The Corpus Approach to the Teaching and Learning of Chinese as an L1 and an L2 in Retrospect

Jiajin Xu

Abstract The use of corpora in the teaching and learning of Chinese has a history of nearly a century. Pedagogically oriented Chinese corpus studies have originated on a solid methodological footing before computers were available. The creation of concordances and character/word lists, coupled with quantitative analyses of sentence patterns, have offered fascinating insights into Chinese textbook compilation and syllabus design. Such corpus findings have illuminated what lexical items and grammatical patterns should be taught, and in what order vocabulary and grammar points should be presented. Over the last few decades, the craze for Chinese interlanguage corpora has been largely motivated by China's growing global influence. The lexico-grammatical performance in the spoken and written production of Chinese as a second language (CSL) learners has been systematically investigated. Both corpus-based L1 and L2 Chinese studies have been fairly successful in terms of the description of the Chinese (inter)language, but there is still much room for pedagogical implementation, that is, to transform the research into classroom friendly teaching materials.

1 Preamble

A corpus is now commonly understood as a large collection of a representative sample of natural texts, based on which language studies, theoretical or applied, can be conducted with the aid of computer tools (Biber et al. 1998; Hunston 2002). It is safe to say that corpora and corpus methodology have secured a solid niche in present-day linguistics and applied linguistics. The main appeal of the corpus approach to language studies is characterised by the potential of quantitative profiling of actual language use.

J. Xu (✉)
Beijing Foreign Studies University, Beijing, China
e-mail: xujiajin@bfsu.edu.cn

X. Lu and B. Chen (eds.), *Computational and Corpus Approaches to Chinese Language Learning*, Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-13-3570-9_3

The term 'corpus linguistics' appeared as early as 1959 (Voegelin 1959: 216), 'but its roots go way back, unless we restrict the term to the use of texts in electronic form' (Johansson 2011: 115). The early non-machine readable corpora are variously called 'pre-computer corpora', 'pre-electronic corpora', or 'corpora B.C.[1]' (Francis 1992). Thus, the concept of 'corpus' and its related 'corpus method(ology)' and 'corpus approach' in this chapter will refer broadly to both pre-electronic and electronic language databases and their respective theoretical constructs.

Before the advent of computers, the idea of a 'corpus methodology' or a 'corpus approach' has long been experimented and practiced in applied linguistics. For example, around 1820 John Freeman (1843) compiled a frequency list of English words based on approximately 20,000 words to teach adults to read. In 1838,[2] Issac Pitman devised the alphabetic and numerical arrangements of frequent words based on 10,000 words taken from 20 books, 500 from each. Pitman's word list was meant to facilitate the learning of stenography, the practice of writing English words in shorthand. Similar stenography-oriented corpus work was published by German scholar Fredrick Kaeding (1897). Early corpus work better known to the field of language education is Thorndike's (1921, 1931, 1944) corpus-based word books for language teachers.

In this review chapter, both pre-electronic and computerised corpus work will be considered with due reverence. Corpus methodology in pre-computer age may well account for the notion of 'innovation' in language teaching and learning. Admirable to contemporary scholars, early corpus workers tallied individual occurrences of language items all by hand, and formulated probabilistic claims about language use and applied them to pedagogic praxis.

Previous reviews on Chinese corpus linguistics aimed at a comprehensive introduction to the construction of corpora and the corpus research of all kinds (e.g. Feng 2006; McEnery and Xiao 2016; Xu 2015), in which the account on corpus-based Chinese language teaching and learning were cursory. This article, however, will mainly focus on the corpus approach to language teaching and learning of Chinese, both as a mother tongue and a second language. Moreover, pedagogically informed Chinese corpus research and practice in Chinese mainland, Hong Kong, Taiwan and overseas will all be addressed *passim* in this chapter.

[1] 'B.C.' here is Nelson Francis' play on words, meaning 'Before Computer' rather than its literal sense 'Before Christ'.

[2] Issac Pitman's word frequency list was published in 1843 in *The Phonotypic Journal*, but the research was done in 1838 (Pitman 1843: 161).

2 Key Corpus Projects in Teaching and Learning Chinese as an L1

2.1 First Concordance Projects for Chinese Classics Exegesis

The earliest Chinese corpus projects commenced in China around the 1920s and were motivated by classic exegesis and basic literacy. Back to the Qing Dynasty (A.D. 1644–1911), mainstream scholars worked on Chinese literary and metaphysical canons. After the downfall of China's last feudal dynasty, some elite Chinese scholars committed themselves to collating the myriad of Chinese classics. Stemming from dissatisfaction with the hundreds of contradicting and confusing annotations and interpretations of Chinese philosophical canon *Lao Tzu's Tao-Te-Ching*, the Western-educated Admiral 蔡廷干 Ting-Kan Tsai (1861–1935) compiled probably the first ever Chinese concordance 老解老 *Laojielao 'A Synthetic Study of Lao Tzu's Tao-Te-Ching in Chinese'* (1922). The text of *Tao-Te-Ching*, in this case, was considered a corpus, based on which a frequency list of all characters was created. The frequency count of a character and the original sentences and chapters in which it appeared were presented (as illustrated in Fig. 1 where the sixth character 無 occurs 102 times). One of the primary purposes of the concordance was meant to be a learning aid for younger generations to understand *Tao-Te-Ching*, especially when such key philosophical terms as 無 wu 'nothingness' was considered elusive and ambiguous given limited context.

Tsai's concordance initiative was acclaimed and shortly followed up by William Huang's (1893–1980) *Harvard-Yenching Institute Sinological Index Series* from the early 1930s (Hung 1931, 1932: 9–10). It was a gigantic enterprise which published a total of 64 titles in 77 volumes of concordances of Chinese classics from 1931 to 1950 (Huang 1962–1963: 8), such as the complete concordances of 易经 *Yijing* 'Book of Changes', 礼记 *Liji* 'The Classic of Rites', 论语 *Lunyu* 'The Analects', 孟子 *Mengzi* 'The Works of Mencius', and the like. The first volume of the index series is 说苑引得 *Shuoyuan Yinde* 'Index to Shuo Yuan' (Hung 1931), and the fourth volume (Hung 1932)—引得说 *Yinde Shuo* 'On Indexing'—of the series is a theoretical work in which Chinese concordance method 中国字庋擷 *Zhongguo zi guixie* 'Chinese character based retrieving' was formulated. Hung (ibid.: 8) explains that 引得 is the transliteration of English terminology 'index', and also known as 堪靠灯 *kenkaodeng* whose English equivalent is 'concordance'. Hung prefers the term 'index' to its synonymous counterpart 'concordance'; the two are not exactly the same though. The Index Series has become a key resource for learners and researchers of Chinese canons.

