# Corpus-based Chinese studies

## A historical review from the 1920s to the present

Jiajin Xu
Beijing Foreign Studies University

This article reviews corpus-based Chinese studies, both applied and theoretical, from the 1920s to the present. It will be shown that, while corpus-based Chinese studies have been gaining momentum for only the last couple of decades, the roots of Chinese corpus linguistics go all the way back to the beginning of the 20th century. Today the bulk of corpus-based Chinese studies is oriented toward applied linguistics, with the compilation of frequency character/word lists and interlanguage Chinese studies being the most popular types of research. In addition to applied linguistic studies, this overview also highlights some innovative corpus studies on lexical and grammatical aspects of both classical and modern Chinese, as well as studies of sociolinguistic variation and discourse pragmatics. Overall, important groundwork in Chinese corpus linguistics is acknowledged and future directions are discussed.

**Keywords:** Chinese, corpus linguistics, lexical frequency studies, interlanguage studies, syntactic/grammatical studies, discourse-pragmatics, classical/historical Chinese studies

**关键词:** 汉语, 语料库语言学, 字/词频研究, 中介语研究, 句法研究, 话语语用研究, 古汉语研究

### 1. Introduction

The field of corpus research can be described as 'a tale of many C's': corpus, corpora, collection (of texts), collocation, colligation, concgram, concordance, context, computerised; the list of C-initialed keywords could certainly go on. We can further divide these C-words into two types: 'corpus as data' (i.e. corpus, corpora, collection of texts, computerised, etc.) and 'corpus as method' (i.e. collocation, concordance, concgram, etc.). In the current linguistics literature, 'corpus' is mainly

*Chinese Language and Discourse* 6:2 (2015), **218–244**. DOI 10.1075/cld.6.2.06xu
ISSN 1877-7031 / E-ISSN 1877-8798 

associated with data sources and/or research tools or analytical methods, despite a small group of Birmingham-influenced corpus linguists' view of it as a theoretical stance (cf. Tognini-Bonelli 2001; Teubert 2005; McEnery and Hardie 2012: chapter six). When dealing with Chinese corpora, we can add another C-word to the mix, Chinese, which is a major language of the world. For this review, I will discuss Chinese-based corpus studies in terms of both evolvement and accomplishment.

In this review, corpus-based Chinese studies include all treatments of Chinese with corpora as data, regardless of the nationality of the scholars who conduct the study. My discussion will focus on six major areas of study: current major corpora, general lexical frequency studies, syntactic/grammatical studies, interlanguage studies, discourse pragmatic studies, and classical/historical Chinese studies.

## 2. Current major corpora

Chinese linguists are now blessed with a myriad of large and publicly available Chinese corpora, often accessible on the internet. As for scholars of many other languages (Teubert 2005: 1), corpus data are nowadays considered the default resource for many Chinese linguists. Chinese corpus resources can be described in terms of general purpose corpora, interlanguage corpora, and domain-specific corpora.

### 2.1 General purpose corpora

Three of the most widely used Chinese general purpose corpora include the CCL (Centre for Chinese Linguistics, Peking University) corpus, and the CNC (Chinese National Corpus, also known as *Guojia Yuwei Yuliaoku* 'The State Language Commission Corpus'), the BCC (BLCU Chinese Corpus). Given its size and accessibility, the CCL corpus, with a collection of texts totalling over 470 million Chinese characters, is probably the most cited one. However, skepticism about the CCL database as a corpus is concerned with its skewed proportion of text collection. For instance, the overwhelming majority of the corpus consists of literary texts, and it contains an enormous amount of translated Chinese texts and an unnecessarily great number of linguistics research articles, and so forth. The CNC has become increasingly popular over the last few years because of its relatively balanced sampling and corpus size (i.e. 100 million characters). The BCC Corpus is the first ever balanced Chinese corpus of over ten billion characters. In addition to these mega-corpora, a couple of smaller corpora of one million or several million characters/words have been the empirical basis for many Chinese studies as well. The Academia Sinica corpus (Version 3) of five million words is characteristic

of Chinese published in Taiwan. LCMC (Lancaster Corpus of Mandarin Chinese), the UCLA Corpus of Written Chinese, and ToRCH2009 (Texts of Recent CHinese 2009, created at Beijing Foreign Studies University) corpus are three Brown-type, balanced corpora of one million words each.

Large spoken Chinese corpora are extremely rare. The Spoken Chinese Corpus of Situated Discourse (SCCSD), consisting of more than five hundred hours of transcripts and audio- and video-recording, and based on the spontaneous speech of people from all walks of life, has been developed at the Chinese Academy of Social Sciences, but is not publicly available due to human ethics or confidentiality reasons. The Lancaster Los Angeles Spoken Chinese Corpus (LLSCC) is another balanced corpus of spoken Chinese. The corpus is composed of one million words of dialogues and monologues, both spontaneous and scripted. Unfortunately, access to LLSCC is also restricted, due to similar confidentiality reasons.

One of the biggest projects on classical Chinese is the national initiative of digitising *Siku Quanshu* 'Complete Library of the Four Treasuries', which is the largest collection of books in ancient China. The texts used for the project contain approximately 800 million characters. Texts of classical Chinese can also be found at the CCL, CNC and BCC online interfaces.

### 2.2 Interlanguage corpora

The earliest corpus-based interlanguage Chinese studies date back to early 1990s at Beijing Language Institute (BLCU), home to a large population of international students of Chinese (Chu and Chen 1993). The pioneering work at Beijing Language Institute is now being carried on by an ambitious interlanguage Chinese corpus project — the International Corpus of Learner Chinese. The projected corpus size will be 50 million characters, made up of 45 million from written interlanguage Chinese and five million from spoken interlanguage Chinese. Apart from BLCU, Nanjing Normal University, Sun Yat-sen University, and Ji'nan University are also well-known for their interlanguage Chinese corpora.

### 2.3 Specialised corpora

Internet Chinese corpora are becoming the fastest-growing text collections, and will, with no doubt, marshal the next generation of corpus construction. This is both echoed by and overlapping with the world-wide craze for 'Big Data' in the natural language processing field. The ZHTenTen simplified Chinese corpus mounted at Sketch Engine (now 2.1 billion words and which designates a target corpus size of $10^{10}$, or 10 billion, words) is a case in point. ZHTenTen has been grammatically annotated and lends itself to concordancing, collocation and term

extraction, through a user-friendly user interface. It is safe to say that specialised corpora of this kind will become the mainstream general corpora of the near future.

Numerous domain-specific corpora are already readily accessible, such as *Renmin Ribao* 'People's Daily' database (a complete collection of the newspaper from 1946 to present) and the National Broadcast Media Language Resource Corpus of 195,182,188 characters maintained at Communication University of China. One which merits special mention is LIVAC (Linguistic Variation in Chinese Speech Communities), which is a more linguistically-driven synchronous Chinese corpus containing the printed Chinese media of major Chinese communities such as Hong Kong, Taiwan, Beijing, Shanghai, Singapore, and Macau. More than 500 million characters of news media texts have been gathered since 1995.

In addition, a great number of specialised Chinese corpora are archived at LDC (Linguistic Data Consortium, http://www.ldc.upenn.edu), Chinese LDC (http://www.chineseldc.org), ELRA (European Language Resources Association, http://www.elra.info), CHILDES (Child Language Data Exchange System, http://childes.psy.cmu.edu) and other data-hosting bodies. The availability of specialised corpora makes possible empirical studies on interlanguage study, sociolinguistic comparison, and buzzwords in media and the internet and their change over time. The many other in-house corpora for linguistics and/or natural language processing purposes cannot all be described here. Some of them will be mentioned passim in the review, while others can be found in Appendix A.

It is also important to point out that, given the constraint of space and the focus on Chinese proper of the present paper, the voluminous work on English-Chinese parallel corpora, translational Chinese corpora, Chinese dialect corpora and corpora of ethnic languages in China is not reviewed.

