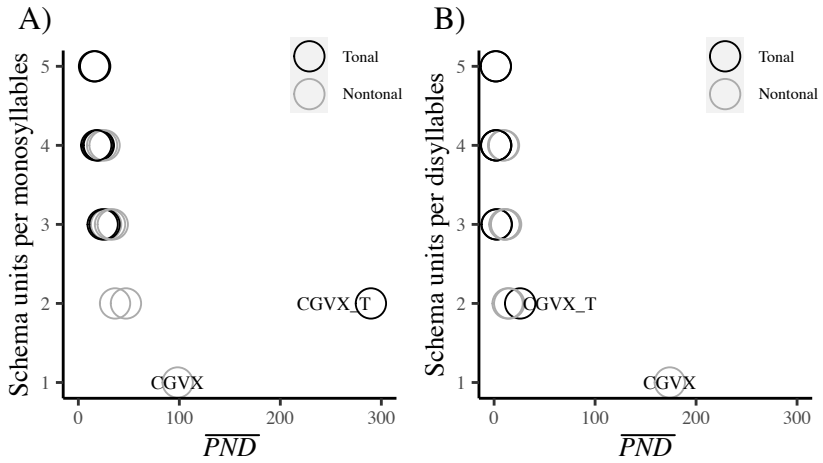

Figure 2. The number of units within each segmentation schema (Schema units), plotted against mean phonological neighborhood density ($\overline{PND}$) for monosyllables **A**) and disyllables **B**)

While the average density of words varies due to the number of units within segmented schemas, a somewhat different story occurs for unsegmented schemas. From the nontonal unsegmented schema (CGVX), shown in Figure 1A, we can see that *niang* is the only monosyllable within a neighborhood of nontonal disyllabic words. The fact that Mandarin is predominantly a disyllabic language explains why the average monosyllable has roughly 100 disyllabic neighbors, as shown in Figure 2A. While monosyllables in the CGVX schema are restricted to disyllabic neighbors, disyllables can pull from all available monosyllables, disyllables, and trisyllables. This

explains why, as illustrated in Figure 2B, when compared to all other schemas, only CGVX shows an increase in density from monosyllabic to disyllabic words. For the tonal unsegmented schema (CGVX_T), shown in Figure 1B, a given monosyllable has neighbors that are only other above-threshold monosyllables with the same lexical tone, resulting in an average of 279 neighbors, as illustrated in Figure 2A. When we look at disyllables, in Figure 2B, we see that average density decreases for CGVX_T. This is because, in contrast to the CGVX schema, all other schemas, including CGVX_T, follow the same pattern noted in European languages: words with more units (i.e., longer words) find fewer neighbors in the lexicon than those with fewer units (i.e., shorter words) (Frauenfelder, Baayen, and Hellwig 1993).

The next variables we introduce pertain to the number of neighbors produced through either of the three phonological edit distance calculations: the addition (Add_PND), deletion (Del_PND), or substitution (Sub_PND) of a segment or tone. Neergaard and Huang (2019), in their examination of monosyllabic spoken phonological associates found that participants used substitution to a greater extent than either addition or deletion. This pattern reflects the nature of segmented schemas. Mean Sub_PND (Figure 3C: $\overline{Sub\_PND}$) has higher values across the segmented schemas than both mean Add_PND (Figure 3A: $\overline{Add\_PND}$) and mean Del_PND (Figure 3B: $\overline{Del\_PND}$). The unsegmented schemas (CGVX and CGVX_T) again show different characteristics. Because unsegmented monosyllables are not comprised of smaller phonological units, they cannot be broken down further. As such, they are the only schemas among monosyllables to have zero deletion neighbors. For neighbors that are identified through the addition of a unit, CGVX has the highest count among the schemas because for that schema monosyllables reside in a network of disyllables. In contrast, CGVX_T has zero addition neighbors because the edit distance metric only allows one phonological unit to differ between words to be considered a neighbor; however, in this schema all words have at least two units. Because of this one-unit difference rule, CGVX_T has a very high $\overline{Sub\_PND}$, such that every monosyllable is a neighbor with all other monosyllables of the same tone. Unlike all other schemas among monosyllables, CGVX has zero substitution neighbors.

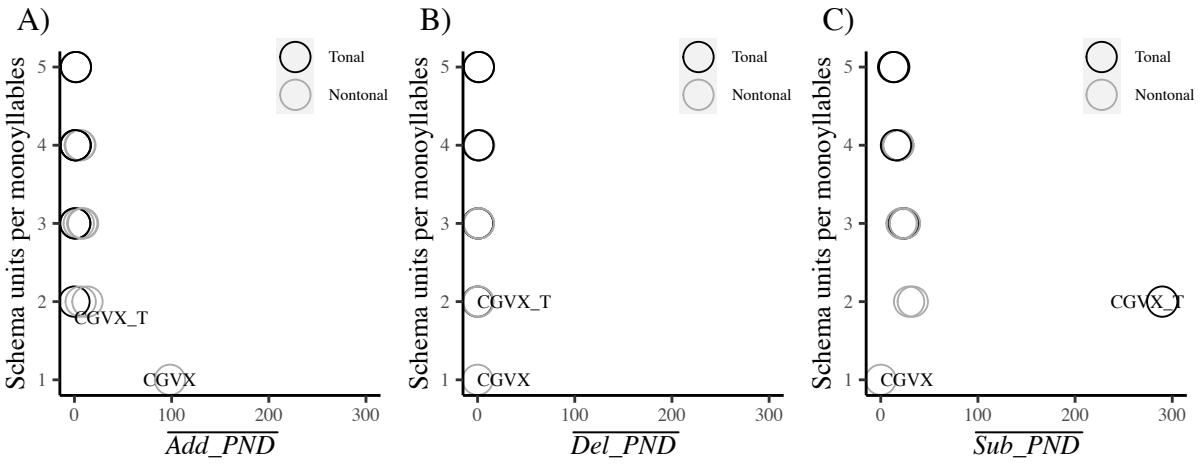

Figure 3. The number of units within each segmentation schema (Schema units), plotted against the means among monosyllables of the three phonological edit distance types: **A)** addition ($\overline{Add\_PND}$), **B)** deletion ($\overline{Del\_PND}$), and **C)** substitution ($\overline{Sub\_PND}$)

The zero count of substitution neighbors for the CGVX schema reveals a modeling decision that altered the traditional edit distance metric. In its normal application the edit distance metric will result in fundamental units being neighbors with other fundamental units, i.e., those that cannot be broken down into smaller units. In the CGVX schema, monosyllables are fundamental units. That meant that all monosyllables above the 30k threshold were neighbors with all other monosyllables above the 30k threshold. According to the traditional measure, *niang* would be a neighbor with other monosyllables that share no common attributes like *si* /sɨ/, *wo* /uo/, *bei* /peɪ/, etc. This matching of all monosyllables to all monosyllables accordingly inflated the PND counts by roughly 300 words for each monosyllable. Because the edit distance metric is at the root of all neighborhood and network variables, that meant that all values would be similarly inflated for monosyllables of the CGVX schema. While this issue was most notable with the CGVX schema, it was also an issue with all nontonal schemas. Thus, to be consistent across all nontonal schemas we disallowed fundamental units from being neighbors of each other. While this decision removed all substitution neighbors for monosyllables of the CGVX schema, it only slightly affected monosyllables of the other nontonal schemas that had fewer fundamental units.

The final phonological neighborhood measure, neighborhood frequency (NF.SCHEMA), involves the mean frequency of all phonological neighbors per a given lexical item. To explain why nontonal schemas have on average higher $\overline{NF}$ than tonal schemas, as illustrated in Figures 4, we must consider the difference between tonal and nontonal phonological words. Phonological words are made by collapsing all orthographically homophonic words into a single word form. The lexical frequency of tonal phonological words is thus the sum of each orthographic entry corresponding to the same pronunciation. The lexical frequency however for nontonal phonological words is the sum of all the frequencies of orthographic words that share the same pronunciation without considering tone. For instance, the highly homophonic word, *yi4*, with 10 homophones, has a lexical frequency of 206,140. Its nontonal counterpart, *yi*, with 71 homophones, has a lexical frequency of 233,934, which is an increase of roughly 28,000 occurrences. $\overline{NF}$ is thus higher for nontonal schemas because it is a mean of higher baseline frequencies.

