Behavior Research Methods manuscript No.
(will be inserted by the editor)

Chinese Lexical Database (CLD)

A large-scale lexical database for simplified Mandarin Chinese

Ching Chu Sun · Peter Hendrix · Jianqiang Ma · Rolf Harald Baayen

Received: date / Accepted: date

Abstract We present the Chinese Lexical Database (CLD): a large-scale lexical database for simplified Chinese. The CLD provides a wealth of lexical information for 3,913 one-character words, 34,233 two-character words, 7,143 three-character words, and 3,355 four-character words, and is publicly available through `http://www.chineselexicaldatabase.com`. For each of the 48,644 words in the CLD, we provide a wide range of categorical predictors, as well as an extensive set of frequency measures, complexity measures, neighborhood density measures, orthography-phonology consistency measures, and information-theoretic measures. We evaluate the explanatory power of the lexical variables in the CLD in the context of experimental data through analyses of lexical decision latencies for one-character, two-character, three-character and four-character words, as well as word naming latencies for one-character and two-character words. The results of these analyses are discussed.

C.C. Sun
Eberhard Karl's Universität Tübingen
E-mail: ching.chu.sun@uni-tuebingen.de

P. Hendrix
Eberhard Karl's Universität Tübingen
E-mail: peter.hendrix@gmail.com

J. Ma
Eberhard Karl's Universität Tübingen
E-mail: jianqiang.ma@uni-tuebingen.de

R. H. Baayen
Eberhard Karl's Universität Tübingen
E-mail: harald.baayen@uni-tuebingen.de

Keywords lexical database · Mandarin Chinese · simplified Chinese · Chinese Lexical Database · CLD

1 Introduction

Over the last decades, the wealth of experimental research in the psycholinguistic literature has been complemented with large-scale lexical resources. The most well known lexical database, perhaps, is CELEX (Baayen et al., 1995), which contains a large amount of lexical information for English, German, and Dutch. Language-specific lexical databases have been developed as well. The MRC psycholinguistic database (Coltheart, 1981), Lexique (New et al., 2001, 2004, 2007), dlexDB (Heister et al., 2011), and EsPal (Duchon et al., 2013), for instance, are examples of lexical resources for English, French, German, and Spanish. In recent years, lexical databases have also been developed for less well-studied languages, including Modern Greek (Kyparissiadis et al., 2017; Ktori et al., 2008), Modern Arabic (Boudelaa and Marslen-Wilson, 2010), and Malay (Yap et al., 2010).

Another less-studied language is Mandarin Chinese, which is also referred to as 普通话 ("common language"). Mandarin Chinese is part of the Sino-Tibetan language family. It is the official language of China and has, according to recent estimates, nearly a billion (935 million) native speakers (Parkvall, 2007). Mandarin Chinese is a tonal language. The basic phonological unit is the syllable. Each syllable consists of vowels and consonants in a (C)V(C) structure at the segmental level and a tone at the suprasegmental level (cf. Sun, 2006). In the writing system, syllables correspond to 汉字 (Hanzi, literal translation: "Chinese characters").

According to the Table of General Standard Chinese Characters, there are about 8,100 Chinese characters, of which 6,500 are commonly used (Ministry of Education of the People's Republic of China, 2013). Due to the complicated nature of traditional Chinese characters, learning to read or write in Chinese is a difficult task. To master the language, one has to memorize and rehearse thousands of characters. To improve literacy, the Chinese government decided to simplify over 2,200 characters in the 1950s (Honorof and Feldman, 2006). The resulting writing system is referred to as simplified Chinese and is the standard writing system in modern-day China.

Due to the historical focus of psycholinguistic research on Germanic and – to a lesser extent – Roman languages, lexical processing is less well-studied in Mandarin Chinese than it is in English, German, or Dutch. Concomitantly, relatively few lexical resources exist for Mandarin Chinese. For traditional Chinese, which has remained the standard writing system in Taiwan, Taiwan Sinica has compiled a large-scale lexical database (Chang et al., 2016), with naming latencies and 12 numerical variables for 3,314 one-character words. For simplified Chinese, four lexical resources have recently been developed. First, Liu et al. (2007) released a lexical database that contains word naming latencies and 15 lexical predictors for 2,423 one-character words. Second, SUBTLEX-CH provides a collection of character and word frequency counts based on a corpus of movie subtitles in simplified Chinese (Cai and Brysbaert, 2010). Third, the Chinese Lexicon Project (CLP) contains lexical decision latencies for 2,500 characters and is released without lexical variables (Sze et al., 2014).[1] Fourth, MELD-SCH (Tsang et al., 2017) provides lexical decision latencies for 1,020 one-character, 10,022 two-character, 949 three-character, and 587 four-character words, as well as 10 numerical predictors for these words.

Thus far, lexical resources available for Mandarin Chinese have primarily been developed for one-character words. According to Honorof and Feldman (2006), no more than 34% of the word tokens in Mandarin Chinese are single character words. Frequency counts based on a large-scale corpus of simplified Chinese, the Simplified Chinese Corpus of Webpages (henceforth SCCoW; Shaoul et al., 2016) yielded a similar estimate of 35%. An overwhelming majority of the remaining words consist of two characters. In the SCCoW, 59% of all word tokens are two-character words. A further 5% of all word tokens in the SCCoW are three-character words, whereas less than 1% of all word tokens consist four or more characters. The addition of multi-character words, therefore, is an equally straightforward and crucial extension of existing lexical resources for simplified Chinese.

Useful sets of lexical predictors have been provided by existing lexical resources. In particular, Liu et al. (2007) provided an extensive set of lexical predictors, including measures of orthographic and phonological frequency, visual complexity, phonological neighborhood density, character combinability, and the consistency of the orthography-phonology mapping. Furthermore, subjective ratings were included for age of acquisition, familiarity, concreteness and imageability. In addition, Tsang et al. (2017) made available word and character frequency counts, as well as stroke counts, character combinability counts, and pronunciation counts for a large semi-random subset of one-character, two-character, three-character, and four-character words in SUBTLEX-CH (Cai and Brysbaert, 2010).

Nonetheless, the set of lexical predictors provided by existing lexical resources offers some room for improvement. Most notably, existing resources typically provide a single lexical variable for each theoretical concept. Concepts like frequency, visual complexity, or consistency of the orthography-phonology mapping, however, can be expressed through a variety of lexical variables defined in different units or at different grain sizes. The possibility to compare different measures of a theoretical concept allows for more in-depth analyses of experimental data, as well as for more precise formulations of theoretical or computational models of lexical processing.

Furthermore, the importance of the combinatorial properties of characters increases in the context of multi-character words. To accurately capture the influence of these properties on lexical processing, counts of the number of words in which characters occur may not be sufficient. Additional measures of the combinational properties of the characters in multi-character words therefore need to be developed. For this purpose, information-theoretic measures of the association between the characters in multi-character words, and of the predictability of a character given the other characters in a word are particularly promising (cf. C. C. Sun, 2016).

Here, we present a new lexical database for simplified Chinese: the Chinese Lexical Database (henceforth CLD). The CLD provides a wealth of lexical information for the set of one-character, two-character, three-character and four-character words that occur in

[1] In the context of the Chinese Lexicon Project, Tse et al. (2016) recently provided lexical decision latencies for two-character words as well. In this study, however, stimuli were presented in traditional Chinese.

both SUBTLEX-CH (Cai and Brysbaert, 2010) and in the Leiden Weibo Corpus (Van Esch, 2012), with the exclusion of proper nouns and words that contain traditional (rather than simplified) Chinese characters. It contains 3,914 one-character words, 34,233 two-character words, 7,143 three-character words, and 3,355 four-character words, for a total of 48,644 words. These 48,644 words consist of 4,895 unique characters. Below, we first introduce the CLD and the lexical information it comprises. Next, we evaluate the explanatory power of the lexical variables in the CLD for both lexical decision data and word naming data.

2 Chinese Lexical Database (CLD)

The CLD is released under the GNU General Public License. The database is available via the library of Eberhard Karl's Universität Tübingen (`http://dx.doi.org/10.15496/publikation-21197`), as well as through `http://www.chineselexicaldatabase.com`. Furthermore, the database is provided in the supplementary materials for this paper. On `http://www.chineselexicaldatabase.com`, we provide two options to access the data in the CLD. First, the database can be downloaded in .txt and .csv format (61.4 mb, zipped: 19.8 mb), in .pdf format (33.7 mb), or as a data frame for the statistical software R (13.7 mb). Second, the CLD can be accessed through a search interface. Users have the option to search the full database, or to submit lists of words, characters or radicals for which lexical information should be displayed. Similarly, either the full set of variables can be shown or the user may select a subset of variables in which she is interested.

For the categorical variables that describe the structure, type, and tone of a character (see our description of the lexical variables in the CLD below), factor levels that should be included in the output can manually be selected (by default all factor levels are included). For numerical variables, minimum and maximum values can be set to limit the range of a variable in the output. The result of a search can either be viewed in the browser or e-mailed to the user.

Below, we provide a description of the information contained in the Chinese Lexical Database (CLD). The lexical variables in the CLD can be divided into six conceptual classes: categorical predictors, frequency measures, complexity measures, neighborhood density measures, orthography-phonology consistency measures, and information-theoretic measures. Below, we describe the lexical variables in each of these classes. A schematic overview of the classes of lexical variables is provided in Table 1.

2.1 Categorical variables

The CLD contains a number of categorical variables. The first five categorical variables are the word (*Word*) and its component characters (*C1*, *C2*, *C3*, and *C4*; henceforth we refer to identical variables for the first, second, third, and fourth character through the abbreviation *C1-C4*). Words and characters have pronunciations as well. Pronunciations for both words and characters are provided both in Pinyin (*Pinyin*, *C1-C4 Pinyin*) and in IPA format (*IPA*, *C1-C4 IPA*). Pinyin literally means "spell the sound". Pinyin transcriptions translate Chinese characters into a romanized form based on their pronunciation. The Pinyin variables in the CLD are based on a publicly available Pinyin annotator developed by Xiao (2015).

Pronunciations in IPA format are based on the Pinyin transcriptions of a word and were obtained by applying a set of Pinyin-to-IPA conversion rules to the Pinyin transcriptions (Wikipedia, 2016). To allow users to control for the influence of phonetic differences on the registration of the onset and offset of a response in naming tasks, we furthermore included the initial phoneme and the final phoneme of each word as lexical variables in the CLD (*Initial Phoneme*, *Final Phoneme*). Finally, we provide the tone (1 (high-level); 2 (high-rising); 3 (low/dipping); 4 (high-falling); 5 (neutral)) of each character (*C1-C4 Tone*).

In addition to the above-mentioned measures that describe the pronunciation of a word and its characters, we included a set of categorical measures that describe the orthographic structure of each character (*C1-C4 Structure*). Each character in the CLD is encoded as one of six different structures: Left-Right (e.g., 明 "brightness" or "tomorrow"), Left-Right-Bottom (e.g., 边 "side"), Up-Down (e.g., 草 "grass"), Circle (e.g., 回 "return"), Half Circle (e.g., 区 "area"), or Single (e.g., 开 "open"). More complicated structures were approximated as best as possible through this set of six structures. The character 涮, for instance, was encoded as left-right (氵 + 刷), whereas the character 烹 (亨 + 灬) was encoded as up-down.

The final categorical predictor at the character level is character type (*C1-C4 Type*). Chinese characters can be divided into six different types, which, in the Chinese linguistic literature, are called the "six writings" (cf. Yip, 2000). Hsieh (2006) argues that there are four basic types of character constructions among these "six writings": pictographic, pictologic, pictosynthetic, and pictophonetic. The other two types, "phonetic loan character" and "cognate", he argues, are extensions of the four basic character types. In our encoding of the character type measures in the CLD, we follow Hsieh (2006)

Table 1 Overview of the (classes of) predictors in the Chinese Lexical Database. Abbreviations: C = character, SR = semantic radical, PR = phonetic radical, OLD = orthographic Levenshtein distance, PLD = phonological Levenshtein distance.