## 3. General lexical frequency studies

### 3.1 Heqin Chen[1] and his *Yutiwen Yingyong Zihui* 'Characters Used in Vernacular Chinese'

Parallel to, or probably directly influenced by, early corpus work in the West, corpus-based Chinese studies started very early, with word counting or word list[2]

1. Chen is considered the founding father of child psychology in China.

2. Most likely, word lists in a Chinese context are character lists. The term 'word list' used in this article refers to both character and tokenised word lists, for easy comparison to word lists in English and other languages.

creation, in China. *Characters Used in Vernacular Chinese* compiled by Heqin Chen and his associates in the early 1920s was a milestone in corpus-based Chinese studies.[3]

Chen's team collected six categories of texts (namely, children's book, newspapers, magazines, extra-curricular writings by school pupils, fictions, and miscellany[4]) and built a corpus of 554,478 characters, out of which a radical-sorted character frequency list and a frequency-sorted character list were created by hand. The first character list is sorted by number of radicals in a character from least to most; the second one is arranged by ascending order of frequency counts. They are similar to English word lists sorted by alphabetical and frequency order. A brief report about the character list was first published in the journal of *New Education* in 1922,[5] and the complete version of the list was published as a booklet by The Commercial Press in 1928. This new list was based on an enlarged corpus of 902,678 characters.[6] A more accessible version of the list is the reprint in the edited volume of *The Complete Works of Heqin Chen* (2008: 55–114).

Some major contributions of Chen's Character List to Chinese corpus linguistics and its application can be highlighted as follows.

---

**3.** In some review articles (e.g. Liu 2009: 63; Hai 2011: 2), Jinxi Li was regarded as the earliest Chinese frequency list creator in China. However, in his four-page-long discussion article, Li (1922), he did not mention a single word about his corpus building and methodology of word list building. Li, instead, asked a few crucial questions. For example, how many characters should be taught to school pupils? And how can the characters be graded for different levels of pupils? According to Li, both a frequency-based word list (e.g. 合同, 聪明, 便宜 etc.) and a character list were needed for the purpose of syllabus design. However, Li did not provide any answers to the questions.

**4.** The sampling frame coincides with that of Thorndike's (1921: iii) work. Thorndike gathered texts from children's literature, the Bible and English classics, elementary school textbooks, books about cooking, sewing, farming, the trades, and the like, daily newspapers, and correspondence. From autumn 1917 to 15 August 1919, Heqin Chen was pursuing his Master and PhD at Teachers College, Columbia University where Edward Thorndike was professor of Educational Psychology (cf: Chen and Chen 2008: 573–574).

**5.** The 1922 report appeared in *New Education* was reprinted in the inaugural issue (June 2014) of the new journal *Yuliaoku Yuyanxue* 'Corpus Linguistics' published by Foreign Language Teaching and Research Press.

**6.** The total number of character tokens of the 1928 corpus should be 902,658 (i.e. 554,478 characters of the 1922 corpus plus the added 348,180 characters) (cf. Chen 1928: 6); however, it was mistakenly reported as 902,678 (Chen 1928: 'General points' page after the inside cover). The miscalculation was corrected in the reprinted 2008 edition of *Characters Used in Vernacular Chinese*.

1. The project was one of the first Chinese corpus projects in the modern sense, because it followed principled sampling considerations and collected a sizeable quantity of authentic texts.
2. Chen (1922: 994) observed the overall tendency of the inverse proportionality between relative frequency and character rank as recognised by George Zipf (1935) based on empirical work on English, Latin, and Chinese in the 1930s.
3. The project was consciously motivated by pedagogical ends (for school pupils, adults, language assessment, dictionary making and materials development), and used as vocabulary control criterion in the Chinese textbook *Pingmin Qianzi Ke* 'Early Chinese Lessons for Illiterates' compiled by Tao and Zhu (1923).
4. The list played a significant role in promoting vernacular Chinese in China immediately after the May Fourth Movement in 1919.
5. Chen explicitly called for, in vocabulary teaching, a focus on the frequent over the rare (Chen 1922: 987, 994).

Heqin Chen is much lauded in that his methodology has not been superseded even 90 years later, in the computer age, for the frequency list of radical usage in Chen's Character List cannot be generated with Chinese concordancers even today. Despite the pioneering work and its research excellence and originality, Chen's corpus work was unfortunately not followed up on for decades. The character list and related research were to a large extent shelved until they were later reclaimed as a ground-breaking corpus project in some historical accounts (e.g. Feng 2006, 2012) of Chinese corpus work around the turn of the 21st century.

### 3.2 Frequency-based Chinese word list projects since the late 1970s

Heqin Chen (1922) aside, corpus-based Chinese word list projects were largely unheard of until the 1970s.[7] Hai (2011: 3) mentions six character lists compiled between 1950 and 1965. No direct evidence, however, could verify that the lists were corpus-based or frequency-driven.

Eric Shen Liu (1973) compiled a frequency dictionary of 3,000 tokenised words based on approximately 250,000 words of dramatic, fictional, essayistic, periodical, and technical literature published mainly in Shanghai and Taipei from 1921 to 1960, which is probably the earliest electronic corpus based Chinese word list. Unfortunately, the list did not seem to play a significant role in Chinese language research and pedagogy either in the Chinese mainland or Taiwan.

7. Ao (1929a, 1929b) slightly augmented Chen's corpus and made an enlarged character list in 1929.

I review, below, a number of published and publicly available lists produced in the 1970s onwards and which have made a positive contribution to Chinese corpus studies. However, it is not possible to exhaust every frequency-based Chinese character lists compiled thus far.

### 3.2.1 *The character list of Project 748*

In August 1974, the project *Hanzi Xinxi Chuli Xitong Gongcheng* 'Information Processing System of Chinese Characters' (also known as Project 748) was proposed and funded in September the same year by the then Ministry of Planning of the PRC. One of the objectives of the key national project was to count the total number of Chinese characters in actual use. More than 1,500 people were involved in counting the characters used in a corpus of 21,657,039 characters. Texts of science and technology, literary and arts, political classics, and news reports were collected for the purpose. A frequency-based character list was built in 1977, and later published as *Hanzi Pindu Tongji* 'Frequency Calculation of Chinese Characters' (Bei and Zhang 1988). The 5,991 character types were categorised into five levels on the basis of their frequency counts. The (average) number of strokes used for characters at each level was also provided. The character list of Project 748 was the first general service list of Chinese characters after the founding of the People's Republic of China in 1949.

### 3.2.2 *Early frequency books of Chinese words compiled at Beijing Language Institute*

In 1986, the Institute of Language Teaching Research at Beijing Language Institute (now called Beijing Language and Culture University) published its corpus-based *Xiandai Hanyu Pinlu Cidian* 'Frequency Dictionary of Chinese Words' (1988). It is a dictionary for tokenised words with frequency information. Separate lists for monosyllabic, disyllabic, and multi-syllabic words were provided. A balanced corpus of 1,807,398 characters was constructed for the frequency dictionary. The corpus consists of newswire texts, popular science discourse, plays, comic cross-talks, spontaneous spoken discourse, and literary texts. The dictionary provides raw frequency, relative frequency, text distribution (number of texts), genre distribution, and combined metric of word use. A subset of the corpus (containing elementary and secondary school textbooks totalling approximately 520,000 characters) was published separately as *Hanyu Cihui de Tongji yu Fenxi* 'The Statistics and Analysis of Chinese Words' (1985a). In the meantime, a booklet called *Changyong Zi he Changyong Ci* 'Frequently Used Characters and Words' (1985b) with 1,000 high frequency characters and 3,817 high frequency words, was also published. As stated explicitly in the introductions to the three frequency lists, the shared objective

of the projects was to explore how many words should be taught to international learners of Chinese at different proficiency levels.

### 3.2.3 *Early frequency lists developed at Beihang University*

From the early 1980s onward, Yuan Liu's team at Beihang University undertook two frequency list projects. They were published as Liu et al. (1990) and China State Language Commission and China State Bureau of Standards (1992). The 1990 list was a tokenised word list based on 25 million characters of texts covering politics, history, philosophy, sports and life, news, literature, construction and transportation, forestry, husbandry, basic science and so on. They divided the text categories into two overarching genres: social sciences and natural sciences. About half of the corpus (ca. 11,080,000 characters) was used for a genre-based character list published in 1992.

### 3.2.4 *Chinese National Corpus and its frequency lists*

The construction of the Chinese National Corpus (CNC, the State Language Commission Corpus) was initiated in 1991, and its current size is about 100 million characters. Following Liu et al. (1990) and China State Language Commission and China State Bureau of Standards (1992), the sampling categories of the CNC cover social sciences, natural sciences and miscellany. The sub-categories of the texts are almost the same as those of the 1990 and the 1992 projects. The latest national Chinese character list, *Tongyong Guifan Hanzi Biao* 'A General Service List of Chinese Characters',[8] was released in August 2013, and contains three graded character lists: 3,500 basic characters as Level One, 3,000 common but less frequent characters as Level Two, and Level Three with 1,605 proper nouns, technical, domain-specific and archaic Chinese characters. The list, especially the first two levels of it, is based on the frequency counts of the CNC[9] and a couple of other large corpora (ibid. 7–8).

