Huang, Chu-Ren, Lan Li and Xin-Chun Su. 2016. Lexicography in the contemporary period. In Sin-Wai Chan (Ed.), *The Routledge Encyclopedia of the Chinese Language*. 545-562. London: Routledge. [Pre-publication draft]

Lexicography in the Contemporary Period

Huang Chu-Ren and *Li Lan*

Hong Kong Polytechnic University, Hong Kong, China

Su Xinchun

Jiageng College of Xiamen University, China

1 Introduction: Historical Background of Chinese Lexicography

The history of Chinese lexicography can be traced back to nearly nineteen hundred years ago to Xu Shen's 许慎 *Shuowen Jiezi* 《说文解字》. Xu (131 CE) established a semantics-driven orthography-based framework for lexicography. He analyzed Chinese characters and found that component parts encoding semantic concepts, called bu4shou3 (部首, *radical*), can be used to identify and classify related characters. In *Shuowen Jiezi,* each Chinese character (an orthographic unit and an equivalent of a conventionalized sociological word in Chinese) is given an entry according to the radical it contains (and hence its conceptual classification). The entry contains a rough definition of its meaning, often in relation to the basic meaning of the radical; the character composition according to its components (部件 *bujian*); and very often also gives hint on its pronunciation. Although *Erya*《尔雅》is often claimed to be an even earlier collection of Chinese 'words' in different categories, it is important to note that *Erya* is a taxonomic collection of terms without linguistic information. Most crucially, the 540-radical system of *Shuowen Jiezi* has been adopted by all major Chinese dictionaries for nearly two millennia with adaptation and simplification. Indeed, we may conclude that Chinese lexicography started with and has been dominated by *Shuowen Jiezi*.

Although many Chinese dictionaries in the modern era still retain a reorganized and reduced set of radicals as either its primary structure or secondary index the field of lexicography did undergo drastic changes. The vernacular language movement in early twentieth century created an environment for Chinese lexicographers to focus more on the commonly spoken language and lexical words which may contain one or more characters. Hence the first change is the emergence of word-based dictionary 辞典 *cidian* or 词书 *cishu* versus the traditional character-based dictionary 字典 *zidian* (literally *character dictionary*). In addition, different systems are invented to give phonological representation of the pronunciation so that dictionaries can be organized according to how words are spoken daily instead of written according to the long literary tradition. The earlier phonetic alphabet system of 注音符号 *zhuyin fuhao* is still used in Taiwan while Pinyin Romanization adopted by mainland China has become the international standard. Based on different phonological transcription conventions, the second change is the emergence of lexicographic system organized according to alphabetic order. The alphabetical order of English is adopted for Pinyin, while ordering according to articulatory location is conventionalized for the phonetic alphabet system (hence the popular name of *bopomofo*). These two emergent changes in the early part of the twentieth century brought Chinese lexicography to a shared convention with modern lexicography. It also laid the foundation for more recent developments driven by the computerization of the Chinese writing system and the easy accessibility of digital content. In what follows, we will focus on these recent developments.

2 Fundamental Issues: Between character 字 *zi* and word 詞 *ci*, and from character encoding to word segmentation

The identification of a lexical unit is the fundamental issue of lexicography. The commonly held (but also often challenged) assumption that the linguistic word should be the most basic lexical unit (e.g. Hartmann 2003; Bloomfield 1926) does not translate into an executable procedure in Chinese lexicography due to its lack of conventionally marked word boundaries (e.g. Huang and Xue 2012) and confusion caused by the competing concepts of character and word (字 *zi* and 詞 *ci* respectively in Chinese.) By adopting the neutral term 'lexical unit', the ISO 24613:2008 standard for electronic lexicon incorporated a word-like concept in its formal definition and was successfully implemented for a wide range of languages in the world (Francopoulo 2013) including Asian languages (Francopoulo and Huang 2014). Although this result suggests that it is possible to have a common conceptual lexical unit for different languages the character vs. word competition has been, and still is, one of the most critical issues driving research and development in Chinese lexicography in the contemporary period.