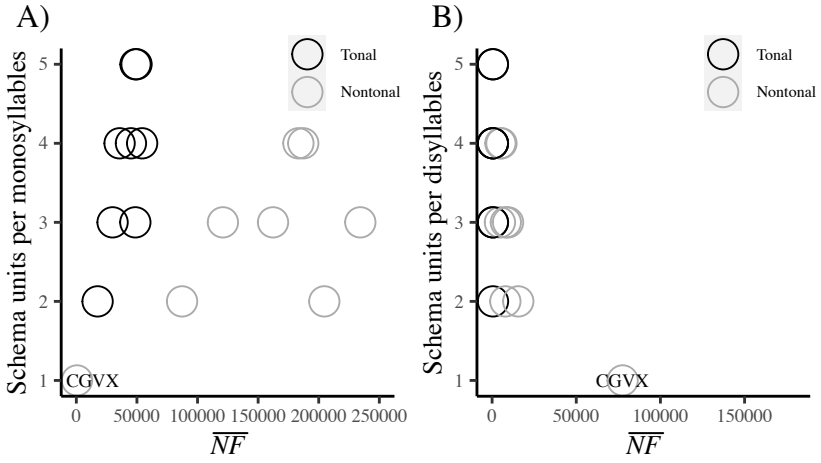

Figure 4. The number of units within each segmentation schema (Schema units), plotted against mean neighborhood frequency ($\overline{NF}$) for monosyllables **A**) and disyllables **B**)

The one schema that does not behave as the others is CGVX. For monosyllables in the CGVX schema, as shown in Figure 4A, CGVX is low in $\overline{NF}$ because monosyllables in this schema have only disyllabic neighbors. Disyllables on average have lower lexical frequency than monosyllables. Inverse to monosyllables, disyllables in the CGVX schema have the highest $\overline{NF}$. The reason being that disyllables in this schema feature high-frequency monosyllabic neighbors.

2.4.9 Network science measures

Researchers have recently begun to employ network science measures to the study of phonological processing. Thus far in the literature, phonological networks have been built upon the premise of similarity, wherein phonological words are nodes, and the links between words (i.e., the network's edges) are based on the relational parameter used to define phonological neighbors (i.e., the addition, deletion or substitution of a single phonological unit) (Vitevitch 2008). As such, the word-level value of PND is the same as the commonly used network measure known as degree (i.e., the number of edges per node). Topological features extracted from phonological networks have been used to analyze both participant-level verbal productions (Neergaard, Luo, et al. 2019), and whole-vocabularies (Arbesman et al. 2010b, 2010a; Brown et al. 2018; Dautriche et al. 2017; Neergaard and Huang 2019; Shoemark et al. 2016; Siew 2013; Siew and Vitevitch 2019; Stella et al. 2018; Stella and Brede 2015; Turnbull and Peperkamp 2016; Vitevitch 2008). Meanwhile, word-level network values extracted from whole-vocabularies, have given insight into phonological processes through several network measures.

DoWLS-MAN offers five variant categories of word-level network measures calculated using the R package 'igraph' (Csárdi and Nepusz 2006). The first of the network measures we will present is clustering coefficient (CC.SCHEMA). It is a measure, ranging between 0 and 1, that reflects the interconnectedness of a word's neighbors (Carlson et al. 2014; Chan and Vitevitch 2009, 2010; Goldstein and Vitevitch 2014). As can be seen in Figure 5, CC can be measured independently of PND. The two example monosyllables, taken from the C_G_V_X_T schema, show that word-level networks of an equal number of phonological neighbors can vary in CC. CC is calculated by dividing the number of attested triangles connected to a given node by the number of possible triangles. Thus, while *quan4* has a possible 28 neighbors, only 6 are actual neighbors, (e.g., *quan1~quan2, quan1~quan3, quan2~quan3, juan4~xuan4, juan4~yuan4, xuan4~yuan4*), resulting in: 6 ÷ 28 = 0.214.

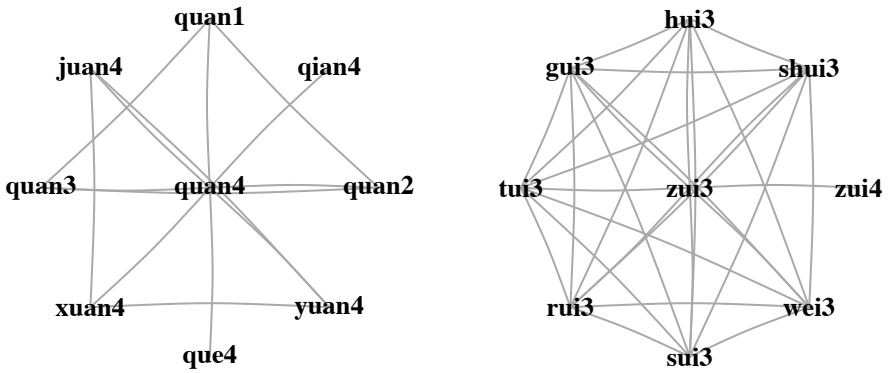

Figure 5. Example word-level networks from the C_G_V_X_T schema. While *quan4* /tɕʰyɛn^{41}/ (劝, "to advise/urge") and *zui3* /tsueɪ^{214}/ (嘴, "mouth") have an equal number of phonological neighbors (PND = 8), they vary in clustering coefficient (*quan4*: CC = 0.214; *zui3*: CC = 0.750)

Similar to the neighborhood measures, CC varies due to segmentation, albeit not to the same extent. In Figure 6 we see mean CC ($\overline{CC}$) plotted against the number of units within each segmentation schema. When considering monosyllables, illustrated in Figure 6A, there appears to be no defining trend. CGVX_T stands out from the rest due to the relationship between monosyllables of that network. As illustrated above in Figure 1B, above-threshold monosyllables in that schema are neighbors with all other above-threshold monosyllables that share the same tone. $\overline{CC}$ among disyllables, however, reveals a higher average for nontonal unsegmented schemas. This illustrates that when lexical tone is stripped from the lexicon not only do phonological neighbors increase (e.g., Figure 2), but so does the interconnectedness between them.

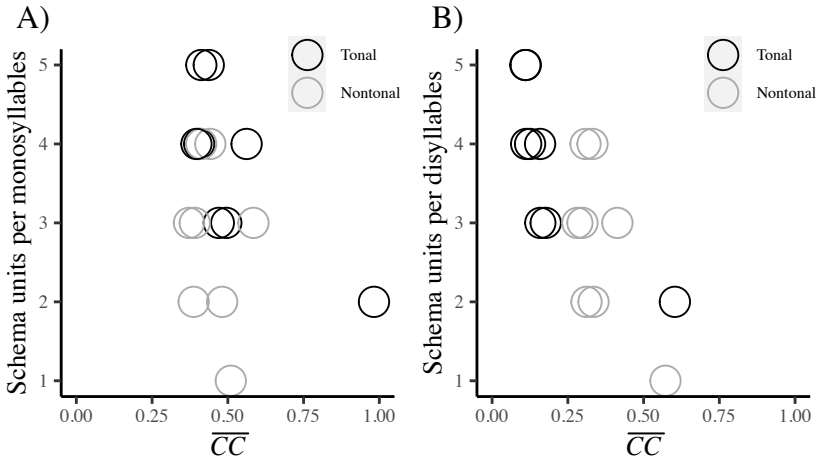

Figure 6. The number of units within each segmentation schema (Schema units), plotted against mean clustering coefficient ($\overline{CC}$) for monosyllables **A**) and disyllables **B**)

CC is unique among the network variables because it was calculated for every word in the database, rather than for just the above-threshold items. For below-threshold words and nonwords, CC values were calculated one at a time in regards to above-threshold items. This was done because CC is calculable with just the information pertaining to immediate neighbors and does not need information about its relation to nodes at a greater distance or within a given component.

The remaining network variables were calculated from only above-threshold words. This was done out of necessity because of the static nature of graphic networks.

The next network category is CS.SCHEMA, which entails the component sizes to which each node belongs within its network (Castro et al. 2017; Siew and Vitevitch 2015; Stella 2018). A component is a subgraph wherein at least two words share an edge. A node that does not share an edge with another node is called a hermit and can be identified as having a CS.SCHEMA value of 1. As we have previously shown in Figure 2 with $\overline{PND}$, the relative density of words is affected by segmentation and lexical tone. Due to the varying densities of the phonological networks, a component will emerge that is proportionally larger than all other components; what is commonly referred to the network's giant component. In Figure 7 we illustrate how for tonal segmented networks, their giant components are smaller than the giant components of nontonal segmented networks. Meanwhile, the unsegmented network belonging to CGVX_T, with its increased density

relative to tonal segmented schemas, has a giant component comparable to the nontonal segmented networks. Finally, the CGVX schema, which showed the highest $\overline{PND}$ among disyllables, has a giant component that outranks all other networks.

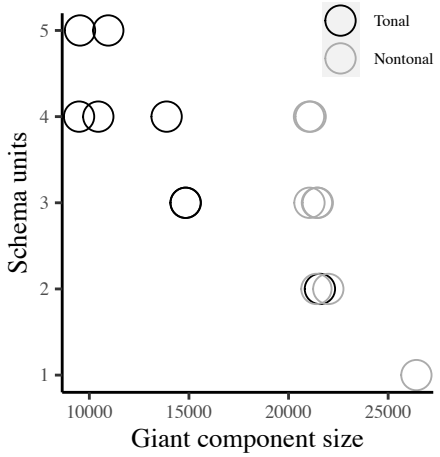

Figure 7. The number of units within each segmentation schema (Schema units), plotted against the size of each network's giant component

DoWLS-MAN provides three centrality measures. The first of which, betweenness centrality (Btw.SCHEMA), reports the number of times a given word is a bridge between the shortest paths of two other words within the same component (Stella et al. 2017). As we can see in Figure 8A, greater density between words penalizes the score attributed to a given word. This is noted by the fact that the lowest mean Btw values are attributed to the giant components of networks built from nontonal schemas and the two unsegmented schemas. This implies that within a phonological network greater sparsity increases a word's betweenness because it increases the likelihood that communication between distant words would pass through a given word.

