Chapter 32
Chinese Language Resources: A Comprehensive Compendium

Anran Li, Weidong Zhan, Jia-Fei Hong, Zhao-Ming Gao, and Chu-Ren Huang

Abstract This chapter will present a collective effort to compile a comprehensive repository of accessible Chinese language resources that can be used online, licensed for use, or accessed in published form. The compendium will be presented in three parts according to each language resource's type of accessibility, which is a direct consequence of the type of relevant information provided for each resource. Within each accessibility type, the resources were then further divided according to the following resource types: integrated resources, corpora, lexical resources, and wordnet/ontology. We believe that this four-way classification system will facilitate intuitive searches. However, this design will make it difficult to search for a resource within the same class due to having to rely on the alphabetic order of the titles of the resources. Lastly, it is important for our readers to bear in mind that such a repository is bound to be incomplete given the scale and distributional nature of the resources and the productivity of new resource construction. We plan to post this compendium online to allow easier access and provide updates in the future.

Keywords Repository of resources · Corpora · Mandarin · Languages in China

A. Li (✉) · C.-R. Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China
e-mail: 17903045r@connect.polyu.hk; churen.huang@polyu.edu.hk

W. Zhan
Department of Chinese Language and Literature, Peking University, Beijing, China
e-mail: zwd@pku.edu.cn

J.-F. Hong
Department of Chinese as a Second Language, National Taiwan Normal University, Taipei, Taiwan
e-mail: jiafeihong@ntnu.edu.tw

Z.-M. Gao
Department of Foreign Languages and Literatures, National Taiwan University, Taipei, Taiwan
e-mail: zmgao@ntu.edu.tw

C.-R. Huang et al. (eds.), *Chinese Language Resources*, Text, Speech and Language Technology 49, https://doi.org/10.1007/978-3-031-38913-9_32

32.1 Online Resources

32.1.1 Integrated Resources

Resource title	Developer and maintainer/author/host	Web sites	Notes
Adventures in Wen-Land 文國尋寶記—中小學語文知識網路	Institute of Linguistics, Academia Sinica, 中央研究院語言學研究所/ Chu-Ren Huang, Feng-Ju Lo et al. 黃居仁, 羅鳳珠 等	http://wen.ling.sinica.edu.tw/ http://cls.lib.ntu.edu.tw/wen	This is an integrated resource, including corpora of elementary school textbooks, lexica, classical Chinese literature, and references such as dictionaries, as well as language learning games and tools. This resource is intended to be used by advanced learners as well as teachers (for the preparation of teaching materials). As this is an integrated resource, some external links do not function now but the two mirror sites have preserved some unique functions
Audio Media Center of the National Language Resource Monitoring and Research Center 国家语言资源监测与研究中心有声媒体分中心	Communication University of China 中国传媒大学	http://ling.cuc.edu.cn	This platform has eight language resources and/or language tools: homophone auto-generation software, parallel corpus retrieval software (CUC_ParaConc), multilingual corpus processing software (HyConc), resources for neology research, charts for diachronic changes in media language, a national public opinion database of language and characters, a media language corpus, media language corpus segmentation, and an annotation system

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Buddhist Electronic Texts Integration of the Chinese Buddhist Electronic Texts Association (CBETA) 中華電子佛典協會電子佛典集成	Chinese Buddhist Electronic Texts Association 中華電子佛典協會	http://www.cbeta.org/ http://cbetaonline.dila.edu.tw/	The CBETA aims to digitalize and share all Chinese Buddhist texts. The CBETA database can be accessed online, freely downloaded after registration, or obtained on a CD. The online search tools meet state-of-the-art corpus linguistic requirements and are essential resources for religion, culture, and language studies, especially in terms of the impact of Buddhism on Chinese. This is also one of the largest historical corpora of translated texts in the world
Chinese Classics on the Web 網路展書讀	Yuan Ze University, 元智大學/Feng-ju Lo 羅鳳珠	http://cls.lib.ntu.edu.tw/	This is the aggregated website of Feng-ju Lo's life-long dedication to digital humanities for classical Chinese literature. The content ranges from the Four great books to Tang and Song poetry, Ming Dynasty plays, the great Chinese novels, and Southern-Min vernacular literature. Each web site explores different technologies and showcases different ways to integrate texts for reading, teaching, and research
Chinese-English Index System of the National Academy for Educational Research (Trial Version) 國家教育研究院華英雙語索引典系統(試用版)	National Academy for Educational Research 國家教育研究院	http://coct.naer.edu.tw/bc/	This corpus contains a collection of articles from the fields of literature, science, finance and economics, the arts, ideology, culture, global, and entertainment over the past

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			20 years. These articles have both Chinese and English versions. The Chinese parts of the articles are shown in traditional Chinese and the means of expression is Taiwan Chinese
Digital Resources Center for Global Chinese Teaching and Learning 全球華語文數位教與學資源中心	Institute of Linguistics, Academia Sinica 中央研究院語言學研究所/ Chin-Chuan Cheng, Chu-Ren Huang et al. 鄭錦全, 黃居仁 等	http://elearning.ling.sinica.edu.tw	This resource center is linked to multiple corpora. The central piece is a platform that provides *Word-Focused Extensive Reading* (一詞泛讀) to guide learners automatically through corpus-generated data in small chunks. It is also linked to the Key Word in Context (KWIC) interfaces of multiple Academia Sinica corpora and generates only three sentences at a time based on several user-designated criteria (such as easiness, synonyms, etc.)
Minority Languages Center of the National Language Resource Monitoring and Research Center 国家语言资源监测与研究中心少数民族语言分中心	Minzu University of China 中央民族大学	http://nmlr.muc.edu.cn/ziyuanzhongxin/	This platform contains several minority language corpora, including Mongol, Tibetan, the Uygur language, the Kazak language, etc.
Overseas Chinese Center of the National Language Resource Monitoring and Research Center 国家语言资源监测与研究中心海外华语研究分中心	Jinan University 暨南大学	https://huayu.jnu.edu.cn/source.aspx	This platform contains many corpora, word lists, language tools, and many other language resources, which are mainly focused on the teaching of Chinese as a foreign language, especially in Southeast Asia's condition

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Print Media Center of the National Language Resource Monitoring and Research Center 国家语言资源监测与研究中心平面媒体分中心	Beijing Language and Culture University 北京语言大学	http://cnlr.blcu.edu.cn/	This platform has five parts: the National Language Resources Dynamic Circulation Corpus (DCC, 10 billion characters from 18 newspapers over the past 10 years); the Traditional Culture Diachronic Corpus (CCC, corpus of ancient books and records); the Semantic Cloud Platform (SCP, enables people to see the collocation conditions of the DCC and has a word-embedding function); the Language Calculation Lab (LC-LAB); and the Green Book of Chinese Language Usage Condition
Scripta Sinica 漢籍全文資料庫計畫	Academia Sinica 中央研究院	hanchi.ihp.sinica.edu.tw	Scripta Sinica is the largest Chinese full-text database and it encompasses an enormous breadth of historical materials, such as almost all the important Chinese classics, especially those related to Chinese history. It started in 1984 as the first major Chinese digital archives project and now contains 1173 titles and more than 665 million characters
SouWenJieZi—A Linguistic KnowledgeNet 搜文解字 - 語文知識網路	Institute of Linguistics, Academia Sinica, 中央研究院語言學研究所/Chu-Ren Huang, Feng-ju Lo et al. 黃居仁, 羅鳳珠 等	http://words.sinica.edu.tw	This platform contains an electronic dictionary, a literature knowledge center, an ancient writing and Chinese characters evolution knowledge base, and several language games. This is

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			part of the Language Archives Project
Taiwan Digital Archives Program: Language Archives (Phase I, II) 語言典藏計畫(第一期、第二期)	Institute of Linguistics, Academia Sinica 中央研究院語言學研究所/ Chu-Ren Huang et al. 黃居仁 等	http://languagearchives.sinica.edu.tw/cht/index.php.html	This is the first integrated language archiving project in greater China. The coverage includes Pre-Qin excavated texts, Classical Chinese, Modern Mandarin, Taiwan Southern-Min and Hakka from historical perspectives, and endangered Formosan languages
Taiwan Southern Min and Hakka Archives 台灣閩客語語言典藏	Institute of Linguistics, Academia Sinica 中央研究院語言學研究所/ Chin-Chuan Cheng, Min-hua Chiang et al. 鄭錦全, 江敏華 等	http://museum02.digitalarchives.tw/ndap/2003/banlamgu	This site contains language resources for both Taiwan Southern Min and Taiwan Hakka. The content includes fieldwork data as well as historical data in searchable corpus format
The Global Database of Events, Language, and Tone Project 事件、语言与音调全球数据库项目	Kalev Leetaru (from Yahoo!) Kalev Leetaru (雅虎) and Georgetown University 乔治城大学	https://www.gdeltproject.org/data.html	This corpus is part of the Google GDELT Project. The 2015 data alone has recorded nearly three quarters of a trillion emotional snapshots and more than 1.5 billion location references, while its total archives span more than 215 years. The corpus enables users to retrieve and analyze tasks