Character encoding: representation and variations

The dichotomy of Chinese dictionaries dictates the definition of lexical entries: characters are lexical entries in a dictionary of characters and words are lexical entries in a dictionary of words. Although orthographic convention has clearly defined character boundaries orthographic variations also pose a challenge to the definition of which forms belong to the same character entry. The encoding of Chinese characters, in fact, was one of the first research issues in computational processing of the Chinese language, which brought the field of Chinese writing system (文字学 *wenzi xue*) to the forefront of recent computational studies (Hsieh 1996).

Note that a lexical unit typically represents what language users perceive as a single minimal form-meaning pair which allows some variations in forms. In Chinese orthography the variations go beyond graphic variations of the same glyph in different (historical, regional, or typographic) conventions. For instance, the concept of ‘peak’, sharing the same phonological form of *feng1* in modern Mandarin can be represented by either 峰 or 峯, two variants with their components composed differently (left-right vs. top-down). They should be free variants in almost all contexts and be treated under one single entry with rare exception for proper names. However, this is not possible given the traditional character-form based approach. This inconsistency in dealing with glyphic variants can be further exemplified by the four homographs 刃刄刃刃 *ren4* ’blade’. The authoritative *Kangxi Zidian* listed 刄 separately from the others and Unicode followed suit by giving it a different code. Close inspection will see that these variants differ only in the position and shape of the dot, which serves to refer to the ‘blade’ by marking its location on a knife 刀 *dao*1. In this case, neither the component parts nor the meaning can be differentiated among these variants. A more complicated example involves three glyphs 冲衝沖 *chong1* ’to charge (ahead)‘ and/or ’to crash (with water)’. In simplified Chinese, the two water-dot 冲 stands for both concepts and will be one lexical unit. For traditional Chinese, the water-based 沖, as well as the non-water related 衝 ‘charge, onslaught’ are different entries. However, for Japanese kanji, the three water dots 沖 forms a single entry, while the same character can also serve as glyph variants to the two dot 冲 for both traditional and simplified Chinese. The complexity of identifying characters is compounded by the need to identify and represent them in a computer. The computational solution by the Intelligent Chinese Character Encoding System (Jhuang et al. 2005) can provide a way to better define characters as lexical entries. This system can decompose each character based on philological principles, orthographic conventions and a string of finite number of component parts. Such ordered sequence can serve as identifiers for characters. Take the 峰 and 峯 variants, for example, they are actually represented by the same unique identifier of 山夂丰 (as 夆 can be further specified as the result of top-down concatenation of two components 夂 and 丰, which cannot be broken down to further components). In addition, variants of the same characters in different historical or regional conventions can be identified by the same sequence. The sequence itself can be taken as instruction on how to realize these variants by combining the component parts using graphs according to the convention. For example, programmes have been developed to render characters in different modern fonts as well as historical conventions such as oracle bones and small seals. In turn, the same encoding sequence can be used to search for different historical orthographic conventions or regional variants. There are two principle ways to generate variants: by instantiating each component in different homographic forms according to the temporal or regional 'font' variations, or by implementing a different combinatory procedure (e.g. left to right or top to down) while still following the top-left first, bottom-right last general constraint. In terms of (computational) lexicography, the encoding system enables similar characters to be searched and compared in ways beyond the traditional *zidian* (character dictionaries) classifications of radicals (部首). For instance, it is now possible to link 勞 *lao* 'labor' to 男 *nan* 'man' as both contains 力 li 'effort'.

Identification of words and words as lexical units

Words as lexical entries have long been the cornerstone of modern lexicography. However, it should also be noted that identification of words is often dependent on orthographic conventions, and hence identification of words in a language which lacks conventions to mark word boundaries, like Chinese, can be challenging (Huang and Xue 2012). However, take either Bloomfield's (1926) definition of 'minimal free form' or the lexicographic definition of 'smallest meaningful unit' (e.g. Jackson 2012; Francopoulo 2013), the main challenge in Chinese remains the lack of a set of operational criteria to define words. For instance, whether compounds or other multi-word units (such as idiom, chunks or proper names of persons and organizations) should be listed as an entry very often depends not on whether they are a word or not, but on the purpose and design criteria of a dictionary. With word dictionaries replacing character dictionaries as the default and more popular form of Chinese dictionaries, a clear operational definition of words as lexical units remains as a critical research topic in Chinese lexicography.