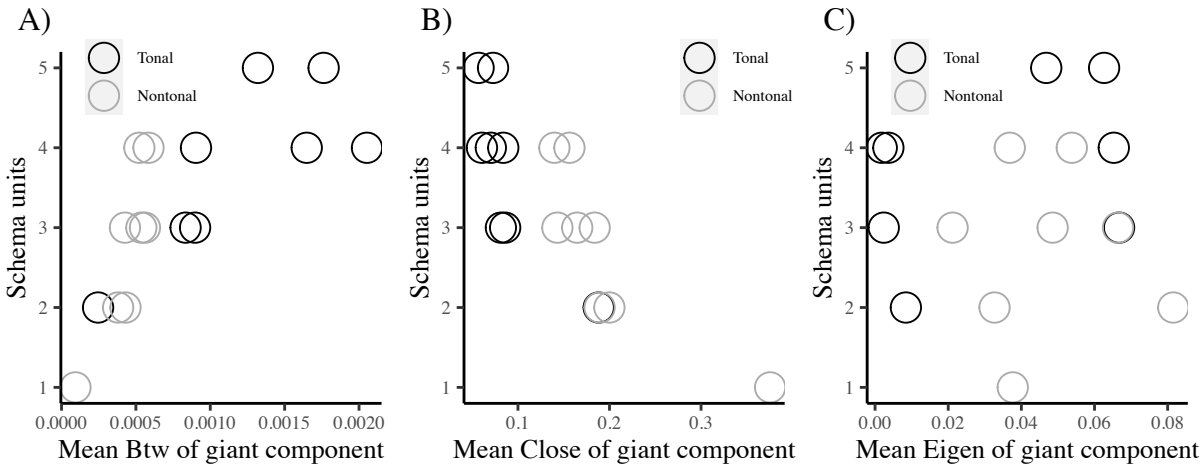

Figure 8. The number of units within each segmentation schema (Schema units), plotted against the mean values of A) Betweenness centrality (Btw), B) Closeness centrality (Close), and C) Eigenvector centrality (Eigen).

Closeness centrality (Close.SCEMA) measures the average shortest path length between a given node and all other nodes within its component (Castro et al. 2017; Goldstein and Vitevitch 2017; Iyengar et al. 2012). As can be seen in Figure 8B, mean Close mirrors the pattern seen with $\overline{PND}$ and giant component sizes, such that tonal segmented schemas reveal lower values than nontonal segmented schemas, who are themselves lower in values than the dense unsegmented schemas.

Eigenvector centrality (Eigen.SCHEMA) measures a word's influence within its component, in lieu of the density of its neighbors. For example, a dense word with 40 neighbors that are each sparsely connected to only a few neighbors would have a lower eigenvector centrality than a dense word with 40 neighbors that are each densely connected to other words. Assessing mean Eigen in Figure 8C in terms of segmentation, it would appear that it is the only measure not to be affected by the density of words brought on by the presence or lack of lexical tone and/or clustering of segmental units. Given that Eigen is the only measure presented in the database that has yet to be explored in the phonological network literature, the current exploration of its distributional features is promising for the identification of an underlying influence on the structure of the lexicon that is distinct from segmentation.

2.5 Reaction time analyses

2.5.1 Threshold model selection

To test whether we can empirically identify an optimal threshold we constructed three versions of the database with thresholds set at 20k, 30k, and 40k words. To contrast the three thresholds we chose the MELD megastudy (Tsang et al., 2018), which consists of orthographic lexical decisions for 12,560 Mandarin Chinese words. While it would be optimal to use a megastudy created from a speech processing task, MELD is the only megastudy that offers reaction times from such a broad sample of words. Meanwhile, a precedent has been set for investigations into phonological neighborhood activation during orthographic processing tasks, namely with English speakers (Grainger et al. 2005; Siew and Vitevitch 2019; Yates 2005). Unique to the current analysis is that in contrast to English, which implements an alphabet meant to linearly model the phonology of the language, the Mandarin speakers that took part in the MELD megastudy judged the lexicality of Chinese characters, which have exceedingly low transparency between character construction and pronunciation. For instance, Zhou (2003) reported that only 3% of existing characters from the 1971 Xinhua Dictionary could be used to reliably predict segment and tonal information.

While the primary goal of the following analysis is to identify an optimal threshold, one side effect is that it will also reveal which phonological segmentation schema best represents the MELD participants' orthographic lexical decisions. Neergaard and Huang (2019) proposed that this type of exploratory method might identify the mental targets activated during the task due to the task's cognitive demands, a hypothesis which lends itself towards certain predictions based on the nature of Chinese characters.

The orthographic lexical decision task entails judging the lexicality of written words, and has been shown for English speakers to be sensitive to readers' inner speech (Abramson and Goldinger 1997). Among readers of Chinese however, there has long been a debate supported by evidence of 1) phonological processing occurring before that of semantics (e.g., Guo, Peng, and Liu 2005; Perfetti and Zhang 1995; Spinks et al. 2000; Tan and Perfetti 1999; Xu et al. 1999), 2) semantics before phonological processing (e.g., Liu et al. 2011; Zhang, Zhang, and Kong 2009), and finally, 3) arguments purporting limited to no phonological influences on reading (e.g., Chen and Shu 2001; Wong, Wu, and Chen 2014; Zhou and Marslen-Wilson 1999, 2000). A limitation to these studies is that phonological activation is always represented by stimuli that are homophonous. This

means that only characters that share a pronunciation have been examined to support this debate. That we are aware of, nowhere in the literature of reading in Chinese have phonological neighbors been investigated. However, there are studies that have examined the orthographic equivalent.

The edit distance metric of addition, deletion or substitution has been used with Chinese characters to create two version of orthographic neighborhood density (OND). These versions differ based on what is considered the smallest unit within the orthographic word or character. In one OND version the smallest unit is the whole character; for example, 果 *guo3* within 水果 *shui3guo3*, 如果 *ru2guo3*, 果园, *guo3yuan2*, etc. (Huang et al., 2006; Li et al., 2015; Tsai et al., 2006; Wu et al., 2013). The other OND method relies on the phonetic radical within complex characters (Bi et al., 2006; Li et al., 2017; Li et al., 2010, 2011; Wang & Zhang, 2011; Wu & Chen, 2003; Yang & Wu, 2014). It produces neighborhoods of characters that are regular in pronunciation, for example, 羊 *yang2* within 洋 *yang2*, 樣 *yang4*, 氧 *yang3*, etc.; and those that are irregular in pronunciation, such as 月 yue4 within 育 yu4, 朋 peng2, and 胡 hu2, etc.

Based on these two OND methods of calculating neighbors, we can estimate which phonological schemas would be closest in content. The character-level OND calculation is similar to the unsegmented schemas that identify neighbors either without lexical tone (CGVX) or with lexical tone (CGVX_T). They differ from the character-level OND measure because in the calculation of phonological neighborhoods, all homophones are collapsed into a single phonological word. Aside from homophony, the character-level OND measure is closest to the CGVX schema because they are both calculated without considering lexical tone. Turning to the phonetic-radical OND calculation, we see an emphasis on the regularity of vowel and rime information, while eschewing tonal information. As such, it is most similar to the nontonal onset/complex-rime schema (C_GVX). However, given that less than 48% of complex characters have the same pronunciation as their phonetic radicals (Zhou, 1978), and that consistency between the phonetic radical and how readers process words is known to effect reading in Chinese (Hsu et al. 2009; Lee et al. 2005, 2009), this method of calculating OND might be a likely candidate amongst words with a consistent ortho/phono mapping, but a less likely candidate over a large number of words that vary in consistency. Accordingly, we assumed that the character-level grain size, reflected in either the CGVX or CGVX_T schemas, would be the optimal schematic representation of phonological processing during orthographic lexical decision.

2.5.2 Methods and discussion

We first filtered the MELD word list by excluding duplicate entries and polyphones (e.g., "分子": *fen1zi3*, *fen4zi3*). Next, words were excluded if they had error rates of 25% or greater. In order to reduce skewed PND values among the tonal segmented schemas, which have high instances of words without neighbors (i.e., PND = 0), we excluded all trisyllables and quadrisyllables. To further improve the PND distributions among tonal segmented schemas, we used PND values from the sparsest schema (C_G_V_X_T) to exclude stimuli based on segment length (SegLen), and high z-scored standard deviations of RTs (zRTSD) from the MELD study. This improved the PND distribution not only for the C_G_V_X_T schema but across all tonal segmented schemas. Similarly, under the premise that SegLen would affect reading latencies we sought to improve its distribution by excluding words from SegLen values that were drastically higher than others. This was again achieved by excluding disyllables with high zRTSD scores in order to achieve a flatter distribution. Finally, we excluded reaction times that were 2.5 standard deviations above or below the mean. Post exclusion, our stimuli consisted of 2,224 words.