Categorical variables
Word, C1-C4, Pinyin, C1-C4 Pinyin, IPA, C1-C4 IPA, Initial Phoneme, Final Phoneme, C1-C4 Tone, C1-C4 Structure, C1-C4 Type, C1-C4 SR, C1-C4 PR, C1-C4 PR Pinyin, C1-C4 PR Regularity
Frequency measures
Orthographic frequency Frequency, C1-C4 Frequency, Frequency SUBTLEX-CH, C1-C4 Frequency SUBTLEX-CH, Frequency LWC, C1-C4 Frequency LWC, C1-C4 Family Size, C1-C4 Family Frequency, C1-C4 SR Frequency, C1-C4 SR Family Size, C1-C4 PR Frequency, C1-C4 PR Family Size
Phonological frequency Phonological Frequency, C1-C4 Phonological Frequency, Mean Phoneme Frequency, C1-C4 Mean Phoneme Frequency, Min Phoneme Frequency, C1-C4 Min Phoneme Frequency, Max Phoneme Frequency, C1-C4 Max Phoneme Frequency, C1-C4 Initial Phoneme Frequency, Mean Diphone Frequency, C1-C4 Mean Diphone Frequency, Min Diphone Frequency, C1-C4 Min Diphone Frequency, Max Diphone Frequency, C1-C4 Max Diphone Frequency, C1-C4 Initial Diphone Frequency, Transitional Diphone 1-3 Frequency
Complexity measures
Length, Strokes, C1-C4 Strokes, C1-C4 Pixels, C1-C4 Picture Size, C1-C4 SR Strokes, C1-C4 PR Strokes, Phonemes, C1-C4 Phonemes
Neighborhood density measures
Phonological N, C1-C4 Phonological N, PLD, C1-C4 PLD, C1-C4 OLD Pixels
Orthography-phonology consistency
Character level C1-C4 Friends, C1-C4 Friends Frequency, C1-C4 Homograph Types, C1-C4 Homograph Tokens, C1-C4 Homographs Frequency C1-C4 Homophone Types, C1-C4 Homophone Tokens, C1-C4 Homophones Frequency
Phonetic radical level C1-C4 PR Friends, C1-C4 PR Friends Frequency, C1-C4 PR Enemies Types, C1-C4 PR Enemies Tokens, C1-C4 PR Enemies Frequency, C1-C4 PR Backward Enemies Types, C1-C4 PR Backward Enemies Tokens, C1-C4 PR Backward Enemies Frequency
Information-theoretic measures
C1 Conditional Probability, C12 Conditional Probability, C123 Conditional Probability, C1 Backward Conditional Probability, C12 Backward Conditional Probability, C123 Backward Conditional Probability, C1 Entropy, C12 Entropy, C123 Entropy, C1 Backward Entropy, C12 Backward Entropy, C123 Backward Entropy, C1-C2 Relative Entropy Pointwise Mutual Information, Position-specific Pointwise Mutual Information, T-Score, Position-specific T-Score, Entropy Character Frequencies

and discern four different types of characters. Characters that do not fit into one of the four basic character types are encoded as "Other". Most commonly, characters encoded as "Other" were simplified to the extent that they no longer belong to one of the four basic character types.

Pictographic characters are the oldest type of Chinese characters and originate from non-linguistic symbolic systems. The orthographic form of these characters corresponds to their meaning (e.g., 川 ("river"); 山 ("mountain")). The idea behind pictologic characters is similar. Pictologic characters, however, typically refer to objects that do not have a concrete, easy to depict shape. Often, a stroke is added to a pictographic character to express a more specific or more abstract concept. The pictologic character 刃 ("blade"), for instance, was derived from the pictographic character 刀 ("knife"). Pictosynthetic characters comprise multiple

pictographic characters that together form a new character. The combination of 日 ("sun") and 月 ("moon"), for instance, results in the new character 明 ("brightness"), which describes a shared semantic property of both characters.

The fourth type of character, pictophonetic, combines two common components of Chinese characters: the semantic radical and the phonetic radical. Semantic radicals typically provide information about the meaning of a character. The semantic radical 氵, for instance, commonly appears in words that describe liquids, such as 河 ("river"), 海 ("ocean"), or 泪 ("tears"). Phonetic radicals provide information about the pronunciation of a character. This information, however, is not always reliable. While the pronunciation of the character 榆 ("elm") and its phonetic radical 俞 are identical ("yu2"), the pronunciation of the character 拓 ("tuo4") and its phonetic radical 石 ("shi2") are entirely different. We return to the inconsistency of the information provided by the phonetic radical below. An example of a pictophonetic character is 清 ("to clean"), which consists of the semantic radical 氵 ("liquid") and the phonetic radical 青. The phonetic radical determines the pronunciation of the character 清, which is identical to the pronunciation of the phonetic radical: "qing1".

Both phonetic and semantic radicals have received considerable attention in the experimental psycholinguistic literature for Mandarin Chinese. We therefore included both the semantic radical (*C1-C4 SR*) and the phonetic radical (*C1-C4 PR*) for each character as lexical variables in the CLD. Both semantic radicals and phonetic radicals were retrieved from the Chinese Character Dictionary on the Mandarin Tools website (Peterson, 2005). While each character has a semantic radical, not every character has a phonetic radical. Of the 4,895 unique character in the CLD, 3,538 contain a phonetic radical (72.28%).

We now return to the issue of the unreliability of the information provided by the phonetic radical. The fact that phonetic radicals do not always provide reliable information about the pronunciation of a character has been shown to influence lexical processing. Seidenberg (1985a), for instance, showed that characters with pronunciations that are identical to the pronunciation of their phonetic radical are named faster than characters for which the pronunciation of the character differs from the pronunciation of the phonetic radical (see also Liu et al., 2007; Hue, 1992). To allow for the investigation of phonetic radical regularity effects, we included the pronunciation of the phonetic radical (*C1-C4 PR Pinyin*) in the CLD. Furthermore, we provide a binary variable that indicates whether (1) or not (0) the pronunciation of a character is identical to the pronunciation of its phonetic radical (*C1-C4 PR Regularity*). This binary variable was set to NA for characters with unpronounceable phonetic radicals. Further numerical measures of the consistency of the information provided by the phonetic radical are introduced in our discussion of orthography-phonology consistency measures below.

2.2 Frequency measures

Frequency measures in the CLD can be divided into two subclasses: orthographic frequency measures and phonological frequency measures. Frequency measures in each of these classes are provided at different grain sizes, ranging from the word level to the radical level for orthographic frequency measures and from the word level to the phoneme level for phonological frequency measures. Below, we describe the orthographic frequency measures and the phonological frequency measures in the CLD. First, however, we discuss how we selected the textual resources that underlie the frequency measures in the CLD.

2.2.1 Corpus selection

A pivotal decision during the construction of a lexical database is to select one or more textual resources on the basis of which word frequency counts are calculated. The quality of word frequency counts has a large influence on the predictive power of the lexical predictors in the database. Word frequency is typically (one of) the strongest predictor(s) for psycholinguistic data sets. Furthermore, a wide range of other lexical variables is calculated on the basis of frequency counts, including frequency counts at smaller grain sizes, family size and family frequency measures, orthography-phonology consistency measures and information-theoretic measures. The quality of these measures, therefore, directly depends on the quality of the word frequency counts.

The predictive power of frequency measures derived from different textual resources depends on the degree to which the content of these resources corresponds to the linguistic experience of the participants in psycholinguistic experiments. A corpus that contains a large collection of bureaucratic documents, for instance, may be less representative of the linguistic experience of the average language user, than a collection of novels.

To find out which corpus yields frequency measures that are most predictive for psycholinguistic data in Mandarin Chinese, we compared frequency measures from six different corpora for seven experimental data sets. The six corpora are SUBTLEX-CH (Cai and Brysbaert, 2010), Chinese Gigaword (Graff and Chen, 2003), the Simplified Chinese Corpus of Webpages (Shaoul

et al., 2016), the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2008), the Leiden Weibo Corpus (Van Esch, 2012), and the Public Weixin Corpus (L. Sun, 2016).

SUBTLEX-CH is a corpus of film subtitles that consists of 33.5 million words. In recent studies, frequency counts from SUBTLEX-CH have been shown to be highly predictive for lexical decision latencies (see Tse et al., 2016; Cai and Brysbaert, 2010), as compared to frequency counts from other textual resources. Chinese Gigaword provides a corpus of newswire texts that consists of no less than 1,118.4 million (1.1 billion) words, whereas the Simplified Chinese Corpus of Webpages (henceforth SCCoW) contains 466.6 million words of text extracted from web pages. The Lancaster Corpus of Mandarin Chinese (LCMC) is a balanced corpus of Mandarin Chinese that comprises written texts from a variety of genres, including press, fictional texts, scientific documents, and more. At 1 million words, however, the size of the LCMC, is limited.

The final two corpora under investigation, the Leiden Weibo Corpus (henceforth LWC) and the Public Weixin Corpus (henceforth PWC), are based on social media. The LWC is a 101.4 million word corpus based on messages posted on Sina Weibo, a social medium that is perhaps best described as a hybrid between Facebook and Twitter (Van Esch, 2012). Sina Weibo has over 300 million monthly active users, which jointly post more than 100 million messages per day. Weixin ("we chat") is a messaging app similar to Snapchat that, similar to Weibo, has over 300 million monthly active users. The PWC is a collection of Weixin messages that comprises 491.2 million words (L. Sun, 2016).

We evaluated the predictive power of word and character frequency measures from the six corpora described above for 7 experimental data sets. The Chinese Lexicon Project (henceforth CLP) contains lexical decision latencies for 2,500 one-character words (Sze et al., 2014), and was recently extended with data for 25,000 two-character words (Tse et al., 2016). In the context of SUBTLEX-CH, Cai and Brysbaert (2010) provide lexical decision latencies for 400 two-character words. Furthermore, MELD-SCH (Tsang et al., 2017) provides lexical decision latencies for 1,020 one-character words and 10,022 two-character words. These five lexical decision data sets are complemented with data from two word naming studies. First, the Traditional Chinese Psycholinguistic Database (TCPD) (Chang et al., 2016) contains naming data for 3,314 characters. Second, in a single-participant study, C. C. Sun (2016) provides naming data for 25,000 two-character words.

Unfortunately, the availability of large-scale data sets for simplified Chinese is limited. Consequently, two of the above-mentioned experimental data sets are based on traditional Chinese, rather than on simplified Chinese. Whereas the lexical decision latencies for one-character words in the CLP were collected from native speakers of simplified Chinese, the lexical decision latencies for two characters words were collected from native speakers of traditional Chinese in Hong Kong. Similarly, the TCPD collected naming latencies from native speakers of traditional Chinese in Taiwan.

The evaluation of frequency measures for simplified Chinese on experimental data obtained from speakers of traditional Chinese is less than optimal. Nonetheless, simplified versions of words and characters tend to retain their original semantic content. Consequently, frequency measures in traditional Chinese are a reasonable proxy for frequency measures in simplified Chinese, and vice versa. Notably, this is not the case for all lexical variables. Character complexities, for instance, differ tremendously between simplified and traditional Chinese.

Frequency effects often level off at high predictor values. To allow for non-linear frequency effects, we evaluated the predictive power of the character and word frequency measures from the different corpora with generalized additive models (henceforth GAMs; Hastie and Tibshirani, 1986). For each corpus, we fitted GAMs with smooths for character frequencies and word frequencies to each data set using version $1.8-16$ of the *mgcv* package (Wood, 2006) for R. The parameter k of the *gam()* function in the *mgcv* package was set to 3 for all smooths to limit the complexity of non-linear predictor effects.

Frequency counts for words that did not appear in a corpus were set to 0 to keep the sample size for each experimental data set constant across the different frequency measures. Words with different forms in simplified Chinese and traditional Chinese were excluded from the CLP and TCPD data sets. Following the recommendations of Box-Cox tests (Box and Cox, 1964), we applied inverse transformations to all dependent variables prior to analysis. For each data set, reaction times further than 2.5 standard deviations from the reaction time mean were removed prior to analysis. We used the deviance explained of the GAMs as a measure of the predictive power of the character and word frequency measures for an experimental data set.

The deviance explained by the GAMs for each corpus-data set combination are presented in Table 2. The average deviance explained (ADE) by the frequency measures from a given corpus across all seven data sets is presented in the last column of Table 2. The performance of the frequency measures from Chinese Gigaword (average deviance explained (ADE): 32.63%)

Table 2 Deviance explained by word and character frequency measures from SUBTLEX-CH (Cai and Brysbaert, 2010), Leiden Weibo Corpus (Van Esch, 2012), Public Weixin Corpus (L. Sun, 2016), Chinese Gigaword (Graff and Chen, 2003), SCCoW (Shaoul et al., 2016), and LCMC (McEnery and Xiao, 2008) for seven experimental data sets (CLP one-character words (Sze et al., 2014), CLP two-character words (Tse et al., 2016), SUBTLEX-CH (Cai and Brysbaert, 2010), MELD-SCH one-character words (Tsang et al., 2017), MELD-SCH two-character words (Tsang et al., 2017), TCPD (Chang et al., 2016), and (C. C. Sun, 2016)).

	CLP 1	CLP 2	SUBTLEX-CH	MELD 1	MELD 2	TCPD	Sun (2016)	mean
n	2,444	9,901	388	1,000	9,792	3,246	24,186	-
SUBTLEX-CH	46.93	36.32	25.93	63.19	41.71	37.23	18.69	38.57
Chinese Gigaword	45.78	25.33	17.20	59.80	28.56	34.01	17.74	32.63
SCCoW	48.33	26.90	19.97	63.09	33.31	34.00	18.25	34.83
LCMC	49.04	24.04	20.36	64.19	29.32	33.78	19.39	34.30
Leiden Weibo Corpus (LWC)	50.39	38.79	27.56	66.50	43.71	38.73	20.73	40.92
Public Weixin Corpus (PWC)	49.81	30.05	22.28	62.93	36.95	35.74	19.67	36.77
SUBTLEX-CH + LWC	50.90	42.22	32.60	66.86	48.25	39.58	20.93	43.05

and SCCoW (ADE: 34.83%) is less than convincing. Presumably, this is a result of the fact that the content in these corpora poorly reflects the reading experience of the average participant in the experimental studies. Chinese Gigaword consists solely of newswire texts, whereas a closer inspection of the SCCoW suggests that bureaucratic texts are substantially overrepresented in this corpus. The frequency measures from the LCMC provided limited explanatory power as well (ADE: 34.30%), presumably due to the limited size of this corpus (1 million words).