### 3.2.5 *Richard Xiao et al. (2009): A frequency dictionary of Mandarin Chinese*

The book presents a list of 5,000 words ordered by frequency in a 50 million word corpus, which is one of the few Chinese word frequency books published outside China. The highlight of Xiao's dictionary is the balanced composition of the corpus, covering spoken data, fiction, news, and academic texts.

It is impossible, in this article, to cite and comment on every frequency list project of Chinese over the time span of nearly one hundred years. What I have

8. The list is downloadable at http://www.gov.cn/gzdt/att/att/site1/20130819/tygfhzb.pdf.

9. A 20 million character sampler of the CNC is available at http://www.cncorpus.org on a registration basis.

reported are the influential ones among those published or publicly (e.g. with online access) available. Among the frequency lists, Heqin Chen's list is groundbreaking; the frequency dictionaries of Chinese words compiled at Beijing Language Institute live up to modern criteria for balanced sampling and sophistication of frequency computation; and *Tongyong Guifan Hanzi Biao* 'A General Service List of Chinese Characters' based on the CNC and other resources is the flagship of modern Chinese character lists.

## 4. Syntactic/grammatical studies

### 4.1 Sentence Pattern Research Group at Beijing Language Institute

Among corpus-based grammatical studies on modern Chinese in the literature, Sentence Pattern Research Group at Beijing Language Institute[10] (1989a, 1989b, 1989c, 1990, 1991) came up as a milestone around 1990. The Research Group published a series of statistical reports on kernel Chinese sentence patterns based on four million characters of texts. What is impressive about the project is the systematicity and exhaustiveness of frequency counting of different types of sentences. General sentence types, such as declarative, interrogative, imperative, exclamatory, negative sentences were counted, and sentences containing verb complement, verb resultative, *ba* construction, *bei* construction, serial verb construction, pivotal sentence, focus construction *shi…de*, and the like were all described statistically. This empirically-based description of sentence usage updates our intuitive knowledge of Chinese sentence patterns and offers a more sound and convincing representation of how sentences are used in authentic Chinese. Such a study of course has had its due impact on the teaching of Chinese in addition to being theoretically relevant in terms of descriptive linguistics.

Hindered by technological and methodological limitations, corpus-based sentential/grammatical level research is practically negligible as compared with corpus based lexical studies. The overwhelming number of corpus-based grammatical studies focus on some particular type of grammatical issue.

How corpus-based language studies can be rightfully valued and acknowledged by the theoretical linguists who do not share a corpus linguistics research agenda is a question that corpus linguists must bear in mind and address. Comprehensive corpus-based description is but the onset of language studies. On the one hand, corpus investigation might give rise to theoretical innovations, as advocated by the so-called corpus-driven linguists (cf. Tognini-Bonelli 2001); on the other hand,

10. The team was headed by Prof. Shuhua Zhao. The project was completed in April 1995.

linguistic theory can inform corpus-based language studies, especially at the data interpretation stage of research. The following two studies can expound the theory formulation and the theory/hypothesis corroboration.

### 4.2 Shaohua Zou's corpus-based Chinese studies

Shaohua Zou is an author who has concentrated his attention on language use and its impact on lexical semantics and morpho-syntactic interpretations. In Zou (2001) and Zou and Ma (2007), Zou and his associates try to convince the readers that frequency is the underlying force that renders positive or negative semantic interpretations of seemingly neutral lexical items, and the preferred interpretation of apparently ambiguous structures. For example, Zou (2001) cited frequency and collocation information from actual language use to explain the negative 'semantic prosody' (Louw 1993) of the Chinese distal demonstrative *nage* construction. Negative collocates of *nage* were found more frequent than those of proximal demonstrative *zhege*. Working along the same methodology and principle, Zou and Ma (2007) continued the study of frequency motivation for the ambiguous grammatical construction VP + NP1 + *de* + NP2, and extended the study to seven complex NP or VP NP constructions. In Zou and Ma (2007), they obtained frequency and co-frequency information from a large dataset instead of obtaining frequency information from published frequency dictionaries (e.g., Institute of Language Teaching Research at Beijing Language Institute 1985a) or text samples collected unsystematically. For example, one database they cited was 11 million characters of *Renmin Ribao* 'People's Daily' texts. What stands out about Zou's work is the attention, from a statistical point of view, to how contextual information of certain lexical items and/or constructions leads us to derive preferred semantic interpretations. Zou is a structuralist linguist of Chinese by training, and did not seem to be influenced by Western (corpus) linguistic theories, yet the statistically-founded contextualism coincides with neo-Firthian contextualism, semantic prosody in particular. Zou's work on preferred interpretation of syntactic ambiguity is also much in line with the current functional linguistic view in that frequency has a role in the 'emergence' of linguistic structure (cf. Bybee and Hopper 2001).

It is a pity that Zou's postulation has not been followed up on by mainstream linguists. If Zou's inquiry into lexical semantics and meaning of constructions can be considered as to a large extent theorising based on real language data, the majority of corpus-based Chinese studies fit in the category of theory-informed language studies.

### 4.3 Siewierska, et al. (2010) on Chinese splittable compounds

Siewierska, et al. (2010) undertook a corpus-based analysis of *liheci* 'Chinese splittable compounds' — a type of disyllabic compound verbs where other words (in particular aspect markers) can occur between the two elements of the compound. The study of *liheci* is at the interface of Chinese morphology and syntax. Whether *liheci* are words or phrases is the major linguistic concern of the study. The study is based on a one-million-word written Chinese corpus and one-million-word spoken Chinese corpus. It draws on prosodic morphology theory (McCarthy and Prince 1995; Feng 2002) on the basis of frequency statistics to reach the following conclusions as to what a word is and what a phrase is in Chinese.

Siewierska et al. propose structural and phonological criteria in light of the general patterning of splittable compounds based on the two mentioned million word Chinese text collections. The quantitative and qualitative combined approach offers a more reliable picture as to what a word is in Chinese no matter whether the component morphemes are in contiguous or discontiguous connection. The theoretical understanding of wordhood in mainstream morphology holds that a word is both a morphological and phonological thing (Matthews 1991). The structural criteria emerge from the statistical results of corpus analysis, which is represented as: Host dependency: head dependence > tail dependence, which can be accounted for as follows:

The host dependency criterion (a > b > c) for judging *liheci* or wordhood in general deems:

a. *liheci* with a clitic-like aspect marker (e.g. the perfective marker *-le*) as compounds instead of phrases;
b. *liheci* with resultative verb complements attached to the main verb quasi-compounds; and
c. other modifiers (classifiers, modifiers, etc.) attached to the first/head morpheme, represented typically by a noun or complement, least likely to be compounds.

Principles of prosodic morphology were taken into account to justifiy the phonological criteria, or 'prosodic word restriction', for wordhood or phrasehood of *liheci*.

The authors propose that the various manifestations of *liheci* define a continuum of phonological conditions as a complement to the grammatical criteria (a > b > c):

a. The combined uses of the first morpheme and the second morpheme are disyllabic compounds;

b. *liheci* in which the verb morpheme and the ending morpheme are separated by one single morpheme under the Trisyllabic Foot Rule are possible compounds; while
c. the first morpheme and the second morpheme separated by multi-syllable units in the form or combination of quantifiers, adjectival modifiers, etc., are phrases.

In addition to comprehensive descriptions of all major aspects of Chinese based on statistical results, corpus-based studies on particular Chinese morpho-syntactic structures have become increasingly popular since the introduction of computer corpora and powerful corpus query tools, for example, corpus-based studies on *liheci* by Wang (2001) and Wang (2011), Xiao and McEnery (2004) on Chinese aspect based on LCMC from a cross-linguistic perspective, Tao (2000) on the argument structure of *chi* 'to eat', and so forth.

## 5. Interlanguage studies

If corpus-based frequency lists of Chinese characters/words are a long-standing tradition of Chinese corpus linguistics, corpus-based interlanguage Chinese studies are probably the most popular burgeoning area of corpus-based Chinese studies in China, increasing over the last couple of decades.