Multiple regression models, each containing four predictor variables, were used per each segmentation schema, per each threshold level, resulting in 48 models. Each model contained one

invariant category and four variant categories. In order to select variables that would not lead to high instances of colinearity within the residuals of the 48 models, combinations of variables were tested and evaluated. The final variables placed in each of the 48 models were SegLen, log10 lexical frequency (FreqDL.SCHEMA), phonological neighborhood density (PND.SCHEMA), homophone density (HD.SCHEMA), and clustering coefficient (CC.SCHEMA). The three centrality measures (Betweeness, Closeness, and Eigenvector), and neighborhood frequency were not placed into the models due to high colinearity. Component size was not used because it had very low variation within multiple schemas, making its inclusion statistically untenable.

An ANOVA analysis revealed a lack of significance between the three thresholds ($F = 0.064$; $p = 0.938$). Mean r^2 values showed a lower mean for the 40k threshold (mean $r^2 = 0.302$), yet no difference between those of 20k (mean $r^2 = 0.308$) and 30k (mean $r^2 = 0.308$). However, as can be seen in Table 4, the tonal unsegmented schema (CGVX_T) was the top performing model per each threshold, reaching a top marginal r^2 of 0.392 for the 30k threshold.

Table 4. R^2 values per each database threshold

Schema	20k	30k	40k
CGVX_T	0.387	0.392	0.389
GC_V_X_T	0.364	0.364	0.357
CG_VX_T	0.364	0.364	0.357
C_GVX_T	0.360	0.360	0.353
C_V_C_T	0.354	0.353	0.345
C_G_V_X_T	0.355	0.353	0.346
C_G_V_C_T	0.353	0.350	0.344
C_G_VX_T	0.349	0.347	0.339
CGVX	0.260	0.291	0.266
CG_VX	0.277	0.273	0.267
CG_V_X	0.268	0.265	0.262
C_G_VX	0.252	0.248	0.245
C_GVX	0.250	0.245	0.243
C_G_V_X	0.246	0.242	0.239
C_V_C	0.245	0.241	0.241
C_G_V_C	0.246	0.241	0.238

As we predicted, an unsegmented schema was represented in the top-performing model. As shown in Table 4, among the 30k schemas, CGVX_T outranked all other tonal schemas, and CGVX outranked all other nontonal schemas. This evidence suggests that participants activated networks of syllable-sized mental representations induced by the grain size of Chinese characters. As to what within the model supported the unsegmented schemas, we can start by making inferences from Table 4, and from inspection of the top performing model. As evident in Table 4, all tonal schemas outranked nontonal schemas. This implies that lexical tone influenced orthographic lexical decisions. In Table 5 we see that the largest portion of available variance, with a partial r^2 value of 0.179, belonged to the facilitative effect of FreqDL.CGVX_T. In this tonal frequency measure, we see a combination of phonology and orthography. Similarly, HD.CGVX_T involves a combination of phonological and orthographic activation of mental representations. Fitting with the literature, tonal homophones led to slower RTs.

Table 5. Model estimates for the top 30k model

	Estimate	SE	t value	p value	r^2
(Intercept)	6.96^{E-04}	2.03^{E-06}	343.50	< 0.001	
SegLen	-6.78^{E-06}	2.01^{E-06}	-3.37	< 0.001	0.005
FreqDL.CGVX_T	-3.62^{E-05}	1.65^{E-06}	-21.97	< 0.001	0.179
PND.CGVX_T	4.60^{E-05}	3.01^{E-06}	15.28	< 0.001	0.095
HD.CGVX_T	3.60^{E-06}	4.69^{E-07}	7.68	< 0.001	0.026
CC.CGVX_T	-3.47^{E-07}	2.33^{E-06}	-0.15	0.881	< 0.001

Interestingly, we see contradictory effects between the two phonological variables SegLen and PND.CGVX_T. While words greater in SegLen facilitated lexicality judgments, words with greater numbers of phonological neighbors were inhibitory to reaction times. In interpreting these results it's important to note that the participants of the MELD study were tasked with judging the lexicality of orthography. If the participants were instead judging auditorily presented stimuli, we would expect SegLen to be inhibitory to recognition based on the simple fact that longer auditory stimuli of equivalent segmental duration take longer to recognize. In the judgment of orthographic lexicality, however, it is likely that SegLen played an opposing role to the influence of PND, i.e., the lexical competition caused by co-activated phonological neighbors. For instance, longer words tend to be less confusable in part because they have fewer neighbors and as such are more distinctive. Thus, the lack of phonological competitors that comes with the increase in segmental units likely led to greater ease in the decision-making process for longer words.

While this conclusion does conveniently fit the evidence, it should be taken with caution. One particular concern that the current exploratory analysis cannot address is that of timing. For instance, do the inhibitory effects of PND and HD take place during the verification of lexicality with words in long-term memory, or are they a result of initial recognition and as such a byproduct of inner-speech? Future research can build on the current exploratory analysis to resolve when the lexical competition actually occurs and further elaborate on the role that segmentation and phonological similarity plays in reading Chinese.

3. Conclusion

In this article we presented the Database of Word-Level Statistics for Mandarin Chinese (DoWLS-MAN). Motivated by the lack of consensus on how syllables are segmented during speech processing, DoWLS-MAN is the first lexical database to offer researchers the ability to source lexical statistics from multiple segmentation schemas (8 with tone and 8 without tone). This flexibility allows researchers the ability to either build stimuli sets that support existing models of Mandarin segmentation, or to test multiple hypotheses of segmentation according to the items' lexical statistics. Due to the presentation of values that differ due to syllable segmentation, DoWLS-MAN provides lexical information of both invariant and variant categories. Among the invariant categories are lexical characteristics such as each item's initial segment, lexical tone, syllable structure, dominant PoS, and syllable, segment and pinyin lengths. Those values of the variant category include subtitle lexical frequency, density measures, such as, phonological neighborhood density, phonological neighborhood frequency, and homophone density, and finally network science measures including clustering coefficient, and measures of centrality (betweenness, closeness, eigenvector). Variant and invariant categories are available for five classes of lexical items, including 92,915 words sourced from SUBTLEX-CH (Cai & Brysbaert, 2010), 1,669 "added" monosyllabic items that correspond to Chinese characters that were not featured in the wordlist, and 4,304 nonword items that belong to the tone gap (740), syllable gap (609) or systemic gap (2,955) categories.

One particular concern that we paid attention to in the construction of DoWLS-MAN was the word list. We began by adapting the Subtlex-CH word list. We identified and disambiguated pronunciations for 4,300 polyphonous words, and reallocated words and lexical frequencies for 8,415 parsing errors. Next we considered the use of a threshold, i.e., a means to restrict phonological neighborhood values to a subset of the word list while simultaneously being able to offer lexical statistics for the full word list. We performed a model selection procedure wherein three versions of the database were created that differed in their respective thresholds (20k, 30k, and 40k). Our analysis of orthographic lexical decisions from the MELD mega-study (Tsang et al. 2018) revealed that the top performing model belonged to a threshold of 30,000 words.

Meanwhile, our analysis also contributed to the literature on the role of phonology during the reading of Chinese characters. We found that a facilitative tonal lexical frequency effect from the tonal unsegmented schema (CGVX_T) was the primary influence on lexical decisions. The secondary influence on lexical decisions came from an inhibitory effect of phonological neighborhood density. The fact that the highest performing model was the CGVX_T schema suggests that participants activated networks of syllable-sized mental representations induced by the grain size of Chinese characters. The exploratory analysis opens up directions for further research into phonological neighborhood effects during reading in Chinese.

Users of the database can obtain lexical characteristics for user-defined lists of items, or generate a list of phonological words and nonwords according to user-defined ranges and categories of lexical characteristics. DoWLS-MAN is freely available for search or download at https://dowls.site.

4. Acknowledgments

Funding for this project was made available in part through the Hong Kong General Research Fund, and the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir programme". We'd also like to give a special thanks to Gwenaël Longo for his excellent work on the website, and Qing Cai for allowing us access to the raw Subtlex-CH corpus.

5. Open Practices Statement

All of the data discussed in this article are available for search and download at https://dowls.site. Meanwhile, data organized according to individual schema are downloadable at https://github.com/karlneergaard/Database_of_word-level_statistics.

6. References

Abramson, Marianne, and Stephen D. Goldinger. 1997. "What the Reader's Eye Tells the Mind's Ear: Silent Reading Activates Inner Speech." *Perception and Psychophysics* 59(7):1059–68.

Alario, Francois Xavier, Laetitia Perre, Caroline Castel, and Johannes C. Ziegler. 2007. "The Role of Orthography in Speech Production Revisited." *Cognition* 102:464–75.

Ao, Benjamin X. P. 1992. "The Non-Uniqueness Condition and the Segmentation of the Chinese Syllable." *Working Papers in Linguistics* 42:1–25.

Arbesman, Samuel, Steven H. Strogatz, and Michael S. Vitevitch. 2010a. "Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish." *Entropy* 12(3):327–37.