The most predictive frequency measures were derived from the corpora that best reflect the experience of modern-day speakers of simplified Chinese: the PWC (i.e., a corpus of chat messages, ADE: 36.77%), and SUBTLEX-CH (i.e., a corpus of film subtitles, ADE: 38.57%) both yielded frequency measures with competitive performance. The most predictive frequency measures, however, were obtained from the microblogging messages in the LWC. The LWC frequency counts proved most predictive for each of the seven experimental data sets and yielded an average deviance explained of 40.92%.

The performance of the character and word frequencies from the LWC indicates that messages on the social networking site Weibo represent an important part of the linguistic experience of contemporary speakers of Chinese. Messages on a social networking site, however, constitute a highly specific genre that does not cover the linguistic input of average language users in its entirety. It could be the case, therefore, that frequency measures from a combination of corpora would outperform the LWC frequency measures. To investigate this possibility, we calculated frequency counts for all 63 possible combinations of the six corpora ($2^6 - 1$; not all corpora can be excluded) and re-ran the above-mentioned GAMs.

Indeed, an improvement upon the performance of the frequency counts from the LWC turned out to be possible. Excellent predictive performance for the seven data sets under investigation was achieved by frequency counts based on both the LWC and SUBTLEX-CH. As can be seen in Table 2, this combined frequency measure increased the average deviance explained from 40.92% to 43.05%. For each individual data set, frequency counts based on both the LWC and SUBTLEX-CH outperformed frequency counts that were solely based on the LWC. Together, social networking messages and the film subtitles thus seem to provide a good representation of the language experience of the average modern-day speaker of simplified Chinese. We therefore decided to use summed frequency counts from the LWC and SUBTLEX-CH in the CLD.

2.2.2 Orthographic frequency

The orthographic frequency measures in the CLD are listed in Table 1. Frequency measures are included at different grain sizes, ranging from the word level to the radical level. Below, we provide a brief conceptual description of each frequency measure.

The unit in which frequency measures in the CLD are provided is frequency per million. For convenience, raw frequency counts are provided at the word and the character level as well. At the word level, the main frequency measure is *Frequency*: the frequency of a word per million words in a meta-corpus that contains the LWC and SUBTLEX-CH. Separate frequencies for the LWC and SUBTLEX-CH are provided as well. Word frequency effects for both one-character words and two-character words have been observed in a large number of studies and across experimental tasks. Both one-character words (Seidenberg, 1985a; Liu et al., 2007; C. C. Sun, 2016) and two-character words (Liu, 1999; C. C. Sun, 2016), for instance, are named faster when they

are more frequent. In addition, lexical decision latencies are shorter for more frequent one-character (Lee et al., 2015; Sze et al., 2014) and two-character words (Zhang and Peng, 1992; Peng et al., 1999).

At the character level, *C1-C4 Frequency* denote the frequency of the first, the second, the third and the fourth character per million characters. Again, separate frequency counts for the LWC and SUBTLEX-CH are provided as well. As was the case for word frequency measures, effects of character frequency measures are abundant in the literature. Zhang and Peng (1992), Taft et al. (1994), and Peng et al. (1999) all reported character frequency effects for lexical decision. Character frequency effects have been observed in other measures of language processing as well. Yan et al. (2006), for instance, found an effect of character frequency on eye fixation durations, whereas Kuo et al. (2003) and Lee et al. (2004) observed character frequency effects in fMRI studies.

At the character-level, we furthermore included measures of family size and family frequency (see e.g., Schreuder and Baayen, 1997): *C1-C4 Family Size* and *C1-C4 Family Frequency*. Family size is defined as the number of words a character occurs in. Family frequency is the summed frequency of all words in which a character occurs. Liu et al. (2007) referred to family size as "number of word formations" and found that characters with greater family sizes were named faster than characters with smaller family sizes (cf., Baayen et al., 2006; Hendrix, 2016, for similar findings in English). By contrast, Huang et al. (2006) observed inhibitory family frequency effects for Mandarin Chinese, with longer naming latencies for characters with higher family frequencies. The exact nature of family size effects in Mandarin Chinese, therefore, remains unclear.

In addition, we included a number of frequency measures below the character level. For both the phonetic radical and the semantic radical, frequency counts (*C1-C4 SR Frequency*, *C1-C4 PR Frequency*), as well as family size measures (*C1-C4 SR Family Size*, *C1-C4 PR Family Size*) are provided. Radical frequency is defined as the summed frequency of all characters in which a radical occurs. Radical family size refers to the number of characters in which a radical occurs. C. C. Sun (2016) observed an inhibitory effect of semantic radical frequency in a word naming task. Conversely, a number of studies found facilitatory effects of both phonetic radical family size (Feldman and Siok, 1997; Taft and Zhu, 1997; Lee et al., 2015) and semantic radical family size (Feldman and Siok, 1997, 1999a,b) in lexical decision. Effects of lexical properties of the phonetic and semantic radical have been interpreted as evidence for compositional processing at the character level. Taft (2006), for instance, proposed a reading model of Chinese in which access to visual components precedes access to characters (see also Taft et al., 1999; Taft and Zhu, 1997). As noted by Feldman and Siok (1999a), this view stands in contrast to theories that assume that the character is the "primary unit of visual recognition (e.g., Cheng, 1981; Hoosain, 1991; Liu, 1988)".

2.2.3 Phonological frequency

As was the case for orthographic frequency, the CLD contains measures of phonological frequency at the word level, at the character level and below the character level. *Phonological Frequency* denotes the frequency of the phonological form of the word, whereas *C1-C4 Phonological Frequency* provide the frequency of the phonological form of the first character, the second character, the third character and the fourth character, respectively.

Below the character level, the CLD contains a number of phonological frequency measures at the phoneme level and at the diphone level. We supply average frequencies, as well as minimum and maximum frequencies for the phonemes and diphones in each word and each character. These measures of phonological frequency are encoded in the following lexical variables: *Mean Phoneme Frequency*, *C1-C4 Mean Phoneme Frequency*, *Min Phoneme Frequency*, *C1-C4 Min Phoneme Frequency*, *Max Phoneme Frequency*, *C1-C4 Max Phoneme Frequency*, *Mean Diphone Frequency*, *C1-C4 Mean Diphone Frequency*, *Min Diphone Frequency*, *C1-C4 Min Diphone Frequency*, *Max Diphone Frequency*, and *C1-C4 Max Diphone Frequency*. Furthermore, we provide the frequency of the initial phoneme and the initial diphone of each character (*C1-C4 Initial Phoneme Frequency*, *C1-C4 Initial Diphone Frequency*).

A final set of phonological frequency measures below the character level encodes the frequency of the diphones that connect the syllables in multi-character words. *Transitional Diphone 1 Frequency* is the frequency of the diphone that connects the first and second syllables, *Transitional Diphone 2 Frequency* is the frequency of the diphone that connects the second and third syllables, and *Transitional Diphone 3 Frequency* is the frequency of the diphone that connects that third and fourth syllable. Together, the phonological frequency measures at the word level, at the character level, and below the character level provide a comprehensive quantification of the frequency of phonological forms for one-character, two-character, three-character, and four-character words in simplified Chinese.

2.3 Complexity measures

The visual complexity of words and characters is a further conceptual construct that predicts lexical processing in Mandarin Chinese. The most basic measure of visual complexity for Chinese is word length in characters (*Length*). At smaller grain sizes, however, more refined measures of visual complexity exist as well. Characters with a greater number of strokes, for instance, have been shown to yield longer reaction times in both lexical decision (Lee et al., 2015) and word naming tasks (Liu et al., 2007; Leong et al., 1987). The CLD therefore provides the number of strokes in the word as a whole (*Strokes*), as well as the number of strokes in each character (*C1-C4 Strokes*). Furthermore, it provides stroke counts for the semantic radical (*C1-C4 SR Strokes*) and the phonetic radical (*C1-C4 PR Strokes*) of the characters in a word. All stroke counts in the CLD are based on information provided by the Chinese Character Dictionary on the Mandarin Tools website (Peterson, 2005).

As noted above, stroke counts have been shown to predict behavioral measures of language processing across a number of tasks. However, alternative measures of visual complexity at the character level can be constructed as well. In the CLD, we included two alternative measures of character complexity: pixel counts and picture size. To calculate these measures, we generated PNG image files for each of the 4,895 unique characters in the CLD (font: SimHei, font size: 80, font color: black, image background: white). We define pixel counts as the number of non-white pixels in these image files (*C1-C4 Pixels*), and picture size as the size of the image files in bytes (*C1-C4 Picture Size*). It will be interesting to see how the predictive power of these alternative measures of visual complexity compares to the predictive power of stroke counts.

Thus far, we discussed measures of the *visual* complexity of a word and its characters. As was the case for frequency, however, complexity can be a *phonological* property of a character or a word as well. In addition to the above-mentioned visual complexity measures, we therefore included a measure of phonological complexity in the CLD. For both words and characters, we provide phoneme counts. The final five complexity measures in the CLD, therefore, are *Phonemes* and *C1-C4 Phonemes*.

2.4 Neighborhood density measures

The fourth group of measures in the CLD consists of neighborhood density measures. In English, orthographic neighborhood density influences reaction times in both lexical decision (see e.g., Andrews, 1989; Forster and Shen, 1996; Grainger, 1992) and word naming (see e.g., Andrews, 1992, 1997; Grainger, 1990; Coltheart et al., 1988). Due to the nature of the writing system, however, it is not trivial to calculate orthographic neighbors for Mandarin Chinese. Calculating orthographic neighbors on the basis of shared letters, for instance, is not possible. Nonetheless, we included a measure of orthographic neighborhood density in the CLD for each character: *C1-C4 OLD Pixels*. OLD stands for orthographic Levenshtein distance and refers to the average distance between a character and its n closest neighbors. Following Yarkoni et al. (2008) and the results of an exploration of the predictive power of the *OLD Pixels* measure across different values of n, we set n to 20 for the *OLD Pixels* measures in the CLD.

The distance between characters was calculated on the basis of the PNG image files for the characters that we mentioned above. For all characters, we defined each pixel as either white or non-white. For a given character, we then calculated the distance between that character and all other characters. The distance between two characters was defined as the number of pixels with a different status (i.e., white for one character and non-white for the other character). Neighbors of a character, then, are characters for which a limited number of pixels has a different status. The orthographic Levenshtein distance for a character is the average number of pixels that differ between that character and its 20 closest neighbors.

Phonological neighborhood density has been shown to influence lexical processing as well. Lexical decision latencies in English, for instance, are shorter for words with more phonological neighbors (see e.g., Yates et al., 2004; Baayen et al., 2006), as are naming latencies (see e.g., Vitevich, 2002; Hendrix, 2016). Defining phonological neighborhood density measures for Mandarin Chinese is much less of a challenge than defining orthographic neighborhood density measures for Mandarin Chinese. As is the case for alphabetic languages, shared phonemes are a solid basis for phonological neighborhood measures. In Chinese, however, phonological forms can differ from each other not only at the segmental level, but also at the suprasegmental level. Pronunciations may differ with respect to their constituent phonemes, but also with respect to tone. Although we acknowledge that segmental and suprasegmental parts of a pronunciation encode fundamentally different types of information, we decided to treat differences in phonemes and differences in tone in the same manner when calculating neighborhood density measures (i.e., both a different phoneme and a different tone result in an increase in distance of 1). We would

be more than willing to provide more refined measures, however, if future research indicates that better measures can be obtained by distinguishing segmental and suprasegmental information when calculating phonological neighborhood density measures.

Two types of phonological neighborhood density measures are included in the CLD. First, we calculated Coltheart's N (Coltheart et al., 1977) at the word level, as well as at the character level. The variables *Phonological N*, and *C1-C4 Phonological N* give the number of words or characters that differ from the target word or character by one phoneme or by one tone. Second, we calculated phonological Levenshtein distances. Analogous to the orthographic Levenshtein distance measure described above, the phonological Levenshtein distance measures *PLD* and *C1-C4 PLD* provide the average distance between the pronunciation of a word or character and its 20 closest neighbors.

2.5 Orthography-phonology consistency

As noted above, Mandarin Chinese has an inventory of about 8,100 characters. These characters are mapped onto a limited set of phonological forms. According to estimations by DeFrancis (1984) (as cited by Chen and Dell, 2006), there are about 1,200 unique syllables when tone is taken into consideration. When tone is ignored, this number is reduced to about 400. A large number of orthographic units thus is mapped onto a relatively small number of phonological forms. As a result, the mapping between orthography and phonology in Mandarin Chinese is less than consistent.

The mapping between orthography and phonology can be inconsistent in both directories. Homography describes the phenomenon of multiple pronunciations being mapped onto the same orthographic unit (i.e., character). Conversely, homophony occurs when the same pronunciation is shared by multiple characters. Thus far, psycholinguistic research on Chinese has primarily focused on homophony. Lee et al. (2015) and Wang et al. (2012) found inhibitory effects of homophony in visual and auditory lexical decision. By contrast, other studies found facilitatory effects of homophony in word naming (Ziegler et al., 2000) and in auditory word recognition (Chen et al., 2009, 2016).

Despite the fact that experimental research has primarily focused on homophony, we included predictors describing the consistency of both the phonology-to-orthography mapping and the orthography-to-phonology mapping. Each character in a word has a specific pronunciation. For a given character with a given pronunciation, a friend is defined as an occurrence of the same character-pronunciation mapping in a different word. For each character, we provide friend counts (*C1-C4 Friends*), as well as friend frequencies (*C1-C4 Friends Frequency*). Friends frequency is defined as the summed frequency of all friends.