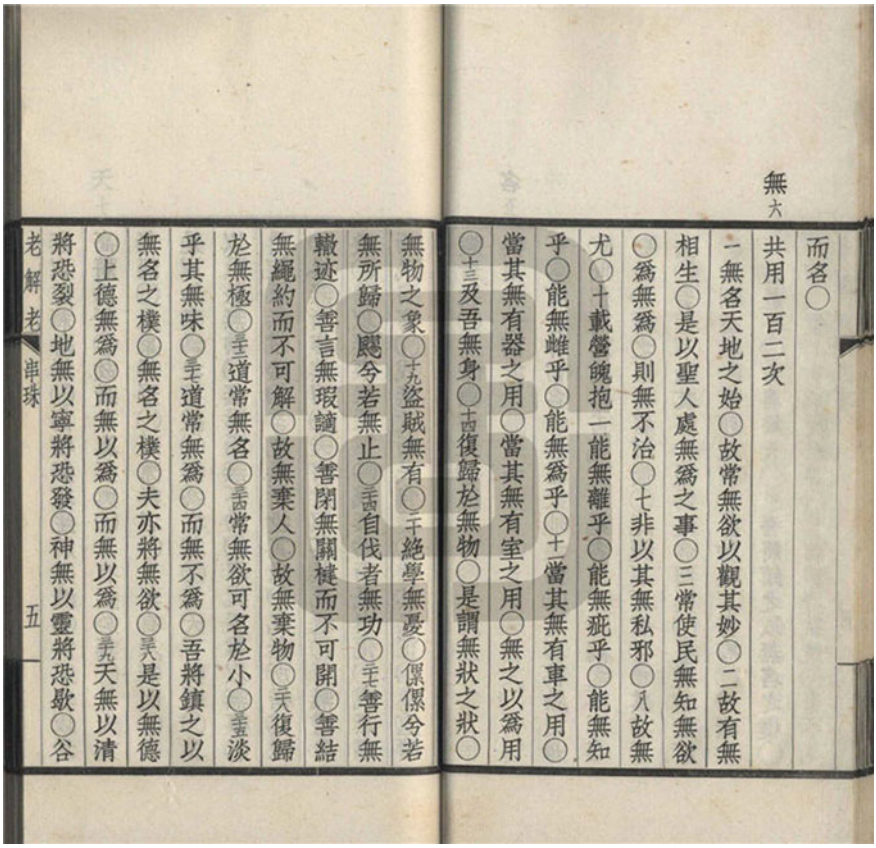
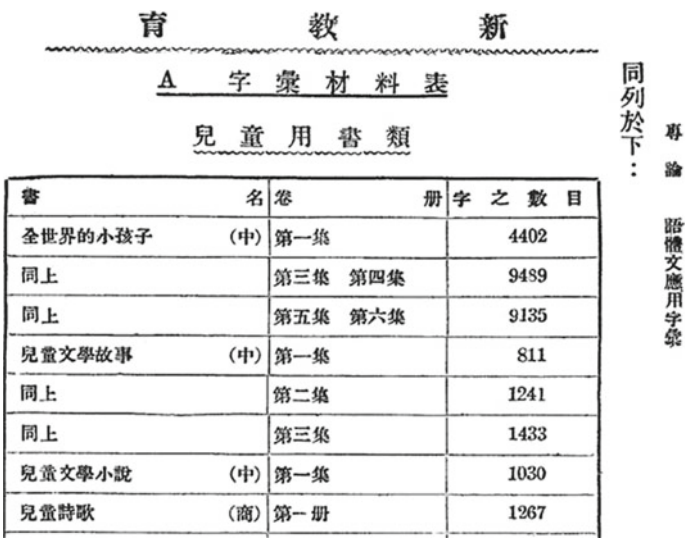


Fig. 1 A snapshot of *Laojielao* '*A synthetic study of Lao Tzu's Tao-Te-Ching in Chinese*' concordances (p. 51) (The two pages shown here present part of the concordance of the 102 occurrences of 無 'Nothingness' in *Tao-Te-Ching*. The circles between the lines serve as the separator of 無 sentences. The smaller font size numerals attached to the initial position of some sentences, such as 一 before 無名天地之始, indicate the chapter number where the sentence can be found in *Tao-Te-Ching*)

2.2 *Pre-computer Chinese Corpus-Based Character Lists and Chinese Textbooks for Basic Literacy*

In the beginning years of the twentieth century, China experienced great political turbulence and instability. Against this backdrop, the massive number of illiterates turned out an imminent problem for the government at the time. The psychologist and educationalist 陈鹤琴 Heqin Chen addressed the critical social issue by virtue of a corpus-based project on compiling a Chinese character list. 'While the data used in Chen was not computerised, his list of basic Chinese characters was nevertheless corpus-based' (McEnery and Xiao 2016: 439). The intended goal of the project was to



新教育

A 字彙材料表

兒童用書類

書名	卷冊	字之數目
全世界的小孩子 (中)	第一集	4402
同上	第三集 第四集	9489
同上	第五集 第六集	9135
兒童文學故事 (中)	第一集	811
同上	第二集	1241
同上	第三集	1433
兒童文學小說 (中)	第一集	1030
兒童詩歌 (商)	第一册	1267

同列於下：

專論 語體文應用字彙

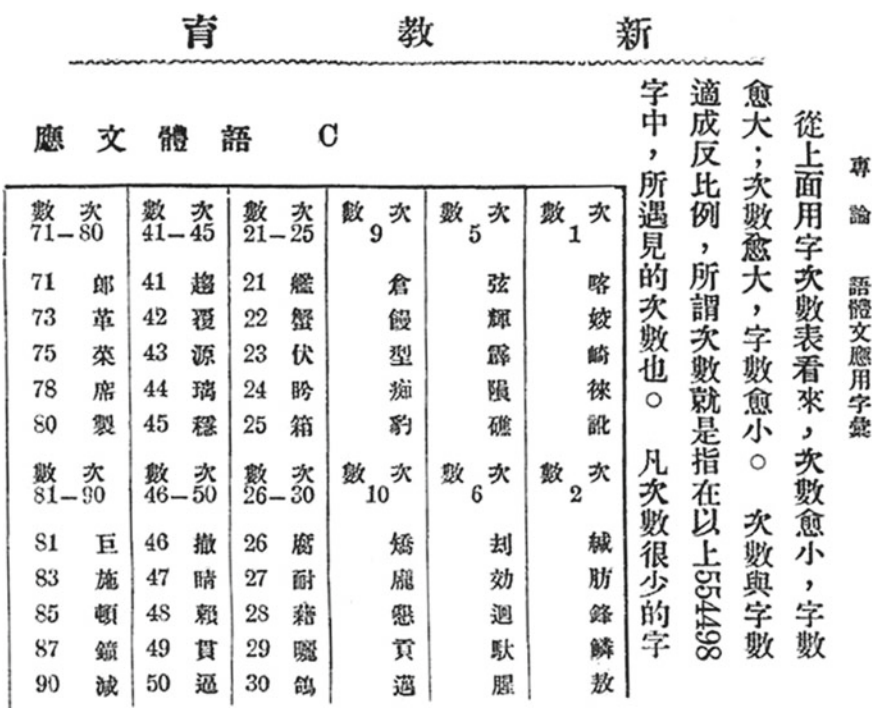
Fig. 2 A snapshot of text sampling in *Yutiwen yingyong zihui* '*Characters used in vernacular Chinese*' (Chen 1922: 990)

survey, from a language education curriculum perspective, how many characters and in what sequence the characters should be exposed to illiterate learners of Chinese.

