Database of Mandarin Neighborhood Statistics

Karl Neergaard, Hongzhi Xu, Chu-Ren Huang
The Hong Kong Polytechnic University
11 Yuk Choi Rd. Hunghom, Hong Kong
E-mail: karlneergaard@gmail.com, hongz.xu@gmail.com, churen.huang@polyu.edu.hk

Abstract

In the design of controlled experiments with language stimuli, researchers from psycholinguistic, neurolinguistic, and related fields, require language resources that isolate variables known to affect language processing. This article describes a freely available database that provides word level statistics for words and nonwords of Mandarin, Chinese. The featured lexical statistics include subtitle corpus frequency, phonological neighborhood density, neighborhood frequency, and homophone density. The accompanying word descriptors include pinyin, ascii phonetic transcription (sampa), lexical tone, syllable structure, dominant PoS, and syllable, segment and pinyin lengths for each phonological word. It is designed for researchers particularly concerned with language processing of isolated words and made to accommodate multiple existing hypotheses concerning the structure of the Mandarin syllable. The database is divided into multiple files according to the desired search criteria: 1) the syllable segmentation schema used to calculate density measures, and 2) whether the search is for words or nonwords. The database is open to the research community at https://github.com/karlneergaard/Mandarin-Neighborhood-Statistics.

Keywords: lexical statistics, phonological neighborhood density, Mandarin, Chinese

1. Introduction

Essential to conducting controlled language experiments is the availability of linguistic tools that provide word level statistics. While a frequency list of word occurrence might be the primary goal of many databases used by psycholinguists, neighborhood density measures have become instrumental in studies of language processing. Phonological neighborhood density (PND), which is the calculation of whole word sound similarity through the addition, deletion or substitution of a single phoneme, has been extensively studied, which explains the large number of resources available, for example, in English (N-Watch: Davis, 2005; IPhod: Vaden et al., 2009); Spanish (BuscaPalabras: Davis & Perea, 2005), French (Lexique: New et al., 2001) and representing multiple European languages (Clearpond: Marian et al., 2012). For Mandarin, while frequency measures have come from large-scale corpora (e.g., Sinica Corpus: Chen et al., 1996; Lancaster Corpus of Mandarin Chinese (LCMC): McEnery & Xiao, 2003; The language corpus system of Modern Chinese Study (LCSMCS): Sun, et al., 1997) or more recently a movie subtitle corpus (Subtlex-CH: Cai & Brysbaert, 2010), there is to date no resource available for neighborhood density measures. The current paper introduces a freely available database of lexical statistics including neighborhood density measures for Mandarin, Chinese.

A principle goal of the current database is to create a centralized tool for the testing of multiple hypotheses on the nature of the Mandarin mental lexicon. The area in which researchers have widely disagreed is in the phonological content of the Mandarin syllable.

Depending on the area of research being explored, researchers will claim that the base syllable consists of either three or four segments. The four-segments approach consists of a maximal syllable of CVVX, wherein the C corresponds to initial consonants, the following V is commonly referred to as the medial glide, the second V allows for monophthongs, and the final X corresponds either to the second vowel in a diphthong, or a final consonant. Table 1 provides examples of how individual segments correspond to this structure.

	C	V	V	X
liang1 (相)	l	i	a	ŋ
yang1 (央)		i	a	ŋ
ya1 (鸭)		i	a	
ang1 (肮)			a	ŋ
ai1 (哀)			a	ɪ
a1 (阿)			a	

Table 1: Example base syllables according to the CVVX syllable structure

Mandarin, Chinese is a tonal language, said to have four principle tones that are written as numerals 1-4. A limited number of syllables are said to be without tone and thus are assigned the numeral 0. For transcription purposes, the tonal information, here represented by a T, can be added to the end of the base syllable, as is the tradition in writing words in pinyin (Mandarin Romanization). Examples of syllable structures plus lexical tone can be seen in Table 3.

While the CVVX approach identifies four segmental units that does not imply that there is agreement as to the number of phonological units within the syllable. Initial proposals did not consider the role of tone in the syllable, but instead debated the constituents of the rime (C_VVX: Xu, 1980; C_V_VX: Cheng, 1966) and the role of the medial glide (CV_VX: Bao, 1990; CV_V_X: Ao, 1992).