For a given character, a homograph is defined as an occurrence of the same character with a different pronunciation. For each character, we provide homograph type (*C1-C4 Homograph Types*) and token (*C1-C4 Homograph Tokens*, i.e., the number of words in which the same character occurs with a different pronunciation) counts, as well as homograph frequencies (i.e., the summed frequency of the homograph tokens; *C1-C4 Homographs Frequency*). A homophone for a given character is defined as an occurrence of a different character with the same pronunciation. As was the case for homographs, we provide homophone type (*C1-C4 Homophone Types*) and token counts (i.e., the number of words in which a different character has the same pronunciation; *C1-C4 Homophone Tokens*), as well as homophone frequencies (i.e., the summed frequency of the homophone tokens; *C1-C4 Homophones Frequency*) for each character.

In our discussion of the categorical variables above, we mentioned that the consistency of the orthography-phonology mapping below the character level has received considerable attention from researchers as well. In particular, the reliability of the information provided by the phonetic radical has been shown to influence lexical processing (see e.g., Seidenberg, 1985a; Liu et al., 2007; Hue, 1992). We therefore decided to include lexical variables analogous to the consistency measures at the character level discussed above for the phonetic radical. Phonetic radical friends are defined as occurrences of the same phonetic radical in a character with the same pronunciation. Phonetic radical enemies, by contrast, are occurrences of the same phonetic radical in a character with a different pronunciation. Finally, phonetic radical backward enemies are occurrences of a different phonetic radical in a character with the same pronunciation.

The calculation of type counts, token counts and frequency measures for phonetic radical friends, phonetic radical enemies and phonetic radical backward enemies resulted in the following orthography-phonology consistency measures: *C1-C4 PR Friends*, *C1-C4 PR Friends Frequency*, *C1-C4 PR Enemies Types*, *C1-C4 PR Enemies Tokens*, *C1-C4 PR Enemies Frequency*, *C1-C4 PR Backward Enemies Types*, *C1-C4 PR Backward Enemies Tokens*, and *C1-C4 PR Backward Enemies Frequency*. Together, these variables provide in-depth information about the reliability of the information provided by the phonetic radical.

2.6 Information-theoretic measures

The last group of variables in the CLD consists of information-theoretic measures. Information theory originates from the seminal work of Claude Shannon (Shannon, 1948) and concerns the study of quantitative properties of communication systems. Key concepts in information theory are uncertainty and information, which are both considered measurable physical quantities. From an information theoretic perspective, language processing is, at its core, a process of uncertainty reduction. Consider, as an example, the recognition of words in an auditory lexical decision experiment. Prior to the onset of the word, the participant is uncertain as to which word she will hear. Once auditory input starts coming in, this uncertainty gradually decreases. At the end of the auditory input (or earlier), the uncertainty is typically reduced to such an extent that the participant is able to provide a correct lexicality judgement. Information-theoretic measures encode the amount of uncertainty in the signal (e.g., entropy, relative entropy) or, conversely, the extent to which uncertainty is reduced by the signal (e.g., conditional probability, association measures). The extent to which uncertainty is reduced by the signal is also referred to as *information*.

Information-theoretic properties of characters and words have received relatively little attention in the experimental literature for Mandarin Chinese. For English, however, information-theoretic measures have been shown to be highly predictive of behavioral measures of lexical processing. For compound processing, for instance, Schmidtke et al. (2016) reported longer lexical decision latencies for high entropy words, whereas Kuperman et al. (2007) observed longer acoustic durations for these words. Similarly, Kuperman et al. (2008a) argued that conditional probabilities play an important role in compound processing. Furthermore, several studies have documented increased processing costs for stimuli with greater relative entropy (Milin et al., 2009a,b; Kuperman et al., 2010; Baayen et al., 2011; Hendrix et al., 2017), whereas association measures such as mutual information have been shown to influence acoustic durations at both the word level (Pluymaekers et al., 2005) and the segment level (Kuperman et al., 2008b). Recently, C. C. Sun (2016) reported effects of information-theoretic measures for Mandarin Chinese as well, with robust effects of entropy and relative entropy in both word naming and phrase reading. Based on these findings, we decided to include a set of information-theoretic measures in the CLD. Given the relative novelty of information-theoretic measures in the context of psycholinguistic research on Mandarin Chinese, we provide more in-depth descriptions of these measures below.

The lexical predictors described thus far were calculated across words of different lengths. The family size of a character, for instance, was defined as the total number of one-character, two-character, three-character, and four-character words that character occurred in. Explorations of the predictive power of information-theoretic measures, however, indicated that information-theoretic measures calculated within words of the same length provided somewhat more explanatory power for behavioral data as compared to information-theoretic measures calculated across word lengths. The information-theoretic measures described below were therefore calculated for words of the same length. The conditional probability of a two-character word given the first character, for instance, was calculated on the basis of all two-character words with the same first character, whereas the conditional probability of a three-character word given the first character was calculated on the basis of all three-character words with the same first character. All information-theoretic measures were calculated on the basis of the word frequencies provided in the CLD.

The first type of information-theoretic measure in the CLD is conditional probability. The conditional probability measures in the CLD encode the probability of the current word given the first character (*C1 Conditional Probability*; defined for two-character words, three-character words, and four-character words only), the first two characters (*C2 Conditional Probability*; defined for three-character words and four-character words only), and the first three characters (*C3 Conditional Probability*; defined for four-character words only). The first character of the two-character word 勒索 ("to extort", frequency: 3.80), for instance, is 勒. This character is the first character in two other two-character words: 勒令 ("to compel", frequency: 0.79) and 勒紧 ("to tighten", frequency: 0.46). *C1 Conditional Probability* for the word 勒索, therefore, is $\frac{3.80}{3.80+0.79+0.46} = 0.75$.

In addition to forward conditional probabilities, we included backward conditional probabilities. These backward conditional probabilities encode the probability of the first character (*C1 Backward Conditional Probability*; defined for two-character words, three-character words, and four-character words only), the probability of the first two characters (*C12 Backward Conditional Probability*; defined for three-character words and four-character words only), and the probability of the first three characters (*C123 Backward Conditional Probability*; defined for four-character words only) given the rest of the characters in a word. The

second character of the two-character word 勒索 ("to extort", frequency: 3.80), for instance, is the second character in 17 other two-character words. The summed frequency of these 17 words is 100.99. *C1 Backward Conditional Probability* for the word 勒索, therefore, is $\frac{3.80}{3.80+100.99} = 0.04$.

Conditional probability describes *predictability* at the word level. By contrast, the second type of information-theoretic predictor in the CLD, entropy, describes *uncertainty* at the word level. Higher values of entropy measures indicate greater uncertainty. For a given word, *C1 Entropy* (defined for two-character words, three-character words, and four-character words only) describes the uncertainty about a word given its first character. Numerically, it is defined as the entropy over the probability distribution for all words with the same first character (and the same number of characters) as the current word. As noted above, the first character of the word 勒索 ("to extort", frequency: 3.80), occurs as the first character in two other two-character words: 勒令 ("to compel", frequency: 0.79) and 勒紧 ("to tighten", frequency: 0.46). Converting the frequency counts for these words into probabilities results in a probability of 0.75 for 勒索, a probability of 0.16 for 勒令, and a probability of 0.09 for 勒紧. The entropy ($-\sum_{i=1}^{n} p_i * \log_2(p_i)$) over this probability distribution is 1.04. *C1 Entropy* for the word 勒索, therefore, is 1.04. Analogously, *C12 Entropy* (defined for three-character and four-character words only) and *C123 Entropy* (defined for four-character words only) encode the uncertainty about a word given its first two characters and first three characters.

As was the case for conditional probability, the CLD provides backward entropy measures as well. The backward entropy measures describe the uncertainty about the first character (*C1 Backward Entropy*; defined for two-character words, three-character words, and four-character words only), the uncertainty about the first two characters (*C12 Backward Entropy*; defined for three-character words and four-character words only), and the uncertainty about the first three characters (*C123 Backward Entropy*; defined for four-character words only) given the rest of the characters in a word. *C1 Backward Entropy* for two-character words, for instance, is defined as the entropy over the probability distribution for all two-character words with the same second character. The second character 翅 of the word 鸡翅 ("chicken wings", frequency: 12.56), for instance, occurs as the second character in two other two-character words: 鱼翅 ("fin", frequency: 1.88) and 展翅 ("to spread wings", frequency: 1.63). Converting the frequencies of these words into probabilities gives a probability of 0.78 for 鸡翅, a probability of 0.12 for 鱼翅, and a probability of 0.10 for 展翅. The entropy over this probability distribution is 0.98. *C1 Backward Entropy* for the word 鸡翅, therefore, is 0.98.

The third type of information-theoretic measure in the CLD is relative entropy, which describes the distance between probability distributions. The relative entropy of two probability distributions, also known as the Kullback-Leibler divergence between the distributions, is defined as:

$$\sum_{i=1}^{n} p_i * \log_2(\frac{p_i}{q_i}) \tag{1}$$

In psycholinguistic research, the relative entropy measure is typically used to describe the distance between the probability distribution of an inflectional or derivational paradigm for a given word and the probability distribution of that paradigm for all words in the language. Milin et al. (2009a), for instance, collected probability distributions for the inflectional paradigms of a large set of Serbian nouns (i.e., probability of nominative case, probability of genitive case, et cetera for each noun) and calculated the relative entropy for each noun on the basis of these probability distributions. They found that lexical decision latencies for nouns with higher values of relative entropy (i.e., nouns with atypical probability distributions) were longer as compared to lexical decision latencies for nouns with lower values of relative entropy (i.e., nouns with typical probability distributions).

In Mandarin Chinese, the relative entropy measure can be applied to the combinatorial properties of characters in two-character words. For *C1 Relative Entropy*, we defined the reference distribution q as the probability distribution of *second* characters across all two-character words. For a given word, p was defined as the probability distribution of second characters across all two-character words with the same first character. Analogously, for *C2 Relative Entropy* the reference distribution q is the probability distribution of *first* characters across all two-character words. For a given second character, p is the probability distribution of first characters across all two-character words with the same second character.

For further clarification, consider the fictive example for *C1 Relative Entropy* in Table 3. The lexicon for this example contains 6 characters that occur as a second character. To calculate *C1 Relative Entropy* for the character 天 ("sky"), we need two sets of frequencies. First, the frequencies of the 6 two-character words in which 天 is the first character are required. Second, we need the frequencies of the 6 second characters across all first characters. Converting both frequency distributions to probabilities yields the probability distribu-

Table 3 Relative entropy: fictive example for *C1 Relative Entropy* for the character 天. Abbreviations: C2 = Character 2; Freq. = Frequency.

Word	Freq.	p	C2	Freq.	q
天气 ("weather")	203	0.67	气	1034	0.31
天使 ("angel")	59	0.19	使	206	0.06
天才 ("genius")	22	0.07	才	175	0.05
天上 ("heaven")	13	0.04	上	1737	0.52
天际 ("skyline")	5	0.02	际	141	0.04
天职 ("duty")	2	0.01	职	52	0.02

tions p (the probability distribution of second characters for the first character 天) and q (the probability distribution of second characters for all first characters). To calculate *C1 Relative Entropy* for the character 天, p and q are entered into Equation 1. For our example, this yields a relative entropy of 1.14.

The more similar the probability distributions p and q, the smaller the relative entropy. A small value for relative entropy therefore indicates that a first or second character combines with second or first characters in a typical way, whereas a large value for relative entropy indicates that a first or second character combines with second or first characters in an atypical way. As such, relative entropy is a measure of the prototypicality of the way in which a character combines with other characters in two-character words. The frequency distribution across two-character words that a character with low relative entropy occurs in tend to be relatively flat. By contrast, the frequency distribution across the two-character words that a character with high relative entropy occurs in tend to be more spiky.

Extension of the relative entropy measures in the CLD to three-character words and four-character words is theoretically possible. *C1 Relative Entropy* for three-character words, for instance, would describe the typicality of the way in which the first character in a three-character word combines with second and third characters. The probability distribution of second and third characters across three-character words, however, is extremely sparse. The same problem applies, to an even larger extent, to four-character words. As a result of this sparseness, the informativeness of the relative entropy measure substantially decreases with word length. We therefore decided to define the relative entropy measures in the CLD for two-character words only.

A set of four association measures constitutes the fourth type of information-theoretic measure in the CLD. These measures describe the strength of the association between the characters in a word. Each measure is based on a comparison on the expected frequency of a multi-character word and its observed frequency. The expected frequency of a two-character word is defined as:

$$\frac{\text{C1 Frequency} * \text{C2 Frequency}}{\text{Total Frequency}} \tag{2}$$

where Total Frequency is the summed frequency of all 2-character words in the CLD. The expected frequency for three-character and four-character words is defined analogously.

For example, the observed frequency of the word 苹果 ("apple") is 107.66. To calculate the expected frequency of the word 苹果 ("apple"), the summed frequency of two-character words that contain *the character* 苹 (summed frequency: 107.66; 苹果 ("apple") is the only two-character word in the CLD that contains the character 苹), the summed frequency of two-character words that contain the character 果 (64 words; summed frequency: 1975.34), as well as the summed frequency of all two-character words in the CLD (34,233 words; summed frequency: 334, 227.90) are required. Plugging these numbers into Equation 2 yields an expected frequency of $\frac{107.66*1975.34}{334,227.90} = 0.64$ for the word 苹果 ("apple").