The earliest corpus-based interlanguage Chinese studies in China were started in the early 1990s at Beijing Language Institute (BLCU) with a great number of international students of Chinese. BLCU has been undertaking an interlanguage Chinese corpus project — International Corpus of Learner Chinese. The projected size of the corpus is 50 million characters of written and spoken interlanguage Chinese data. Corpus expertise at BLCU is based in two institutes: The International R&D Centre for Chinese Education and Institute of Language Information Processing. Over the years, BLCU has developed a whole array of learner corpora, concordancers, annotation tools, and corpus-informed Chinese learning tools.[11] Apart from BLCU, Ji'nan University, Ludong University, Nanjing Normal University, Shanghai Jiao Tong University, Sun Yat-sen University, and Xiamen University are also well-known for their interlanguage Chinese corpus studies.

Corpus-based Chinese interlanguage studies are characterised by lexical and syntactic error analysis in terms of variables such as inherent properties of Chinese, learner factors, and typological differences. For instance, Cui (2005) examined the use of Chinese prepositions by European and American learners. Xiong (1996)

11. The resources can be accessed at http://nlp.blcu.edu.cn/online-systems/.

investigated the *ba* constructions used by international students. Li and Wu (2013) probed into the relativisation in interlanguage Chinese. All these cases draw on the BLCU interlanguage corpus. Cui collated all major Chinese prepositions, Xiong categorised four types of *ba* constructions based on previous literature, and Li and Wu focused on the accessibility, embeddedness, and animacy of Chinese relative clauses. Corpus data provided statistical and probabilistic basis for the overuse, underuse and misuse of the three phenomena. The quantitative analyses were then scrutinised in terms of different learner factors. For example, the use of *ba* constructions was compared between male and female students, among students of different L1 backgrounds, or across different proficiency levels (i.e. starter, intermediate, advanced, or Levels 1 to 3). The errors committed by the learners were very often explained as stemming from typological differences between the L1 of the writers and Chinese. The interlanguage scholars also found that some errors were fairly common, if not universal, among interlanguage produced by learners of different L1 backgrounds. The findings and argumentation bear much resemblance to English interlanguage corpus based studies pioneered by Granger (1996, 1998, 2002, among many others) in terms of research design, data annotation and analysis, though Granger's Contrastive Interlanguage Analysis was scarcely cited by scholars of Chinese interlanguage studies.

Most up-to-date Chinese interlanguage corpus research can be encapsulated by papers presented at the last two national symposia of Chinese interlanguage corpora (Xiao and Zhang 2011; Cui and Zhang 2013).

## 6. Discourse-pragmatic and sociolinguistic studies

Two Chinese corpus studies merit special mention as influential sociolinguistic projects. One is the Dynamic Circulation Corpus (DCC) initiated by Pu Zhang at Beijing Language and Culture University, and the other is the LIVAC corpus project led by Benjamin Tsou at City University of Hong Kong. Both projects make headline news each year and attract attention when releasing their annual new word rosters. The DCC was meant to be a primarily diachronic corpus monitoring general Chinese language use. LIVAC was designed to be a corpus of Chinese varieties across different regions. As the two projects carry on and more texts are added year by year, both corpora have become diachronic. What is most innovative about the DCC is the sampling method, 'degree of circulation,' proposed by Zhang (1999a, 1999b). The text collection adopts a genre model for written language, and situational and demographic criteria for spoken language. The sampling strategy is different from mainstream sampling frames in the West but makes good sense as a way to represent language in actual use. The 'degree of circulation (DC)' model

is represented by the equation: DC = the volume of circulation * the density of circulation * the area of circulation * the frequency of circulation. All major newspapers, magazines, and TV and radio broadcasting services were rated in terms of the four parameters in the equation. The DCC sampling model approximates major language use in public domains. The LIVAC initiative (Tsou et al. 1997; Tsou and You 2007, 2010) takes into account mainstream Chinese newspapers in Beijing, Hong Kong, Macau, Shanghai, Singapore, Taiwan, Shenzhen, Zhuhai, and Guangzhou, which makes it a perfect variationalist sociolinguistic data source for comparing native Chinese and diaspora Chinese. What the two corpora offer so far focuses on such lexical matters as neologisms, and lexical variations across different Chinese speech communities, but they fall short of theoretical linguistic studies in terms of lexical, grammatical, or discoursal issues.

A few spoken Chinese-based discourse studies also exist, but are greatly limited by their quantity of data used, for example, Tao (1996) on the prosodic units in Chinese conversation, Luke and Pavlidou (2002) on telephone calls, Gu's (2009) multimodal analysis of situated spoken discourse, Xu (2009) on discourse markers in spoken Chinese of urban teenagers, Chen and Guo (2010) on Chinese motion expressions, Thompson and Tao's (2010) revisit of word class 'adjective' in spoken Chinese discourse, and Yang (2011) on repairs in Chinese doctor-patient conversions. The list of such studies can go on, but, overall, lexical frequency studies and interlanguage studies significantly outnumber corpus-based Chinese discourse studies.

## 7. Classical/historical Chinese studies

### 7.1 Lexical frequency work on classical Chinese

Comprehensive work on frequency counts based on large collections of classical Chinese corpora has been very rare and relatively recent. Two projects on computing character use in classical Chinese merit special mention. One is *Shisan Jing Zipin Yanjiu* 'The Frequency Study of Thirteen Chinese Canons'[12] (Hai 2011). Hai calculated and graded the characters in the thirteen texts into three bands: high

**12.** *The Thirteen Canons* refer to *Shijing* 'Book of Poetry', *Yijing* 'Book of Changes', *Zhouli* 'The Rites of Zhou', *Liji* 'The Classic of Rites', *Yili* 'Etiquette and Rites', *Chunqiu Zuozhuan* 'The Commentary of Zuo', *Yizhuan* 'The Commentary to the Book of Changes', *Lunyu* 'The Analects', *Erya* 'The Literary Expositor', *Shangshu* 'Book of Documents', *Xiaojing* 'Classic of Filial Piety', *Mengzi* 'The Works of Mencius', *Chunqiu Gongyangzhuan* 'The Commentary of Gongyang', and *Chunqiu Guliangzhuan* 'The Commentary of Guliang'. Such is one version of *The Thirteen Canons*; disagreements about the inclusion of Chinese classics are common.

frequency, moderate frequency, and low frequency. Character frequency variations across texts were also tabulated and analysed.

Another key project on classical Chinese characters is the national initiative of character use in *Siku Quanshu* 'Complete Library of the Four Treasuries', which is the largest collection of books in ancient China. The texts used for the character frequency project *Guji Hanzi Zipin Tongji* 'Character Frequency Calculation of Classical Chinese' (Unihan Digital Technology Co., Ltd. 2008) contain approximately 800 million characters. The frequency list project is a thorough and foundational work for computational linguistics and information systems for classical Chinese.

Recent years have seen many statistical reports of frequency character lists of oracles, bronze script, *qin* scripts on bamboo slips, etc., based on a digitised database of ancient scripts. They are, however, a mere record or display of textual data. More theoretical generalisations have to be made in light of the increasing quantity of digitised classical Chinese texts.

## 7.2 Early concordances to Chinese classics

Concordances to classics in printed format might not be considered corpus studies. However, they are inherently related to present-day computer corpus work. Concordances are particularly helpful in researching classics as they both list citations from the classics under individual entries, and provide the line number, page number and chapter. The first concordance of Chinese canons, *Laojielao* 'A synthetic study of Lao Tzu's Tao Te Ching in Chinese', was compiled in 1922 by Tsai Ting Kan (1861–1935). From the 1920s to 1984, according to Pan (1984), some 500 concordances and concordance-like bibliographic indices have been published. Among the great number of concordances, two massive concordance projects, headed by William Hung at Peking University and Din Cheuk Lau at the Institute of Chinese Studies (ICS) at Chinese University of Hong Kong, are monumental. Peking University's Harvard-Yenching Institute Sinological Index Series had been published from 1930s through 1940s (cf. Hung 1932), and the ICS concordance series around 2000 (cf. Lau, et al. 1992). There are occasional concordances produced by Western scholars, such as *A Concordance of* Baiyujing by Eifring (1992). A more common way of concordance of Chinese classics is an appendix as found at the end of *Lunyu Yizhu* 'Annotations to the Analects' (Yang 1980: 214–316). All of these serve as great reference tools for researching language use in Chinese classics.