32.1.2 Corpora

Resource title	Developer and maintainer/author/host	Web sites	Notes
A Collection of Chinese Corpora and	The University of Leeds 利兹大学		This corpus contains three subcorpuses: the

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Frequency Lists Sharoff 汉语综合语料库		http://corpus.leeds.ac.uk/query-zh.html	Chinese Internet Corpus (280 million words); the Lancaster Corpus of Mandarin Chinese; and the Chinese Business Corpus (30 million words)
Academia Sinica Tagged Corpus of Ancient Chinese 中央研究院上古漢語標記語料庫	Institute of Linguistics, Institute of Information Science, Academia Sinica 中央研究院資訊科學研究所, 語言學研究所/Pei-Chuan Wei, Paul Thompson, Chenghui Liu, Chu-Ren Huang, Keh-Jiann Chen 魏培泉, Paul Thompson, 劉承慧, 黃居仁, 陳克健	http://lingcorpus.iis.sinica.edu.tw/ancient/	This is part of the historical Chinese tagged corpus from Academia Sinica, the first segmented and PoS-tagged corpus of Classical Chinese corpora in the world. The Ancient Chinese Corpus covers Pre-Qin and Western Han texts, which represent some of the oldest (near) vernacular texts of Chinese
Academia Sinica Tagged Corpus of Early Mandarin Chinese 中央研究院近代漢語標記語料庫	Institute of Linguistics, Institute of Information Science, Academia Sinica 中央研究院資訊科學研究所, 語言學研究所/Pei-Chuan Wei, Paul Thompson, Chenghui Liu, Chu-Ren Huang, Keh-Jiann Chen 魏培泉, Paul Thompson, 劉承慧, 黃居仁, 陳克健	http://lingcorpus.iis.sinica.edu.tw/early/	This is part of the historical Chinese tagged corpus from Academia Sinica, the first segmented and PoS-tagged corpus of Classical Chinese corpora in the world. The Middle Chinese Corpus covers the Wei and Jin Dynasties to the Northern and Southern Dynasties and focuses on vernacular texts, including many translated Buddhist texts. This represents the period in which the Chinese language underwent critical changes
Academia Sinica Tagged Corpus of Middle Chinese 中央研究院中古漢語標記語料庫	Institute of Linguistics, Institute of Information Science, Academia Sinica 中央研究院資訊科學研究所, 語言學研究所/Pei-chuan Wei, Paul Thompson,	http://lingcorpus.iis.sinica.edu.tw/middle/	This is part of the historical Chinese tagged corpus from Academia Sinica, the first segmented and PoS-tagged corpus of Classical Chinese

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
	Chenghui Liu, Chu-Ren Huang, Keh-Jiann Chen 魏培泉, Paul Thompson, 劉承慧, 黃居仁, 陳克健		corpora in the world. The Early Mandarin Chinese Corpus covers Yuan to Qing vernacular texts, mostly novels and plays. This represents the period in which Mandarin Chinese was established as the common spoken language for the educated
Asian Scientific Paper Excerpt Corpus-JC 亚洲科技文献摘要语料库	Japan Science and Technology Agency 日本科学技术振兴处 National Institute of Information and Communications Technology, Japan 日本国立情报通信研究所	http://lotus.kuee.kyoto-u.ac.jp/ASPEC/	This platform consists of a Japanese-English paper abstract corpus (ASPEC-JE, three million parallel sentences) and a Japanese-Chinese paper excerpt corpus (ASPEC-JC, 680,000 parallel sentences)
Balanced Corpus of Ancient Chinese (State Language Commission) 国家语言文字工作委员会古籍语料库	State Language Commission 国家语言文字工作委员会	http://www.aihanyu.org/cncorpus/ACindex.aspx	This Ancient Chinese corpus contains nearly 100 million characters, including most of the ancient texts in *Si Ku Quan Shu* (四库全书), which includes different kinds of ancient books and records from the Zhou Dynasty to the Qing Dynasty
Balanced Corpus of Modern Chinese (State Language Commission) 国家语言文字工作委员会现代汉语平衡语料库	State Language Commission 国家语言文字工作委员会	http://www.aihanyu.org/cncorpus/CnCindex.aspx	This Modern Chinese corpus contains 9487 writings. The corpus has 19,455,328 characters (including Chinese characters, alphabets, numbers, punctuation marks, etc.), 12,842,116 tokens (including monosyllabic words, disyllabic words, polysyllabic words, letter words, foreign words, numeric strings, punctuation

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			marks, etc.), and 162,875 types. Among all the 162,875 word forms, 151,300 are Chinese words
BCC Online Corpus 北京语言大学汉语语料库	Beijing Language and Culture University 北京语言大学	http://bcc.blcu.edu.cn/	This Chinese corpus contains about 15 billion characters. It includes writings from many different fields (two billion from newspapers, three billion from literature, three billion from Weibo, three billion from the science and technology fields, two billion from Ancient Chinese, and two billion from other sources)
Bilingual Laws Information System (BLIS) 香港律政司雙語法例資料系統	Department of Justice, The Government of Hong Kong SAR 香港律政司	https://www.elegislation.gov.hk/search	This corpus provides current and past versions of consolidated legislation dating back to June 30, 1997, and PDF copies marked "verified copy" have official legal status
CCL Chinese-English Aligned Parallel Corpus 北京大学中国语言学研究中心汉英对齐平行语料库	Center for Chinese Linguistics, Peking University 北京大学中国语言学研究中心/Weidong Zhan, Rui Guo et al. 詹卫东, 郭锐 等	http://ccl.pku.edu.cn:8080/ccl_corpus/index_bi.jsp	This parallel corpus has 233,589 sentence pairs in 2374 aligned documents. In the corpus, there are 259,425 Chinese sentences and 287,924 English sentences, which cover both written language and spoken language. It also contains practical writings, literature works, and news from different domains, such as politics, science, sports, social culture, industry and commerce, the arts, and movies

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
CCL Online Corpus 北京大学中国语言学研究中心语料库	Center for Chinese Linguistics, Peking University 北京大学中国语言学研究中心/Weidong Zhan, Rui Guo et al. 詹卫东, 郭锐 等	http://ccl.pku.edu.cn:8080/ccl_corpus	The scale of this corpus is 700 million characters. The articles cover different kinds of registers and date from the eleventh century B.C. to the contemporary era. The content is raw materials
Chinese Academic Journal Corpus (Chinese Texts) 中文學術語料庫	National Taiwan Normal University 國立臺灣師範大學	http://140.122.83.220:5566/cqpweb/chineseall/	This corpus contains 1000 articles from the core journals of the humanities and social science fields in Taiwan. Its scale is about nine million characters
Chinese Discourse Annotated Corpus of the Harbin Institute of Technology 哈尔滨工业大学中文篇章关系语料库	Harbin Institute of Technology 哈尔滨工业大学	http://ir.hit.edu.cn/hit-cdtb/	This corpus contains 525 annotated texts. The raw texts are from four kinds of texts in OntoNotes 4.0: "broad news," "magazines," "news wires," and "web." For each of the texts, the corpus can annotate three kinds of discourse relations: clause discourse relations; complex sentence discourse relations; and sentence group discourse relations
Chinese Interlanguage Corpus 華語仲介語語料庫	National Academy for Educational Research 國家教育研究院	http://coct.naer.edu.tw/cqpweb/learners/	The language materials in this corpus mainly came from non-native Chinese speakers' essays (from universities in Taiwan) and a testing corpus that is accredited by the Steering Committee for the Test of Proficiency—Huayu (Taiwan)