Chen and his nine associates, between 1919 and 1921, collected vernacular Chinese (rather than classical Chinese used by intellectuals and aristocrats) texts totalling 554,478 characters from a wide spread of genres, ranging from children's literature, news, magazines and vernacular Chinese fiction. A frequency list of characters (Chen 1922, see Figs. 2 and 3) was created and sorted in both radical and frequency order. Four thousand two hundred and sixty-one distinct characters were identified from the corpus. Chen's list was widely known as 语体文应用字汇 *Yutiwen yingyong zihui* 'Characters used in vernacular Chinese'. It was later expanded by Chen himself with a larger corpus of 902,678 characters (Chen 1928) and significantly updated by Liu (1926) and Ao (1929a, b) with more everyday and practical Chinese writing text samples.

Chen's corpus work was immediately hailed by Chinese educators and adopted in Chinese textbooks (Tao and Zhu 1923; Ao 1929a, b). Graded character lists based on Chen's 4261 characters became the vocabulary selection criterion for many Chinese textbooks. For instance, about one thousand frequent characters served as the word ladder, so to speak, for all series of the *Pingmin*[3] *Jiaoyu* 'Mass Education' textbooks. The series were sold for more than three million (Liu 1926: 1) copies after their publication in about 3 years. This was phenomenal given the population of 350 million Chinese people in the 1920s.

[3] *Pingmin* was defined by the leaders of the Mass Education Movement as illiterates (Tao and Zhu 1923: 43).


新教育

專論 語體文應用字彙

從上面用字次數表看來，次數愈小，字數愈大；次數愈大，字數愈小。次數與字數適成反比例，所謂次數就是指在以上554498字中，所遇見的次數也。凡次數很少的字

C 語體文應用字彙

次數 1		次數 5		次數 9		次數 21—25		次數 41—45		次數 71—80	
次數 2		次數 6		次數 10		次數 26—30		次數 46—50		次數 81—90	
喀	絨	弦	輝	倉	饅	21	繼	41	趨	71	郎
姣	崎	輝	露	饅	型	22	蟹	42	覆	73	革
崎	徠	露	隈	型	痴	23	伏	43	源	75	菜
徠	訛	隈	礁	痴	豹	24	盼	44	璃	78	席
訛	次	礁	次	豹	次	25	箱	45	穆	80	製
次數 2		次數 6		次數 10		次數 26—30		次數 46—50		次數 81—90	
絨	械	次	封	次	矯	26	腐	46	撒	81	巨
械	肪	封	効	矯	龐	27	耐	47	晴	83	施
肪	鋒	効	迴	龐	懇	28	蕃	48	賴	85	頓
鋒	鱗	迴	駄	懇	貢	29	曬	49	貫	87	鎳
鱗	敖	駄	屋	貢	適	30	鴿	50	逼	90	滅

Fig. 3 A snapshot of character counting in *Yutiwen yingyong zihui* '*Characters used in vernacular Chinese*' (Chen 1922: 994)

The use of Chinese textbooks with careful vocabulary control was part of the nationwide mass education movement that attempted to educate the illiterates in an efficient manner. That is to say, the character list enabled instructors to teach the most common Chinese characters to illiterates in cities, in rural areas, and in the army during their limited time after work. Listed below are a few textbooks used at that time.

(1) Book 1–Book 3 of 平民千字课 *Pingmin qianzi ke* 'Foundation characters' by 晏阳初 Y. C. James Yen and 傅葆琛 Daniel C. Fu (1924) of the National Association of Mass Education Movement. About 300 characters were allocated to each book. The series were Romanised and translated into English under the name of '1000 Chinese Foundation Characters' by William White to be used by Western learners of Chinese. The 1000 Chinese characters were divided into four levels and 250 characters for each level. Its international students oriented edition was published by the University of Toronto in 1944.

(2) Book 1–Book 4 of 平民千字课 *Pingmin qianzi ke* 'Early Chinese lessons for illiterates' compiled by 陶知行 Zhixing Tao[4] and 朱经农 Jingnong Zhu. Book 1 was published in 1923 by the Commercial Press.
(3) Book 1–Book 4 of 市民千字课 *Shimin qianzi ke* 'Textbook of one thousand characters for townspeople' compiled by the National Association of Mass Education Movement (1928).
(4) Book 1–Book 4 of 农民千字课 *Nongmin qianzi ke* '1000 Chinese foundation characters for peasants' compiled by the National Association of Mass Education Movement (1922a).
(5) Book 1–Book 4 of 士兵千字课 *Shibing qianzi ke* '1000 Chinese foundation characters for servicemen' compiled by the National Association of Mass Education Movement (1922b).

Amongst them, in addition to the better-known Heqin Chen's (1922) character list and Zhixing Tao and Jingnong Zhu's corpus-informed texts, Y. C. James Yen and his language education projects merit special mention. Yen received his PhD from Yale University. Upon his graduation in 1918 which was the time of World War I, he went to France as a volunteer for the Y.M.C.A. to teach the 20,000 illiterate Chinese labourers to read. Yen developed a basic Chinese vocabulary of about 1300 characters in his interaction with the coolies. His basic Chinese vocabulary was integrated into his Foundation Chinese textbooks. Yen and Chen developed their 1000 basic characters independently, and luckily two character lists share over 80% of the characters (Yen 1922: 1012).

Yen and Tao were close partners in the so-called National Association of Mass Education Movements which was first organised by Yen in Beijing in 1923. In 1921, Zhixing Tao and other educationists founded the China Education Improvement Society in Nanjing, the then capital of China at the time. The two leading educationalists and activists of China at the time joined hand in the national campaign to combat illiteracy. The like-minded twin-star scholars tried their utmost to ameliorate the overall literacy situation on a 'maximum vocabulary and minimum time' basis. The core vocabulary formed the basis of their syllabus of Chinese teaching, and the 1000 characters came from their experimental method of collecting frequently used characters in real life. The underlying idea here is none other than corpus methodology.

Heqin Chen's character list can be regarded as a general purpose common core Chinese vocabulary. Specialised vocabulary was also counted up in existing compilations of Chinese textbooks for different target readerships. For instance, James

[4]陶知行 Zhixing Tao was an early alias of the better-known Chinese educationalist 陶行知 Xingzhi Tao. The difference in name says something about the transformation of his philosophy. Influenced by Chinese Confucianist scholar 王阳明 Wang (1472–1529), Tao (1891–1946) took his name *Zhixing* (meaning knowledge-action) in the 1910s and *Xingzhi* (meaning action-knowledge) in 1934 (Tao 1934: 286–287). Both names, *Zhixing* and *Xingzhi*, showed Tao's identification with Yangming Wang's theory of 知行合一 *zhi xing he yi* 'unity of knowledge and action'. When *Xingzhi* was adopted, Tao seemed to prioritise *xing* (action) over *zhi* (knowledge), which suggested that knowledge was derived from empirical engagement (Boorman 1970: 243–244; Browning and Bunge 2009: 388).