Meanwhile, the three-segments approach consists of a maximal syllable of CVC, wherein all vowels, including medial glide, monophthong, and diphthong are treated as a single unit that is then followed by the three final consonants: /r, n, ŋ/. Examples of the CVC syllable can

be seen in Table 2. The first study to consider the influence of PND on Mandarin speech processing used the CVC syllable structure, while continuing the route of disregarding tone (Tsai, 2007).

	C	V	C
liang1 (相)	l	ia	ŋ
liao3 (了)	l	iaʊ	
yang1 (央)		ia	ŋ
ya1 (鸭)		ia	

Table 2: Example base syllables of CVC syllable structure

More recent speech production studies, both behavioral and those implementing ERP, have since provided evidence of a mental lexicon in which tonal information is phonologically related to lexical access along the lines of segmental or syllable information. Multiple behavioral priming studies have found no priming of syllable onsets, suggesting that the syllable is a single phonological unit (e.g., O'Seaghdha et al., 2010). This proposal is equivalent to a single phonological unit plus lexical tone: CVVX_T. ERP studies however have provided evidence that Mandarin speakers, like English speakers, process words incrementally and segmentally (e.g., Malins & Joannisse, 2012), suggesting that each phone within the syllable, plus the lexical tone, are individual units: C_V_V_X_T.

In order to provide a useful tool to the research community studying phonological phenomena in Mandarin, we felt it necessary to provide a resource for each of the above-mentioned proposals. In Table 3 we illustrate how the current database represents each of the fourteen segmentation schemas. You will notice the underscore signifies a separation between phonological units, while for simplicity sake we have opted to use the numeral version of the lexical tone rather than the more complex IPA symbols.

Without Tone		With Tone	
C_V_C	/l_ia_ŋ/	C_V_C_T	/l_ia_ŋ_3/
C_V_V_X	/l_i_a_ŋ/	C_V_V_X_T	/l_i_a_ŋ_3/
C_V_VX	/l_i_aŋ/	C_V_VX_T	/l_i_aŋ_3/
C_VVX	/l_iaŋ/	C_VVX_T	/l_iaŋ_3/
CV_V_X	/li_a_ŋ/	CV_V_X_T	/li_a_ŋ_3/
CV_VX	/li_aŋ/	CV_VX_T	/li_aŋ_3/
CVVX	/liaŋ/	CVVX_T	/liaŋ_3/

Table 3: Mandarin segmentation schemas

2. Preparation of the Lexicon

The lexicon from which all lexical statistics are calculated is from the Subtlex-CH word frequency list (Cai & Brysbaert, 2010). Subtlex-CH was chosen due to it having the highest correlation with spoken reaction times when compared with LCSMCS and LCMC frequency lists. This word list provided both movie subtitle frequency and part of speech per orthographic word.

Chinese characters from the Subtlex-CH word list were then translated into pinyin using the CKIP Lexicon (Chinese Knowledge Information Processing Group, 1995). Out of vocabulary words were translated manually. Of the roughly 99,000 words present in the word list, more than 4300 words were found to be polyphonous, meaning that more than one pronunciation was ascribed to identical characters either within a multisyllabic word or for individual monosyllabic words. The vast majority of the multisyllabic words' pronunciations were resolved through the help of their PoS assignments. Three native Mandarin speakers annotated the remaining 151 polyphonous words for their pronunciation as given within sentential context from the corpus. Of the polyphonous words, 62 tokens found no annotator agreement and were removed.

Having translated the Chinese characters into their corresponding pinyin, the next step was to apply a phonological system. Pinyin words were translated to an ascii transcription, commonly referred to as sampa, following the syllable inventory in Neergaard & Huang (2016). See Appendix A for a translation of the sampa to IPA with accompanying example Chinese characters and pinyin. This syllable inventory was shown to outperform two other commonly used Mandarin inventories (Lin, 2007; Zhao & Li, 2009) in a word similarity task with a native Mandarin-speaking population.

3. Syllable Segmentation

Users are able to search for lexical statistics according to the segmentation schema/s that constitute their linguistic criteria or experimental hypotheses. Table 3 contains each of the fourteen segmentation schemas for which the database files are named.