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Chinese-English Parallel Corpus 中英雙語平行語料庫	National Taiwan Normal University 國立臺灣師範大學	http://140.122.83.198:7002/	This corpus contains materials from English film subtitles and Hong Kong news. Based on alignment technology, it enables users to retrieve both Chinese and English materials
COCT Spoken Language Corpus COCT 口語語料庫	National Academy for Educational Research 國家教育研究院	http://coct.naer.edu.tw/cqpweb/bl/	This corpus contains a collection of different kinds of programs, including law, politics, military science, finance and economics, current events, science, culture, education, lifestyles, and the arts, from the past 10 years. The language that is used in the programs is limited to Mandarin Chinese in Taiwan
COCT Written Language Corpus COCT 書面語語料庫	National Academy for Educational Research 國家教育研究院	http://coct.naer.edu.tw/cqpweb/yl2016/	This corpus contains articles from many different fields, including philosophy, religion, science, applied science, social sciences, history, geography, language and literature, the arts, finance, and entertainment, from the past 10 years. All these articles came from books with an ISBN code. The corpus also contains news from *United Daily News* and *China Times* from 1999 to 2016. The language that is used in the programs is limited to Mandarin Chinese in Taiwan
Corpus of Political Speeches 政治演講語料庫	Hong Kong Baptist University 香港浸會大學/Kathleen Ahrens	http://digital.lib.hkbu.edu.hk/corpus/index.php	This is a comparable corpus of political speeches from four different jurisdictions: the Corpus of

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			U.S. Presidential Speeches (1789–2015); the Corpus of Policy Addresses by Hong Kong Governors (1984–1996) and Hong Kong Chief Executives (1997–2014); the Corpus of Speeches Given on New Year's Day and Double Tenth Day by Taiwan Presidents (1978–2014); and the Corpus of Reports on the Work of the Government by Premiers of the People's Republic of China (1984–2013)
Early Cantonese Colloquial Texts: A Database 早期粵語口語文獻資料庫	Hong Kong University of Science and Technology 香港科技大學	http://ccl.ust.hk/ccl/useful_resources/useful_resources.html	This corpus contains a collection of seven kinds of Cantonese teaching dictionaries and textbooks compiled by early Western scholars: *Vocabulary of the Canton Dialect*; *Chinese Chrestomathy in the Canton Dialect*; *Cantonese Made Easy* (four editions); and *A Chinese and English Phrase Book in the Canton Dialect*
Early Cantonese Tagged Database 早期粵語標注語料庫	Hong Kong University of Science and Technology 香港科技大學	http://ccl.ust.hk/ccl/useful_resources/useful_resources.html	This corpus contains is a collection of 10 literatures, including *The Gospel According to St. Mark* (in English and Cantonese), *Easy Phrases in the Canton Dialect of the Chinese Language*, *A Chinese and English Phrase Book in the Canton Dialect*,

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			Progressive and Idiomatic Sentences in Cantonese Colloquial, and so on. Its scale is about 160,000 characters
English-Chinese Parallel Concordancer 漢英平行語料庫	The Hong Kong Institute of Education 香港教育學院	http://ec-concord.ied.edu.hk/paraconc	The scale of this corpus is 576,724 Chinese characters (413,823 words). It enables users to search for concordances and see the translation of texts
HSK Learner Corpus of Composition Texts 北京语言大学 HSK 汉语水平考试动态作文语料库	Beijing Language and Culture University 北京语言大学	http://bcc.blcu.edu.cn/hsk	This is an interlanguage corpus collection of 11,569 Chinese essays (about 4.3 million characters) that were written by non-native speakers. All the essays were collected from the essay test of the Chinese Proficiency Test (HSK) from 1992 to 2005
Intelligent Collocation Search Engine 智慧搭配詞搜尋引擎	National Taiwan Normal University 國立臺灣師範大學	http://140.122.83.243:8000/ICE/Index.htm	This corpus contains a collection of articles from the fields of business, journalism, justice, academic research, finance and economics, and travel. It has the function of collocation word retrieval
Korean-Chinese Parallel Corpus 韩汉平行语料库	Korea Advanced Institute of Science and Technology 韩国科学技术学院	http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus	This platform has many corpora, including a Chinese-English-Korean multilingual corpus that contains 60,000 sentences, a DongaKorean-English-Japanese-Chinese multilingual newspaper corpus that

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			contains 1791 files, and so on
Learn Chinese Online via Podcast and MP3 洛杉矶汉语学习中心语音学习资料库	Los Angeles Chinese Learning Center, USA 洛杉矶汉语学习中心	http://chinese-school.netfirms.com/learn-Chinese-online.html	This recorded corpus enables users to learn Chinese online. It contains recordings of pinyin, vocabularies, phrases, and so on
Linguistic Variation in Chinese Speech Communities (LIVAC) 香港城市大學泛華語共時同題語料庫	City University of Hong Kong 香港城市大學	http://www.livac.org/search.php	This is a synchronous Chinese corpus, with a scale of 2.5 billion characters. Six hundred million characters have been processed and analyzed. The texts are from different Chinese communities. It also possesses an ever-expanding Pan-Chinese dictionary of more than two million entries
NICT Japanese-Chinese Parallel Corpus 日本国立情报通信研究所日汉平行语料库	National Institute of Information and Communications Technology, Japan 日本国立情报通信研究所	http://universal.elra.info/product_info.php?cPath=42_43&products_id=2044	This corpus contains 38,383 sentence pairs collected from Japanese newspapers and manually translated into Chinese. Its scale is 947,066 Japanese words and 877,859 Chinese words, all encoded in Unicode. The corpus is aligned at word and phrase levels. The texts are segmented and annotated with part-of-speech tags, morphological structures, and syntactic structures
NTU Multilingual Corpus 南洋理工大学多语语料库	Nanyang Technological University 南洋理工大学	http://compling.hss.ntu.edu.sg/ntumc/	This corpus contains 375,000 words (15,000 sentences) in six languages (English, Chinese, Japanese, Korean, Indonesian, and Vietnamese). It enables

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			users to retrieve words by concepts, word, lemmas, parts-of-speech, etc.
Sinica Corpus 中央研究院現代漢語平衡語料庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, and Institute of Linguistics, Academia Sinica 中央研究院資訊科學研究所, 語言學研究所 中文詞知識庫小組/Keh-Jiann Chen, Chu-Ren Huang 陳克健, 黃居仁	http://asbc.iis.sinica.edu.tw/ http://lingcorpus.iis.sinica.edu.tw/modern/ (five million word version)	The Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus) is the first PoS-tagged balanced corpus of Mandarin Chinese, as well as the first Chinese corpus on the web (since 1997). The current version (4.0) contains more than 10 million words (with more than 14 million characters). The corpus search interface allows KWIC searches (both with or without PoS) and has many collocation calculation tools, such as Mutual Information (MI) calculation
Spoken Language Corpus of Chinese Learners (Spoken Language Test) 華語學習者口語語料庫(口語考試)	National Taiwan Normal University 國立臺灣師範大學	http://140.122.83.243/mp3c/	The spoken language materials in this corpus are from the recorded documents of the spoken language test in the Test of Chinese as a Foreign Language (TOCFL, the test for Teaching Chinese as a Second Language [TCSL] learners) from 2008 to April 2011. Its scale is 770,000 characters. The corpus enables users to retrieve the usage conditions and phonological representations of learners
Taiwan Corpus of Child Mandarin	National Taiwan University, Education	http://tccm.corpus.eduhk.hk/	This corpus contains the CHILDES-style