Yen's collaborator Daniel Fu collected plenty of texts of farming, gardening, contracts, almanacs, invitation letters, and other practical writings by Chinese farmers and discovered more characters which were not on Heqin Chen's and James Yen's lists, when he prepared the Chinese textbooks for farmers. The new characters that Fu found could successfully distinguish general Chinese from farming Chinese, and served the very learning need of illiterate farmers.

To summarise the early, pre-computer corpus-based language education, it is clear to see that 100 years ago in China the most pressing concern was the eradication of illiteracy. The empirically grounded character lists and the corresponding textbooks proved to be extremely effective and helpful to common people of all walks of life. The textbooks had been popular for about two decades, but the momentum was suspended and discontinued due to the power change in 1949.

2.3 Computer Corpus-Based Lexical Studies for the Teaching and Learning of Chinese

Chen and Yen type of corpus-based language teaching and learning endeavour were unheard of until the years after the Cultural Revolution (1966–1978). From 1979 onwards, corpus projects started to grow along a clearly uphill trajectory in different parts of China, and computer corpora played a central role in it.

Since the late 1970s and the early 1980s, an increasing number of corpus projects contributed to the teaching of Chinese language in one way or another. The greatest number of new character lists and tokenised word lists had been made available based on larger updated corpora.

The following list serves as a quick overview of the key corpus-based Chinese lexical frequency lists. Please refer to Xu (2015) for a comprehensive review.

(1) Liu (1973). Frequency Dictionary of Chinese Words.
(2) Bei and Zhang (1988). Hanzi Pindu Tongji [Frequency Calculation of Chinese Characters].
(3) Liu et al. (1990). Xiandai Hanyu Changyong Ci Cipin Cidian [A Dictionary of Frequency of Modern Chinese Words].
(4) China State Language Commission and China State Bureau of Standards. (1992). Xiandai Hanyu Zipin Tongji Biao [A Frequency List of Modern Chinese Characters].
(5) Huang et al. (1996). The Academia Sinica Balanced Corpus for Mandarin Chinese.
(6) Tsou et al. (1997). LIVAC (LInguistic VAriation in Chinese communities) Synchronous Corpus.
(7) Zhang (1999). The Dynamic Circulation Corpus (DCC).
(8) Xiao et al. (2009). A Frequency Dictionary of Mandarin Chinese.

The latest national achievement of the corpus-based Chinese lexical project is 通用规范汉字表 *Tongyong Guifan Hanzi Biao* 'A General Service List of Chinese

Characters'. The character list was compiled by the Chinese State Language Commission and officially released by China's State Council in June 2013. The general service character list is made up of three graded character lists: 3500 basic characters as Level One, 3000 characters as Level Two, and Level Three with 1605 proper nouns, technical, domain-specific and archaic Chinese characters. The lists, especially the first two levels, are based on the frequency counts of the Chinese National Corpus. Other character lists, such as Bei and Zhang (1988) and Zhang (1999), were also integrated into the list.

Once an officially approved character list is in place, language pedagogy professionals, and scholars do not need to build a corpus or create character lists on their own, as the national character is based on a large and balanced corpus of modern Chinese—the Chinese National Corpus.

Most Chinese corpus projects reviewed so far focus on the creation of a lexical frequency list, and some of them produce both frequency lists and their document frequency, namely, the distribution across different genres or text types. However, this is still insufficient from a language curriculum perspective.

2.4 Computer Corpus-Based Grammatical Studies for the Teaching and Learning of Chinese

Corpus-based grammatical studies are far fewer than that of corpus-based lexical studies. During pre-computer time, the quantitative analysis of sentence patterns was seldom, if ever seen. When computer technology is available, the calculation of sentence patterns is still an underdeveloped field. Amongst the very few corpus-based grammatical studies for Chinese pedagogic purposes, Shuhua Zhao's 现代汉语基本句型 *Xiandai Hanyu Jiben Juxing* 'Basic Sentence Patterns of Modern Chinese' (The Sentence Pattern Research Group at Beijing Language Institute 1989a, b, c, 1990, 1991)[5] is a project that deserves special attention. Zhao and her project team made an exhaustive counting of all major sentence patterns in some elementary and secondary school Chinese textbooks as well as in some intensive Chinese reading textbooks used for college students.

The occurrence and distribution across different programme levels of broad sentence types in Chinese textbooks, such as declaratives, general interrogatives, rhetorical questions, imperatives, exclamatory, and negative sentences were systematically tallied and reported. One of their statistical reports is reproduced in Table 1.

In a similar fashion, grammatical categories were computed, for instance, the use of sentences with a lexical verbal phrase predicate, focus constructions 是…的 *shi…de*, existential sentences, 把 *ba*-constructions, 被 *bei*-constructions, serial verb constructions, and so forth.

[5]赵淑华 Shaohua Zhao was the lead scholar and director of the Sentence Pattern Research Group at Beijing Language Institute.

Table 1 An example of Zhao's statistical tables of Chinese sentence patterns (Zhao et al. 1995: 16)

Sent. type	Textbook								
	College Chinese (Elementary)					College Chinese (Intermediate)			
	Bk 1	Bk 2	Bk 3	Subtotal	% of all sent.	Bk 1	Bk 2	Subtotal	% of all sent.
Declarative	415	801	1660	2885	83.45	1837	2854	4691	89.3
General interrogative	148	128	150	426	12.32	67	151	218	4.15
Rhetorical question	1	3	29	33	0.96	52	87	139	2.65
Imperative	38	21	10	69	2.00	46	79	125	2.38
Exclamatory	3	13	28	44	1.27	30	50	80	1.52
Total	605	966	1886	3457		2032	3221	5253	
Negative[a]	42	73	173	290	8.39	257	424	681	12.96

[a]The negative sentence frequency counts were not included in the total number of the sentence types, but placed underneath the Total row. This last row featuring the negative sentences has been transcribed exactly as found in the source table

This comprehensive sentence pattern project proves to be a solid and sound resource for the understanding of the general patterns of Chinese language use at the syntactic level in primary, secondary and tertiary Chinese language textbooks. It is needless to say that Zhao's research is a great resource for Chinese as a second language teaching and learning as well. As a matter of fact, the project was largely meant to serve the purpose of language teaching for CSL in the first place. Essentially, the description of grammatical patterns goes far beyond the teaching of Chinese as a second language (TCSL). As Zhao claimed in the Sentence Pattern Research Group at Beijing Language Institute (1989a), the results of the project would become an important resource for Chinese textbook compilation, Chinese syntactic studies in general, as well as natural language processing applications such as machine translation.