In addition to Mandarin lexical statistics, the current database also provides neighborhood information for 4,404 monosyllabic nonwords. The nonword lists include what are known as tone gap words, i.e., existing syllables in the Mandarin syllable inventory that because of their lexical tone are not considered real words, or do not correspond to an existing Chinese character, and segmental gap nonwords, which are nonwords made from the combination of phones from the Mandarin phoneme inventory to create syllables that do not exist in the Mandarin syllable inventory. Users will find files of the same nonwords for each of the fourteen segmentation schemas.

4. Word Level Statistics

The final manipulations of the lexicon occurred for the purpose of creating neighborhood statistics. It was first necessary to choose the vocabulary size from which all neighborhood calculations would be made. A common range used by multiple sources is around 20,000 words (e.g., Marian et al., 2012). This number comes from a study of the written vocabulary size of American English speaking college students, which found that the average

vocabulary size, excluding morphological variants of a single lemma, was between 17,000-20,000 words (Gouldon et al., 1990). We chose to build each segmentation schema file on the lower end of this range specifically because while English has morphological variants of lemmas, for instance due to number, (peach -> peach**es**), and verb conjugation (walk -> walk**s**, walk**ed**, walk**ing**), Mandarin lacks such morphological variation.

Prior to the calculation of lexical statistics, the database was collapsed according to homophonous words, such that words with identical sampa, i.e., phonological words, were reduced to a single entry, their combined frequencies summed, and dominant PoS assigned to the PoS with the highest frequency. Each phonological word is presented with its corresponding tone number, syllable structure, and pinyin, phoneme, and syllable length.

Phonological words in the database are abbreviated with the title "Pho", as can be seen in Table 4. Columns that feature the phonological word plus its lexical tone are titled "Pho+T", while columns that feature the phonological word without lexical tone are titled "Pho-T". The same naming scheme was used for pinyin words ("PY") with and without tone, i.e., "PY+T" and "PY-T" respectively. Phonological units within a given phonological word were distinguished by the presence of a blank space between phones, such that the C_V_VX_T version of liang3 is: l i aN 3.

Neighborhood statistics were then calculated from the top 17,000 phonological words for all words and nonwords within each of the syllable segmentation schemas. Thus, while a given file will contain 80,000+ orthographic words, each word's homophone density (HD), phonological neighborhood density (PND), and neighborhood frequency (NF) are calculated from only the top 17,000 phonological words.

In order to follow the conventions practiced in similar resources, proper names also needed to be removed. Instead of striking them from the database altogether, we reduced their word frequency to 1, thus barring them from the top 17,000 so that they did not contribute to the neighborhood calculations.

4.1 Homophone Density

Homophone density (HD) refers to the number of words that share the same phonological representation. In a Chinese language, such as Mandarin, this number can reach above 100 if tone is not considered (Duanmu, 2005). In the current database HD was calculated by counting the number of orthographic words that mapped to each of the top 17,000 phonological words. In Table 2 it is noteworthy that the presence or absence of lexical tone within a segmentation schema determines what is or isn't considered a homophonous word. This is because when tone is removed, all possible phonological words that would otherwise differ due to tone, are collapsed into a single item. For example, the monosyllabic phonological word that includes tone, liaU3, has just three orthographic representations (了, 瞭, 蓼). However, when tone is removed from the phonological word two things might happen to alter its HD. The first is that full syllables, including, liaU3, liaU2 and liaU4 collapse to a bare phonological syllable, liaU, which corresponds, in the present database, to sixteen Chinese characters (了, 瞭, 蓼, 聊, 料, 疗, 撂, etc.). Given what is allowed by the specific segmentation schema, the HD of a toneless phonological word will also be altered by the collapsing of multiple syllables, such that liaU, will include an additional four homophones based on disyllabic words for li and aU (利奥, 李奥, 里奥, 丽奥).

4.2 Neighborhood Statistics

Because PND is calculated through the addition, deletion or substitution of a single phoneme, manipulating the size a given phonological unit within a database can radically alter the number a word has of phonologically similar words, also known as neighbors. See Table 4 for an illustration of variation across a selection of segmentation schemas.