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
(TCCM) 台灣兒童語言語料庫	University of Hong Kong 國立臺灣大學 香港教育大學/Hintat Cheung 張顯達		language acquisition corpus Children Learning Mandarin in Taiwan
Taiwan Presidential Corpus 遷台後歷屆總統元旦及國慶文告資料庫	Institute of Linguistics, Academia Sinica 中央研究院語言學研究所/Chu-Ren Huang, Kathleen Ahrens, Weilun Lu 黃居仁, Kathleen Ahrens, 呂維倫	http://140.109.19.114/president/	This corpus consists of all major announcements on New Year's Day, national days, etc. by the first four presidents of Taiwan
The Hong Kong Bilingual Child Language Corpus 香港雙語兒童語言資料庫	Chinese University of Hong Kong 香港中文大學	http://www.cuhk.edu.hk/lin/home/bilingual.htm	This corpus contains longitudinal speech data from six bilingual children exposed to Cantonese and English from birth. These children grew up in a one-parent-one-language environment where each parent was a native speaker of the respective language
The UCLA Written Chinese Corpus 加州大学洛杉矶分校汉语书面语语料库	University of California Los Angeles 加州大学洛杉矶分校 University Centre for Computer Corpus Research on Language of Lancaster University 兰开斯特大学计算机语料库及语言研究中心/Hongyin Tao, Richard Xiao 陶红印, 肖忠华	http://www.lancaster.ac.uk/fass/projects/corpus/UCLA/default.htm	This corpus is the Chinese counterpart of the Freiburg-LOB Corpus of British English (FLOB) and the Brown corpora of British and American English. The samples in the corpus were collected from written Modern Chinese available from the Internet from 2000 to 2012. Its scale is 1,119,930 words
Web-based Chinese Corpus 臺灣網路語料庫	National Taiwan Normal University 國立臺灣師範大學	http://140.122.83.220:5566/cqpweb/chacademicjournal/	Based on the concept of "Web as Corpus," this corpus directly uses web resources as materials. Its scale is 400 million characters
Wikidata Corpus 维基百科语料库	Wikimedia Foundation 维基媒体基金会	https://github.com/Samurais/wikidata-corpus	This corpus trains data from Chinese Wikidata using the

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			word2vec method for word-embedding tasks
Written Language Corpus of Chinese Learners (Writing Practice and Writing Test) 華語學習者書面語語料庫(寫作練習與 寫作考試)	National Taiwan Normal University 國立臺灣師範大學	http://kitty.2y.idv.tw/~hjchen/cwrite-mtc/main.cgi http://kitty.2y.idv.tw/~hjchen/cwrite/main.cgi	These two corpuses contain Chinese essays written by non-native Chinese learners. The first one is a collection of the practice writing essays of non-native Chinese learners in National Taiwan Normal University from 2010 to 2012. Its scale is two million characters. The second one is comprised of materials from the writing test of the TOCFL (the test for TCSL learners) from 2006 to 2012. The scale of the corpus is 1.5 million characters. Both corpuses enable users to perform online retrieval

32.1.3 Lexical Resources

Resource title	Developer and maintainer/author/host	Web sites	Notes
Data Bank of Common First and Last Characters of Chinese Words 常用詞首、詞尾字資料庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學研究所 中文詞知識庫小組/Keh-Jiann Chen, Chu-Ren Huang 陳克健, 黃居仁	http://140.109.19.103/affix/	This knowledge base is a collection of 4025 commonly used first and last characters of nouns and verbs from the Sinica Corpus. All the characters are provided with senses, categories in *Tong Yi Ci Ci Lin* (for nouns), word formation rules (for verbs), and examples

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Lexicon of Pre-Qin Oracle, Bronze Inscription and Bamboo Scripts 先秦甲骨金文簡牘詞彙庫	Institute of History and Philology, Academia Sinica 中央研究院歷史語言研究所/Chao-Jung Chen, Bo-Sheng Jhong, Guo-Hua Yuan, Ming-chorng Hwang 陳昭容, 鍾柏生, 袁國華, 黃銘崇	http://inscription.asdc.sinica.edu.tw/	This is an online lexicon of words from original excavated Pre-Qin scripts written on oracle bones, bronze inscription, and bamboo. Actual graphic forms, variants, and PoS are presented. This is one of the Language Archives projects
Online Dictionary of Taiwan Sign Language 台灣手語線上辭典	The Taiwan Sign Language Research Group, Institute of Linguistics, National Chung Cheng University 國立中正大學語言學研究所 臺灣手語研究小組/James H.-Y. Tai and S. C. Jane Tsay	http://tsl.ccu.edu.tw/web/chinese/	This digital dictionary contains video files of Taiwan Sign Language words
Word Index of the Balanced Corpus of Modern Chinese (State Language Commission) 国家语言文字工作委员会现代汉语平衡语料库字词索引	State Language Commission 国家语言文字工作委员会	http://www.aihanyu.org/cncorpus/WDindex.aspx	This is a word list extracted from the Balanced Corpus of Modern Chinese (State Language Commission). Each of the items has information on part-of-speech and word frequency

32.1.4 Wordnet/Ontology

Resource title	Developer and maintainer/author/host	Web sites	Notes
Chinese Open WordNet 汉语开放词网	Nanyang Technological University 南洋理工大学	http://compling.hss.ntu.edu.sg/cow/	This wordnet is based on the concepts of the Princeton WordNet and the Global WordNet Grid. It contains 42,315 synsets, 79,812 senses, and 61,536 unique words

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Chinese WordNet 中文詞彙網路	Institute of Linguistics, Academia Sinica 中央研究院語言學研究所 National Taiwan University 臺灣大學/Chu-Ren Huang, Shu-Kai Hsieh 黃居仁, 謝舒凱	http://cwn.ling.sinica.edu.tw/ http://lope.linguistics.ntu.edu.tw/cwn/ Version 2.0. http://lope.linguistics.ntu.edu.tw/cwn2/	This wordnet, based on the idea of English WordNet, aims to offer integrated materials to distinguish Chinese word senses. Based on the construct of lexical semantics and ontology, it is an effective reference for linguistics researches. It has 5600 word forms and 13,160 senses
CoreNet 核心词网	Korea Advanced Institute of Science and Technology 韩国科学技术学院	http://semanticweb.kaist.ac.kr/home/index.php/CoreNet_Corpus	This is a net of words based on their semantics. It contains a Word-to-Concept System that includes 23,938 Korean words with 58,985 senses and 34,409 Chinese words with 39,352 senses. It also contains a Predicate Case Frame that includes 973 Korean words with 1909 senses and 368 Chinese words
E-HowNet Ontology 廣義知網知識本體架構	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen 陳克健	http://ehownet.iis.sinica.edu.tw/index.php	This knowledge base connects more than 90,000 entries in the Chinese Knowledge Information Processing (CKIP) Chinese Lexical Knowledge Base to HowNet nodes. It aims to build a vocabulary knowledge base that can express the relationships between

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			concepts and the relationships among the properties of the concepts
Emotion Ontology of Chinese Words 情感词汇本体库	Dalian University of Technology 大连理工大学	http://ir.dlut.edu.cn/EmotionOntologyDownload	This ontology describes Chinese words and phrases from many different aspects, including part-of-speech, emotion category, emotion intensity, polarity, etc. The ontology is based on Ekman's emotion classification system (six basic emotions: anger, disgust, fear, happiness, sadness, and surprise). It classifies emotions into seven main classes and 21 subclasses
Hantology 漢字知識本體	Institute of Linguistics, Academia Sinica 中央研究院語言學研究所/Ya-Min Chou, Chu-Ren Huang 周亞民, 黃居仁	http://hantology.ling.sinica.edu.tw/index.htm	This ontology enables users to search for Chinese characters by semantic symbol or input a Chinese character to get the main semantic symbol, and it is composed of semantic symbols and their original senses
The Academia Sinica Bilingual Ontological WordNet (Sinica BOW) 中央研究院中英雙語知識本體詞網	Institute of Linguistics, Institute of Information Science, Academia Sinica 中央研究院語言學研究所, 資訊科學研究所/Chu-Ren Huang 黃居仁	http://bow.ling.sinica.edu.tw	This platform integrates three main resources, which are WordNet, Suggested Upper Merged Ontology (SUMO), and the English-Chinese Translation Equivalents Database (ECTED). Sinica

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			BOW functions both as an English-Chinese bilingual wordnet and bilingual lexical access to SUMO

32.1.5 Treebanks

Resource title	Developer and maintainer/author/host	Web sites	Notes
Sinica Treebank Version 3.0 中文句結構樹資料庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen, Chu-Ren Huang et al. 陳克健, 黃居仁 等	http://turing.iis.sinica.edu.tw/treesearch/	This database includes six documents, 61,087 Chinese tree graphs, and 361,834 words. All the language materials were extracted from the Sinica Corpus. The tree graphs were automatically generated and manually amended. These graphs show the syntactic and semantic information of sentences

32.1.6 Chinese Information Processing Tools

Resource title	Developer and maintainer/author/host	Web sites	Notes
Chinese Segmentation System 中文斷詞系統	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen 陳克健	http://ckipsvr.iis.sinica.edu.tw/	This language tool can extract unknown words from inputted texts and segment the texts (including unknown words). It not only shows the segmentation results and the unknown word list but also shows the operational procedures of the program