The sentence pattern database has been chiefly employed to look for and sort out frequent and less frequent, typical simple and complex sentence patterns for language educators and language teachers. As it is commonly observed, grammatical change is far less dramatic than it is in the case of lexis. This means that research conducted around 1990 is still of currency and validity for language education today.

2.5 L1 Chinese Production Corpora: Collections of School Pupils' Essays

The Chinese corpus projects discussed so far can be understood as the 'input corpus' (Sugiura 2002: 316; Nesselhauf 2004: 146)[6] work. The language input here refers to Chinese language teaching textbooks and/or newspapers, fiction, essays etc. that learners are likely to read in real life. The corpus compliers can gather texts from different input sources, make statistical claims about Chinese characters, words, phrases and sentences, and eventually turn them into teaching and learning resources. However, the language output, or production, of native Chinese speakers has not been attended to in real earnest in corpus-based Chinese studies. The Chinese school pupil's essay corpus developed at the Institute of Modern Educational Technology, Beijing Normal University was an important undertaking of L1 Chinese output corpus (Wei et al. 2008). In the project report, as of August 2007, the corpus contained over 11 million characters of Chinese texts from elementary and secondary school pupils of five grades (i.e. first to fifth graders) across China. 162 schools (148 elementary schools and 14 secondary schools) were involved, and seven cities were covered (namely Beijing, Fengning, Dalian, Guangzhou, Shenzhen, Xiamen and Hong Kong). To a large extent, the databank was developmental in the sense that Chinese essays of the same groups of pupils were archived and arranged in a chronological order (Table 2).

[6]The term 'input corpus' is used by some learner corpus linguists meaning the collection of learners' language exposures such as teachers' talk in class as well as the written texts that the learners are confronted with in learning. In this article, input corpora mainly refer to the written texts, textbooks in particular.

Table 2 Some key information of the school pupil essays

Categories	Statistics
Number of essays	79,244
Chinese characters	11,456,403
Number of pupils	2164

A query system was developed for the corpus; both standard queries and queries with sociolinguistic variables (e.g. region, school type, year of entering grade one, date of writing, grade, etc.) were enabled. All the texts were tagged for part of speech and annotated for syntactic patterns; therefore, word-class and grammatical category based queries were also available. Unfortunately, the corpus is not publicly accessible. The construction of learner production corpora of this kind should be encouraged, and it would be very much desirable to promote the sharing of corpus resources among language researchers and practitioners at large.

Quantitative analysis of Chinese language proudly emerged nearly a century ago at a very high standard in terms of the quality of research and the theoretical and practical impact they had on Chinese language education. Over the years, the majority of pedagogically oriented Chinese corpus research has been on the building of lexical frequency lists, viz. character lists and tokenised word lists. The corpus texts focus more on Chinese language input, e.g. Chinese textbooks or newspapers, magazines, popular readings and fiction. Less attention was directed to native Chinese learners' L1 production. Besides, corpora of Chinese for specific and academic purposes need special attention (cf. Chen and Tao this volume).

3 Key Corpus Projects in Teaching and Learning Chinese as an L2

The last two decades have seen a development boom in corpus-based Chinese studies for Teaching Chinese as a Second Language due to China's growing global influence. As such, the construction of Chinese interlanguage corpora has become very popular in the wake of the augmented enrolment of international learners of Chinese. Learner corpora started to emerge in the West around 1993, according to Granger (2015: 7). The first Chinese learner corpus project began at Beijing Language and Culture University also in 1993 independently without any scholarly communication with the Western corpus linguists. Unlike the development of L1 Chinese corpus research, L2 Chinese corpus research, from its onset, prioritised learner production data. Chinese L2 input corpora only caught up at a later stage. Note that most L1 Chinese lexical frequency list projects, including some recent ones such as Xiao et al. (2009), may well be adopted for the teaching and learning of Chinese as an L2.

3.1 L2 Learner Chinese Corpora

The earliest corpus-based interlanguage Chinese studies date back to 1993 at the Beijing Language Institute, now Beijing Language and Culture University (BLCU) (Chu and Chen 1993). The corpus, consisting of the first Chinese interlanguage dataset, was described at length in Chen (1996). The overall corpus size was about 3.5 million characters, and a re-sampled core L2 Chinese learner corpus was about one million characters. 23 textual and sociolinguistic variables were marked up. All the student essays in the Chu and Chen's corpus were sentence segmented, tokenised and tagged for part of speech categories (Table 3).

BLCU's Chinese interlanguage corpus construction was followed up in the late 1990s and the early 2000s by the hitherto most frequently cited L2 Chinese learner corpus, *HSK (Hanyu Shuiping Kaoshi) Dongtai Zuowen Yuliaoku 'Chinese Proficiency Test Dynamic Essay Corpus'* (Zhang 2003). The first release of the HSK corpus contained more than 20,000 essays written by HSK test-takers starting in 1992, and as the corpus keeps growing, the modifier 'dynamic' is added to the corpus name.

BLCU's L2 Chinese learner corpus work is now being upgraded to a globally-oriented interlanguage Chinese corpus project—the International Corpus of Learner Chinese. The projected corpus size will be 50 million characters, including 45 million written interlanguage Chinese and five million spoken interlanguage Chinese (Cui and Zhang 2011).

BLCU has been the leader in L2 Chinese learner corpus research. Other research teams in the field, however, have developed their own distinctive Chinese interlanguage corpus projects, such as learner corpora with intensive annotation on character misspelling as well as more balanced corpora of spoken and written learner Chinese.

The writing of Chinese characters is supposedly the hardest part of Chinese learning. The L2 Chinese learner corpus developed at the National Taiwan Normal University (Teng et al. 2007) is arguably the earliest Chinese interlanguage corpus which has a specific focus on (traditional) Chinese character writing errors. In Phase I of the project, 2457 instances of misspelling were collected from 72 learners of Chinese from 22 countries. An additional 52 learners' data were archived and 1858 misspellings were annotated for Phrase II. All the misspellings were classified into one of nine error types (i.e. *quesheng* 'omission', *zengbu* 'addition', *daihuan*

Table 3 Information on the first interlanguage Chinese corpus by Chu and Chen (1993)

Attributes (partial)	Value
L1 background	59 countries
Age range	16–35
Male/female	50.93%/47.93% (remaining unstated)
Task types	Homework essays (63.45%), exam essays (15.31%), writing after reading or listening (19.21%)

'substitution', *fenhe* 'division/combination', *cuowei* 'misplacement', *chutou* 'cross-the-border', *jingxiang* 'flip', *bianxing* 'transformation', *hezi* 'blending', *jianhuazi* 'simplification', *xingsizi* 'deja vu'). The image files of the errors were stored alongside each entry in the corpus. The National Taiwan Normal University corpus (Teng et al. 2007) might be disqualified as a corpus because its size is too small, and only individual characters, rather than running texts, were recorded in the database.