The current database provides measures for all addition, substitution, and deletion calculations. Users will also find information for the total number of neighbors and the neighbors of a given phonological word represented as sampa transcribed phonological words.

#Units	Schema	Pho	HD	PND	NF
5	C_V_V_X_T	l i a U 3	3	13	11,7225
4	C_V_V_X	l i a U	20	19	544,604
4	C_V_C_T	l iaU 3	3	17	141,167
3	C_V_C	l iaU	16	25	530,714
4	CV_V_X_T	li a U 3	3	26	210,398
3	CV_V_X	li a U	16	32	1,202,860
3	CV_VX_T	li aU 3	3	28	233,947
2	CV_VX	li aU	20	36	1,248,573
3	C_VVX_T	l iaU 3	3	26	198,555
2	C_VVX	l iaU	16	47	932797
2	CVVX_T	liaU 3	3	262	5,622,750
1	CVVX	liaU	20	440	20,391,422

Table 4: Statistical variation across schemas for the phonological word liaU3

The final neighborhood statistic, neighborhood frequency (NF), is calculated by summing the subtitle word frequencies of all of a given phonological word's neighbors. Table 2 illustrates the tendency for the number of phonological units a phonological word is constituted of to effect the number of neighbors a phonological word has, which in turn effects the size of the NF measure.

5. Acknowledgements

This work was partially supported by the VariAMU project (project code: 1.70.xx.99QX), which is funded by the Initiative d'excellence Aix-Marseille (AiMidex).

6. Conclusion

In this article we introduce a freely available database for the creation of Mandarin stimuli for words and

nonwords. Its organization is made such that multiple hypotheses can be tested according to a researcher's experimental design and working premise of Mandarin syllable segmentation. The use of Subtlex-CH (Cai & Brysbaert, 2010) subtitle word frequency allows the database to offer the current best fit in word level statistics for speech processing experiments and other applied and clinical purposes. The searchable word characteristics, such as, sampa, syllable, phoneme, and pinyin length, tone number, syllable structure, dominant PoS, and homophone density will assist researchers in the creation of controlled stimuli. Most importantly the database is the first to provide the research community with Mandarin neighborhood statistics: PND, and NF.

The database is organized according to twenty-eight files: fourteen that provide lexical statistics for Mandarin words according to a given segmentation schema (7 with tone and 7 without tone), and fourteen files that use the segmentation schemas to provide lexical statistics for nonwords. It is our intention to provide a centralized tool to the research community so as to establish a unified baseline for comparisons across future studies investigating neighborhood effects amongst Mandarin speakers. The database can be accessed here: https://github.com/karlneergaard/Mandarin-Neighborhood-Statistics.

7. Main References

Bao, Z. M. (1990). Fan-Qie Languages and Reduplication. *Linguistic Inquiry 21*.

Cai, Q., Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*(6): e10729. Retreived March 10, 2010 from http://crr.ugent.be/programs-data/subtitle-frequencies/subtlex-ch

Chen, K-J., Huang, C-R., Chang, L-P., & Hsu, H-L. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation* (PACLIC 11). pp. 167-176. Seoul, South Korea. Retrieved March 10, 2016 from http://app.sinica.edu.tw/cgi-bin/kiwi/mkiwi/kiwi.sh?language=1

Cheng, Robert L. (1966). Mandarin phonological structure. *Journal of Linguistics 2.2*: 135-158.

Chinese Knowledge Information Processing Group. (1995). *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior research methods*, *37*(1), 65-70. Retrieved March 10, 2016, from http://www.pc.rhul.ac.uk/staff/c.davis/Utilities/

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*(4), 665-671. Retrieved March 10, 2016 from http://www.pc.rhul.ac.uk/staff/c.davis/Utilities/

Duanmu, S. (2005). Chinese (Mandarin), Phonology of. In *Encyclopedia of Language and Linguistics*, 2nd edition, pp. 1-8, Elsevier Publishing House.

Goulden R., Nation P., Read J. (1990) How large can a receptive vocabulary be? *Applied Linguistics, 11*, 341–363. doi:10.1093/applin/11.4.341.

Lin, Y. H. (2007). *The Sounds of Chinese with Audio CD* (Vol. 1). Cambridge University Press.