As words are basic units of a Chinese dictionary, two issues have received attention in recent lexicographic studies: the syllabicity of the Chinese language and the emergence of romanised words. First, although the earlier fallacy that Chinese is a monosyllabic language has been debunked the debate on whether a typical Chinese word should be mono- or di-syllabic has continued (e.g. Su 2001). It is important to note that the percentage of mono-syllabic words is limited by the number of characters, while there is no such constraint on the number of multi-syllabic words. The corpus-based study of Huang et al. (2002) shed light on this complex issue. They showed that mono- and di-syllabic words account for more than 90% of all instances of words in Chinese; and while there are more disyllabic words (in terms of word types) mono-syllabic words tend to have higher frequency. Based on the 5-million-word POS-tagged and balanced Sinica Corpus they found that mono-syllabic and disyllabic words each contribute to over 45% token frequency in Chinese. In terms of word types, however, disyllabic words compose of over 46% or all word types (and mono-syllabic words less than 3%, since there are only 6,000 or so commonly used mono-syllabic words). In sum, the distributional strength of these two types of words differs in terms of word types (di-syllabic words) and word frequency (mono-syllabic words) hence either can be considered as the dominant prototype of Chinese.

Second, it is crucial for modern lexicographers to recognize that not all Chinese words are rendered as characters. In fact, by different counts, there are at least 100 words in Chinese that are, typically, or only, written with alphabetic characters or a combination of alphabetic and Chinese characters. Examples are CCTV (China Central TV station 中央电视台), 阿Q (a fatalistic protagonist of Lu Xun's novel meant to be a prototype Chinese person from the past, now referring to all people with that characteristic), and AA制 ('to go Dutch'). Most modern Chinese dictionaries now include alphabetic words although they (except for those starting with Chinese characters) are typically put in a separate section and not listed together with the character-represented words. The lexicographic treatment of alphabetic words in Chinese remains an open research issue.

Selection of Lexical Entries

Once the issue of what is a lexical unit is determined the selection of lexical entries poses yet another challenge. The selection of lexical entries for character dictionaries (*zidian*), is different from that in word dictionaries (*cidan*). However radically different they could be in content or format, they do not differ fundamentally in terms of entry selection. This shows that the inventory and use of Chinese characters, is highly conventionalized, just like English dictionaries. A survey of entries in word dictionaries in Chinese (*cidian*), however by Huang (1998), showed a very different story. They compared five different Chinese dictionaries or lexica compiled between 1993 and 1997 in China and Taiwan (but converted to common character representation), ranging from less than 40,000 entries to over 156,000 entries. They showed that the mutual coverage between any two dictionaries, defined by the mean of the entry coverage of dictionary A over dictionary B and vice versa, range from 49% to 68%. In particular, what is surprising is that bigger dictionaries do not have better coverage over smaller dictionaries. The 156,710 entry Revised Mandarin Chinese Dictionary (RMCD, from Taiwan) covers only 68.58% of the 70,325 entry ABC Chinese Dictionary (ABC, from the USA). And the 56,162 entry Xiandai Hanyu Cidian (XH) covers only 72.48% of the 39,459 entry Standard Segmentation Lexicon (GB), even though they are both from China and the compilation of GB consulted XH. Overall, they found that only 21,655 entries were shared by all five dictionaries and that corpus frequency alone was not a good predictor of these entries. Such variations, especially the failure of much larger dictionaries to include entries from much smaller dictionaries, underline the issue that Chinese lexicography has yet to develop a set of commonly accepted and inter-operable criteria for lexical entry selection.

3 Chinese Dictionaries and Corpora

By the sixth century AD two major frameworks of Chinese lexicographic arrangement have been established. The first is the radical-stroke system following *Shuowen Jiezi*, and the second is the rhyming system following 陆法言 Lu Fayan's rhyme book 切韵 *Qieyun* (Lu 601 CE). The traditional 反切 *fan3qie1* system segments the monosyllabic pronunciation of a character to two parts: initial (consonant) and rhyme. For instance, the pronunciation of 党 *dang3* is represented as 多朗切, which means that it takes the initial ('*d*' in modern Mandarin) from 多 ('*duo*' in modern Mandarin) to combine with the rhyme '*ang*' (in modern Mandarin) of 朗 *lang*, forming *dang* in modern Mandarin. This system is an innovative way to represent phonological awareness without inventing a new set of symbols. It is also surprisingly robust since the initial-rhyme mapping can largely be preserved in spite of sound changes (as sound changes tend to apply to all members of the same class). What *Qieyun* contributed to the *fanqie* is the explicit classification and naming of the rhyme classes. The *Shouwen Jiezi* based radical-stroke system is more popular and has persisted down to the present because its use requires only the basic knowledge of the Chinese writing system. The *Qieyun* system, however, requires expert knowledge of rhyme groups (as philologist or as poet) hence is limited to use by scholars.