All *four* association measures are positive when the observed frequency is greater than the expected frequency and negative when the observed frequency is smaller than the expected frequency. The first two association measures, *Pointwise Mutual Information* (see Myers and Gong, 2002, for an application of mutual information in the context of language processing in Mandarin Chinese) and *Position-specific Pointwise Mutual Information*, are based on the (logged) ratio between observed and expected frequencies. *Pointwise Mutual Information* is defined as:

$$\log_2 \left(\frac{\text{observed frequency}}{\text{expected frequency}} \right) \tag{3}$$

The pointwise mutual information for the word 苹果 ("apple") therefore is $\log_2 \left(\frac{107.66}{0.64} \right) = 7.39$.

Likewise, Position-specific pointwise mutual information can be calculated using Equation 2 and Equation 3. However, for position-specific pointwise mutual information the character frequencies are position-specific. That is, instead of using the overall frequencies of both characters across all two-character words, the frequency of character one is defined as the frequency of character one in the first position in two-character words and the frequency of character two is defined as the frequency of character two in the second position of two-character words.

For example, for the word 苹果 ("apple"), the frequency of two-character words with 苹 as the first character (frequency: 107.66) is the same as the overall frequency of the character 苹 in two-character words. The

frequency of two-character words with the 果 as the second character (frequency: 1,788.38), however, is somewhat lower than the overall frequency of the character 果 in two-character words (frequency: 1975.34). For the word 苹果, the position-specific expected frequency thus is $\frac{107.66*1{,}788.38}{334{,}227.90} = 0.58$. This results in a position-specific pointwise mutual information of $\log_2\left(\frac{107.66}{0.58}\right) = 7.54$.

As pointed out by Gries (2010), "pointwise MI is known to return very high association scores for low-frequency words [here: characters] as well as for technical terms or other expressions that exhibit very little or no variation. On the other hand, the *t*-score returns high association scores to word pairs [here: words] with high co-occurrence frequencies and provides a better measure of the non-randomness of the co-occurrence" (p. 14; cf. Evert, 2009). The *t*-score measure is defined as:

$$\frac{\text{observed frequency} - \text{expected frequency}}{\sqrt{\text{expected frequency}}} \tag{4}$$

As before, we included both general *t*-score and position-specific *t*-score as lexical variables in the CLD. For the word 苹果 ("apple"), *T-Score* is $\frac{107.66-0.64}{\sqrt{0.64}} = 133.78$, whereas *Position-specific T-Score* is $\frac{107.66-0.58}{\sqrt{0.58}} = 140.60$.

The final information-theoretic measure in the CLD is *Entropy Character Frequencies*. *Entropy Character Frequencies* is the entropy over the probability distribution of both characters in a two-character word. For the word 鲨鱼 ("shark"), for instance, the frequency of the first character 鲨 is 9.86 and the frequency of the second character 鱼 is 304.50. Converting these frequencies to probabilities gives a probability of 0.03 for the first character and 0.97 for the second character. The entropy over the character frequencies for the word 鲨鱼, therefore, is $-\sum_{i=1}^{n} p_i * \log_2(p_i) = 0.20$. Values of *Entropy Character Frequencies* are higher when the frequencies of the characters in a two-character word are more similar.

3 Evaluation

Above, we introduced the categorical and numerical variables provided by the CLD. In this section, we evaluate the explanatory power of these variables in the context of lexical decision and word naming data. As noted in our discussion of frequency measures above, the availability of large-scale data sets for simplified Chinese is limited. Large-scale multi-participant word naming data for multi-character words, for instance, do not exist. Similarly, the available of lexical decision data for three-character and four-character words is limited. Compromises, therefore, were necessary when selecting the data sets used for the evaluation of the lexical predictors in the CLD.

For lexical decision, we opted to use the data for one-character and two-character words in the Chinese Lexicon Project (Sze et al., 2014; Tse et al., 2016), as well as the data for one-character, two-character, three-character and four-character words in MELD-SCH (Tsang et al., 2017). As noted above, the lexical decision latencies for one-character words in the CLP were obtained in simplified Chinese, whereas the lexical decision latencies for two-character words were obtained in traditional Chinese. For two-character words, we restricted the evaluation of the CLD measures to the subset of the two-character words in the CLP for which the written form is identical in simplified and traditional Chinese. This ensures that visual complexity measures in the CLD are appropriate for the data under investigation.

The lexical decision data from MELD-SCH are not entirely problem-free either. The primary concern with respect to the MELD-SCH data is statistical power. Whereas lexical decision latencies are provided for no less than 10,022 two-character words, the number of words for which data were collected is substantially lower for one-character words ($n = 1{,}020$), three-character words ($n = 949$), and four-character words ($n = 587$). As a result, it will be be difficult to detect subtle predictor effects in the MELD-SCH data for one-character, three-character and four-character words. Nonetheless, the MELD-SCH data may provide valuable information about the most prominent predictor effects for three-character and four-character words. Despite the problems mentioned above, we therefore expect the evaluation of the explanatory power of the lexical variables in the CLD for the lexical decision latencies in the CLP and in MELD-SCH to provide a reasonable overview of the types of effects one could expect in the lexical decision task in simplified Chinese.

For word naming, we use the word naming data for 4,710 one-character and 25,935 two-character words provided by C. C. Sun (2016). For these word naming data, C. C. Sun (2016) reported the standard effects of frequency and visual complexity measures, as well as effects of the information-theoretic measures entropy and relative entropy. The word naming data in C. C. Sun (2016) were obtained from a single participant. C. C. Sun (2016) demonstrated, however, that qualitatively and quantitatively similar predictor effects were obtained for a second participant (see also Pham and Baayen, 2015). The word naming study in C. C. Sun (2016) was carried out in simplified Chinese. The partic-

ipant was a highly educated 30-year old native speaker of Mandarin Chinese from mainland China.

Henceforth, we will refer to the above-mentioned data sets as LD1 CLP (lexical decision for one-character words from the CLP, data: Sze et al., 2014), LD1 MELD-SCH (lexical decision for one-character words from MELD-SCH, data: Tsang et al., 2017), LD2 CLP (lexical decision for two-character words, from the CLP, data: Tse et al., 2016), LD2 MELD-SCH (lexical decision for two-character words, from MELD-SCH, data: Tsang et al., 2017), LD3 MELD-SCH (lexical decision for three-character words, from MELD-SCH, data: Tsang et al., 2017), LD4 MELD-SCH (lexical decision for four-character words, from MELD-SCH, data: Tsang et al., 2017), NAM1 (naming data for one-character words, data: C. C. Sun, 2016), and NAM2 (naming data for two-character words, data: C. C. Sun, 2016). Below, we first describe the methodology for the statistical analysis of these four data sets. Next, we discuss the results for each type of data set: lexical decision for one-character words (LD1 CLP, LD1 MELD-SCH), lexical decision for two-character words (LD2 CLP, LD2 MELD-SCH), lexical decision for three-character words(LD3 MELD-SCH), lexical decision for four-character words (LD4 MELD-SCH), word naming for one-character words (NAM1), and word naming for two-character words (NAM2).

3.1 Analysis

We fit linear regression models to the lexical decision and word naming data sets using version $1.8-23$ of the *mgcv* package (Wood, 2006, 2011) for the statistical software package R. Although the generalized-additive models (henceforth GAMs, Hastie and Tibshirani, 1986) provided by the *mgcv* package allow for non-linear predictor effects through the use of smooths, we imposed linearity on all predictor effects for simplicity. All predictor effects were therefore modeled through parametric terms, with the exception of the effects of the multi-level categorical predictors *Initial Phoneme* and *Final Phoneme*, which were modelled through random effect smooths.

Following the recommendation of Box-Cox tests (Box and Cox, 1964), we applied inverse transformations to all dependent variables prior to analysis ($f(x) = \frac{-1000}{x}$). To increase the uniformity of the predictor distributions, power transformations were applied to predictors as well. Based on the distributional properties of a predictor, one of the following transformations was applied: $f(x) = \frac{1}{x^2}$, $f(x) = \frac{1}{x}$, $f(x) = \log(x)$, $f(x) = \sqrt{x}$, $f(x) = x$ (identity transformation), or $f(x) = x^2$.

We removed outliers further than 3 standard deviations from the mean for each dependent variable. To prevent further data loss, we did not remove outliers for predictors prior to analysis. For the reported models, however, we verified that all reported predictor effects were quantitatively and qualitatively similar when predictor outliers were removed from the model.

The reported models were constructed using forward selection, using an α-level of 0.0001. Due to the fact that predictors in the CLD often describe closely related concepts, the data sets under investigations suffer from extreme collinearity. The simultaneous inclusion of highly correlated predictors in a regression model can lead to misinformed conclusions about the qualitative and quantitative nature of predictor effects (see e.g., Friedman and Wall, 2005; Wurm and Fisicaro, 2014). We verified that all reported predictor effects are robust through post-hoc analyses based on a principal components analysis with varimax rotation. The principal components analysis was carried out using version 1.6.9 of the *psych* package for R (Revelle, 2016). Unless indicated otherwise, the effects of the principal components corresponding to lexical predictors were quantitatively and qualitatively similar to the predictor effects reported below.

3.2 Lexical decision: one-character words

The results for the linear regression analysis of the lexical decision latencies for one-character words in the CLP and in MELD-SCH are presented in Table 4 and Table 5, respectively. For each significant predictor effect, we report estimated β-coefficients, standard errors and t-values. All corresponding p-values are < 0.0001. The model fit to the lexical decision latencies in the CLP explained 44.36% of the variance for $2,223$ one-character words. The data for the one-character "words" MELD-SCH contain a substantial number of characters that do not occur as independent words in simplified Chinese (e.g., "亥", "迂", "沱"). These words were not included as one-character words in the CLD, and were therefore removed prior to analysis. The regression model fit to the lexical decision latencies for the remaining 777 one-character words explained 61.51% of the variance.

Character frequency (*C1 Frequency*, CLP: $t = -10.530$; MELD-SCH: $t = -4.530$), as well the frequency of the character as an independent word (*Frequency*, CLP: $t = -4.675$; MELD-SCH: $t = -7.850$) showed significant effects on the lexical decision latencies for one-character words from both the CLP and MELD-SCH. More frequent characters that occurred more often as independent words were responded to

Table 4 Results for a linear regression model fit to the (inverse transformed) lexical decision latencies ($f(x) = -1000/x$) for single character words in Sze et al. (2014). Adjusted R^2 of the model: 0.441, deviance explained: 44.36%. Provided are β coefficients, standard errors (S.E.), and t-values. All corresponding p-values are < 0.0001.

parametric terms	β	S.E.	t-value
Intercept	-1.522	0.040	-38.016
C1 Frequency	-0.033	0.003	-10.530
Frequency	-0.009	0.002	-4.675
C1 Strokes	0.038	0.005	7.060
C1 Friends	-0.025	0.004	-6.715
C1 SR Family Size	0.013	0.002	5.499
C1 Tone: 1	-0.143	0.034	-4.202
C1 Tone: 2	-0.147	0.034	-4.324
C1 Tone: 3	-0.152	0.034	-4.446
C1 Tone: 4	-0.143	0.034	-4.235

faster. The facilitatory effects of character frequency (Zhang and Peng, 1992; Taft et al., 1994; Peng et al., 1999) and word frequency (see e.g., Lee et al., 2015; Sze et al., 2014) are qualitatively similar to frequency effects reported in previous lexical decision studies

We furthermore found an effect of visual complexity. In line with previous findings (see e.g., Lee et al., 2015), lexical decision latencies in the CLP were longer for words with more strokes (*C1 Strokes*, $t = 7.060$). For the MELD-SCH data, however, we did not find an effect of stroke count (*C1 Strokes*, $t = 0.923$, $p = 0.356$). A potential explanation for this discrepancy comes from the fact non-words were constructed in a different way for the CLP and MELD-SCH. The non-words in the CLP were created by replacing the semantic radical in a character with a different semantic radical. By contrast, the non-words for one-character words in MELD-SCH were constructed either through non-existing combinations of radicals, or through the addition or deletion of strokes in real characters. The different nature of the non-words in the CLP and MELD-SCH could lead to different processing strategies. It is not immediately clear, however, why the effect of stroke count would be present for the type of non-words in the CLP, but not for the type of non-words in MELD-SCH.

Alternatively, the absence of a visual complexity effect for one-character words in the lexical decision data from MELD-SCH could be due to reduced statistical power. The analysis for the one-character words in the CLP is based on $2,223$ words, whereas the analysis for the one-character words in MELD-SCH is based on 777 words only. The size of the set of one-characters words in MELD-SCH, therefore, may be insufficient to observe an effect of stroke count.

In addition to the effects of frequency and visual complexity, we found three further predictor

Table 5 Results for a linear regression model fit to the (inverse transformed) lexical decision latencies ($f(x) = -1000/x$) for one-character words in Tsang et al. (2017). Adjusted R^2 of the model: 0.614, deviance explained: 61.51%. Provided are β coefficients, standard errors (S.E.), and t-values. All corresponding p-values are < 0.0001.

parametric terms	β	S.E.	t-value
Intercept	-1.103	0.010	-107.202
C1 Frequency	-0.019	0.004	-4.530
Frequency	-0.021	0.003	-7.850
C1 Friends	-0.039	0.005	-7.877

effects. First, we found an effect of the consistency of the orthography-to-phonology mapping. The greater the number of words with the same character-pronunciation mapping (*C1 Friends*), the shorter the lexical decision latencies in both the CLP ($t = -6.715$) and MELD-SCH ($t = -7.877$). This effect of orthography-to-phonology consistency is in line with previous findings in the experimental literature for Mandarin Chinese. Lee et al. (2015) and Wang et al. (2012), for instance, found that characters with more pronunciations (i.e., higher homophone counts) were responded to slower in visual and auditory lexical decision, respectively.