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
Chinese Synonyms for Natural Language Processing and Understanding 中文近义词工具包	Hai Liang Wang, Hu Ying Xi	https://github.com/huyingxi/Synonyms	This toolkit can extract Chinese synonyms automatically and calculate the similarity between two Chinese words or sentences
CKIP Chinese Parser 中文剖析系統	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen 陳克健	http://parser.iis.sinica.edu.tw/	This is an online parser that can segment and parse inputted sentences automatically. It can also automatically label the semantic roles of the sentence components
FudanNLP 复旦大学自然语言处理工具包	Fudan University 复旦大学	https://github.com/FudanNLP/fnlp	This toolkit has the functions of Chinese word segmentation, part-of-speech tagging, named entity recognition, keyword extraction, dependency grammar analysis, text categorization, and so on
HanLP: Han Language Processing 汉语言处理系统	Shanghai Linrun Information Technology Ltd. 上海林原信息科技有限公司	http://hanlp.linrunsoft.com/	This is an open-source language tool that can perform multiple tasks, including Chinese word segmentation, part-of-speech tagging, named entity recognition, keyword extraction, auto-abstraction, dependency grammar analysis, text categorization, and so on
ictclas4j Segmenter ictclas4j 中文分词系统	Ying Jiang 姜赢	https://code.google.com/archive/p/ictclas4j/	This segmenter is an open-source project based on FreeICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)
Language Cloud of the Harbin Institute of Technology 哈尔滨工业大学语言云	Harbin Institute of Technology 哈尔滨工业大学	http://www.ltp-cloud.com/	This is a language processing platform based on the “Language Technology Platform” (LTP) of the

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			Harbin Institute of Technology. It offers plenty of NLP technologies such as Chinese word segmentation, part-of-speech tagging, named entity recognition, dependency parsing, and semantic role labeling. It comes in the Python version and the Docker version at https://github.com/HIT-SCIR/pyltp and https://github.com/HIT-SCIR/ltp, respectively
Lingpipe	Alias-i	http://alias-i.com/lingpipe/demos/tutorial/chineseTokens/read-me.html	This is a toolkit for processing text using computational linguistics methods. It can be used to perform tasks such as finding the names of people, organizations, or locations in the news; carrying out classification tasks for Twitter search results automatically; and giving spelling suggestions for queries
NAER Segmentor 國家教育研究院中文分詞系統	National Academy for Educational Research 國家教育研究院	https://github.com/naernlp/Segmentor	This segmentor uses the part-of-speech marking system of Sinica. It can segment traditional Chinese text very quickly, but users cannot use their own dictionaries
Natural Language Toolkit (NLTK) 自然语言处理工具包	Department of Computer and Information Science, University of Pennsylvania 宾夕法尼亚大学计算机与信息科学系/Steven Bird, Edward Loper	http://www.nltk.org/install.html	This is a suite of libraries and programs for symbolic and statistical natural language processing that is written in Python. It has the functions of classification, tokenization, stemming, tagging, parsing,

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
			and semantic reasoning. It contains Chinese data (treebank, segmenter, parser, etc.) from the CKIP group of Academia Sinica
NiuParser 1.3.0 中文句法语义分析系统	Northeastern University, China 东北大学	http://www.niuparser.com	This platform has the functions of Chinese word segmentation, part-of-speech tagging, named entity recognition, machine translation, public opinion analysis, dependency grammar analysis, semantic role labeling, automatic writing, knowledge graphs, and a question-answering system.
NiuTrans 东北大学统计机器翻译系统	Natural Language Processing Laboratory of Northeastern University, China 东北大学自然语言处理实验室	http://www.niutrans.com/niutrans/NiuTrans.ch.html	This is an open-source statistical machine translation system that is written in C++
NLPIR-ICTCLAS Chinese Lexical Analysis System NLPIR-ICTCLAS 汉语分词系统	Hua-Ping Zhang 张华平	http://ictclas.nlpir.org	This segmenter has the functions of Chinese and English word segmentation, part-of-speech tagging, named entity recognition, new word recognition, keyword extraction, and Weibo analysis. It permits users to use their own dictionaries
Polyglot	Rami Al-Rfou	https://pypi.python.org/pypi/polyglot	This is a natural language pipeline that supports massive multilingual applications, including tokenization, language detection, named entity recognition, part-of-speech tagging, sentiment analysis, word embedding, morphological analysis, and transliteration

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes
SnowNLP: Simplified Chinese Text Processing 简体中文文本处理工具包		https://github.com/isnowfy/snownlp	This is a class lib for Python. It offers the functions of Chinese word segmentation, part-of-speech tagging, emotion analysis, text categorization, keyword and abstract extraction, and so on
The Stanford Parser 斯坦福句法分析器	Stanford University 斯坦福大学	https://nlp.stanford.edu/software/lex-parser.html	This probabilistic parser gains knowledge from training sets and produces the most likely analysis of new sentences. This is a multilingual parser that can parse different languages, including Chinese, English, German, French, and so on
The Stanford Word Segmenter 斯坦福分词系统	Stanford University 斯坦福大学	https://nlp.stanford.edu/software/segmenter.shtml	This segmenter can perform Chinese word segmentation tasks automatically

32.2 Licensable Resources

32.2.1 Integrated Resources

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese LDC 中文语言资源联盟	Chinese Information Processing Society of China 中国中文信息学会	http://www.chineseldc.org/resource_list.php	This platform creates and collects systematic speech data that can be used in lexicon, language corpus, and instrumental reference researches. It can distribute existing data to departments for education, scientific research, governmental purposes, and the development of industrial technology	Apply and pay

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese Lexicons 中文詞知識庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen 陳克健	http://ckip.iis.sinica.edu.tw:8080/license/	This platform has several language resources, including Chinese Parser, Sinica Treebank, a Chinese segmentation system, E-HowNet, Chinese word sketches, a public opinion analysis system, a Chinese word-embedding corpus, a Chinese segmentation corpus, and a Chinese news corpus	Apply
Linguistic Data Consortium (LDC) 语言资源联盟	University of Pennsylvania 宾夕法尼亚大学	https://www.ldc.upenn.edu/language-resources	This platform contains a great deal of different language resources that can meet different requirements of users	Apply and pay
Natural Language Toolkit (NLTK) Corpora 自然语言处理工具包数据平台	Natural Language Toolkit (NLTK) 自然语言处理工具包	http://www.nltk.org/nltk_data/	This platform contains dozens of corpora and trained models that can help users to use the Natural Language Toolkit more efficiently	Apply

32.2.2 Corpora

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese Gigaword Corpus Edition 5.0 中文亿词语料库第五版	Linguistic Data Consortium, University of Pennsylvania 宾夕法尼亚大学语言数据联盟/Robert Parker et al.	https://catalog.ldc.upenn.edu/LDC2011T13	The Chinese Gigaword Corpus is the largest Chinese Corpus in the world collected by the LDC. Each new edition is larger than the previous one. Edition 5.0 contains over eight billion (8000 million) characters from *Agence France Presse*, the Central News	Apply and pay

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
			Agency (Taiwan), China News Service, *Guangming Daily*, *People's Daily*, *People's Liberation Army Daily*, the Xinhua News Agency, and *Zaobao Newspaper* (Singapore). The only version of the Chinese Gigaword Corpus is Version 2.0	
Chinese Speech Corpus 中文語音語料庫	National Taiwan Normal University 國立臺灣師範大學	http://140.122.83.243/ac/query.php	This corpus contains materials from several Chinese teleplays, which users can retrieve online	Register
CORPUS Program 中文新聞語料庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen, Chu-Ren Huang et al. 陳克健, 黃居仁 等	http://www.aclclp.org.tw/use_cp_c.php	This corpus has 14 million characters. The material was collected from *United Daily News*, *China Times*, *Liberty Times*, and *CommonWealth Magazine*	Apply and pay
Standard Segmentation Corpus 中文分詞語料庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Chu-Ren Huang, Keh-Jiann Chen 黃居仁, 陳克健	http://www.aclclp.org.tw/use_ssc_c.php	This corpus has two million words, which are segmented only; the words do not have part-of-speech tags.	Apply and pay
Tagged Chinese Gigaword Corpus Version 2.0 中文億詞標注語料庫第二版	Linguistic Data Consortium, University of Pennsylvania 賓夕法尼亞大學語言資料聯盟/Chu-Ren Huang et al. 黃居仁 等	https://catalog.ldc.upenn.edu/LDC2009T14	The Tagged Chinese Gigaword Corpus Version 2.0, processed and PoS tagged by the Academia Sinica team, is the largest tagged	Apply and pay