The commonly accepted phoneme based representational framework in lexicography was not widely applied to the Chinese language before the vernacular language (白话 *baihua*) movement in the late 1920's. The clear need to represent the language as spoken arose and several different strategies were introduced to ‘write down’ the way people speak. Two of the systems are still commonly used today in both language teaching and in lexicography, as mentioned earlier in the introductory section. They are the phonetic alphabet system of 注音符号 *zhuyin fuhao,* which relies on a set of invented symbols to represent Mandarin Chinese phonemes, and the Pinyin Romanization which relies on the Roman alphabet. As mentioned before, they each lead to a different lexicographic system for indexing: *bopomofo* relies on location of articulation while Pinyin relies on English alphabetic order. It should be noted that there were several popular systems of romanisation for Chinese before Pinyin especially among missionaries and second language learners. The most prominent among them are the Wade-Giles system and the Yale system. Earlier bilingual dictionaries adopting these two Romanization systems also typically follow English alphabetical orders for organization of their content.

Contemporary Chinese dictionaries are normally organized in four ways:

(1) By Pinyin Romanization. Most modern Chinese dictionaries use the Pinyin system for all the characters which makes it possible for dictionaries to be arranged in alphabetic order. Pinyin arrangement is essential for beginning first and second language learners do not recognize Chinese characters yet. Since the users need to know the pronunciation of a word before using the dictionary, pinyin only organization can be challenging when a user is trying to look up the pronunciation for a new character/word. Hence it is not uncommon for such dictionaries to come with a stroke or radial based index.

(2) By radical. Instead of the 540 radicals from *Shuowen Jiezi*, most modern Chinese dictionaries adopt the 214 radicals system based on the *Kangxi Zidian* (《康熙字典》) published in 1716. A radical of a Chinese character is the part that indicates the meaning as related to the basic concept represented by that radical. The radical component has a conventional location in a character accordingly to its radical and its instantiated variant form, most often at left, right, top, or bottom part etc. of the character. In the case of a simple character, the word itself is a radical. The order of radicals is arranged in the Radical Index, usually at the front of a dictionary, according to the number of strokes constituting them, i.e. 木 (4 strokes) *mu*, which means ‘wood’ or ‘tree’. All the characters having the same radical are then ordered according to the number of strokes. For a word dictionary, the convention is to organize all words with the same initial character as one major entry than order the words with same initial character according to

a secondary organization (usually by stroke number, or pronunciation, with words with identical second character clusters together, and so on and so forth.) It is important to note that even though most characters in this group have the radical at the left, such as 杆 (3 strokes in addition to radical) *gan* 'bar'; and 柱 (5 strokes) *zhu* 'pillar'; there are also characters in this group with the referential mark on the character, such 本 *ben* 'root'; at bottom, such as 柔 *rou* 'flexible'; and on top, such as 查 *cha* 'to check'. Similarly, since radial based dictionaries are difficult to use without knowing how to write them or the number of strokes, a modern radical based dictionary typically has a stroke or pronunciation based secondary index.

(3) By stroke number. The stroke method refers to the total number of strokes that make up a single character. A monolingual Chinese dictionary usually has a list of 'difficult words' in the front matter, arranged in ascending number of strokes. The characters in this list usually have many strokes and it is difficult to determine their radicals. In order to count the strokes correctly it is essential to learn the correct stroke order of Chinese characters. Stroke counting is rarely used as the main organization method of Chinese dictionaries any more.