Arbesman, Samuel, Steven H. Strogatz, and Michael S. Vitevitch. 2010b. "The Structure of

Phonological Networks Across Multiple Languages." *International Journal of Bifurcation and Chaos* 20(3):679–85.
Arutiunian, Vardan, and Anastasiya Lopukhina. 2020. "The Effects of Phonological Neighborhood Density in Childhood Word Production and Recognition in Russian Is Opposite to English." *Journal of Child Language* 1–19.
Baayen, Rolf Harald, Petar Milin, and Michael Ramscar. 2016. "Frequency in Lexical Processing." *Aphasiology* 30(11):1174–1220.
Baayen, Rolf Harald, R. Piepenbrock, and L. Gulikers. 1995. "The CELEX Lexical Data Base on CD-ROM." *Linguistic Data Consortium* (January 1995).
Baayen, Rolf Harald, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. 2017. "The Cave of Shadows: Addressing the Human Factor with Generalized Additive Mixed Models." *Journal of Memory and Language* 94:206–34.
Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. "The English Lexicon Project." *Behavior Research Methods* 39(3):445–59.
Bao, Zhiming. 1990. "Fan-Qie Languages and Reduplication." *Linguistic Inquiry* 21.
Bi, Hong-Yan, W. Hu, and X. C. Weng. 2006. "Orthographic Neighborhood Effects in the Pronunciation of Chinese Words." *Acta Psychologica Sinica* 38(6):7–11.
Borleffs, Elisabeth, Ben A. M. Maassen, Heikki Lyytinen, and Frans Zwarts. 2019. "Cracking the Code: The Impact of Orthographic Transparency and Morphological-Syllabic Complexity on Reading and Developmental Dyslexia." *Frontiers in Psychology* 9(Article 2534):1–19.
Brennan, Christine, Fan Cao, Nicole Pedroarena-Leal, Chris McNorgan, and James R. Booth. 2013. "Reading Acquisition Reorganizes the Phonological Awareness Network Only in Alphabetic Writing Systems." *Human Brain Mapping* 34(12):3354–68.
Brown, Kevin S., Paul D. Allopenna, William R. Hunt, Rachael Steiner, Elliot Saltzman, Ken Mcrae, and James S. Magnuson. 2018. "Universal Features in Phonological Neighbor Networks." *Entropy* 20(526):e20070526.
Brysbaert, Marc, and Boris New. 2009. "Moving beyond Kučera and Francis : A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English." *Behavior Research Methods* 41(4):977–90.
Brysbaert, Marc, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. "How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age." *Frontiers in Psychology* 7(Article 1116):1–11.
Cai, Qing, and Marc Brysbaert. 2010. "SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles." *PLoS ONE* 5(6):e10729.
Cai, Xiao, Yulong Yin, and Qingfang Zhang. 2020. "The Roles of Syllables and Phonemes during Phonological Encoding in Chinese Spoken Word Production: A Topographic ERP Study." *Neuropsychologia* 140(107382).
Carlson, Matthew T., Morgan Sonderegger, and Max Bane. 2014. "How Children Explore the Phonological Network in Child-Directed Speech: A Survival Analysis of Children's First Word Productions." *Journal of Memory and Language* 75:159–80.
Castro, Nichol, Kristin M. Pelczarski, and Michael S. Vitevitch. 2017. "Using Network Science Measures to Predict the Lexical Decision Performance of Adults Who Stutter." *Journal of*

Speech Language and Hearing Research 60(7):1–8.
Chan, Kit Ying, and Michael S. Vitevitch. 2009. "The Influence of the Phonological Neighborhood Clustering Coefficient on Spoken Word Recognition." *Journal of Experimental Psychology: Human Perception and Performance* 35(6):1934–49.
Chan, Kit Ying, and Michael S. Vitevitch. 2010. "Network Structure Influences Speech Production." *Cognitive Science* 34:685–97.
Chang, Ya-ning, Chun-hsien Hsu, Jie-li Tsai, Chien-liang Chen, and Chia-ying Lee. 2015. "A Psycholinguistic Database for Traditional Chinese Character Naming." *Behavior Research Methods* 48(1):1–11.
Chen, Hsuan Chih, and Hua Shu. 2001. "Lexical Activation during the Recognition of Chinese Characters: Evidence against Early Phonological Activation." *Psychonomic Bulletin and Review* 8(3):511–18.
Chen, Jenn-Yeu, Train-Min Chen, and Gary S. Dell. 2002. "Word-Form Encoding in Mandarin Chinese as Assessed by the Implicit Priming Task." *Journal of Memory and Language* 46(January):751–81.
Chen, Jenn-Yeu, Wei-Chun Lin, and Ludovic Ferrand. 2003. "Masked Priming of the Syllable in Mandarin Chinese Speech Production." *Chinese Journal of Psychology* 45(1):107–20.
Chen, K. J., Chu-Ren Huang, L. P. Chang, and H. L. Hsu. 1996. "Sinica Corpus: Design Methodology for Balanced Corpora." Pp. 167–76 in *11th Pacific Asia Conference on Language, Information and Computation (PACLIC 11)*. Seoul, South Korea.
Chen, Train-Min, and Jenn-Yeu Chen. 2013. "The Syllable as the Proximate Unit in Mandarin Chinese Word Production: An Intrinsic or Accidental Property of the Production System?" *Psychonomic Bulletin & Review* 20:154–62.
Chen, Wei-Fan, Pei-Chun Chao, Ya-Ning Chang, Chun-Hsien Hsu, and Chia-Ying Lee. 2016. "Effects of Orthographic Consistency and Homophone Density on Chinese Spoken Word Recognition." *Brain and Language* 157–158:51–62.
Cheng, Robert L. 1966. "Mandarin Phonological Structure." *Journal of Linguistics* 2(02):135–58.
Chinese Knowledge Information Processing Group. 1995. *Technical Report No. 95-02, the Content and Illustration of Sinica Corpus of Academia Sinica.*
Csárdi, Gábor, and Tamás Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal Complex Systems* 1695:1–9.
Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. "Words Cluster Phonetically beyond Phonotactic Regularities." *Cognition* 163:128–45.
Davis, Colin J. 2005. "N-Watch: A Program for Deriving Neighborhood Size and Other Psycholinguistic Statistics." *Behavior Research Methods* 37:65–70.
Davis, Colin J., and Manuel Perea. 2005. "BuscaPalabras : A Program for Deriving Orthographic and Phonological Neighborhood Statistics and Other Psycholinguistic Indices in Spanish." *Behavior Research Methods* 37(4):665–71.
Dell, Gary S. 1986. "A Spreading-Activation Theory of Retrieval in Sentence Production." *Psychological Review* 93(3):283.
Dell, Gary S., Kristopher D. Reed, David R. Adams, and Antje S. Meyer. 2000. "Speech Errors, Phonotactic Constraints, and Implicit Learning: A Study of the Role of Experience in Language Production." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 26(6):1355–67.

Ding, Yi, Ru-de Liu, Catherine A. Mcbride, Chung-hau Fan, Le Xu, and Jia Wang. 2018. "Pinyin and English Invented Spelling in Chinese-Speaking Students Who Speak English as a Second Language." *Journal of Psycholinguisic Research* 47(4).

Duanmu, San. 2007. *The Phonology of Standard Chinese*. 2nd ed. Oxford: Oxford University Press.

Duanmu, San. 2011. "Chinese Syllable Structure." Pp. 1–14 in *The Blackwell companion to phonology*. Wiley.

Duchon, Andrew, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí, and Manuel Carreiras. 2013. "EsPal: One-Stop Shopping for Spanish Word Properties." *Behavior Research Methods* 45:1246–58.

Dufour, Sophie, and Ronald Peereman. 2003. "Lexical Competition in Phonological Priming: Assessing the Role of Phonological Match and Mismatch Lengths between Primes and Targets." *Memory and Cognition* 31(8):1271–83.

van Esch, Daan. 2012. "Leiden Weibo Corpus." Retrieved May 22, 2019 (http://lwc.daanvanesch.nl/).

Feng, Chen, Yuan Yue, and Qingfang Zhang. 2019. "Syllables Are Retrieved before Segments in the Spoken Production of Mandarin Chinese: An ERP Study." *Scientific Reports* 9(11773):1–9.

Frauenfelder, Ulrich H., Rolf Harald Baayen, and Frauke M. Hellwig. 1993. "Neighborhood Density and Frequency Across Languages and Modalities." *Journal of Memory and Language* 32(6):781–804.

Goldstein, Rutherford, and Michael S. Vitevitch. 2014. "The Influence of Clustering Coefficient on Word-Learning: How Groups of Similar Sounding Words Facilitate Acquisition." *Frontiers in Psychology* 5(November):1307.

Goldstein, Rutherford, and Michael S. Vitevitch. 2017. "The Influence of Closeness Centrality on Lexical Processing." *Frontiers in Psychology* 8(1683).

Goulden, Robin, Paul Nation, and John Read. 1990. "How Large Can a Receptive Vocabulary Be?" *Applied Linguistics* 11(4):341–63.

Grainger, Jonathan, Mathilde Muneaux, Fernand Farioli, and Johannes C. Ziegler. 2005. "Effects of Phonological and Orthographic Neighbourhood Density Interact in Visual Word Recognition." *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology* 58(6):981–98.

Guo, Taomei, Danling Peng, and Ying Liu. 2005. "The Role of Phonological Activation in the Visual Semantic Retrieval of Chinese Characters." *Cognition* 98(2).