Second, lexical decision latencies increased as a function of the number of characters in which a semantic radical occurs. This effect of semantic radical family size was limited to the CLP (*C1 SR Family Size*, CLP: $t = 5.933$; MELD-SCH: $t = 0.285$, $p = 0.775$). The inhibitory effect of semantic radical family size for in the CLP is opposite to the facilitatory effects of semantic radical family size observed by Feldman and Siok (1997, 1999a,b). A potential explanation for this discrepancy comes from the different nature of the non-words in the CLP and the non-words in the Feldman and Siok studies.

The non-words in Feldman and Siok (1999a) (see also Feldman and Siok, 1997, 1999b) were constructed "[...] either by taking real characters and changing one or more strokes or by combining two components that did not co-occur". The use of two types of non-character formations avoided "[...] having participants focus either on strokes or on the appropriateness of a particular combination" (p. 565). By contrast, and as noted above, the non-words in the CLP were created through a single mechanism: the replacement of the semantic radical in a character with a different semantic radical (Sze et al., 2014).

By definition, all words contain a valid semantic radical. All non-words in the CLP contained a valid semantic radical as well. The presence of a valid semantic radical, therefore, did not provide information

about the lexical status of a stimulus. In the Feldman and Siok studies, however, not all non-words contained a valid semantic radical. The presence of a valid semantic radical therefore provided probabilistic information about the lexical status of a stimulus. To be precise, it increased the probability that the current stimulus was a word.

More frequent lexical items can be accessed faster than less frequent lexical items. Semantic radicals that occur in a large number of characters tend to be more frequent than semantic radicals that occur in fewer characters. Semantic radicals with large families, therefore, can be accessed faster than semantic radicals with smaller families. As such, the information that a valid semantic radical is present becomes available more rapidly when the family size of the semantic radical is large. As noted above, this information helped determine the lexical status of the stimuli in the Feldman and Siok studies. Consequently, participants were able to respond faster to characters that contain semantic radicals with a large family size.

By contrast, the presence of a valid semantic radical did not provide information about the lexical status of a stimulus in the lexical decision paradigm adopted by the CLP. Instead, the lexical status of a character critically depended on the legitimacy of the combination of the semantic radical and the rest of the character. To successfully complete the lexical decision task in the CLP it was sufficient to decide whether or not the semantic radical of a character was the correct semantic radical.

The principles of associative learning help provide an estimation of the difficulty of this decision. Consider the lexical decision task in the CLP in the context of an event in associative learning, with the semantic radical as the cue and the character as the outcome. Associative learning theory states that the strength of the association between a cue and an outcome is inversely proportional to the number of outcomes the cue occurs with (see Baayen et al., 2011). A semantic radical that occurs in a limited number of characters thus is strongly associated with each of these characters. By contrast, a semantic radical that occurs in a large number of characters is less strongly associated with each of these characters.

To decide whether or not the semantic radical of a character is the correct semantic radical, it is necessary to verify that the semantic radical is a valid cue for the character. The stronger the association between the semantic radical and the character, the less time this verification takes. As such, it is easier to establish the correctness of a semantic radical when the semantic radical occurs in fewer characters. Consequently, participants in the CLP were able to respond faster to characters that contain semantic radicals with a limited family size. The nature of the semantic radical family size effect in lexical decision studies, therefore, may to some extent depend on the properties of the non-word stimuli in the experiment.

Third, we observed an effect of tone. As was the case for the effects of visual complexity and semantic radical family size, this effect was present in the CLP, but not in MELD-SCH, presumably due to differences in statistical power (the qualitative pattern of results in MELD-SCH was similar to the qualitative pattern of results in the CLP). Post-hoc pairwise comparisons indicated that lexical decision latencies were longer for tone 5 as compared to tones 1 through 4. No significant differences were found between tones 1 through 4. Tone 5 (0.76% of the one-character words in the CLP) is much less frequent than tone 1 (23.07%), tone 2 (22.67%), tone 3 (16.37%), and tone 4 (37.11%). In total, no more than 17 data points were available for tone 5. The effect of *C1 Tone*, therefore, may not be robust.

3.3 Lexical decision: two-character words

Table 6 shows the results for the lexical decision latencies for the two-character words in Tse et al. (2016), whereas Table 7 show the results for the two-character words in Tsang et al. (2017). After the removal of reaction time outliers and two-character words with distinct word forms in simplified Chinese and traditional Chinese, the data set from the CLP contains lexical decision latencies for $8,005$ two-character words. The model fit to these lexical decision latencies explained 39.84% of the variance in the data. The linear regression model for the MELD-SCH data explained 48.50% of the variance and was fit to the lexical decision latencies for $9,763$ words.

As was the case for one-character words, we found significant effects of the frequency of both the first (*C1 Frequency*, CLP: $t = 5.949$; MELD-SCH: $t = 9.373$) and the second character (*C2 Frequency*, CLP: $t = 3.950$; MELD-SCH: $t = 7.940$), as well as of the frequency of the word as a whole (*Frequency*, CLP: $t = -63.411$; MELD-SCH: $t = -107.290$). Consistent with previous findings, the word frequency effect in both data sets is facilitatory in nature: more frequent words are responded to faster than less frequent words (Zhang and Peng, 1992; Peng et al., 1999). Contrary to expectation, however, the effects of the frequency of both the first and the second character are inhibitory rather than facilitatory in nature.

Table 6 Results for a linear regression model fit to the (inverse transformed) lexical decision latencies ($f(x) = -1000/x$) for two-character words in Tse et al. (2016). Adjusted R^2 of the model: 0.398, deviance explained: 39.84%. Provided are β coefficients, standard errors (S.E.), and t-values. All corresponding p-values are < 0.0001.

parametric terms	β	S.E.	t-value
Intercept	-1.468	0.006	-242.933
C1 Frequency	0.008	0.001	5.949
C2 Frequency	0.005	0.001	3.950
Frequency	-0.050	0.001	-63.411
C1 Entropy	-0.012	0.001	-7.980
C1 Backward Entropy	-0.010	0.001	-8.023
C1 RE	-0.008	0.001	-7.194
C2 RE	-0.007	0.001	-6.213

Table 7 Results for a linear regression model fit to the (inverse transformed) lexical decision latencies ($f(x) = -1000/x$) for two-character words in Tsang et al. (2017). Adjusted R^2 of the model: 0.484, deviance explained: 48.50%. Provided are β coefficients, standard errors (S.E.), and t-values.

parametric terms	β	S.E.	t-value
Intercept	-1.339	0.012	-107.290
C1 Frequency	0.011	0.001	9.373
C2 Frequency	0.010	0.001	7.940
Frequency	-0.061	0.001	-84.781
C1 Strokes	0.009	0.002	3.860[2]
C2 Strokes	0.010	0.002	4.257
C1 Entropy	-0.011	0.001	-8.771
C1 Backward Entropy	-0.009	0.001	-7.743
C1 RE	-0.005	0.001	-5.551
C2 RE	-0.006	0.001	-5.866

Given the well-established facilitatory nature of character frequency effects (see e.g., Zhang and Peng, 1992; Peng et al., 1999), we suspected that the inhibitory nature of the character frequency effects in our models might be an artifact of collinearity between the predictors in the CLD. Post-hoc analyses using principal components rather than raw variables as predictors (see above) confirmed this suspicion. The effects of the principal components corresponding to *C1 Frequency* (CLP: $t = -18.006$, loading *C1 Frequency* = 0.971; MELD-SCH: $t = -18.540$, loading *C1 Frequency* = 0.969) and *C2 Frequency* (CLP: $t = -16.739$, loading *C2 Frequency* = 0.971; MELD-SCH: $t = -14.840$, loading *C1 Frequency* = 0.971) were in the expected direction for both data sets, with lexical decision latencies being shorter for words that contain more frequent characters.

In addition to the effects of frequency, we observed an effect of visual complexity, albeit only for the lexical decision latencies in MELD-SCH. As expected, a greater number of strokes in the first character (*C1 Strokes*, MELD-SCH: $t = 3.860$; CLP: $t = -2.012$, $p = 0.04$), as well as a greater number of strokes in the second character (*C2 Strokes*, MELD-SCH: $t = 4.257$; CLP: $t = 1.187$, $p = 0.235$) resulted in longer reaction times. The presence of a visual complexity effect for two-character words in MELD-SCH suggests that the absence of such an effect in the MELD-SCH data for one-character words (see above) may be a result of insufficient statistical power. Whereas the data set for two-character words consisted of $8,005$ two-character words, the data set for one-character words contained no more than 777 words.

The absence of an effect of visual complexity for the CLP data stands in contrast to observations by Tse et al. (2016), who found significant effects of the number of strokes in both characters for the CLP lexical decision data for two-character words. It should be noted, however, that the amount of variance explained by a linear regression model including only stroke counts for the first and second character as independent variables in Tse et al. (2016) was no more than 1.5%. This observation confirms that the explanatory power of stroke counts for the lexical decision data in the CLP is limited and explains why we did not observe stroke count effects for the two-character words in the CLP at an α-level of 0.0001.

Furthermore, we observed robust effects of two information-theoretic measures: entropy (*C1 Entropy*, CLP: $t = -7.980$; MELD-SCH: $t = -8.771$ and *C2 Entropy*, CLP: $t = -8.023$; MELD-SCH: $t = -7.743$) and relative entropy (*C1 Relative Entropy*, CLP: $t = -7.194$; MELD-SCH: $t = -5.551$ and *C2 Relative Entropy*, CLP: $t = -6.213$; MELD-SCH: $t = -5.866$). Consistent with the findings of C. C. Sun (2016) for word naming, greater uncertainty about one character given the other character leads to shorter lexical decision latencies. Similarly, lexical decision latencies are shorter when the frequency distribution of one character given the other character is unlike the frequency distribution of the former in the language as a whole.

The inhibitory effects of entropy and relative entropy are in the opposite direction to the entropy effects typically observed in English (see e.g., Milin et al., 2009a,b; Kuperman et al., 2010; Baayen et al., 2011; Hendrix et al., 2017). This is not a result of collinearity: similar effects of entropy and relative entropy were

[2] The p-value for *C1 Strokes* in the analysis of the lexical decision latencies for two-character words in the Tsang et al. (2017) was exactly 0.0001. The effect of *C1 Strokes* was therefore technically not significant at an α-level of 0.0001. A posthoc principal components analysis, however, confirmed suggested the presence of an effect of the stroke count of the first character ($t = 5.189$, $p < 0.0001$). We therefore decided to nonetheless include the effect of *C1 Strokes* in the model for the two-character lexical decision latencies from Tsang et al. (2017).

observed in a post-hoc principal components regression. To us, it is not immediately clear why entropy effects in simplified Chinese seem to be facilitatory, while entropy effects in English are inhibitory. Further research that takes a closer look at the distributional properties of both languages may shed further light on this issue. The robust effects of entropy and relative entropy in the lexical decision latencies for both the CLP and MELD-SCH, however, do establish the necessity of taking the combinatorial properties into account when investigating lexical processing of two-character words in simplified Chinese.

3.4 Lexical decision: three-character words

The results for the linear regression analysis of the lexical decision latencies for three-character words in MELD-SCH are presented in Table 8. After removal of words that are not in the CLD and reaction time outliers, lexical decision latencies for 864 words remained. Despite the limited size of the data set under investigation, however, the effects of three predictors reached significance at an α-level of 0.0001. The model fit to the lexical decision latencies for three-character words explained 38.71% of the variance.

Consistent with the findings for one-character and two-character words reported above, we found a highly significant effect of word frequency (*Frequency*, $t = -21.881$). As expected, reaction times were shorter for more frequent words. For two-character words, we furthermore observed effects of both character frequencies. For the three-character words in MELD-SCH, however, a significant effect of character frequency was present for the final character only (*C3 Frequency*, $t = -4.734$). Words with more frequent final characters were responded to faster as compared to words with less frequent final characters.

The third predictor that significantly influenced lexical decision latencies for three-character words was the conditional probability of the third character given the first two characters (*C12 Conditional Probability*, $t = 7.152$). Surprisingly, lexical decision latencies were longer when the probability of the third character given the first two characters was higher. As a similar inhibitory effect of conditional probability was observed in a post-hoc principal components analysis, the unexpected direction of the conditional probability effect was not due to collinearity. As was the case for the facilitatory effects of entropy and relative entropy for two-character words, further research will be necessary for a better understanding of the inhibitory effect of conditional probability. Nonetheless, the effect of conditional probability once more highlights the potential of information-theoretic measures in the context of lexical processing in simplified Chinese.

3.5 Lexical decision: four-character words

Table 9 shows the results for the analysis of the lexical decision data for four-character words in MELD-SCH. As was the case for three-character words, the number of four-character words in MELD-SCH is limited. After removal of words that are not in the CLD and reaction time outliers, lexical decision latencies for 421 words remained. Nonetheless, we found significant effects of three predictors at an α-level of 0.0001. The model fit to the lexical decision latencies for four-character words explained 28.64% of the variance.