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
			Chinese Corpus in the world. Collected by the LDC, version 2.0 is from the Central News Agency (Taiwan), the Xinhua News Agency, and *Zaobao Newspaper* (Singapore). This version contains more than 1100 million characters and more than 831 million words	
The Babel English-Chinese Parallel Corpus Babel 英汉平行语料库	Lancaster University 兰开斯特大学/ Richard Xiao 肖忠华	http://www.lancaster.ac.uk/fass/projects/corpus/babel/babel.htm	This corpus contains 327 English articles and their translation in Mandarin Chinese. These articles are from the *World of English* and *Time*. The scale of this corpus is 544,095 words (253,633 English words and 287,462 Chinese tokens). The corpus is annotated with part-of-speech tags. Sentence alignment was performed automatically and corrected by hand	Apply
The Lancaster Corpus of Mandarin Chinese 兰开斯特大学中文语料库	Lancaster University 兰开斯特大学/ Tony McEnery, Richard Xiao	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC	This corpus is the Chinese version of the FLOB. Its contents are segmented and annotated with part-of-speech tags	Apply
The Lancaster Los Angeles Spoken Chinese Corpus 兰开斯特洛杉矶汉语口语语料库	Lancaster University 兰开斯特大学/ Richard Xiao 肖忠华	http://www.lancaster.ac.uk/fass/projects/corpus/LLSCC/	This is a corpus of spoken Mandarin Chinese. Its scale is 1,002,151 words from dialogues and monologues, with 73,976 sentences and 49,670 paragraphs. The materials are from	Apply

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
			conversations, telephone calls, play and movie transcripts, TV talk show transcripts, debate transcripts, oral narratives, and edited oral narratives	
The PDC2000 Corpus of Chinese News Text 2000 年《人民日报》全年语料库	Lancaster University 兰开斯特大学/ Richard Xiao 肖忠华	http://www.lancaster.ac.uk/fass/projects/corpus/pdc2000/	This corpus contains a whole years' worth of data from the *People's Daily* (2000). Its scale is about 15 million words in 366 files. Each file consists of a corpus header and the corpus text proper. The content of the corpus is annotated with part-of-speech tags	Apply

32.2.3 Lexical Resources

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese Electronic Dictionary 中文詞庫(八萬目詞)	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen, Chu-Ren Huang 陳克健, 黃居仁	http://www.aclclp.org.tw/use_ced_c.php	This electronic dictionary has 80,000 entries, which include common words, commonly used proper nouns, idioms, idiomatic phrases, derivates, heterographies, terms, and Ancient Chinese words. Each entry includes phonetic notation, frequency, part-of-speech, and semantic class.	Apply and pay

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Reference Lexicon for Segmentation Standard Dictionary 中文分詞詞庫	Chinese Language and Knowledge Processing Group, Academia Sinica 中央研究院中文詞知識庫小組/Chu-Ren Huang 黃居仁	http://www.aclclp.org.tw/use_rlssd_c.php	This word list was extracted from the Standard Segmentation Corpus. It has 42,138 entries, with their frequencies.	Apply and pay
Sinica Chinese Core Vocabulary 中央研究院中文核心詞彙表	Chinese Language and Knowledge Processing Group, Academia Sinica 中央研究院中文詞知識庫小組/Chu-Ren Huang et al. 黃居仁等	http://www.aclclp.org.tw/use_sccv_c.php	This word list includes 1121 high-frequency Chinese words. These words are the intersection of the first 2000 high-frequency words in the Sinica Corpus and the Modern Chinese Dialogue Speech Corpus. Each word includes part-of-speech, frequency (in both corpuses), frequency rank (in both corpuses), English translation, and Chinese and English example sentences.	Apply and pay
The Grammatical Knowledge-base of Contemporary Chinese 北京大学现代汉语语法信息词典	Peking University 北京大学/Shi-Wen Yu et al. 俞士汶 等	http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/EDQWIL	This dictionary is based on Prof. Zhu De-xi's theories and contains 73,000 items. For each item, it offers homographs, pinyin, senses, multi-category conditions, and much more syntactic information. It aims to analyze and generate Chinese sentences automatically.	Apply
Word List with Accumulated Word Frequency in the Sinica Corpus 中央研究院平衡語料庫詞集及詞頻統計	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央	http://www.aclclp.org.tw/use_wlawf_c.php	This word list was extracted from the Sinica Corpus. It shows the part-of-speech, word frequency, and	Apply and pay

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
	研究院資訊科學所中文詞知識庫小組/Keh-Jiann Chen, Chu-Ren Huang 陳克健, 黃居仁		cumulative frequency of each item.	

32.2.4 Wordnet/Ontology

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
HowNet 知网	Language Knowledge Department, Computer and Language Information Research Centre, Chinese Academy of Sciences 中国科学院计算机语言信息中心语言知识研究室/Zhendong Dong 董振东	http://www.keenage.com/html/e_index.html	This is an online common-sense knowledge base of interconceptual relations and interattribute relations of concepts as connoted in Chinese lexicons and their English equivalents	Apply and pay
The Academia Sinica Bilingual Ontological Database 中英雙語知識本體資料庫	Institute of Linguistics, Academia Sinica, 中央研究院/Chu-Ren Huang 黃居仁	http://www.aclclp.org.tw/use_bd_c.php	This database contains a bilingual ontology of about 110,000 Chinese words. The ontology is based on the Institute of Electrical and Electronics Engineers (IEEE)-approved SUMO. It offers not only the bilingual ontology of concepts but also the infrastructure of knowledge management, which can transfer knowledge from different sources into interoperable information	Apply and pay

32.2.5 Treebanks

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese Syntactic Treebank of Tsinghua University 清华大学汉语句法树库	Tsinghua University 清华大学	http://cslt.riit.tsinghua.edu.cn/~qzhou/papers/TCTScheme.pdf	This treebank contains parsed texts from literature, academic fields, news, and practical writings. Its scale is about one million words	Apply and pay
CoNLL X Shared Task Chinese Treebank CoNLL X 中文句結構樹資料庫	Chinese Language and Knowledge Processing Group, Institute of Information Science, Academia Sinica 中央研究院資訊科學研究所中文組實驗室中文詞知識庫小組	http://www.aclclp.org.tw/use_conll_c.php	The materials in this treebank are from the Chinese CoNLL X corpus competition (2006). The tree graphs from the Sinica Treebank are presented in dependency-tree form. The test corpus is free to download, but the training corpus requires payment to use	Apply and pay

32.2.6 Sketch Engine

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese Word Sketch Engine 中文詞彙特性速描系統	Institute of Linguistics, Academia Sinica 中央研究院/Chu-Ren Huang, Adam Kilgarriff et al.	http://wordsketch.ling.sinica.edu.tw/	This system is connected to the LDC Chinese Gigaword Corpus (1.4 billion characters). It can provide users with word sketch information, grammatical relations, synonym analysis, and other lexical and syntactic knowledge of inputted words.	Register
Word Sketch Engine 词汇特征素描系统	Lexical Computing Limited, UK 英国词汇计算有限公司	https://www.sketchengine.co.uk/	Word Sketch Engine contains 500 corpora, each of which contains up to 30 billion words, in more than 90 languages. These corpora show users the grammatical and collocational behavior of words.	Free for 30 days, then pay

32.2.7 Evaluation Resources

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
Chinese Information Retrieval Benchmark Version 3.0 中文資訊檢索標杆測試集第三版	Department of Library and Information Science, National Taiwan University 國立臺灣大學圖書資訊學系/Chen Kuang-Hua 陳光華	http://www.aclclp.org.tw/use_cir_c.php	The material in this corpus was collected according to information retrieval evaluation theories. The corpus aims to be a reliable testing resource for Chinese information retrieval. It includes three parts: documents set, topics (questions) set, and answers set	Apply and pay
DoLWS-MAN: Database of Word-level Statistics (Mandarin)	The Hong Kong Polytechnic University 香港理工大學/Karl David Neergaard, Hongzhi Xu, Chu-Ren Huang, 許洪志, 黃居仁	Available soon from the LDC, University of Pennsylvania	This database provides the lexical characteristics of a descriptive and statistical nature for Mandarin Chinese words and non-words. It was designed for researchers who are particularly concerned with the language processing of isolated words. The database is basically a set of phonological neighborhood data in Mandarin Chinese collected through experiments and fitted with statistical models, and it is searchable by Speech Assessment Methods Phonetic Alphabet (SAMPA), characters, or pinyin	To be determined