Hao, Meiling, Youyi Liu, Hua Shu, Ailing Xing, Ying Jiang, and Ping Li. 2015. "Developmental Changes in the Early Child Lexicon in Mandarin Chinese." *Journal of Child Language* 42(3):505–37.

Hao, Meiling, Hua Shu, Ailing Xing, and Ping Li. 2008. "Early Vocabulary Inventory for Mandarin Chinese." *Behavior Research Methods* 40(3):728–33.

Holliday, Jeffrey J., Rory Turnbull, and Julien Eychenne. 2017. "K-SPAN: A Lexical Database of Korean Surface Phonetic Forms and Phonological Neighborhood Density Statistics." *Behavior Research Methods* 49(5):1939–50.

Hsu, Chun-hsien, Jie-li Tsai, Chia-ying Lee, and Ovid J. L. Tzeng. 2009. "Orthographic Combinability and Phonological Consistency Effects in Reading Chinese Phonograms : An Event-Related Potential Study." *Brain and Language* 108:56–66.

Huang, Hsu-Wen, Chia-Ying Lee, Jie-Li Tsai, Chia-Lin Lee, Daisy L. Hung, and Ovid J. L.

Tzeng. 2006. “Orthographic Neighborhood Effects in Reading Chinese Two-Character Words.” *Neuroreport* 17(10):1061–65.
Iyengar, S. R. Sudarshan, C. E. Veni Madhavan, Katharina A. Zweig, and Abhiram Natarajan. 2012. “Understanding Human Navigation Using Network Analysis.” *Topics in Cognitive Science* 4:121–34.
Keuleers, Emmanuel, Marc Brysbaert, and Boris New. 2010. “SUBTLEX-NL: A New Measure for Dutch Word Frequency Based on Film Subtitles.” *Behavior Research Methods* 42(3):643–50.
Kučera, F., and W. Francis. 1967. “Computational Analysis of Present Day American English.”
Lee, Chia-ying, Hsu-Wen Huang, Wen-Jui Kuo, Jie-li Tsai, and Ovid J. L. Tzeng. 2009. “Cognitive and Neural Basis of the Consistency and Lexicality Effects in Reading Chinese.” *Journal of Neurolinguistics* 23(1):10–27.
Lee, Chia-Ying, Jie-li Tsai, Erica Chung-I. Su, Ovid J. L. Tzeng, and Daisy L. Hung. 2005. “Consistency, Regularity, and Frequency Effects in Naming Chinese Characters.” *Language and Linguistics* 6(1):75–107.
Lété, B., L. Sprenger-Charolles, and P. Colé. 2004. “Manulex: A Grade-Level Lexical Database from French Elementary-School Readers.” *Behavior Research Methods* 36:156–66.
Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Levelt, Willem J. M., Ardi Roelofs, and Antje S. Meyer. 1999. “A Theory of Lexical Access in Speech Production.” *Behavioral and Brain Sciences* 22(1):1–38; discussion 38-75.
Li, Bo. 2011. “Research on Chinese Word Segmentation and Proposals for Improvement.” Roskilde University.
Li, Chuchu, and Min Wang. 2017. “The Influence of Orthographic Experience on the Development of Phonological Preparation in Spoken Word Production.” *Memory & Cognition* 45:956–73.
Li, Chuchu, Min Wang, and Joshua A. Davis. 2015. “The Phonological Preparation Unit in Spoken Word Production in a Second Language.” *Bilingualism: Language and Cognition* 20(2):351–66.
Li, Meng-Feng, Xin-Yu Gao, Tai-Li Chou, and Jei-Tun Wu. 2017. “Neighborhood Frequency Effect in Chinese Word Recognition: Evidence from Naming and Lexical Decision.” *Journal of Psycholinguistic Research* 46:227–45.
Li, Meng-Feng, Wei-Chun Lin, Tai-Li Chou, Fu-Ling Yang, and Jei-Tun Wu. 2015. “The Role of Orthographic Neighborhood Size Effects in Chinese Word Recognition.” *Journal of Psycholinguisic Research* 44:219–36.
Li, Ping, Zhen Jin, and Li-Hai Tan. 2004. “Neural Representations of Nouns and Verbs in Chinese: An FMRI Study.” *NeuroImage* 21:1533–41.
Li, Qing-Lin, Hong-Yan Bi, Tong-Qi Wei, and Bao-Guo Chen. 2011. “Orthographic Neighborhood Size Effect in Chinese Character Naming: Orthographic and Phonological Activations.” *Acta Psychologica* 136(1):35–41.
Li, Qing-Lin, Hong-Yan Bi, and J. X. Zhang. 2010. “Neural Correlates of the Orthographic Neighborhood Size Effect in Chinese.” *European Journal of Neuroscience* 32(5):1–7.
Li, Xiaoqing, Chengqing Zong, and Keh-yih Su. 2015. “A Unified Model for Solving the OOV Problem of Chinese Word Segmentation.” *ACM Transactions on Asian Low-Resource Language Information Processing* 14(3):12:1--12:29.
Li, Yu, Linjun Zhang, Zhichao Xia, Jie Yang, Hua Shu, and Ping Li. 2017. “The Relationship

between Intrinsic Couplings of the Visual Word Form Area with Spoken Language Network and Reading Ability in Children and Adults." *Frontiers in Human Neuroscience* 11(June).

Lin, Yen Hwei. 1989. "Autosegmental Treatment of Segmental Processes in Chinese Phonology." University of Texas at Austin.

Lin, Yen Hwei. 2007. *The Sounds of Chinese with Audio CD*. Cambridge University Press.

Liu, Baolin, Zhixing Jin, Zhao Qing, and Zhongning Wang. 2011. "The Processing of Phonological, Orthographical, and Lexical Information of Chinese Characters in Sentence Contexts: An ERP Study." *Brain Research* 1372:81–91.

Liu, Youyi, Hua Shu, and Ping Li. 2007. "Word Naming and Psycholinguistic Norms: Chinese." *Behavior Research Methods* 39(2):192–98.

Luce, Paul A., and David B. Pisoni. 1998. "Recognizing Spoken Words: The Neighborhood Activation Model." *Ear and Hearing* 19(1):1–36.

Ma, Bosen, Xiaoyun Wang, and Degao Li. 2016. "The Processing of Visual and Phonological Configurations of Chinese One- and Two-Character Words in a Priming Task of Semantic Categorization." *Frontiers in Psychology* 6(Article 1918):1–14.

Ma, Jianqiang, Chunyu Kit, and Dale Gerdemann. 2012. "Semi-Automatic Annotation of Chinese Word Structure and Linguistics." *Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CIPS-SIGHAN 2012)* 0:9–17.

Malins, Jeffrey G., Danqi Gao, Ran Tao, James R. Booth, Hua Shu, Marc F. Joanisse, Li Liu, and Amy S. Desroches. 2014. "Developmental Differences in the Influence of Phonological Similarity on Spoken Word Processing in Mandarin Chinese." *Brain and Language* 138:38–50.

Malins, Jeffrey G., and Marc F. Joanisse. 2010. "The Roles of Tonal and Segmental Information in Mandarin Spoken Word Recognition: An Eyetracking Study." *Journal of Memory and Language* 62(4):407–20.

Malins, Jeffrey G., and Marc F. Joanisse. 2012. "Setting the Tone: An ERP Investigation of the Influences of Phonological Similarity on Spoken Word Recognition in Mandarin Chinese." *Neuropsychologia* 50(8):2032–43.

Mandera, Paweł, Emmanuel Keuleers, Zofia Wodniecka, and Marc Brysbaert. 2015. "Subtlex-PL: Subtitle-Based Word Frequency Estimates for Polish." *Behavior Research Methods* 47(2):471–83.

Marian, Viorica, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. "Clearpond: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities." *PLoS ONE* 7(8).

McClelland, James L., and Jeffrey L. Elman. 1986. "The TRACE Model of Speech Perception." *Cognitive Psychology* 18(1):1–86.

McEnery, Tony, and Richard Xiao. 2003. *The Lancaster Corpus of Mandarin Chinese (LCMC) By*.

Meyer, Antje S., and H. Schriefers. 1991. "Phonological Facilitation in Picture-Word Interference Experiments: Effects of Stimulus Onset Asynchrony and Types of Interfering Stimuli." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17(6):1146–60.

Myers, James, and Jane Tsay. 2005. "The Processing of Phonological Acceptability Judgments." *Proceedings of Symposium on 90-92 NSC Projects* 26–45.