Liu, Y. N., Shu, H., &Wei, J. H. (2006). Spoken word recognition in context: Evidence from Chinese ERP analyses. *Brain and Language, 96*, 37–48.

Malins, J. G., Joanisse, M. F. (2012). Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia, 50*, 2032-2043.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS One, 7*(8), e43230. Retrieved March 10, 2016 from http://clearpond.northwestern.edu

McEnery, A. M., & Xiao, R. Z. (2003). The Lancaster Corpus of Mandarin Chinese. Retrieved March 10, 2016 from http://www.lancaster.ac.uk/fass/projects/corpus/

Neergaard, K., & Huang, C-R. (2016). Phonological neighborhood density in a tonal language: Mandarin neighbor generation task. *90th Annual Meeting of the Linguistic Society of America*, Washington, DC.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE™//A lexical database for contemporary French: LEXIQUE™. *L'année Psychologique, 101*(3), 447–462. Retrieved March 10, 2016 from http://www.lexique.org

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

Sun, H. L., Huang, J. P., Sun, D. J., Li, D. J., & Xing, H. B. (1997). Introduction to language corpus system of modern Chinese study. In M. Y. Hu (Ed.), Paper collection for the Fifth World Chinese Teaching Symposium. Beijing: Peking University Publisher.

Tsai, P. T. (2007). The effects of phonological neighborhoods on spoken word recognition in Mandarin Chinese. Unpublished Master's thesis, University of Maryland.

Vaden, K.I., Halpin, H.R., Hickok, G.S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from http://www.iphod.com

Webster's seventh new collegiate dictionary. (1967). Los Angeles: Library Reproduction Service.

Zhao, J., Guo, J., Zhou, F., & Shu, H. (2011). Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia*, *49*(7), 1761-1770.

Zhao, X., & Li, P. (2009). An online database of phonological representations for Mandarin Chinese. *Behavior research methods*, *41*(2), 575-583.

Xu, Shirong. 1980. *Putonghua yuyin zhishi* [Phonology of Standard Chinese]. Beijing: Wenzi Gaige Chubans.

Appendix

	IPA	Sampa	Pinyin word	Sampa word	Ortho word		IPA	Sampa	Pinyin word	Sampa word	Ortho word
Vowels	a	a	b**a**3	p**a**3	把	Plosives	p	p	**b**u4	**p**u4	不
	ə	@	sh**e**4	S**@**4	蛇		p^h	P	**p**ao3	**P**aU3	跑
	e	e	g**e**i3	k**e**I3	给		k	k	**g**e0	**k**@0	个
	ɛ	E	y**e**3	i**E**3	也		k^h	K	**k**e4	**K@**4	课
	ɨ	1	zh**i**1	Z**1**1	之		t	t	**d**ou1	**t**oU1	都
	i	i	d**i**4	t**i**4	第		t^h	T	**t**a1	**T**a1	他
	ɪ	I	su**i**4	sue**I**4	岁	Fricatives	s	s	**s**uo3	**s**uo3	所
	o	o	ru**o**4	ru**o**4	若		f	f	**f**ang4	**f**aN4	放
	ʊ	U	cho**u**3	Co**U**3	丑		x	x	**h**ui4	**x**ueI4	会
	u	u	**w**o3	**u**o3	我	Affricates	ʂ	S	**sh**i4	**S**14	是
	y	y	**yu**an2	**y**En2	元		ɕ	X	**x**ia4	**X**ia4	下
Nasals	m	m	m**a**1	m**a**1	妈		tɕ	J	**j**iu4	**J**ioU4	就
	n	n	**n**eng2	**n**@N2	能		$tɕ^h$	Q	**q**ing3	**Q**iN3	请
	ŋ	N	xia**ng**3	Xia**N**3	想		ts^h	c	**c**ong2	**c**oN2	从
Liquids	l	l	**l**ie4	**l**iE4	列		$tʂ^h$	C	**ch**u1	**C**u1	出
	r	r	**r**ang4	**r**aN4	让		ts	z	**z**i4	**z**14	字
							tʂ	Z	**zh**e	**Z**@4	这

Appendix: IPA, pinyin, and sampa conversion chart with example pinyin and sampa syllables and Chinese characters