The extremely labour-intensive hand-made concordances have now been superseded by computer concordancers, which enable concordance generation with just a few clicks (see Appendix B). Online concordancing is all the more handy and

popular today. For instance, the Chinese Text Project (http://ctext.org/) is probably the best open access online concordance collection of Chinese classic texts, in which a large portion of the classical Chinese texts are aligned with their modern Chinese translation and English translation.

## 7.3 Halliday's quantitative work on colloquial Chinese of the Yuan Dynasty

Michael Halliday's PhD thesis published in 1959 as *The Language of Chinese "Secret History of the Mongols"* is in many senses a corpus-based work. The 12-chapter *Secret History* is said to be a personal biography of Genghis Khan, originally written in Mongolian in AD1240 and later published with Chinese translation and interlinear Chinese gloss at the end of fourteenth century (Halliday 1959: 1; Lu, et al. 2000: 224). Halliday's work under the guidance of J. R. Firth is a comprehensive and exhaustive description of the lexico-grammar of the historical narrative. All major lexico-grammatical (even textual) categories, including word classes, clause types (e.g. conditional and genitival), mood types (e.g. interrogative, imperative, and neutral), aspect types (e.g. perfective and non-perfective), voice types (e.g. passive, ergative, and active), sentence types (e.g. compound, simple), characters, words, paragraphs and chapters were systematically counted and analysed (Halliday 1959: 47).

In Halliday's discussion, frequency information is one of the major sources of evidence in support of his linguistic description and conceptualisation. For example, in his discussion about one aspectual feature of Chinese, he observed (ibid. 84) that the imperfective particle *jo* (着) is frequently found (47 out of a total of 328 occurrences) in a prepositive complex group modifying a prepositive verb. He was also interested in the co-frequency of linguistic items. The appendix section of the book contains 12 statistical tables summarising the total occurrences (as well as occurrences per chapter of *The Secret History*) of the categories mentioned above. All these show that the work is a solid empirical investigation of the lexico-grammar of Chinese.

The theoretical and methodological significance of the study is three-fold.

1. The study is an exemplary case of probabilistic treatment of linguistic structures.
2. The study is located in broad 'complementarities' (Halliday 2008) of syntagmatic and paradigmatic relations, lexis and grammar, and speech and writing. The empirically-grounded description of the grammatical categories in Chinese became the starting point of his later theorising of 'categories of the theory of grammar' in general and influential Systemic Functional Grammar as well.

3. Halliday placed a particular emphasis on the quantitative analyses based on a corpus of spoken Chinese when he chose *The Secret History*, which is known as 'vernacular and slangy' (Lu et al. 2000: 221), despite that it is a written text. When Halliday studied Chinese under the supervision of Li Wang at Lingnan University in 1949, he was interested in studying the grammar of vernacular and colloquial Cantonese dialect, and put together a corpus of Cantonese sentences (Halliday 1992: 61) as well.

Similar exhaustive frequency-based descriptions of lexico-grammatical features of ancient Chinese are rare. Zhou (2007) on *Soushenji* 'In Search of the Supernatural' is apparently one of the most solid of the very few corpus-based classical Chinese studies. All major types of word classes (e.g. noun, verb, adjective, numeral, pronoun, adverb, preposition, conjunction, auxiliary and particle) have been systematically investigated in *Soushenji*. All the concurrences of each important grammatical categories (e.g. tense and aspect expressions, negation), word class, and typical words were counted. Words and semantic sub-categorisations of major grammatical categories were statistically described. For example, verb directional constructions were sub-categorised as outbound or inbound, and upward, downward, circular, or in an unspecified direction. Frequency counts for the sub-categorisations and typical sentences in *Soushenji* were all presented. The study serves as an important descriptive grammar of the Wei-Jin Period (AD220 — AD420) popular Chinese.

## 7.4 Liu (2009) on historical and regional variation of Chinese character construction

Liu (2009) conducted a study on compositional motivations based on a set of historical data. He adopted a frequency approach to investigating the potential change of classical Chinese character construction from more pictogrammatic to more ideogrammatic and phonogrammatic, which was corroborated by the historical variation from *jiaguwen* 'Oracles' of the Shang Dynasty to *jinwen* 'Bronze script' of the Western Zhou Dynasty and regional variation of *chujianbo* 'writing on Chu Bamboo Slips and Silk Manuscripts' and *qinjian* 'Scripts on Bamboo Slips excavated in ancient Qin territory'. The entire project was based on the tally of *liushu*[13] 'Six Principles of Chinese Character Construction' representation in different historical texts.

13. The Six Principles of Chinese Character Construction are the cornerstones of Chinese characters. They are: (1) *xiangxing* 'pictograms, images of the reality'; (2) *zhishi* 'indicators, a small stroke to indicate the locality or directionality as of a pictogram, e.g. 刃 and 上; (3) *huiyi* 'ideograms'; (4) *xingsheng* 'phonograms, combinations of an image/meaning and sound radical'; (5) *zhuanzhu* 'transformed cognates'; and (6) *jiajie* 'borrowings'.

**Table 1.** Statistical comparison of *sishu* between Chu Bamboo Slips and Silk Manuscripts and Qin Bamboo Slips

| | Chu Bamboo Slips and Silk Manuscripts | | | | Qin Bamboo Slips | | | |
|---|---|---|---|---|---|---|---|---|
| | **Types** | **Type%** | **Tokens** | **% out of *sishu*** | **Types** | **Type%** | **Tokens** | **% out of *sishu*** |
| **pictograms** | 201 | 4.56% | 15735 | 25.15% | 171 | 11.75% | 10386 | 30.81% |
| **indicators** | 51 | 1.16% | 8231 | 13.16% | 40 | 2.75% | 3814 | 11.31% |
| **ideograms** | 549 | 12.45% | 13861 | 22.16% | 357 | 24.54% | 9247 | 27.43% |
| **phonograms** | 3610 | 81.84% | 24733 | 39.53% | 887 | 60.96% | 10266 | 30.45% |

Table 1 highlights the regional variation of four key methods of Chinese character composition between Chu, a then remote region from Central China, and Qin during the Warring States Period. The statistics show that the Chu script is more liberal, as the Chu script uses fewer pictograms than Qin slips (4.56% vs. 11.75%) do, and more phonograms (81.84% vs. 60.96%).

This finding is of great interest because it suggests that the evolution of Chinese language, or at least of the configuration of Chinese characters, might have been driven by vernacular Chinese communities (e.g. Chu) rather than Chinese speakers or writers close to the main ruling centres (e.g. Qin in this case) in highly centralised feudal China.

Liu (2009) on Chinese *sishu* is a typical descriptive study of Chinese character configuration based on classical Chinese texts, while a study by Wang (1983)[14] on the lexico-semantic change of classical Chinese is even more theoretically innovative and intellectually stimulating.

## 7.5 Wang (1983) on word frequency and historical character differentiation

In the paper *Ci de Pinlu he Zi de Fenhua* 'Word frequency and character differentiation', Wang (1983: 8) argues that the key to character differentiation is the frequency of the variants of the same base word (i.e. homophones, homographs, or cognates). According to Wang, the more frequent form of different variants, or allomorph (e.g. 娶, 其, 腰), of the same base word (e.g. 取, 其, 要) tend to reserve the earlier/original form of the word. For instance, between the original meaning 'to acquire, to get' of 取 and the later developed meaning 'to marry a woman' (as

**14.** The paper was presented at the Second Annual Conference of Chinese Linguistics Society in Hefei, Anhui in May 1983. On the last page of the mimeographed handout, Wang mentioned that the manuscript was written up in 1960, and was shelved during the Cultural Revolution, and not made publicly available until 23 years later.