In line with our findings for three-character words, we found effects of two frequency measures. Reaction times were shorter for more frequent words (*Frequency*, $t = -10.490$, as well as for words with more frequent final characters (*C4 Frequency*, $t = -4.883$). In addition, we found an effect of the visual complexity of the initial character: the higher the stroke count for the first character, the longer the reaction times for four-character words (*C1 Strokes*, $t = 4.637$).

Due to the limited size of the data sets for three-character words and four-characters words, the current analyses lack the statistical power to provide detailed

Table 8 Results for a linear regression model fit to the (inverse transformed) lexical decision latencies ($f(x) = -1000/x$) for three-character words in Tsang et al. (2017). Adjusted R^2 of the model: 0.385, deviance explained: 38.71%. Provided are β coefficients, standard errors (S.E.), and t-values. All corresponding p-values are < 0.0001.

parametric terms	β	S.E.	t-value
Intercept	-1.342	0.021	-62.560
Frequency	-0.057	0.003	-21.881
C3 Frequency	-0.014	0.003	-4.734
C12 Conditional Probability	0.093	0.013	7.152

Table 9 Results for a linear regression model fit to the (inverse transformed) lexical decision latencies ($f(x) = -1000/x$) for four-character words in Tsang et al. (2017). Adjusted R^2 of the model: 0.281, deviance explained: 28.64%. Provided are β coefficients, standard errors (S.E.), and t-values. All corresponding p-values are < 0.0001.

parametric terms	β	S.E.	t-value
Intercept	-1.290	0.025	-51.190
Frequency	-0.044	0.004	-10.490
C4 Frequency	-0.014	0.003	-4.883
C1 Strokes	0.034	0.007	4.637

insights into the nature of lexical processing for three-character and four-character words in the lexical decision task. Nonetheless, the analyses reported here suggest that the frequency of multi-character words as a whole as well as the frequency of the final character in these words co-determine lexical decision latencies to a considerable extent. The current analyses paint a less clear picture of the role of information-theoretic measures. While we found an effect of conditional probability for three-character words, no information-theoretic measures reached significance for four-character words.

3.6 Word naming: one-character words

The results for a linear model fit to the naming latencies for one-character words in C. C. Sun (2016) are presented in Table 10. This linear model explains 36.36% of the variance in the naming latencies for 3,368 words. The model contains a significant random effect smooth for the initial phoneme of the word (*Initial Phoneme*, $F = 13.033$). This random effect smooth controls for the variance introduced by the different sensitivity of the response detection algorithm to different phonemes.

Furthermore, the analysis of the word naming data for one-character words revealed effects of three lexical variables. As was the case for the lexical decision latencies for one-character words, both the frequency of the character across all words (*C1 Frequency*, $t = -5.974$) and the frequency of the character as an independent word (*Frequency*, $t = -4.583$) reached significance. As was the case in previous word naming studies, more frequent words were named faster (Seidenberg, 1985a; Liu et al., 2007; C. C. Sun, 2016). In addition, the number of strokes in the character showed an effect consistent with previous research (see e.g., Liu et al., 2007; Leong et al., 1987). The greater the number of strokes, the slower the response (*C1 Strokes*, $t = 3.013$).

As was the case for the lexical decision latencies for one-character words, we furthermore observed an effect of *C1 Friends* ($t = -10.580$). Naming latencies for one-character words were shorter when the same orthography-to-phonology mapping occurred in a larger number of other (multi-character) words. Finally, a post-hoc analysis for the subset of one-character words that contained a phonetic radical ($n = 2,236$) revealed an effect of the regularity of the phonetic radical. Consistent with previous word naming studies (Liu et al., 2007; Seidenberg, 1985b; Hue, 1992), one-character words for which the pronunciation of the character was identical to the pronunciation of the phonetic radical were named faster as compared to one-character words for which the pronunciation of the character and the pronunciationof the phonetic radical were different (*C1 PR Regularity*, $t = -4.136$).

Table 10 Results for a linear regression model fit to the (inverse transformed) naming latencies ($f(x) = -1000/x$) for one-character words in C. C. Sun (2016). Adjusted R^2 of the model: 0.358, deviance explained: 36.36%. Provided are β coefficients, standard errors (S.E.) and t-values for parametric terms, and estimated degrees of freedom (edf), reference degrees of freedom (ref. df) and F-values for smooth terms.

smooth terms	edf	ref. df	F-value
Initial Phoneme (bs="re")	25.450	31.000	13.033
parametric terms	β	S.E.	t-value
Intercept	-1.770	0.035	-51.278
C1 Frequency	-0.029	0.005	-5.974
Frequency	-0.014	0.003	-4.583
C1 Strokes	0.025	0.008	3.013[3]
C1 Friends	-0.058	0.005	-10.580

3.7 Word naming: two-character words

Table 11 presents the results for the naming latencies for two-character words in C. C. Sun (2016). As was the case for one-character words, the model includes a highly significant random effect smooth for the initial phoneme (*Initial Phoneme*, $F = 98.750$). For 24,100 two-character words, the model explains 32.33% of the variance in the naming data.

As was the case for the lexical decision latencies for two-character words, we found significant effects of the frequency of both characters (*C1 Frequency*, $t = -16.864$; *C2 Frequency*, $t = -7.634$). As expected, naming latencies were shorter for words with more frequent characters. In line with previous findings by Liu (1999) and C. C. Sun (2016), we furthermore found a facilitatory effect of word frequency (*Frequency*, $t = -26.629$).

The lexical decision data from MELD-SCH for two-character words furthermore revealed effects of visual complexity. Similarly, we found effects of the number of strokes of both the first character (*C1 Strokes*, $t = 21.532$) and the second character (*C2 Strokes*, $t = 4.061$) for word naming. As was the case for the character frequency effects, the effect of visual complexity was stronger for the first character than for the second character.

[3] The p-value for *C1 Strokes* in the analysis of the naming latencies for one-character words in C. C. Sun (2016) was 0.0026 and therefore not significant at an α-level of 0.0001. A posthoc principal components analysis, however, showed a robust effect of stroke count ($t = 6.575$, $p < 0.0001$). We therefore decided to nonetheless include the effect of *C1 Strokes* in the model for the naming latencies for one-character words.

Table 11 Results for a linear regression model fit to the (inverse transformed) naming latencies ($f(x) = -1000/x$) for two-character words in C. C. Sun (2016). Adjusted R^2 of the model: 0.322, deviance explained: 32.33%. Provided are β coefficients, standard errors (S.E.) and t-values for parametric terms, and estimated degrees of freedom (edf), reference degrees of freedom (ref. df) and F-values for smooth terms. All corresponding p-values are < 0.0001.

smooth terms	edf	ref. df	F-value
Initial Phoneme (bs="re")	27.925	30.000	98.750
parametric terms	β	S.E.	t-value
Intercept	-1.843	0.019	-95.242
C1 Frequency	-0.025	0.001	-16.864
C2 Frequency	-0.012	0.002	-7.634
Frequency	-0.021	0.001	-26.629
C1 Strokes	0.063	0.003	21.532
C2 Strokes	0.011	0.003	4.061
C1 Entropy	-0.027	0.002	-17.311
C2 Entropy	-0.018	0.001	-12.381
C1 Relative Entropy	-0.010	0.001	-8.735
C2 Relative Entropy	-0.008	0.001	-6.660
C1 SR Family Size	0.004	0.001	3.927

In parallel to the lexical decision data, we observed effects of the information-theoretic measures entropy (*C1 Entropy*, $t = -17.311$; *C2 Entropy*, $t = -12.381$) and relative entropy (*C1 Relative Entropy*, $t = -8.735$; *C1 Backward Relative Entropy*, $t = -6.660$) as well. As before, greater values of entropy and relative entropy lead to shorter response times.

Furthermore, the data revealed an effect of a lexical variable that is defined below the character level. For the first character, an increased family size of the semantic radical (*C1 SR Family Size*, $t = 3.927$) results in longer naming latencies. This inhibitory effect of family size is consistent with the inhibitory effect of semantic radical family frequency reported by C. C. Sun (2016) for word naming, but in the opposite direction of the facilitatory effects of semantic radical family size that are typically found in lexical decision experiments Feldman and Siok (1997, 1999a,b) (see, however, our discussion of the semantic radical family size effect in the CLP data for one-character words above).

The opposite pattern of results for semantic radical family size in lexical decision and word naming may suggest that while semantic radicals that occur in a large number of characters help determine the lexical status of a character (i.e., real Chinese character or not), they do not provide much information for the identification of a specific character. As a reviewer pointed out, however, it should be noted that the semantic radical family size effects for lexical decision were observed for one-character words, whereas the semantic radical family size effects for word naming were observed for two-character words. A direct comparison of the semantic radical family size effects in lexical decision and word naming is therefore not possible.

Finally, a post-hoc analysis for the subset of two-character words for which the first character contained a phonetic radical ($n = 11,749$) revealed an effect of the number of friends of the phonetic radical of the first character. Naming latencies for two-character words where shorter when more characters with the same phonetic radical were pronounced in the same manner (*C1 PR Friends*, $t = -4.439$). This effect of phonetic radical friends fits well with the effect of phonetic radical regularity that we reported for one-character words above.

3.8 Discussion

The evaluation of the lexical variables in the CLD for lexical decision latencies and word naming data for one-character and two-character words yielded a number of interesting results. We observed a number of effects that are well-documented in the psycholinguistic literature on Mandarin Chinese, including word and character frequency effects, visual complexity effects and orthography-to-phonology consistency effects. More frequent words and characters give rise to shorter response times, whereas more complex words and characters lead to less efficient processing and longer response times. In addition, response times were shorter for words with more consistent orthography-to-phonology mappings.

Effects were observed at different grain sizes. At the word and character level we found the above-mentioned effects of frequency, visual complexity, and orthography-to-phonology consistency. Below the character level we observed an effect of orthography-to-phonology consistency as well. Naming latencies for single-character words were shorter when the pronunciation of the character was identical to the pronunciation of the phonetic radical. In addition, we found effects of the family size of the semantic radical (i.e., the number of characters in which a semantic radical appears) in both lexical decision and word naming. The greater the family size of the semantic radical of a character, the longer the response time to words containing that character (see, however, Feldman and Siok, 1997, 1999a,b, for facilitatory effects of semantic radical family size in lexical decision).

Furthermore, we found effects related to the way in which characters combine to form words. The information-theoretic measures entropy and relative entropy had substantial explanatory power for two-character words in both lexical decision and word nam-

ing. Surprisingly, a greater uncertainty about one character given the other character led to shorter reaction times. In addition, response latencies were shorter when the frequency distribution of a character given the other character did not resemble the frequency distribution of that character in the language as a whole. In addition, the analysis of lexical decision latencies for three-character words revealed an effect of conditional probability. These effects of information-theoretic measures highlight the importance of taking the combinatorial properties of characters into account when investigating lexical processing above the character level.

Finally, we would like to clarify that the current analyses are by no means intended to provide an exhaustive description of the predictive power of the lexical information provided by the CLD for the experimental data sets under investigation. Above, we focused on the predictor effects that were most prominent in simple linear regression analyses. More complicated relationships between lexical decision latencies and word naming latencies on the one hand, and the lexical variables in the CLD on the other hand are likely to exist. These more complicated relationships include, but are not limited to, non-linear predictor effects and linear or non-linear interactions between predictors. The absence of an effect for a specific predictor in the current analyses, therefore, does not imply the absence of a predictive relationship between that predictor and the dependent variables under investigation.

4 Conclusions

We presented the Chinese Lexical Database (CLD). The CLD is a large-scale lexical database for one-character and two-character words in simplified Chinese. It comprises 3,913 one-character words, 34,233 two-character words, 7,143 three-character words, and 3,355 four-character words, for a total of 48,644 words. The 48,644 words in the CLD consist of 4,895 unique characters. For each of these words and characters, the CLD provides a wealth of lexical information. Categorical variables provide information about the orthographic and phonological form of a word, its characters and the semantic and phonetic radicals in each character. Numerical variables contain in-depth information about the frequency, the complexity, the neighborhood density, the orthography-phonology consistency and the information-theoretic properties of linguistic units at different grain sizes.

The CLD contains an unmatched amount of lexical information for simplified Chinese. Nonetheless, the lexical information provided by the CLD is by no means exhaustive. The current version of the CLD, for instance, does not contain semantic information about words and their characters. In future updates to the CLD, we plan to provide categorical semantic information, as well as numerical semantic measures derived from subjective ratings and distributional semantic models. Similarly, the current version of the CLD does not provide information about the grammatical status of linguistic units. The addition of such information, too, will have high priority in the further development of the CLD.

For now, however, the CLD is the largest lexical resource for simplified Chinese by a substantial margin. An evaluation of the lexical information provided by the CLD for large-scale lexical decision and word naming data demonstrated the potential of the lexical variables in the CLD for uncovering hitherto unobserved effects in experimental data by unveiling robust effects of information-theoretic measures of the combinatorial properties of characters. Furthermore, the information provided by the CLD allows for an objective analysis, re-evaluation, and comparison of predictor effects across studies and experimental paradigms. We hope, and we believe, therefore, that the CLD will prove to be a valuable resource for psycholinguistic research on simplified Chinese.

References

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15:802–814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 18:234–254.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, 4:439–461.