Hanzi Pianwu Biaozhu de Hanyu Lianxuxing Zhongjieyu Yuliaoku 'The Continuity Corpus of Chinese Interlanguage of Character-error System' developed at the Sun Yat-sen University is another important corpus that deals with the misspelled (simplified) Chinese characters. The interlanguage writing samples were complete texts, and all tokenised and part of speech tagged. The misspelled characters were inserted into the texts using the Microsoft Windows Font Creator Program—a True-Type Chinese character/font editing application. More importantly, all the originally hand-written essays were scanned as image files. Users of the corpus can view the scanned essays for more contextualised analysis of what the non-native Chinese writers actually wrote (Zhang 2017).

Some recent Chinese interlanguage corpus projects aim at a more balanced design covering both spoken and written interlanguage. Jinan University Chinese learner corpus (JCLC) (Wang et al. 2015) and Guangwai-Lancaster Chinese Learner Corpus[7] are two cases in point. The JCLC written corpus contains 5.91 million Chinese characters across 8739 texts, and the spoken part is composed of 350,000 characters. Guangwai-Lancaster Chinese Learner Corpus is more balanced in terms of spoken and written data proportion. It is a 1.2-million-word corpus of interlanguage Chinese with a spoken (621,900 tokens, 48%) and a written (672,328 tokens, 52%) part, covering a variety of task types and topics. The entire corpus is tagged for errors as well.

A few other Chinese interlanguage corpora cited in the literature include those developed at Ludong University (Hu and Xu 2010), the University of Hong Kong (Tsang and Yeung 2012), Nanjing Normal University (Xiao and Zhou 2014), and many other universities.

The construction of L2 Chinese learner corpora has become the empirical basis for many doctoral theses and research articles and monographs. The usage patterns of interlanguage Chinese at morphosyntactic and textual levels have been investigated. However, the overwhelming focus of Chinese interlanguage corpus studies has been on error analysis. This might be accounted for by the strong impetus of language praxis.

Another indicator of the progress of L2 Chinese corpus research is that a biennial Chinese learner corpus research conference series 'The International Symposium on the Construction and Application of Chinese Interlanguage Corpora' has been in place since 2010.

[7]The corpus is freely available at https://www.sketchengine.co.uk/guangwai-lancaster-chinese-learner-corpus/.

3.2 L2 Chinese Learners' Input Corpora

Apart from the corpus-based Chinese interlanguage studies, there are some projects on L2 learners' input corpora. For instance, the corpus of Chinese textbooks for international students (CSL textbooks hereafter) developed at Xiamen University (Su 2010) has been made publicly available online,[8] which has become a useful resource for researchers and teachers of Chinese. Eleven types of Chinese textbooks published between 1992 and 2006 were digitised, and modelled into a corpus format. The total running characters of the corpus is 771,350. Besides, Sun Yat-sen University has constructed an updated CSL textbook corpus, which includes 1752 textbooks (54.5% of the total 3212 textbooks) published after 2006. 1802 out of 3212 textbooks (56.1%) were published outside China (Zhou et al. 2017). The two corpus teams have conducted a series of research on the coverage of vocabulary and grammar points across different CSL textbooks.

3.3 Applications of Data-Driven Teaching and Learning of Chinese

Chu's (2004) ChineseTA is probably one of the best known Chinese teaching software packages that integrate corpus linguistics functionalities. For instance, it can compute the occurrences and distribution of characters and words for the loaded Chinese teaching materials, and identify new words against the built-in level lists (e.g. HSK lists) as well as new word proportion. Such corpus-based data-driven features provide quantitative measures for Chinese texts that teachers can adopt for students of certain proficiency levels (Fig. 4).

Kilgarriff et al. (2015) demonstrate how Chinese teaching and learning can benefit from the data-driven methods with the assistance of the online system Sketch Engine. The system provides both Chinese corpora (together with a large number of corpora in other languages) and corpus analysis tools (e.g. concordance, word sketch, sketch difference, thesaurus, etc.). Word sketches—a different name for collocations—are key to the tool. Sketch diff (the shortened form on the system interface for sketch difference) is an often cited feature applicable to meaning distinction of ambiguous near-synonyms. Figure 5 shows the different collocational patterns between 形成 *xingcheng* 'to form' and 造成 *zaocheng* 'to cause' computed by Sketch Engine. Apparently, *xingcheng* tends to co-occur with neutral words like 共识 *gongshi* 'consensus', while *zaocheng* tends to co-occur with negative words, such as 伤亡 *shangwang* 'casualties' and 损失 *sunshi* 'loss'.

The data-driven learning trials on Chinese are still scarce. More collaboration between language practitioners, materials developers, publishers and corpus linguists should be encouraged to produce some corpus-informed computer-assisted language learning tools and mobile or cloud-enhanced learning applications.

[8] http://ncl.xmu.edu.cn/shj/Default.aspx. Chinese textbooks for native Chinese students were also collected alongside the CSL textbook data.

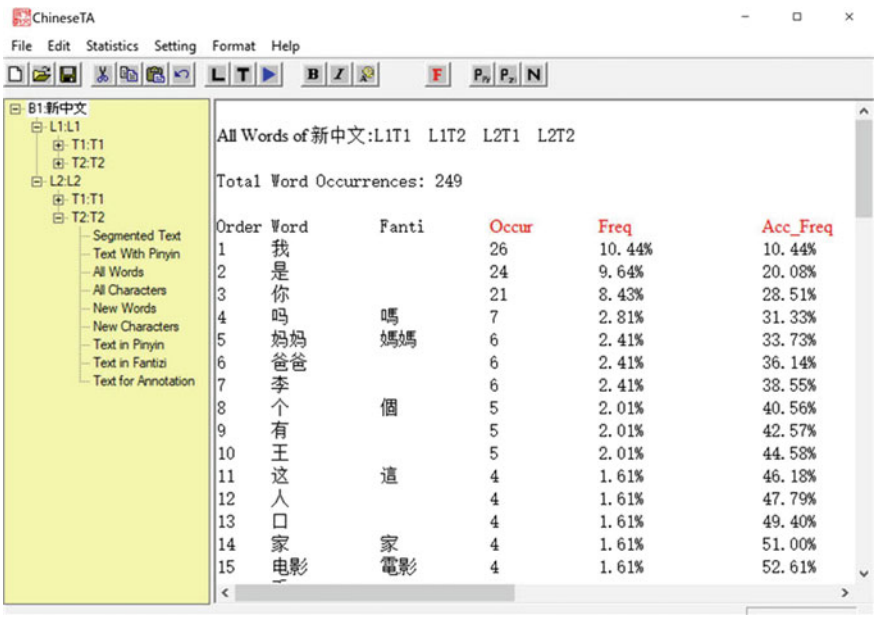


Fig. 4 The frequency list feature of ChineseTA

Fig. 5 The sketch diff feature of Sketch Engine: an example of 形成 versus 造成

4 Concluding Remarks

The use of corpora in the teaching and learning of Chinese has a history of nearly a century. Pedagogically oriented Chinese corpus studies originated on a solid methodological foundation prior to computer use. The creation of concordances, character/word lists, and the quantitative analyses of sentence patterns have offered fascinating insights into Chinese textbook compilation and syllabus design, the range and types of lexical items and grammatical patterns that should be taught, as well as the order in which vocabulary and grammar points should be presented.