(continued)

Resource title	Developer and maintainer/author/host	Web sites	Notes	How to use
EVALution and EVALution-MAN	The Hong Kong Polytechnic University 香港理工大學/Enrico Santus, Hongchao Liu, Chu-Ren Huang et al. 劉洪超, 黃居仁 等	Available soon from the LDC, University of Pennsylvania	This repository contains different versions of EVALution, a dataset containing Semantic Relations and Metadata for Training and Evaluating Distributional Semantic Models. EVALution contains English data while EVALution-MAN contains Mandarin Chinese data	To be determined
SemTransCNC 1.0: Semantic Transparency Dataset of Chinese Nominal Compound	The Hong Kong Polytechnic University 香港理工大學/Shichang Wang, Chu-Ren Huang et al. 王世昌, 黃居仁 等	To be available soon from LDC, University of Pennsylvania	This dataset was built using a series of Mechanical Turk-based experiments. It consists of the overall and the constituent semantic transparency (OST and CST, respectively) data of 1176 dimorphemic Chinese nominal compounds that consist of free morphemes and have mid-range frequencies	To be determined
SIGHAN Bakeoff 2012 SIGHAN 繁體中文剖析資料集 2012 版	Academia Sinica 中央研究院	http://www.aclclp.org.tw/use_bakeoff_c.php	This database has three parts: training set, dry-run set, and testing set. The first two sets are extended from the Sinica Treebank, while the testing set has 1000 new Chinese sentences (annotated), which can be used as testing materials for parsing and Semantic Role Labeling tasks	Apply and pay

32.3 Published Resources/Technical Reports[1]

32.3.1 Segmentation and Part-of-Speech Analysis

1. Chinese Language and Knowledge Processing Group 中央研究院資訊科學所中文詞知識庫小組. 1993. Technical report: Chinese part of speech analysis 詞庫小組技術報告 - 中文詞類分析. In *CKIP technical report* 中文詞知識庫小組技術報告. Taipei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/9305_2013%20revision.pdf
2. Computational Linguistics Laboratory of the Institute of Applied Linguistics 教育部语言文字应用研究所计算语言学研究室. 2001. Standardized set of Chinese POS markers for computational uses 信息处理用现代汉语词类标记集规范. *Applied Linguistics* 语言文字应用. 03:16–20.
3. Huang, Chu-Ren, Keh-Jiann Chen, and Chinese Language and Knowledge Processing Group 黃居仁, 陳克健, 中央研究院資訊科學所中文詞知識庫小組. 1996. Technical report: SouWenJieZi—Chinese word boundary research and word segmentation specification for information processing 詞庫小組技術報告 -「搜」文解字-中文詞界研究與資訊用分詞標準. In *CKIP technical report* 中文詞知識庫小組技術報告. Taipei: Academia Sinica. https://www.researchgate.net/publication/301699745_souwenjiezi-_zhongwencijieyanjiuyuzixunyongfencibiaozhun
4. Huang, Chu-Ren, Shu-Kai Hsieh, and Keh-Jiann Chen. 2017. *Mandarin Chinese words and parts of speech: A corpus-based study*. Routledge.
5. Huang, Chu-Ren, and Nianwen Xue. 2012. Words without boundaries: Computational approaches to Chinese word segmentation. *Language and Linguistics Compass* 6(8):494–505.
6. Institute of Computational Linguistics of Peking University (ed.) 北京大学计算语言学研究所 编制. 1999. *Contemporary Chinese corpus processing specification—Word segmentation and part-of-speech tagging* 现代汉语语料库加工规范——词语切分与词性标注. Beijing: Institute of Computational Linguistics of Peking University (technical report, unpublished). http://www.docin.com/p-1074544403.html
7. Jin, Guang-Jing, Hang Xiao, and Li Fu 靳光瑾, 肖航, 富丽. 2005. Standardized set of Chinese POS markers for computational uses (revised ed.) 信息处理用现代汉语词类标记规范(修订). In *Proceedings of the 4th National Applied Linguistics Workshop* 第四届全国语言文字应用学术研讨会论文集, 9. Beijing: Institute of Applied Linguistics, Ministry of Education.
8. Liu, Yuan, Qiang Tan, and Xu-Kun Shen 刘源, 谭强, 沈旭昆. 1994. *Contemporary Chinese language word segmentation specification and automatic word*

[1] For papers and reports that were written in English, the titles and publication information are listed in English only. For papers and reports that were written in Chinese, the titles and publication information are listed in both English and Chinese.

segmentation methods for information processing 信息处理用现代汉语分词规范及自动分词方法. Beijing: Tsinghua University Press.
9. Yu, Shi-Wen, Hui-Ming Duan, Xue-Feng Zhu, Bin Sun, and Bao-Bao Chang 俞士汶, 段慧明, 朱学锋, 孙斌, 常宝宝. 2003. Corpus processing specification of Peking University: Segmentation, part-of-speech tagging and phonetic notation 北大语料库加工规范: 切分· 词性标注· 注音. *Journal of Chinese Language and Computing* 汉语语言与计算学报 13(2):121–158.

32.3.2 Word List/Dictionary

1. Huang, Chu-Ren, Keh-Jiann Chen, Zhao-Ming Gao, Feng-Yi Chen, and Jheng-Jhong Shen. 1998. Technical report: Accumulated frequency of Sinica Corpus. In *CKIP technical report.* Tai Pei: Academia Sinica.
2. Huang, Chu-Ren, Keh-Jiann Chen, Zhao-Ming Gao, Feng-Yi Chen, and Jheng-Jhong Shen 黃居仁, 陳克健, 高照明, 陳鳳儀, 沈正中. 1998. Technical report: Word frequency dictionary 詞庫小組技術報告 - 詞頻詞典. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/9801_2013.pdf
3. Ji, Chun-Sing 紀春興. 1995. Technical report: Mandarin Chinese character frequency list based on national phonetic alphabets 詞庫小組技術報告 - 注音檢索現代漢語字頻表. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/9501_2013.pdf
4. McEnery, Tony, Richard Xiao, and Paul Rayson. 2015. *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. Routledge.
5. Shen, Jheng-Jhong, Zhao-Ming Gao, and Chu-Ren Huang 沈正中, 高照明, 黃居仁. 1998. Technical report: An English-Chinese glossary of NLP and CL related terms 詞庫小組技術報告 - 自然語言處理及計算語言學相關術語中英對譯表. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica.
6. Wei, Pei-Chuan, Cheng-Huei Liou, Chu-Ren Huang, and Syueh-Ru Wu 魏培泉, 劉承慧, 黃居仁, 吳雪如. 2000. Technical report: Verb-complement word list of novels of Ming and Qing Dynasties 詞庫小組技術報告 - 明清小說動補詞語表. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica.
7. Wei, Pei-Chuan, Cheng-Huei Liou, Pu-Sen Tan, and Chu-Ren Huang 魏培泉, 劉承慧, 譚樸森, 黃居仁. 1994. Technical report: Character frequency list of Ancient Chinese 詞庫小組技術報告 - 古漢語字頻表. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica.
8. Wei, Pei-Chuan, Cheng-Huei Liou, Pu-Sen Tan, and Chu-Ren Huang 魏培泉, 劉承慧, 譚樸森, 黃居仁. 1997. Technical report: Word frequency list of Ancient Chinese 詞庫小組技術報告 - 古漢語詞頻表(甲). In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica.