**Table 2.** Frequency counts across different texts as of character differentiation

| | | 论语 | 墨子 | 孟子 | 庄子 | 荀子 | 杜诗 | 例句 |
|---|---|---|---|---|---|---|---|---|
| 取 | 取 | 11 | 60 | 58 | 37 | 81 | 58 | 青取之于蓝《荀》 |
| | 同娶 | 1 | 3 | 0 | 2 | 1 | 0 | 取妻身迎《墨》 |
| | 娶 | 0 | 0 | 6 | 0 | 0 | 1 | 舜不告而娶为无后也《孟》 |
| 其 | 其 | 254 | 1338 | 560 | 1232 | 1111 | 109 | 其为人也孝悌《论》 |
| | 箕 | 0 | 1 | 2 | 3 | 7 | 4 | 箕踞鼓盆而歌《庄》 |
| 要 | 要 | 1 | 9 | 8 | 12 | 34 | 44 | 总天下之要《荀》 |
| | 同腰 | 0 | 3 | 0 | 2 | 2 | 0 | 夫子曲要磬折《庄》 |
| | 腰 | 0 | 1 | 0 | 0 | 1 | 22 | 行人弓箭各在腰《杜》 |
| | 其他 | 1 | | | | | | 同“约” |

in 娶妻), the significantly higher frequency of the original meaning secures the original form of 取 to the meaning ‘to acquire, to get’. The meaning of ‘to marry a woman’ has to take an additional radical to differentiate it from ‘to acquire, to get’. This is also true in the cases of 其 vis-à-vis 箕, and 要 vis-à-vis 腰 (see Table 2). The schism of the two usages in relation to two morphological forms has taken place over a considerably long time in the history of lexical change. Thus, there could be a fairly long period of mixing or *gongju* ‘cohabitating’ (in Wang’s terms) in between.

Wang’s postulation of frequency-induced lexical differentiation predates the claims by Western grammaticalisation scholars who had been aware of the interaction between frequency and the emergence of linguistic structure (cf. Bybee and Hopper 2001). Wang observes the implicit frequency effect that governs classical Chinese character differentiation (Wang 1983: 21). Wang’s theoretical argument might well complement the current generalisations about the word-level frequency effects (e.g. lexical diffusion theory) based on the examinations of Indo-European language evidence.

Wang (1983) outdoes, to some extent, other frequency based descriptive studies on classical Chinese characters in that it attempts to explain the mechanism for the regular patterns between form and meaning, and between the original form and the present-day one of lexical items.

## 8. Summary

Corpus-based Chinese studies have been gaining momentum for the last couple of decades, but, as was mentioned in our historical overview, the roots of corpus

linguistics in China go far back (unless we restrict corpus linguistics to the use of texts in electronic form only). The field of Chinese corpus linguistics, nevertheless, has not thus far come into existence. Similar to corpus-based English studies, corpus-based Chinese studies are to a large extent a popular methodology. The sad truth in China, however, is that corpus-based Chinese studies have not been playing a role at the centre stage of linguistics and applied linguistics.

The bulk of corpus-based Chinese studies are applied linguistics-oriented in terms of the quantity of projects and publications. Lexical frequency studies, learner Chinese corpus construction, corpus-assisted machine translation system development, and language use monitoring are among the typical foci of Chinese corpus-based applied linguistics research. Chinese corpus-based theoretical linguistics studies are scarce and by no means the mainstream. The original studies by Wang (1983), Zou (2001) and some Western theory-informed Chinese studies such as those by Tao (2000), Xiao and McEnery (2004), and Siewierska, et al. (2010), and Li and Wu (2013) are a smalll number of notable exceptions.

The use of Chinese corpora for machine translation, natural language processing and other computational linguistic areas has seen exciting achievements. However, they are not the focus of attention for the present review. Please refer to Feng (2006, 2012) for the research in such areas.

Corpus-based English studies are currently working along two lines of linguistic inquiry. The first group, mainly influenced by John Sinclair, view language as largely a phraseological phenomenon, and argue that collocation is the cornerstone in the search for units of meaning (Sinclair 2004). '[C]ollocation-via-concordance' (McEnery and Hardie 2012:126) has been the most prominent methodology in addressing various lexical, grammatical and discoursal issues. The interest, along with the related corpus linguistic theoretical assumptions, in collocation has not drawn much attention in Chinese language description or the study of interlanguage Chinese.[15] The other popular line of corpus research in English publications is concerned with discourse and sociolinguistic issues, critical discourse analysis in particular. Not many Chinese corpus studies have been done in this second field either.

Therefore, we can see a massive amount of fruitful research of Chinese studies along the above two lines, in addition to the work we have reviewed in the paper. There is also, for example, a critical need for analyses of spoken and multimodal corpora. Web-based exploitation of Chinese texts, both web-based concordancing and web-based corpus construction in the age of Big Data, should prove to be another appealing area of Chinese corpus research.

**15.** Li (2011) on the semantic prosody of some Chinese lexical items is an exception.

It is hoped that corpus methodology, if not corpus linguistics as an independent discipline, will take a centre stage in Chinese linguistics. To get closer to this goal, Chinese corpus linguists need to be data gatherers, software users and also those who do the theorising. Therefore, a corpus linguist has to be an expert in SLA, cognitive linguistics, functional linguistics, discourse analysis, etc., as well as a half computer scientist.

All the corpus-based Chinese works reviewed in this article deserve our special respect. It is such foundational research that has shaped Chinese corpus linguistics into what it is today.

## Acknowledgements

This research is partially supported by China's National Social Sciences Foundation grant (ref. 12CYY060) and by the National Research Centre for Foreign Language Education (MOE Key Research Institute of Humanities and Social Sciences at Universities), Beijing Foreign Studies University. The author would like to thank the referees and editors for their valuable comments and to Professor Guangqian Zhang and Jori Lindley for proofreading the manuscript.

## References

Ao, Hongde. 1929a. "Yutiwen Yingyong Zihui Yanjiu Baogao: Chen Heqin Shi Yutiwen Yingyong Zihui zhi Xu [A study of characters used in vernacular Chinese: Extending Chen's character list]." *Jiaoyu Zazhi* [Journal of Education] 21 (2): 77–101.

Ao, Hongde. 1929b. "Yutiwen Yingyong Zihui Yanjiu Baogao (Xu): Chen Heqin Shi Yutiwen Yingyong Zihui zhi Xu [A Study of Characters Used in Vernacular Chinese: Extending Chen's Character List (Continued)]." *Jiaoyu Zazhi* [Journal of Education] 21 (3): 97–113.

Bei, Guiqin,Xuetao Zhang and . 1988. *Hanzi Pindu Tongji* [Frequency calculation of Chinese characters]. Beijing: Publishing House of Electronics Industry.

Bybee, Joan, and Paul Hopper (eds). 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/tsl.45

Chen, Heqin. 1922. "Yutiwen Yingyong Zihui [Characters used in vernacular Chinese]." *Xin Jiaoyu* [New Education] 5 (5): 987–995.

Chen, Heqin. 1928. *Yutiwen Yingyong Zihui* [Characters used in vernacular Chinese]. Shanghai: The Commercial Press.

Chen, Heqin. 2008. "Yutiwen Yingyong Zihui [Characters used in vernacular Chinese]." In *Chen Heqin Quanji (Di Liu Juan)* [The complete works of Heqin Chen (Volume 6)], ed. by Xiuyun Chen and Yifei Chen, 55–114. Nanjing: Jiangsu Education Press.

Chen, Liang, and Jiansheng Guo. 2010. "From Language Structures to Language Use: A Case from Mandarin Motion Expression Classification." *Chinese Language and Discourse* 1 (1): 31–65. DOI: 10.1075/cld.1.1.02che

China State Language Commission and China State Bureau of Standards. 1992. *Xiandai Hanyu Zipin Tongji Biao* [A frequency list of modern Chinese characters]. Beijing: Language and Culture Press.

Chu, Chengzhi, and Xiaohe Chen. 1993. "Jianli Hanyu Zhongjieyu Yuliaoku Xitong de Jiben Shexiang [The initial considerations of creating a Chinese interlanguage corpus system]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 7 (3): 199–205.

Cui, Xiliang. 2005. "Oumei Xuesheng Hanyu Jieci Xide de Tedian ji Pianwu Fenxi [The acquisition of Chinese prepositions by European and American learners and analysis of their errors]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 19 (3): 83–95.

Cui, Xiliang, and Baolin Zhang (eds.). 2013. *Dier Jie Hanyu Zhongjieyu Yuliaoku Jianshe yu Yingyong Guoji Xueshu Taolunhui Lunwen Xuanji* [Proceedings of the second international symposium on the construction and application of Chinese interlanguage corpora]. Beijing: Beijing Language and Culture University Press.

Eifring, Halvor. 1992. *A Concordance to* Baiyujing. Oslo: Solum Forlag.

Feng, Shengli. 2002. *The Prosodic Syntax of Chinese*. Muenchen: Lincom Europa.

Feng, Zhiwei. 2006. "Evolution and Present Situation of Corpus Research in China." *International Journal of Corpus Linguistics* 11 (2): 173–207. DOI: 10.1075/ijcl.11.2.03fen

Feng, Zhiwei. 2012. *Ziran Yuyan Chuli Jianming Jiaocheng* [A concise course of natural language processing]. Shanghai: Shang Foreign Language Education Press.