The adoption of a corpus approach to the teaching and learning of Chinese is innovative in the sense that it relies on a quantitative methodology to look at Chinese. The corpus approach highlights the exhaustive account of all lexical, grammatical and textual features. This methodological innovation, nonetheless, should not be reduced to a technological advancement alone. According to Willis (1990a, b), what underlies the corpus approach to language teaching and learning is a descriptivist view of language, an inductive way of learning, and a task-based lexical syllabus. The notions have been best materialised in data-driven learning (Johns 1991). Corpus-based L1 and L2 Chinese studies have been fairly successful in terms of the description of the Chinese (inter)language, but there is still much room for pedagogical implementation, that is, to transform the research findings into classroom friendly teaching materials.

In this regard, we can find notable examples in English, such as corpus-based learner dictionaries, e.g. *Collins COBUILD English Dictionary* (Sinclair 1987), pedagogical grammar books, e.g. *Longman Student Grammar of Spoken and Written English* (Biber et al. 2002) and *Real Grammar: A Corpus-Based Approach to English* (Conrad and Biber 2009), English textbooks, e.g. *Collins COBUILD English Course* series (Willis 1990a, b) and Cambridge University Press' *Touchstone* and *Viewpoint* series (McCarthy and McCarten 2012; McCarthy et al. 2004), ESP and EAP teaching and learning materials e.g. *Academic Vocabulary in Use* (McCarthy and O'Dell 2008), classroom concordancing or data-driven learning tasks and activities, e.g. Tribble and Jones (1990), and the theorising about corpus-based language teaching, e.g. *The Lexical Syllabus* (Willis 1990a, b) and *The Lexical Approach* (Lewis 1993).

Acknowledgements The research was supported in part by the key project of the National Research Centre for Foreign Language Education (MOE Key Research Institute of Humanities and Social Sciences at Universities) (Ref No.: 17JJD740003) at Beijing Foreign Studies University. The author gratefully acknowledges the funding of the Fulbright Visiting Scholar grant during the writing up of the article. The author would also like to thank the referees and editors for their constructive comments, Professor Chengzhi Chu for providing helpful resource of ChineseTA, and Professor Eniko Csomay and Dr. Lu Lu for proofreading the manuscript.

References

Ao, H. (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, H. (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, G., & Zhang, X. (1988). *Hanzi pindu tongji [Frequency calculation of Chinese characters]*. Beijing: Publishing House of Electronics Industry.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. London: Longman.

Boorman, H. (1970). *Biographical dictionary of republican China* (Vol. 3). New York: Columbia University Press.

Browning, D., & Bunge, M. (Eds.). (2009). *Children and childhood in world religions: Primary sources and texts*. New Brunswick: Rutgers University Press.

Chen, H. (1922). Yutiwen yingyong zihui [Characters used in vernacular Chinese]. *Xin Jiaoyu [New Education], 5*(5), 987–995.

Chen, H. (1928). *Yutiwen yingyong zihui [Characters used in vernacular Chinese]*. Shanghai: The Commercial Press.

Chen, X. (1996). Hanyu zhongjie yu yuliaoku xitong jieshao [Introducing the Chinese interlanguage corpus system]. In the *proceedings of the 5th International Chinese Language Teaching conference* (pp. 450–458). Beijing.

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, C. (2004). *ChineseTA (1.0). Stanford university and the silicon valley language technologies*. San Jose, CA: LLC.

Chu, C., & Chen, X. (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.

Conrad, S., & Biber, D. (2009). *Real grammar: A corpus-based approach to English*. New York: Pearson.

Cui, X., & Zhang, B. (2011). Quanqiu hanyu xuexizhe yuliaoku jianshe fangan [A proposal for the building of the International Learner Corpus of Chinese]. *Yuyan Wenzi Yingyong [Applied Linguistics], 19*(2), 100–108.

Feng, Z. (2006). Evolution and present situation of corpus research in China. *International Journal of Corpus Linguistics, 11*(2), 173–207.

Francis, N. (1992). Language corpora B.C. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 17–32). Berlin: Mouton de Gruyter.

Freeman, J. (1843). On grammalogues: To the editor of the Phonotypic Journal. *The Phonotypic Journal, 2*(24), 170–171.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.

Hu, X., & Xu, X. (2010). Mianxiang zhongwen dianhua jiaoxue de hanguo liuxuesheng hanyu zhongjieyu yuliaoku de kaifa yu jianshe [The development of a computer-assisted Chinese language teaching oriented Korean students' interlanguage Chinese corpus]. In the *Proceedings of the Seventh International Conference on Computer-assisted Chinese Language Teaching*.

Huang, C, Chen, K., & Chang, L. (1996). Segmentation standard for Chinese natural language processing. In *Proceedings of the 1996 International Conference on Computational Linguistics*. Copenhagan: Denmark.

Huang, W. (1962–1963). An annotated, partial list of the publications of William Hung. *Harvard Journal of Asiatic Studies, 24*, 7–16.

Hung, W. (1931). *Shuoyuan yinde [Index to Shuo Yuan]*. Peiping: Harvard-Yenching Institute Sinological Index Series, Peking University Library.

Hung, W. (1932). *Yinde shuo [On indexing]*. Peiping: Harvard-Yenching Institute Sinological Index Series, Peking University Library.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Johansson, S. (2011). A multilingual outlook of corpora studies. In V. Viana, S. Zyngier, & G. Barnbrook (Eds.), *Perspectives on corpus linguistics* (pp. 115–129). Amsterdam: John Benjamins.

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *ELR Journal, 4,* 27–45.

Kaeding, F. (1897). *Häufigkeitswörterbuch der Deutschen Sprache*. Berlin: Self-publication.

Kilgarriff, A., Keng, N., & Smith, S. (2015). Learning Chinese with the Sketch Engine. In B. Zou, M. Hoey, & S. Smith (Eds.), *Corpus linguistics in Chinese contexts* (pp. 63–73). Basingstoke: Palgrave Macmillan.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove: LTP.

Liu, D. (1926). Pingjiao zonghui 'gaibian qianzi ke' jianzi gongzuo de jingguo [The process of vocabulary selection for the *Revised Edition of 1000 Foundation Characters* by the National Association of Mass Education Movement]. *Jiaoyu Zazhi [Journal of Education], 18*(12), 1–14.

Liu, E. (1973). *Frequency dictionary of Chinese words*. The Hague: Mouton.