Baayen, R. H., Feldman, L., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53:496–512.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–482.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM).*

Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Boudelaa, S. and Marslen-Wilson, W. D. (2010). Aralex: a lexical database for Modern Standard Arabic. *Behavior Research Methods*, 42(2):481–487.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statisistical Society*, 26:211–246.

Cai, Q. and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, 5(6).

Chang, Y. N., Hsu, C. H., Chen, C. L., and Lee, C. Y. (2016). A psycholinguistic database for traditional Chinese character naming. *Behavior Research Methods*, 48(1):112–122.

Chen, H. C., Vaid, J., and Wu, J. T. (2009). Homophone density and phonological frequency in Chinese word recognition. *Language and Cognitive Processes*, 24(7-8):967–982.

Chen, J. Y. and Dell, G. S. (2006). Word-form encoding in Chinese speech production. In Li, P., Tan, L. H., Bates, E., and Tzeng, O. J. L., editors, *The Handbook of East Asian Psycholinguistics*, volume 1, pages 165–174. Cambridge University Press, New York.

Chen, W. F., Chao, P, C., Chang, Y. N., and Hsu, C. H. (2016). Effects of orthographic consistency and homophone density on Chinese spoken word recognition. *Brain and Language*, 157-158.

Cheng, C. M. (1981). Perception of Chinese characters. *Acta Psychologica Taiwanica*, 23:137–153.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

Coltheart, M., Davelaar, E., Jonasson, J. T., and Besner, D. (1977). Access to the internal lexicon. In Dornick, S., editor, *Attention and performance*, volume VI, pages 535–556. Erlbaum, Hillsdale, New Jersey.

Coltheart, V., Laxon, V. J., and Keating, C. (1988). Effects of word imageability and age of acquisition on children's reading. *British Journal of Psychology*, 79:1–12.

DeFrancis, J. (1984). *The Chinese Language: Facts and Fantasy*. University of Hawaii Press, Honolulu.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., and Carreiras, M. (2013). EsPal: one-stop shopping for Spanish word properties. *Behavior Research Methods*, 45(4):1246–1258.

Evert, S. (2009). Corpora and collocations. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: An International Handbook*, volume 2, pages 1212–1248. De Gruyter Mouton, Berlin, Boston.

Feldman, L. B. and Siok, W. W. T. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning,Memory and Cognition*, 23:778–781.

Feldman, L. B. and Siok, W. W. T. (1999a). Semantic radicals contribute to the visual identification of Chinese characters. *Journal of Memory and Language*, 40:559–576.

Feldman, L. B. and Siok, W. W. T. (1999b). Semantic radicals in phonetic compounds: Implications for visual character recognition in Chinese. In Wang, J., Inhoff, A., and Chen, H. C., editors, *Reading Chinese Script: A Cognitive Analysis*. Erlbaum, Hillsdale, NJ.

Forster, K. and Shen, D. (1996). No enemies in the neighborhood: absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:696–713.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, 59:127–136.

Graff, D. and Chen, K. (2003). Chinese Gigaword LDC2003T09.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, 29:228–244.

Grainger, J. (1992). Orthographic neighborhoods and visual word recognition. In Katz, L. and Frost, R., editors, *Orthography, Phonology, Morphology & Meaning*, pages 131–146. Elsevier, Amsterdam.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In Sánchez, A. and Almela, M., editors, *A Mosaic of Corpus Linguistics: Selected Approaches*, pages 269–291. Peter Lang, Frankfurt am Main.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, 1(3):297–318.

Heister, J., Würzner, K., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., and Kliegl, R. (2011). dlexDB - eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychologische Rundschau*, 62:10–20.

Hendrix, P. (2016). *Experimental explorations of a discrimination learning approach to language processing*. PhD thesis, Eberhard Karl's Universität, Tübingen.

Hendrix, P., Bolger, P., and Baayen, R. H. (2017). Distinct ERP signatures of word frequency, phrase frequency and prototypicality in speech production. *Journal of Experimental Psychology: Language, Memory and Cognition*, 3(1):128–149.

Honorof, D. N. and Feldman, L. (2006). The Chinese character in psycholinguistic research: form, struc-

ture and the reader. In Li, P., Tan, L. H., Bates, E., and Tzeng, O. J. L., editors, *The Handbook of East Asian Psycholinguistics*, volume 1, pages 195–217. Cambridge University Press, New York.

Hoosain, R. (1991). *Psycholinguistic Implications for Linguistic Relativity: A Case Study of Chinese*. Erlbaum, Hillsdale, NJ.

Hsieh, S. K. (2006). *Hanzi, Concept and Computation: A Preliminary Survey of Chinese Characters as a Knowledge Resource in NLP (Doctoral Dissertation)*. Eberhard Karls Universität Tübingen.

Huang, H. W., Lee, C. Y., Tsai, J. L., Lee, C. L., Hung, D. L., and Tzeng, O. J. L. (2006). Orthographic neighborhood effects in reading Chinese two-character words. *Neuroreport*, 17(10):1061–1065.

Hue, C. (1992). Recognition processes in character naming. In Chen, E. and Tzeng, O., editors, *Language Processing in Chinese*, pages 93–107. North-Holland, Amsterdam.

Ktori, M., Van Heuven, W. J. B., and Pitchford, N. J. (2008). Greeklex: A lexical database of Modern Greek. *Behavior Research Methods*, 40(3):773–783.

Kuo, W. J., Yeh, T. C., Lee, C. Y., Wu, Y., Chou, C. C., Ho, L. T., Hung, D. L., Tzeng, O. J. L., and Hsieh, J. C. (2003). Frequency effects of Chinese character processing in the brain: an event-related fMRI study. *NeuroImage*, 18(3):720–730.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008a). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23:1089–1132.

Kuperman, V., Bertram, R., and Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, 62:83–97.

Kuperman, V., Ernestus, M., and Baayen, R. H. (2008b). Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. *JASA*, 124:3897–3908.

Kuperman, V., Pluymaekers, M., and Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America*, 121(4):2261–2271.

Kyparissiadis, A., Van Heuven, W. J. B., and Pitchford, N. J. (2017). GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PLoS ONE*, 12(2).

Lee, C. Y., Hsu, C. H., Chang, Y. N., Chen, W. F., and Chao, P. C. (2015). The feedback consistency effect in Chinese character recognition: Evidence from a psycholinguistic norm. *Language and Linguistics*, 16(4):535–554.

Lee, C. Y., Tsai, J. L., Kuo, W. J., Yeh, T. C., Wu, Y. T., Ho, L. T., Hung, D. L., Tzeng, O. J. L., and Hsieh, J. C. (2004). Neuronal correlates of consistency and frequency effects in Chinese character naming: an event-related fMRI study. *NeuroImage*, 23(4):1235–1245.

Leong, C. K., Cheng, P. W., and Mulcahy, R. (1987). Automatic processing of morphemic orthography by mature readers. *Language and Speech*, 30(2):181–196.

Liu, I. M. (1988). Context effects on word/character naming: Alphabetic versus logographic languages. In Liu, I. M., Chen, H. C., and Chen, M. J., editors, *Cognitive Aspects of the Chinese Language*, pages 81–92. Asian Research Service, Hong Kong.

Liu, I. M. (1999). Character and word recognition in Chinese. In Wang, J., Inhoff, A. W., and Chen, H. C., editors, *Reading Chinese Script: A Cognitive Analysis*, pages 173–187. Lawrence Erlbaum Associates, Hillsdale, NJ.

Liu, Y., Shu, H., and Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods*, 39(2):192–198.

McEnery, T. and Xiao, R. (2008). The lancaster corpus of mandarin chinese. Downloaded from `http://www.lancaster.ac.uk/fass/projects/corpus/LCMC`.

Milin, P., Filipović Durdević, D., and Moscoso del Prado Martín, F. (2009a). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, pages 50–64.

Milin, P., Kuperman, V., Kostić, A., and Baayen, R. (2009b). Paradigms bit by bit: an information-theoretic approach to the processing of paradigmatic structure in inflection and derivation. In Blevins, J. P. and Blevins, J., editors, *Analogy in Grammar: Form and Acquisition*, pages 214–252. Oxford University Press, Oxford.

Ministry of Education of the People's Republic of China (2013). 通用规范汉字表 *[Table of General Standard Chinese Characters]*.

Myers, J. and Gong, S. P. (2002). Cross-morphemic predictability and the lexical access of compounds in Mandarin Chinese. *Folia Linguistica*, 26(1-2):65–96.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28:661–677.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods*, 36:516–524.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: Lexique. *L'Année Psychologique*, 101:447–462.

Parkvall, M. (2007). Världens 100 största sprak [the world's largest 100 languages]. In *Nationalencyklope-*

din. NE Nationalencyklopedin AB, Malmö.

Peng, D. L., Liu, Y., and Wang, C. M. (1999). How is access representation organized? The relation of polymorphemic words and their components in Chinese. In Wang, J., Inhoff, A. W., and Chen, H. C., editors, *Reading Chinese Script: A Cognitive Analysis*, pages 65–89. Lawrence Erlbaum Associates, Hillsdale, NJ.

Peterson, E. (2005). Mandarin tools: Chinese character dictionary. Available through `http://www.mandarintools.com/chardict.html`.

Pham, H. and Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience*, 30.

Pluymaekers, M., Ernestus, M., and Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62:146–159.

Revelle, W. (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.6.9.

Schmidtke, D., Kuperman, V., Gagné, and Spalding, T. (2016). Competition between conceptual relations affects compound recognition: the role of entropy. *Psychonomic Bulletin & Review*, 23(2):556–570.

Schreuder, R. and Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, 37:118–139.

Seidenberg, M. (1985a). The time course of phonological code activation in two writing systems. *Cognition*, 19:1–30.

Seidenberg, M. S. (1985b). The time course of phonological code activation in two writing systems. *Cognition*, 19:1–30.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shaoul, C., Sun, C. C., and Ma, J. Q. (2016). The Simplified Chinese Corpus of Webpages (SCCoW). *Manuscript*.

Sun, C. (2006). *Chinese: A Linguistic Introduction*. Cambridge University Press, United Kingdom.

Sun, C. C. (2016). *Lexical processsing in simplified Chinese: an investigation using a new large-scale lexical database*. PhD thesis, Eberhard Karl's Universität, Tübingen.

Sun, L. (2016). Public Weixin Corpus. Downloaded from `https://github.com/nonamestreet`.

Sze, W., Rickard Liow, S. J., and Yap, M. J. (2014). The Chinese Lexicon Project: a repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, 46(1):263–273.

Taft, M. (2006). Processing of characters by native Chinese readers. In Li, P., Tan, L. H., Bates, E., and Tzeng, O. J. L., editors, *The Handbook of East Asian Psycholinguistics*, volume 1, pages 237–249. Cambridge University Press, New York.

Taft, M., Huang, J., and Zhu, X. (1994). The influence of character frequency on word recognition responses in Chinese. In Chang, H. W., Huang, J. T., Hue, C. W., and Tzeng, O. J. L., editors, *Advances in the study of Chinese language processing*, volume 1, pages 59–73. Department of Psychology, National Taiwan University, Taipei.

Taft, M., Liu, Y., and Zhu, X. (1999). Morphemic processing in reading Chinese. In Wang, J., Inhoff, A. W., and Chen, H. C., editors, *Reading Chinese Script: A Cognitive Analysis*, pages 91–114. Lawrence Erlbaum Associates, Hillsdale, NJ.

Taft, M. and Zhu, X. (1997). Submorphemic processing in reading Chinese. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23(3):761–775.

Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. F., Wang, S., and Chen, H. C. (2017). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods (doi: 10.3758/s13428-017-0944-0)*.

Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., and Lin, D. (2016). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, 49:1503–1519.

Van Esch, D. (2012). Leiden weibo corpus. Downloaded from `http://lwc.daanvanesch.nl`.

Vitevich, M. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28(4):735–747.

Wang, W., Ning, N., and Zhang, J. X. (2012). The nature of the homophone density effects: an ERP study with Chinese spoken monosyllabic homophones. *Neuroscience Letters*, 516(1):67–71.

Wikipedia (2016). Pinyin — Wikipedia, the free encyclopedia. [Online; accessed 17-January-2018].

Wood, S. (2006). *Generalized Additive Models*. Chapman & Hall/CRC, New York.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wurm, L. H. and Fisicaro, S. A. (2014). What residualizing predictors in regression analysis does (and what it does not do). *Journal of Memory and Language*, 72:37–48.

Xiao, H. (2010-2015). 汉语拼音标注工具. `http://www.cncorpus.org/`.

Yan, G., Tian, H., Bai, X., and Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology*, 97:259–268.

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., and Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42(4):992–1003.

Yarkoni, T., Balota, D. A., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

Yates, M., Locker, L., and Simpson, G. B. (2004). The influence of phonological neighborhood on visual word recognition. *Psychonomic Bulletin & Review*, 11:452–457.

Yip, P. C. (2000). *The Chinese Lexicon: A Comprehensive Survey*. Routledge, New York.

Zhang, B. Y. and Peng, D. L. (1992). Decomposed storage in the Chinese lexicon. In Chen, H. C. and Tzeng, O. J. L., editors, *Language Processing in Chinese*, pages 131–149. North-Holland, Amsterdam.

Ziegler, J., Tan, L. H., Perry, C., and Montant, M. (2000). Phonology matters: the phonological frequency effect in written Chinese. *Psychological Science*, 11(3):234–238.