Neergaard, Karl David. 2018. "Phonological Segmentation Neighborhoods." The Hong Kong

Polytechnic University.
Neergaard, Karl David, James Britton, and Chu-Ren Huang. 2019. “Neighborhood in Decay: Working Memory Modulates Effect of Phonological Similarity on Lexical Access.” Pp. 2447–53 in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, edited by A. K. Goel, C. M. Seifert, and C. Freksa. Montreal: Cognitive Science Society.
Neergaard, Karl David, and Chu-Ren Huang. 2016. “Graph Theoretic Approach to Mandarin Syllable Segmentation.” in *Proceedings of The 15th International Symposium on Chinese Languages and Linguistics (IsCLL-15)*. Taipei, Taiwan.
Neergaard, Karl David, and Chu-Ren Huang. 2019. “Constructing the Mandarin Phonological Network: Novel Syllable Inventory Used to Identify Schematic Segmentation.” *Complexity* (Article ID 6979830):1–21.
Neergaard, Karl David, and Chu-Ren Huang. 2021. “Mandarin Chinese Syllable Structure and Phonological Similarity: Perception and Production Studies.” in *Cambridge Handbook of Chinese Linguistics*, edited by Y. H. Lin and C.-R. Huang. Cambridge University Press.
Neergaard, Karl David, Jin Luo, and Chu-Ren Huang. 2019. “Phonological Network Fluency Identifies Phonological Restructuring through Mental Search.” *Scientific Reports* 9(15984):1–12.
Neergaard, Karl David, Hongzhi Xu, and Chu-Ren Huang. 2016. “Database of Mandarin Neighborhood Statistics.” Pp. 4032–4036 in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*. Portorož, Slovenia.
New, Boris, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. “Lexique 2 : A New French Lexical Database.” *Behavior Research Methods* 36(3):516–24.
Norris, Dennis. 1994. “Shortlist - a Connectionist Model of Continuous Speech Recognition.” *Cognition* 52(3):189–234.
O’Séaghdha, Pádraig G., Jenn-Yeu Chen, and Train-min Chen. 2010. “Proximate Units in Word Production: Phonological Encoding Begins with Syllables in Mandarin Chinese but with Segments in English.” *Cognition* 115(2):282–302.
Peng, Shu-Chen, Hui-Ping Lu, Nelson Lu, Yung-Song Lin, Mickael L. D. Deroche, and Monita Chatterjeee. 2017. “Processing of Acoustic Cues in Lexical-Tone Identification by Pediatric Cochlear-Implant Recipients.” *Journal of Speech, Language, and Hearing Research* 60:1223–35.
Perfetti, Charles A., and Sulan Zhang. 1995. “Very Early Phonological Activation in Chinese Reading.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21(1):24–33.
Qi, Zhenghan, Michelle Han, Keri Garel, Ee San Chen, and John D. E. Gabrieli. 2015. “White-Matter Structure in the Right Hemisphere Predicts Mandarin Chinese Learning Success.” *Journal of Neurolinguistics* 33:14–28.
Qu, Qingqing, Markus F. Damian, and Nina Kazanina. 2012. “Sound-Sized Segments Are Significant for Mandarin Speakers.” *Proceedings of the National Academy of Sciences of the United States of America* 109(35):14265–70.
Schiller, Niels Olaf. 2000. “Single Word Production in English: The Role of Subsyllabic Units During Phonological Encoding.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(2):512–28.
Schiller, Niels Olaf. 2004. “The Onset Effect in Word Naming.” *Journal of Memory and Language* 50:477–90.
Sereno, Joan A., and Hyunjung Lee. 2015. “The Contribution of Segmental and Tonal

Information in Mandarin Spoken Word Processing." *Language and Speech* 58(2):131–51.
Shattuck-Hufnagel, Stephanie. 1979. "Speech Errors as Evidence for a Serial-Ordering Mechanism in Sentence Production." Pp. 295–342 in *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*, edited by W. E. Cooper and E. C. T. Walker. Hillsdale, NJ: Erlbaum.
Shoemark, Philippa, Sharon Goldwater, James Kirby, and Rik Sarkar. 2016. "Towards Robust Cross-Linguistic Comparisons of Phonological Networks Towards Robust Cross-Linguistic Comparisons of Phonological Networks." Pp. 110–20 in *the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Berlin, Germany.
Shuai, Lan, and Jeffrey G. Malins. 2016. "Encoding Lexical Tones in JTRACE : A Simulation of Monosyllabic Spoken Word Recognition in Mandarin Chinese." *Behavior Research Methods*.
Siew, Cynthia S. Q. 2013. "Community Structure in the Phonological Network." *Frontiers in Psychology* 4:553.
Siew, Cynthia S. Q., and Michael S. Vitevitch. 2015. "Spoken Word Recognition and Serial Recall of Words from Componenets in the Phonological Network." *Journal of Experimental Psychology: Learning, Memory and Cognition, and Cognition*.
Siew, Cynthia S. Q., and Michael S. Vitevitch. 2019. "The Phonographic Language Network: Using Network Science to Investigate the Phonological and Orthographic Similarity Structure of Language." *Journal of Experimental Psychology: General* 148(3):475–500.
Siew, Cynthia S. Q., Dirk U. Wulff, Nicole M. Beckage, and Yoed N. Kenett. 2019. "Cognitive Network Science : A Review of Research on Cognition through the Lens of Network Representations , Processes , and Dynamics." *Complexity* (Article 2108423):1–24.
Spinks, John A., Ying Liu, Charles A. Perfetti, and Li Hai Tan. 2000. "Reading Chinese Characters for Meaning: The Role of Phonological Information." *Cognition* 76(1):1–11.
Stella, Massimo. 2018. "Cohort and Rhyme Priming Emerge from the Multiplex Network Structure of the Mental Lexicon." *Complexity* (6438702):1–14.
Stella, Massimo, Nicole M. Beckage, and Markus Brede. 2017. "Multiplex Lexical Networks Reveal Patterns in Early Word Acquisition in Children." *Scientific Reports* 7(46730):1–10.
Stella, Massimo, Nicole M. Beckage, Markus Brede, and Manlio De Domenico. 2018. "Multiplex Model of Mental Lexicon Reveals Explosive Learning in Humans." *Scientific Reports* 8(2259).
Stella, Massimo, and Markus Brede. 2015. "Patterns in the English Language: Phonological Networks, Percolation and Assembly Models." *Journal of Statistical Mechanics: Theory and Experiment* 5.
Storkel, Holly L., Jonna Armbruster, and Tiffany P. Hogan. 2006. "Differentiating Phonotactic Probability and Neighborhood Density in Adult Word Learning." *Journal of Speech, Language, and Hearing Research* 49(December):1175–92.
Storkel, Holly L., and Jill R. Hoover. 2010. "An Online Calculator to Compute Phonotactic Probability and Neighborhood Density on the Basis of Child Corpora of Spoken American English." *Behavior Research Methods* 42(2):497–506.
Strand, Julia F. 2013. "Phi-Square Lexical Competition Database (Phi-Lex): An Online Tool for Quantifying Auditory and Visual Lexical Competition." *Behavior Research Methods* 46(1):148–58.
Sun, Chaofen. 2006. *Chinese: A Linguistic Introduction*. Cambridge University Press.

Sun, Ching Chu, Peter Hendrix, Jianqiang Ma, and Rolf Harald Baayen. 2018. "Chinese Lexical Database (CLD): A Large-Scale Lexical Database for Simplified Mandarin Chinese." *Behavior Research Methods* 50(6):2606–29.

Sun, H. L., J. P. Huang, D. J. Sun, D. J. Li, and H. B. Xing. 1997. "Introduction to Language Corpus System of Modern Chinese Study." in *Paper collection for the Fifth World Chinese Teaching Symposium*, edited by M. Y. Hu. Beijing: Peking University Publisher.

Sze, Wei Ping, Susan J. Rickard Liow, and Melvin J. Yap. 2014. "The Chinese Lexicon Project: A Repository of Lexical Decision Behavioral Responses for 2,500 Chinese Characters." *Behavior Research Methods* 46:263–73.

Tan, Li-Hai, and Charles A. Perfetti. 1999. "Phonological Activation in Visual Identification of Chinese Two-Character Words." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25(2):382–93.

Tan, Li-Hai, Min Xu, Chun Qi Chang, and Wai-Ting Siok. 2013. "China's Language Input System in the Digital Age Affects Children's Reading Development." *Proceedings of the National Academy of Sciences of the United States of America* 110(3):1119–23.

Treffers-Daller, Jeanine, and James Milton. 2013. "Vocabulary Size Revisited: The Link between Vocabulary Size and Academic Achievement." *Applied Linguistics Review* 4(1):151–72.

Tsai, Jie-Li, Chia-Ying Lee, Ying-Chun Lin, Ovid J. L. Tzeng, and Daisy L. Hung. 2006. "Neighborhood Size Effects of Chinese Words in Lexical Decision and Reading." *Language and Linguistics* 7(3):659–75.

Tsai, Pei-Tzu. 2007. "The Effects of Phonological Neighborhoods on Spoken Word Recognition in Mandarin Chinese." Universty of Maryland, College Park.

Tsang, Yiu-Kei, Jian Huang, Ming Lui, Mingfeng Xue, Yin-Wah Fiona Chan, Suiping Wang, and Hsuan-Chih Chen. 2018. "MELD-SCH: A Megastudy of Lexical Decision in Simplified Chinese." *Behavior Research Methods* 50:1763–77.

Turnbull, Rory, and Sharon Peperkamp. 2016. "What Governs a Language's Lexicon? Determining the Organizing Principles of Phonological Neighbourhood Networks." Pp. 83–94 in *Proceedings of the 5th International Workshop on Complex Networks and their Applications (COMPLEX NETWORKS 2016)*, edited by H. Cherifi, S. Gaito, W. Quattrociocchi, and A. Sala. Springer, Cham.