Liu, Y., Liang, N., Wang, D., Zhang, S., Yang, T., Jie, C., et al. (1990). *Xiandai hanyu changyong ci cipin cidian [A dictionary of frequency of modern Chinese words]*. Beijing: Astronautic Publishing House.

McCarthy, M., & O'Dell, F. (2008). *Academic vocabulary in use*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2004). *Touchstone (Student's Book 1)*. Cambridge: Cambridge University Press.

McCarthy, M., & McCarten, J. (2012). *Viewpoint (Level 1 Student's Book)*. Cambridge: Cambridge University Press.

McEnery, T., & Xiao, R. (2016). Corpus-based study of Chinese. In S. Chan (Ed.), *The Routledge encyclopedia of the Chinese language* (pp. 438–451). London: Routldge.

National Association of Mass Education Movement. (1922a). *Nongmin qianzi ke [1000 Chinese foundation characters for peasants]*. Shanghai: The Commercial Press.

National Association of Mass Education Movement. (1922b). *Shibing qianzi ke [1000 Chinese foundation characters for servicemen]*. Shanghai: The Commercial Press.

National Association of Mass Education Movement. (1928). *Shimin qianzi ke [Textbook of one thousand characters for townspeople]*. Shanghai: The Commercial Press.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–152). Amsterdam: John Benjamins.

Pitman, I. (1843). List of words from which grammalogues may be selected. *The Phonotypic Journal, 2*(23), 161–163.

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 (Xu yi) [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 (Xu er) [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 (Xu san) [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 (Xu si) [Basic sentence patterns of modern Chinese (Continued IV)]. *Shijie Hanyu Jiaoxue [Chinese Teaching in the World], 5*(1), 23–29.

Sinclair, J. (1987). *Collins COBUILD English dictionary*. London: Collins.

Su, X. (2010). Jiaocai yuyan tongji yanjiu de duoweidu gongneng [The multi-dimensional function of the statistical research on textbook language]. In *Proceedings of the Innovation of International Chinese Teaching Theories and Models Conference*. Xiamen.

Sugiura, M. (2002). Collocational knowledge of L2 learners of English: A case study of Japanese learners. In T. Saito, J. Nakamura, & S. Yamazaki (Eds.), *English corpus linguistics in Japan* (pp. 303–323). Amsterdam: Rodopi.

Tao, X. (1934). Xing zhi xing [Action knowledge action]. *Shenghuo Jiaoyu [Life Education], 1*(11), 286–287.

Tao, Z., & Zhu, J. (1923). *Pingmin qianzi ke [Early Chinese lessons for illiterates]*. Shanghai: The Commercial Press.

Teng, S. Hong, Y. Chang, W. & Lu, C. (2007). Huayuwen xuexizhe hanzi pianwu shuju ziliaoku jianli ji pianwu leixing fenxi [The construction of Chinese learners' character writing error databse and the analysis of error types]. In *Proceedings of 2007 National Linguistics Conference* (pp. 313–325). Tainan: National Cheng Kung University.

Thorndike, E. (1921). *The teacher's word book*. New York: Teachers College, Columbia University.

Thorndike, E. (1931). *A teacher's word book of the twenty thousand words found most frequently and widely in general reading for children and young people*. New York: Teachers College, Columbia University.

Thorndike, E., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Tribble, C., & Jones, G. (1990). *Concordances in the classroom: A resource book for teachers*. London: Longman.

Tsai, T. (1922). *Laojielao [A synthetic study of Lao Tzu's Tao-Te-Ching in Chinese]*. Beijing: Self-publication.

Tsang, W., & Yeung, Y. (2012). The development of the Mandarin Interlanguage Corpus (MIC): A preliminary report on a small-scale learner database. *JALT Journal, 34*(2), 187–208.

Tsou, B., Lin, H., Chan, T., Hu, J., Chew, C., & Tse, J. (1997). A synchronous Chinese language corpus from different speech communities: Construction and application. *International Journal of Computational Linguistics and Chinese Language Processing, 2*(1), 91–104.

Voegelin, C. (1959). The notion of arbitrariness in structural statement and restatement I: Eliciting. *International Journal of American Linguistics, 25*(4), 207–220.

Wang, M., Malmasi, S. & Huang, M. (2015). The Jinan Chinese Learner Corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 118–123). Denver, CO: The Association for Computational Linguistics.

Wei, S., Zhao, P., Yang, X., & Chen, L. (2008). Daxing zhongguo xiaoxuesheng zuowen yuliaoku de shengcheng [The construction of a large-scale Chinese pupils' written language corpus]. *Modern Educational Technology, 18*(12), 45–48.

Willis, D. (1990a). *The lexical syllabus: A new approach to language teaching*. London: Collins ELT.

Willis, J. (1990b). *Collins COBUILD English course (First lessons, Student's edition)*. London: Collins ELT.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. London: Routledge.

Xiao, X., & Zhou, W. (2014). Hanyu zhongjieyu yuliaoku biaozhu de quanmianxing ji leibie wenti [The exhaustiveness and taxonomy of Chinese interlanguage corpus annotation]. *Shijie Hanyu Jiaoxue [Chinese Teaching in the World], 28*(3), 368–377.

Xu, J. (2015). Corpus-based Chinese studies: A historical review from the 1920s to the present. *Chinese Language and Discourse, 6*(2), 218–244.

Yen, J. (1922). Pingmin jiaoyu xin yundong [A new movement of mass education]. *Xin Jiaoyu [New Education], 5*(5), 1007–1026.

Yen, J., & Fu, D. (1922). *Pingmin qianzi ke 'Foundation characters' (Books 1–3)*. Dingxian: National Association of Mass Education Movement.

Yen, J., & Fu, D. (1924). *Foundation characters* (2nd revised edition). Shanghai: National Committee of Y.M.C.A. of China.

Zhang, B. (2003). HSK (Hanyu Shuiping Kaoshi) dongtai zuowen yuliaoku jianjie [Introducing Chinese proficiency test dynamic essay corpus]. *Ceshi Yanjiu [Assessment Research], 1*(4), 37–38.

Zhang, P. (1999). 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.

Zhang, R. (2017). Hanyu zhongjieyu yuliaoku zhong de hanzi pianwu chuli yanjiu [The character errors in Chinese interlanguage corpora]. *Yuliaoku Yuyanxue [Corpus Linguistics], 3*(2), 50–59.

Zhao, S., Liu, S., & Hu, X. (1995). Beijing Yanyan Xueyuan xiandai hanyu jingdu jiaocai zhu kewen juxing tongji baogao [The BLCU report of the sentence patterns of the main texts of Modern Chinese Intensive Reading]. *Yuyan Jiaoxue yu Yanjiu [Language Teaching and Linguistic Studies], 17*(2), 11–26.

Zhou, X., Bo, W., Wang, L., & Li, Y. (2017). Guoji hanyu jiaocai yuliaoku de jianshe yu yingyong [The construction and application of international Chinese textbook corpus]. *Yuyan Wenzi Yingyong [Applied Linguistics], 25*(1), 125–135.