The Chinese Lexical Database (CLD)

Ching Chu Sun
University of Tübingen, Germany

Abstract

We present the Chinese Lexical Database CLD. The CLD is a new large-scale lexical database for Mandarin Chinese that provides over 150 descriptive and lexical-distributional variables for more than 30,000 words in simplified Chinese. The information in the CLD can be used for the construction of experimental stimuli and the analysis of experimental data in psycholinguistic research on simplified Chinese. The CLD can be downloaded for free, and an online search interface is provided at `http://www.chineselexicaldatabase.com`.

Keywords: lexical database, Mandarin Chinese, simplified Chinese, Chinese Lexical Database, CLD

Introduction

Lexical databases are a valuable research tool for psycholinguistic research. They provide researchers with descriptive and lexical-distributional information about a large number of words in a language. This information can be used to efficient construction of experimental stimuli, as well as for a thorough analysis of experimental data. Furthermore, patterns of results for existing data sets can be directly compared through a re-analysis using the predictors in the database.

Large-scale lexical databases have been released for a number of languages in the last decades. Lexical databases are now available for most well-studied languages, including English (see, e.g., Balota et al., 2007; Coltheart, 1981; Keuleers et al., 2012; Baayen et al., 1995), German (see, e.g., Heister et al., 2011; Baayen et al., 1995), French (see, e.g. New et al., 2001, 2004, 2007; Ferrand et al., 2010), and Dutch (Brysbaert et al., 2016; Baayen et al., 1995). However, for less-studied languages, relatively few lexical resources have been developed.

For Mandarin Chinese, a few databases with lexical information exist. Taiwan Sinica, for instance, constructed a lexical database of with 12 numerical predictors for 3,300 commonly used characters. For simplified Chinese, at least two lexical resources currently exist. Liu et al. (2007) provide 15 lexical variables for nearly 2,500 characters that can be used as independent words, whereas Cai and Brysbaert (2010) released character frequencies and word frequencies based on a large collection of subtitles from movies. Nonetheless, compared to the resources available for other languages these databases are limited in size. Here, we present the Chinese Lexical Database (CLD), a new large-scale lexical resource for simplified Chinese.

The Chinese Lexical Database (CLD)

Below, we briefly introduce the Chinese Lexical Database (CLD). For a comprehensive description of the CLD, the interested reader is referred to Sun (2016). The Chinese Lexical Database (CLD) contains lexical information for 30,645 one-character and two-character words in simplified Chinese. The number of unique characters in these words is 5,242. Not all characters can be used as words. Therefore, the total number of one-character words in the CLD is 4,710, whereas the total number of two-character words is 25,935. As such, the number of words in the CLD is substantially larger than the number of words in existing resources for Chinese.

The CLD contains a wealth of lexical information about each of the 30,645 it contains. In total, 23 categorical variables and 141 numerical variables are available for each word. The categorical variables provide information about the characters in a word, pronunciations, Pinyin transcriptions (a translation of characters into a romanized form based on pronunciations), tones, character types and structures, and phonetic and semantic radicals.

The numerical predictors in the CLD describe lexical-distributional properties of characters and words, and were calculated on the basis of the SCCoW (Shaoul et al., 2016): a corpus of webpages in simplified Chinese that consists of over 450 million words and 750 million characters. Numerical predictors in the CLD include measures of the orthographic and phonological frequency of words, characters, radicals and phonemes, as well as measures of the visual complexity of words and characters at various grain sizes. Orthographic and phonological neighbourhood characteristics are included as well, as are lexical variables that describe the orthography-to-phonology consistency of characters and phonetic radicals. Finally, the CLD provides a number of information-theoretic measures that describe combinatorial properties of characters. Sun (2016) recently documented the importance of such measures for understanding language processing in the word naming task.

The CLD is released under the GNU General Public License. A database dump is freely available with this publication in `.txt` format. Downloads of the CLD in other formats, as well as an online search interface are available at `http://www.chineselexicaldatabase.com`. The online search interface allows users to view subsets of the words and lexical variables in the CLD, and to receive the result of search queries by e-mail. Future versions of the CLD will be released under a new handle in the online repository of the University of Tübingen, as well as on `http://www.chineselexicaldatabase.com`.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM).* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. I., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *JEP:HPP*, *42*(3), 441–458.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6). Available from e10729.doi:10.1371/journal.pone.0010729

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

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., et al. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488-496.

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

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

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

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

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

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

Ramscar, M., Dye, M., & McCauley, S. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4), 760–793.

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

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