Granger, Sylviane. 1996. "From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora." In *Languages in Contrast: Text-based cross-linguistic studies*, ed. by Karin Aijmer, et al, 37–51. Lund: Lund University Press.

Granger, Sylviane (ed.). 1998. *Learner English on Computer*. London: Longman.

Granger, Sylviane. 2002. "A Bird's-eye View of Learner Corpus Research." In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, ed. by Sylviane Granger, et al, 3–33. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/lllt.6.04gra

Gu, Yueguo. 2009. "From Real-life Situated Discourse to Video-Stream Data-mining." *International Journal of Corpus Linguistics* 14 (4): 433–466. DOI: 10.1075/ijcl.14.4.01gu

Hai, Liuwen. 2011. *Shisan Jing Zipin Yanjiu* [The frequency study of the thirteen Chinese canons]. Beijing: Higher Education Press.

Halliday, Michael. 1959. *The Language of the Chinese "Secret History of the Mongols"*. Oxford: Basil Blackwell.

Halliday, Michael. 1992. "Language as System and Language as Instance: The Corpus as a Theoretical Construct." In *Directions in Corpus Linguistics: Proceedings of Nobel symposium 82*, ed. by Jan Svartvik, 61–77. Berlin: Mouton de Gruyter.

Halliday, Michael. 2008. *Complementarities in Language*. Beijing: The Commercial Press.

Hung, William. 1932. *Yinde Shuo* [On indexing]. Peking: Harvard-Yenching Institute Sinological Index Series, Peking University Library.

Institute of Language Teaching Research at Beijing Language Institute. 1985a. *Hanyu Cihui de Tongji yu Fenxi* [The statistics and analysis of Chinese words]. Beijing: Foreign Language Teaching and Research Press.

Institute of Language Teaching Research at Beijing Language Institute. 1985b. *Changyong Zi he Changyong Ci* [Frequently used characters and words]. Beijing: The Publishing House of Beijing Language Institute.

Institute of Language Teaching Research at Beijing Language Institute. 1988. *Xiandai Hanyu Pinlu Cidian* [Frequency dictionary of Chinese words]. Beijing: The Publishing House of Beijing Language Institute.

Lau, Din Cheuk, Ho Che Wah, and Chen Fong Ching (eds.). 1992. *A Concordance to* Shuoyuan *No. 1* (ICS Ancient Chinese Texts Concordance Series). Hong Kong: The Commercial Press.

Li, Fanglan. 2011. *Xiandai Hanyu Yuyiyun de Lilun Tansuo yu Xide Yanjiu: Yuliaoku Yuyanxue Shijiao* [A theoretical exploration into semantic prosody and its acquisition of modern Chinese: A corpus linguistics perspective]. Unpublished PhD thesis. Minzu University of China.

Li, Jinman, and Fuyun Wu. 2013. "Leixingxue Gaikuo yu Eryu Xuexizhe Hanyu Guanxi Congju Chanchu Yanjiu [Typological generalisations and the study on the production of Chinese relative clauses by second language learners]." *Waiyu Jiaoxue yu Yanjiu* [Foreign language teaching and research] 45 (1): 80–92.

Li, Jinxi. 1922. "Guoyu zhong Jiben Yuci de Tongji Yanjiu [Statistical considerations of basic ocabulary in Chinese]." *Guowen Xuehui Congkan* [Journal of Chinese language society] 1 (1): 81–84.

Liu, Eric Shen. 1973. *Frequency Dictionary of Chinese Words*. The Hague: Mouton.

Liu, Yuan, Nanyuan Liang, Dejin Wang, Sheying Zhang, Tieying Yang, Chunyu Jie, and Wei Sun. 1990. *Xiandai Hanyu Changyong Ci Cipin Cidian* [A dictionary of frequency of modern Chinese words]. Beijing: Astronautic Publishing House.

Liu, Yun. 2009. "Hanyu Cihui Tongji Yanjiu Shuping [A review of Chinese vocabulary statistical studies]." *Hanyu Xuexi* [Chinese Language Learning] 30 (1): 62–69.

Liu, Zhiji. 2009. "Zipin Shijiao de Gu Wenzi Sishu Fenbu Fazhan Yanjiu [Research on the distribution and development of four categories of character construction in ancient writings from the isual angle of character frequency]." *Gu Hanyu Yanjiu* [Research in ancient Chinese Language] 22 (4): 2–11.

Louw, Bill. 1993. "Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies." In *Text and Technology: In honour of John Sinclair*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 157–176. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/z.64.11lou

Lu, Wu, Fuyin Nan, and Shan Chen (eds.). 2000. *Yuanchao Mishi Jiaozhu* [Collated and annotated secrect history of the Mongols]. Jinan: Qilu Publishing House.

Luke, Kang-kwong, and Theodossia-Soula Pavlidou (eds.). 2002. *Telephone Calls: Unity and Diversity in the Structure of Telephone Conversations across Languages and Cultures*. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/pbns.101

Matthews, Peter. 1991. *Morphology* (2nd Edition). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139166485

McCarthy, John, and Alan Prince. 1995. "Prosodic Morphology." In *Handbook of Phonology*, ed. by John Goldsmith, 318–366. Oxford: Blackwell.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Pan, Shuguang. 1984. *Guji Suoyin Gailun* [Indexing of Chinese classics: A general introduction]. Beijing: Catalogs and Documentations Publishing House.

Sentence Pattern Research Group at Beijing Language Institute. 1989a. "Xiandai Hanyu Jiben Juxing [Basic sentence patterns of modern Chinese]." *Shijie Hanyu Jiaoxue* [Chinese teaching in the world] 3 (1): 26–35.

Sentence Pattern Research Group at Beijing Language Institute. 1989b. "Xiandai Hanyu Jiben Juxing (Xuyi) [Basic sentence patterns of modern Chinese (Continued I)]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 3 (3): 144–148.

Sentence Pattern Research Group at Beijing Language Institute. 1989c. "Xiandai Hanyu Jiben Juxing (Xuer) [Basic sentence patterns of modern Chinese (Continued II)]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 3 (4): 211–219.

Sentence Pattern Research Group at Beijing Language Institute. 1990. "Xiandai Hanyu Jiben Juxing (Xusan) [Basic sentence patterns of modern Chinese (Continued III)]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 4 (1): 27–33.

Sentence Pattern Research Group at Beijing Language Institute. 1991. "Xiandai Hanyu Jiben Juxing (Xusi) [Basic sentence patterns of modern Chinese (Continued IV)]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 5 (1): 23–29.

Siewierska, Anna, Jiajin Xu, and Richard Xiao. 2010. "*Bang-le Yi Ge Da Mang* (Offered a Big Helping Hand): A Corpus Study of the Splittable Compounds in Spoken and Written Chinese." *Language Sciences* 32 (4): 464–487. DOI: 10.1016/j.langsci.2009.08.002

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Tao, Hongyin. 1996. *Units in Mandarin Conversation: Prosody, Discourse, and Grammar*. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/sidag.5

Tao, Hongyin. 2000. "Cong 'Chi' Kan Dongci Lunyuan Jiegou de Dongtai Tezheng ['Eating' and emergent argument structure]." *Yuyan Yanjiu* [Language research] 20 (3): 21–38.

Tao, Zhixing, and Jingnong Zhu. 1923. *Pingmin Qianzi Ke* [Early Chinese lessons for illiterates]. Shanghai: The Commercial Press.

Teubert, Wolfgang. 2005. "My Version of Corpus Linguistics." *International Journal of Corpus Linguistics* 10 (1): 1–13. DOI: 10.1075/ijcl.10.1.01teu

Thompson, Sandra, and Hongyin Tao. 2010. "Conversation, Grammar, and Fixedness: Adjectives in Mandarin Revisited." *Chinese Language and Discourse* 1 (1): 3–30. DOI: 10.1075/cld.1.1.01tho

Thorndike, Edward. 1921. *The Teacher's Word Book*. New York City: Teachers College, Columbia University.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/scl.6

Tsai, Ting Kan. 1922. *Laojielao* [The interpretation of Dao De Jing based on Dao De Jing texts]. Beijing: Self-publication. A synthetic study of LaoTzu's *TaoTeChing* in Chinese

Tsou, Benjamin, and Rujie You. 2007. '*21 Shiji Huayu Xin Ciyu Cidian*' Bianzhu Ganyan [Reflections on compiling *'The Dictionary of Chinese Neologisms for the 21st Century'*]. *Cishu Yanjiu* [Lexicographical Studies] 29 (6): 123–128.