Verdonschot, Rinus Gerardus, Mariko Nakayama, Qingfang Zhang, Katsuo Tamaoka, and Niels Olaf Schiller. 2013. "The Proximate Phonological Unit of Chinese-English Bilinguals: Proficiency Matters." *PLoS ONE* 8(4).

Vitevitch, Michael S. 2002. "The Influence of Phonological Similarity Neighborhoods on Speech Production." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(4):735–47.

Vitevitch, Michael S. 2008. "What Can Graph Theory Tell Us about Word Learning and Lexical Retrieval?" *Journal of Speech, Language, and Hearing Research* 51:408–22.

Vitevitch, Michael S., and Paul A. Luce. 2016. "Phonological Neighborhood Effects in Spoken Word Perception and Production." *Annual Review of Linguistics* 8(2):75–94.

Vitevitch, Michael S., and Eva Rodríguez. 2004. "Neighborhood Density Effects in Spoken Word Recognition in Spanish." *Journal of Multilingual Communication Disorders* 3(1):64–73.

Vitevitch, Michael S., and Mitchell S. Sommers. 2003. "The Facilitative Influence of Phonological Similarity and Neighborhood Frequency in Speech Production in Younger and

Older Adults." *Memory & Cognition* 31(4):491–504.
Vitevitch, Michael S., and Melissa K. Stamer. 2006. "The Curious Case of Competition in Spanish Speech Production." *Language and Cognitive Processes* 21(6):760–70.
Vitevitch, Michael S., and Melissa K. Stamer. 2009. *The Influence of Neighborhood Density (and Neighborhood Frequency) in Spanish Speech Production: A Follow-up Report*.
Vitevitch, Michael S., Melissa K. Stamer, and Douglas Kieweg. 2012. "Short Research Note: The Beginning Spanish Lexicon: A Web-Based Interface to Calculate Phonological Similarity among Spanish Words in Adults Learning Spanish as a Foreign Language." *Second Language Research* 28(1):103–12.
Wan, I. Ping. 2006. "A Psycholinguistic Study of Postnuclear Glides and Coda Nasals in Mandarin." *Journal of Language and Linguistics* 5(2):158–76.
Wang, Jie, Andus Wing-kuen Wong, Suiping Wang, and Hsuan-chih Chen. 2017. "Primary Phonological Planning Units in Spoken Word Production Are Language-Specific: Evidence from an ERP Study." *Scientific Reports* 7(5815):1–8.
Wang, Quanhong, and Jiewei Zhang. 2011. "N400 Solution Effect of Chinese Character Fragments: An Orthographic Neighborhood Size Effect." *Brain Research Bulletin* 86(3–4):179–88.
Wang, Wenna, Xiaojian Li, Ning Ning, and John X. Zhang. 2012. "The Nature of the Homophone Density Effect: An ERP Study with Chinese Spoken Monosyllable Homophones." *Neuroscience Letters* 516(1):67–71.
Wang, Yun, and Jeanine Treffers-Daller. 2017. "Explaining Listening Comprehension among L2 Learners of English: The Contribution of General Language Proficiency, Vocabulary Knowledge and Metacognitive Awareness." *System* 65:139–50.
Wen, W. 1980. "Cong Yingwen de Tongxingci Lai Kan Hanyu Pinyin Wenzi de Tongyinci [A Study of Chinese Homophones from the View of English Homographs]." *Yuwen Xiandaihua [Modernizing Our Language]* 2:120–24.
Wiener, Seth, and Rory Turnbull. 2016. "Constraints of Tones, Vowels and Consonants on Lexical Selection in Mandarin Chinese." *Language and Speech* 59(1):59–82.
Wong, Andus Wing-Kuen, Y. Wu, and H. C. Chen. 2014. "Limited Role of Phonology in Reading Chinese Two-Character Compounds: Evidence from an ERP Study." *Neuroscience* 256:342–51.
Wu, Fei, and Michael Kenstowicz. 2015. "Duration Reflexes of Syllable Structure in Mandarin." *Lingua* 164:87–99.
Wu, Jei-Tun, and Hsin-Chin Chen. 2003. "Chinese Orthographic Priming in Lexical Decision and Naming." *Chinese Journal of Psychology* 45(1):75–95.
Wu, Jei-Tun, Fu-Ling Yang, and Wei-Chun Jin. 2013. "Beyond Phonology Matters in Character Recognition." *Chinese Journal of Psychology* 55(3):289–318.
Xia, Quansheng, Lan Wang, and Gang Peng. 2016. "Nouns and Verbs in Chinese Are Processed Differently: Evidence from an ERP Study on Monosyllabic and Disyllabic Word Processing." *Journal of Neurolinguistics* 40:66–78.
Xu, Shirong. 1980. *Putonghua Yuyin Zhishi [Phonology of Standard Chinese]*. Beijing: Wenzi Gaige Chubans.
Xu, Yaoda, Alexander Pollatsek, and Mary C. Potter. 1999. "The Activation of Phonology During Silent Chinese Word Reading." *Journal of Experimental Psychology: Learning Memory and Cognition* 25(4):838–57.
Yang, Fu-Ling, and Jei-Tun Wu. 2014. "Orthographic Inhibition between Characters with

Identical Semantic Radicals in Primed Character Decision Tasks." *Chinese Journal of Psychology* 56(1):49–63.
Yao, Yao, and Bhamini Sharma. 2017. "What Is in the Neighborhood of a Tonal Syllable? Evidence from Auditory Lexical Decision in Mandarin Chinese." Pp. 1–14 in *Proceedings of the Linguistic Society of America*. Vol. 2. Austin, Texas.
Yates, Mark. 2005. "Phonological Neighbors Speed Visual Word Processing: Evidence from Multiple Tasks." *Journal of Experimental Psychology: Learning Memory and Cognition* 31(6):1385–97.
Yeh, Li-li, Bill Wells, Joy Stackhouse, and Marcin Szczerbinski. 2015. "The Development of Phonological Representations in Mandarin-Speaking Children: Evidence from a Longitudinal Study of Phonological Awareness." *Clinical Linguistics & Phonetics* 29(4):266–75.
You, Wenping, Qingfang Zhang, and Rinus Gerardus Verdonschot. 2012. "Masked Syllable Priming Effects in Word and Picture Naming in Chinese." *PLoS ONE* 7(10).
Yu, Mengxia, Ce Mo, You Li, and Lei Mo. 2015. "Distinct Representations of Syllables and Phonemes in Chinese Production: Evidence from FMRI Adaptation." *Neuropsychologia* 77:253–59.
Yu, Mengxia, Ce Mo, and Lei Mo. 2014. "The Role of Phoneme in Mandarin Chinese Production: Evidence from ERPs." *PLoS ONE* 9(9):e106486.
Zhang, Qin, John X. Zhang, and Lingyue Kong. 2009. "An ERP Study on the Time Course of Phonological and Semantic Activation in Chinese Word Recognition." *International Journal of Psychophysiology* 73(3):235–45.
Zhang, Qingfang, and Markus F. Damian. 2019. "Syllables Constitute Proximate Units for Mandarin Speakers: Electrophysiological Evidence from a Masked Priming Task." *Psychophysiology* 1–15.
Zhang, Yang, Jan Niehues, and Alex Waibel. 2016. "Integrating Encyclopedic Knowledge into Neural Language Models." in *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, The United States of America.
Zhao, Ping, and Xiaoli Ji. 2018. "Validation of the Mandarin Version of the Vocabulary Size Test." *RELC Journal* 49(3):308–21.
Zhao, Xiaowei, and Ping Li. 2009. "An Online Database of Phonological Representations for Mandarin Chinese." *Behavior Research Methods* 41(2):575–83.
Zhou, Wei, Veronica P. Y. Kwok, Mengmeng Su, Jin Luo, and Li Hai Tan. 2020. "Children's Neurodevelopment of Reading Is Affected by China's Language Input System in the Information Era." *Npj Science of Learning* 5(1):1–9.
Zhou, Xiaolin, and William Marslen-Wilson. 2000. "The Relative Time Course of Semantic and Phonological Activation in Reading Chinese." *Journal of Experimental Psychology: Learning Memory and Cognition* 26(5):1245–65.
Zhou, Xiaolin, and William D. Marslen-Wilson. 1999. "Phonology, Orthography, and Semantic Activation in Reading Chinese." *Journal of Memory and Language* 41(4):579–606.
Zhou, Youguang. 1978. "To What Degree Are the 'Phonetics' of Present Day Chinese Characters Still Phonetic?" *Zhongguo Yuwen* 3:172–77.
Zhou, Youguang. 2003. *The Historical Evolution of Chinese Languages and Scripts. Pathways to Advanced Skills Series, Volume 8.* Columbus, Ohio: Ohio State University National East Asian Languages Resource Center.
Ziegler, Johannes C., Mathilde Muneaux, and Jonathan Grainger. 2003. "Neighborhood Effects

in Auditory Word Recognition: Phonological Competition and Orthographic Facilitation." *Journal of Memory and Language* 48(4):779–93.