9. Wei, Pei-Chuan, Cheng-Huei Liou, Pu-Sen Tan, and Chu-Ren Huang 魏培泉, 劉承慧, 譚樸森, 黃居仁. 1997. Technical report: Word frequency list of "The Analects of Confucius" 詞庫小組技術報告 - 論語詞頻表. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica.
10. Yu, Shi-Wen, Xue-Feng Zhu, Elisabeth Kaske, and Zhi-Wei Feng 俞士汶, 朱学锋, Elisabeth Kaske, 冯志伟. 1996. *English-Chinese lexicon of computational linguistics* 英汉对照计算语言学词语汇编. Beijing: Peking University Press.
11. Yu, Shi-Wen, Xue-Feng Zhu, Hui Wang, Hua-Rui Zhang, Yun-Yun Zhang, De-Xi Zhu, Jian-Ming Lu, and Rui Guo 俞士汶, 朱学锋, 王惠, 张化瑞, 张芸芸, 朱德熙, 陆俭明, 郭锐. 2003. *The grammatical knowledge-base of contemporary Chinese—A complete specification* (version 2) 现代汉语语法信息词典详解 (第二版). Beijing: Tsinghua University Press.
12. Yu, Shi-Wen, Xue-Feng Zhu, Hui Wang, and Yun-Yun Zhang 俞士汶, 朱学锋, 王惠, 张芸芸. 1998. *The grammatical knowledge-base of contemporary Chinese—A complete specification* 现代汉语语法信息词典详解. Beijing: Tsinghua University Press.

32.3.3 Corpus Construction

1. Chen, Keh-Jiann, and Chu-Ren Huang. 2017. Modern Chinese balanced corpus of Academia Sinica. In *Encyclopedia of Chinese language and linguistics*, ed. Rint Sybesma. Leiden: Brill Publishers. https://doi.org/10.1163/2210-7363_ecll_COM_000191
2. Chinese Language and Knowledge Processing Group 中央研究院資訊科學所中文詞知識庫小組. 1998. Technical report: The content and instruction of Sinica Corpus 詞庫小組技術報告 - 中央研究院平衡語料庫的內容與說明. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/9804_2013.pdf
3. Huang, Chu-Ren 黃居仁. 2016. Corpus and language resources construction in Taiwan 臺灣語料庫與語言資源建設. In *The language situation in China (2016)* 中國語言生活狀況報告 *(2016)*, ed. Department of Language Information Management of Ministry of Education 教育部語言文字資訊管理司 編撰, 259–267. Beijing: Commercial Press.
4. Huang, Chu-Ren, Keh-Jiann Chen, and Zhao-Ming Gao 黃居仁, 陳克健, 高照明. 2016. Language processing research and language resources construction motivated by linguistic characteristics of Chinese 兼顧漢語語言特色的語言資訊化建設研究. *The Journal of Chinese Sociolinguistics* 中國社會語言學 02: 13–25.
5. Tseng, Shu-Jyuan, and Yi-Fen Liou 曾淑娟, 劉怡芬. 2002. Technical report: Instruction of spoken Mandarin Chinese corpus annotation system 詞庫小組技術報告 - 現代漢語口語對話語料庫標注系統說明. In *CKIP technical report* 中文詞知識庫小組技術報告. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/0201_2013.pdf

6. Yu, Shi-Wen, Hui-Ming Duan, Xue-Feng Zhu, and Bin Sun 俞士汶, 段慧明, 朱学锋, 孙斌. 2002. The basic processing of Contemporary Chinese corpus at Peking University SPECIFICATION 北京大学现代汉语语料库基本加工规范. *Journal of Chinese Information Processing* 中文信息学报. 16(5):49–64.

32.3.4 Semantic Representation

1. Chinese Language and Knowledge Processing Group. 2009. Technical report: Lexical semantic representation and semantic composition—An introduction to E-HowNet. In *CKIP technical report*. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/200901_2016b.pdf
2. Huang, Chu-Ren 黃居仁. 2007, 2006, 2005, 2004. Technical report: Meaning in sense in Mandarin Chinese version 4.0/3.0/2.0/1.0 詞庫小組技術報告 - 中文的詞義小辭典 4.0/3.0/2.0/1.0 版. In *CKIP technical report* 中央研究院文獻語料庫與詞庫小組技術報告. Tai Pei: Academia Sinica.
3. Huang, Chu-Ren (ed.) 黃居仁 主編. 2007, 2006, 2005, 2004. Technical report: Differentiation and description principles of Chinese lexical meaning version 4.0/3.0/2.0/1.0 詞庫小組技術報告 - 中文的詞彙意義的區辨與描述原則 4.0/3.0/2.0/1.0版. In *CKIP technical report* 中央研究院文獻語料庫與詞庫小組技術報告. Tai Pei: Academia Sinica.
4. Huang, Chu-Ren, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang 黃居仁, 謝舒凱, 洪嘉馡, 陳韻竹, 蘇依莉, 陳永祥, 黃勝偉. 2010. Chinese Wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing 中文詞彙網路:跨語言知識處理基礎架構的設計理念與實踐. *Journal of Chinese Information Processing* 中文資訊學報 24(2):14–23.
5. Huang, Shu-Ling, Su-Chu Lin, and Keh-Jiann Chen. 2014. Technical report: Sense representations for extended modalities in E-HowNet. In *CKIP technical report*. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/201401-tech_report_modality.pdf
6. Huang, Shu-Ling, Su-Chu Lin, and Keh-Jiann Chen. 2014. Technical report: The interactions among syntax, semantics, and morphology— How lexical structures affect verbal semantics and syntax. In *CKIP technical report*. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/201402-tech_report_morphology.pdf
7. Huang, Shu-Ling, Su-Chu Lin, Wei-Yun Ma, and Keh-Jiann Chen. 2015. Technical report: Semantic roles and semantic role labeling. In *CKIP technical report*. Tai Pei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/20151215-final-tech_report_semantic%20roles%20and%20semantic%20role%20labeling.pdf
8. Mei, Jia-Ju, Yi-Ming Zhu, Yun-Qi Gao, and Hong-Xiang Yin 梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 1983. *TongYiCi CiLin* 同义词词林. Shanghai: The Commercial Press.

9. Xue, Nian-Wen, and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering* 15(1):143–172.
10. Yuan, Yu-Lin 袁毓林. 2014. The description system and usage cases of qualia structure of the Chinese nouns 汉语名词物性结构的描写体系和运用案例. *Contemporary Linguistics* 当代语言学. 16(01):31–48, 125.

32.3.5 Treebank Construction

1. Chen, Keh-Jiann, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, and Zhao-Ming Gao. 2003. Sinica treebank: Design criteria, representational issues and implementation. In *Treebanks: Building and using parsed corpora*, ed. Anne Abeillé, 231–248. Dordrecht and Boston: Kluwer Academic Publishers.
2. Chen, Keh-Jiann, Chi-Ching Luo, Zhao-Ming Gao, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, and Chu-Ren Huang. 1999. The CKIP Chinese treebank: Guidelines for annotation. In *Proceedings of ATALA Workshop–Treebanks*, 85–96. Paris, France.
3. Chinese Language and Knowledge Processing Group 中央研究院資訊科學所中文組實驗室中文詞知識庫小組. 2013. Technical report: Semantic roles in Sinica Treebank 詞庫小組技術報告 - 句結構樹中的語義角色. In *CKIP Technical Report* 中文詞知識庫小組技術報告. Taipei: Academia Sinica. http://ckip.iis.sinica.edu.tw/CKIP/tr/201301_20140813.pdf
4. Huang, Chu-Ren, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao, and Kuang-Yu Chen. 2000. Sinica treebank: Design criteria, annotation guidelines, and on-line interface. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 29–37. Sapporo, Japan.
5. Huang, Chu-Ren, and Keh-Jiann Chen. 2017. Sinica treebank. In *Handbook of linguistic annotation*, ed. Nancy Ide and James Pustejovsky, 641–657. Dordrecht: Springer.
6. Xue, Nian-Wen, and Fei Xia. 2000. The bracketing guidelines for the Penn Chinese treebank (3.0). *IRCS Technical Reports Series* 39.
7. Xue, Nian-Wen, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering* 11(2):207–238.
8. Zhan, Weidong 詹卫东. 2009. Contemporary Chinese treebank processing specification (version 1.0) 现代汉语树库加工规范 (version 1.0). Department of Chinese Language and Literature of Peking University (technical report, unpublished). http://www.docin.com/p-475935862.html
9. Zhan, Weidong 詹卫东. 2009. Frequently asked questions of Contemporary Chinese treebank annotation 现代汉语树库标注常见问题举例. Department of Chinese Language and Literature of Peking University (technical report, unpublished). http://www.docin.com/p-469628749.html

10. Zhou, Qiang 周强. n.d. Technical report of Tsinghua Chinese treebank 清华大学汉语树库构建技术报告. Department of Computer Science and Technology of Tsinghua University (technical report, unpublished). http://www.docin.com/p-598594922.html