Tsou, Benjamin, and Rujie You. 2010. *Quanqiu Huayu Xin Ciyu Cidian* [An international dictionary of Chinese neologisms]. Beijing: The Commercial Press.

Tsou, Benjamin, Hing-Lung Lin, Terence Chan, Jerome Hu, Ching-hai Chew, and John K. P. Tse. 1997. "A Synchronous Chinese Language Corpus from Different Speech Communities: Construction and Application." *International Journal of Computational Lingusitics and Chinese Language Processing* 2 (1): 91–104.

Unihan Digital Technology Co., Ltd. 2008. *Guji Hanzi Zipin Tongji* [Character frequency calculation of classical Chinese]. Beijing: The Commercial Press.

Wang, Chunxia. 2001. *Jiyu Yuliaoku de Lihe Ci Yanjiu* [*A corpus-based study of splittable sompounds*]. M.A. dissertation, Beijing Language and Culture University.

Wang, Fengyang. 1983. *Ci de Pinlu he Zi de Fenhua* [Word frequency and character differentiation]. Paper presented at the *Second Annual Conference of Chinese Linguistics Society*. Hefei, Anhui, May 1983.

Wang, Haifeng. 2011. *Xiandai Hanyu Liheci Lixi Xingshi Gongneng Yanjiu* [A functional study of the split forms of splittable compounds in Modern Chinese]. Beijing: Peking University Press.

Xiao, Richard, and Tony McEnery. 2004. *Aspect in Mandarin Chinese: A Corpus-based Study*. Amsterdam: John Benjamins Publishing Company. DOI: 10.1075/slcs.73

Xiao, Richard, Paul Rayson, and Tony McEnery. 2009. *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners*. London: Routledge.

Xiao, Xiqiang, and Wangxi Zhang (eds.). 2011. *Shoujie Hanyu Zhongjieyu Yuliaoku Jianshe yu Yingyong Guoji Xueshu Taolunhui Lunwen Xuanji* [Proceedings of the first international symposium on the construction and application of Chinese interlanguage corpora]. Beijing: World Publishing Corporation.

Xiong, Wenxin. 1996. "Liuxuesheng Ba Zi Jiegou de Biaoxian Fenxi [An Analysis of the Performance of Ba Constructions by International Students]." *Shijie Hanyu Jiaoxue* [Chinese Teaching in the World] 10 (1): 80–87.

Xu, Jiajin. 2009. *Qingshaonian Hanyu Kouyu zhong Huayu Biaoji de Huayu Gongneng Yanjiu* [The use of discourse markers in spoken Chinese of urban teenagers]. Beijing: Foreign Language Teaching and Research Press.

Yang, Bojun. 1980. *Lunyu Yizhu* [Annotations to the Analects]. Beijing: Zhonghua Book Company.

Yang, Shiqiao. 2011. *Jiyu Yuliaoku de Hanyu Yihuan Huihua Xiuzheng Yanjiu* [A corpus based study of repair in Chinese doctor–patient conversations]. Unpublished PhD thesis. Shanghai: Shanghai International Studies University.

Zhang, Pu. 1999a. "Guanyu Daguimo Zhenshi Wenben Yuliaoku de Jidian Lilun Sikao [Some theoretical thoughts about the large-scale corpora of authentic texts]." *Yuyan Wenzi Yingyong* [Applied Linguistics] 8, 1, 34–43.

Zhang, Pu. 1999b. "Guanyu Yugan yu Liutongdu de Sikao [On Language sense and degree of circulation]." *Yuyan Jiaoxue yu Yanjiu* [Language Teaching and Linguistic Studies] 21 (2): 83–96.

Zhou, Shengya. 2007. *Soushenji Yuyan Yanjiu* [A linguistic study of *Soushenji*]. Beijing: China Renmin University Press.

Zipf, George. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Mifflin Company.

Zou, Shaohua, and Biao Ma. 2007. *Qiyi de Qingxiangxing Yanjiu* [Studies of preferred interpretations of morpho-syntactic ambiguities]. Beijing: China Social Sciences Press.

Zou, Shaohua. 2001. *Yuyong Pinlu Xiaoying Yanjiu* [Studies in frequency effects of language use]. Beijing: The Commercial Press.

## Appendix A. Some publicly available Chinese corpora

| Corpus name | Corpus creator | Resource URL |
|---|---|---|
| Academia Sinica Balanced Corpus of Modern Chinese | Academia Sinica | http://www.sinica.edu.tw/SinicaCorpus/ |
| BCC (BLCU Chinese Corpus) | Endong Xun, Beijing Language and Culture University | http://bcc.blcu.edu.cn |
| Chinese National Corpus | China State Language Commission | http://www.cncorpus.org |
| HSK Dynamic Corpus of Essays | HSK Testing Service of Beijing Language and Culture University | http://202.112.195.192:8060/hsk/login.asp |
| LCMC (Lancaster Corpus of Mandarin Chinese) corpus | Richard Xiao, Lancaster University, UK | http://124.193.83.252/cqp/ Texts downloadable at http://ota.oucs.ox.ac.uk/scripts/download.php?otaid=2474 |
| LIVAC (Linguistic Variation in Chinese Speech Communities) | The Hong Kong Institute of Education's Research Centre on Linguistics and Language Information Sciences/Chilin (HK) Ltd. | http://www.livac.org |
| National Broadcast Media Language Resources Online | Communication University of China | http://ling.cuc.edu.cn/RawPub/ |
| Sheffield Corpus of Chinese | Xiaoling Hu, Nigel Williamson, and Jamie McLaughlin, Sheffield University | http://ota.ahds.ac.uk/headers/2481.xml |
| Spoken Learner Chinese Corpus of Ji'nan University | College of Chinese Language and Culture, Ji'nan University | http://www.globalhuayu.com/corpus5/Default.aspx |
| The UCLA Corpus of Written Chinese | Hongyin Tao, UCLA | http://124.193.83.252/cqp/ |
| ToRCH2009 (Texts of Recent Chinese 2009) | Jiajin Xu, Beijing Foreign Studies University | http://124.193.83.252/cqp/ Texts downloadable at http://www.bfsu-corpus.org/channels/corpus |
| Written Learner Chinese Corpus of Ji'nan University | College of Chinese Language and Culture, Ji'nan University | http://www.globalhuayu.com/corpus3/Search.aspx |

## Appendix B. Publicly available corpus tools which support the processing of Chinese text concordancers

| Concordancer | Free/commercial | Institution | URL |
|---|---|---|---|
| AntConc | Freeware | Laurence Anthony, Waseda University | http://www.antlab.sci.waseda.ac.jp/software.html |
| BFSU PowerConc | Freeware | Jiajin Xu, Maocheng Liang and Yunlong Jia, Beijing Foreign Studies University | http://www.bfsu-corpus.org/static/PowerConc.html |
| HyConc | Freeware | Nanchang Cheng, Communication University of China | http://ling.cuc.edu.cn/chs/download/HyConcV3.9.6.zip |
| WordSmith Tools | Commercial | Mike Scott, Liverpool University/Aston University | http://www.lexically.net/wordsmith/ |
| Xaira | Freeware | Lou Burnard, Oxford University | http://xaira.sourceforge.net/ |

Chinese word tokenisers/POS taggers/parser/semantic tagger

| Tool | Institution | URL |
|---|---|---|
| ICTCLAS | Institute of Computing Technology, Chinese Academy of Science/Beijing Institute of Technology | http://www.ictclas.org/ictclas_download.aspx |
| MySegTag | China State Language Commission | http://www.cncorpus.org |
| Stanford Word Segmenter | The Stanford NLP Group | http://nlp.stanford.edu/software/segmenter.shtml |
| The Stanford Parser | The Stanford NLP Group | http://nlp.stanford.edu/software/lex-parser.shtml |
| UCREL Chinese Semantic Tagger | University Centre for Computer Corpus Research on Language (UCREL), Lancaster University | http://phlox.lancs.ac.uk:8080/ucrel/semtagger/chinese |

*Author's address*

Jiajin Xu
National Research Centre for Foreign Language Education
Beijing Foreign Studies University
Beijing 100089
China

xujiajin@bfsu.edu.cn