(4) By Four-Corner method (四角號碼, *sijiao haoma*). The Four-Corner method was invented by Wang Yun-Wu 王雲五 in 1920's and the first dictionary by Four-Corner method was published in 1928 (Wang 1928). This method is based on the fact that Chinese characters each can be considered as a glyph in a square. The assumption is that 10 features each from each corner of the square will be enough to represent all frequently used characters. The corners are numbered 0 to 9 in relation to their shapes. For instance, 木 has the number of 4090, 杆 4194, 柱 4091, and 及 1724. The advantage of the system is its economy and independence from either knowing how to write or pronounce the character. The disadvantage is, of course, that this new system needs to be learned and memorized independent of learning the language itself. It was popularly used for coding Chinese telegraphs and in early computation but rarely used nowadays.

Chinese character dictionaries (zidian)

The modern paradigm of Chinese character dictionary was set by *Kangxi Zidian* (《康熙字典》*, The Imperial Character Dictionary of Kangxi*), which was commissioned by the Qing Emperor Kangxi (1662-1722) and launched in 1716. It contains 47,035 characters arranged in a system with 214 radicals and number of stokes. Up to now the dictionary is still one of the most authoritative and widely used dictionaries of the Chinese language and has been used as the source of nearly all Chinese character dictionaries.

The four-volume *Zhonghua Da Zidian*《中华大字典》(*Comprehensive Chinese Dictionary of Characters*), can be regarded as a decedent of *Kangxi Zidian*. It was edited by Xu, Y., Lu, F. and Auyang, P. and published by Zhonghua Publishing House in Shanghai in 1915. The dictionary added more than 1,000 new characters and corrected over 4,000 mistakes in the original dictionary *Kangxi Zidian.* The 48,000 entries were organised under 214 Kangxi radicals. For characters with the same pronunciation, or homophones, different meanings were listed in the same entry. Although the dictionary has been updated separately a couple of times in China and in Taiwan, its acceptance has not surpassed *Kangxi Zidian.* The reason is probably due to its difficulties of use: index by the number of strokes and the lack of radical index in the body text (Teng and Biggerstaff 1971: 131).

The mostly widely used Chinese character dictionary is *Xinhua Zidian* 《新华字典》, (*New China Character Dictionary*), which was published by the People's Education Press in Beijing in 1953 with Wei Jiangong as the chief editor. The pocket-sized dictionary was compiled for education purposes: to promote Pinyin, Putonghua and simplified Chinese characters. The book is the best-selling Chinese dictionary, an authoritative reference for the Chinese language and a compulsory tool for children to learn Chinese in primary education. Its 11th edition was published in 2011 by the Commercial Press in Beijing. Organised in Pinyin with radical index, the dictionary is regarded as very convenient to use. However, its pocket size has limited the inclusion of characters with only 3500 characters. Its online version is much bigger, with all the contents of a Chinese dictionary may have: meaning, Pinyin, grammar, sense demarcation, usage, and different search methods with Pinyin, radicals or number of strokes.

Chinese word dictionaries (cidian)

Ciyuan (《辞源》), *Sources of Words*, edited by Lu Erkui, was published in 1915. It was a groundbreaking effort in Chinese lexicography, regarded as the first word dictionary in Chinese with emphasis on literary, historical and classical terms (Hartmann 2003:16). The book has been updated several times later and has had many editions and prints. The latest edition of *Ciyuan* contains 12,980 head entries, under which are 84,134 definitions of words and phrases, totaling 11.3 million characters. The four volumes are arranged by radicals with a Pinyin index at the end of the dictionary. *Ciyuan* has been positioned as a reference work for researchers and students of pre-modern Chinese.

Cihai (《辞海》), *Sea of Words,* is another major dictionary of words in Chinese. Its main feature is encyclopedic coverage of history, philosophy, law, medicine and science. It was first compiled by Shu Xincheng in 1938 and published by Zhonghua Book Company. The dictionary has been regularly revised afterwards by Shanghai Lexicographical Publishing House. The American sinologist George A. Kennedy (1953: 131) regards *Cihai* as the basis for sinological studies and the principal value of the dictionary lies in its explanations for compound expressions and its citations illustrating the use of words and expressions

The most authoritative dictionary of modern Chinese is *Xiandai Hanyu Cidian* (《现代汉语词典》), *The Contemporary Chinese Dictionary.* The project started in 1958 in the Institute of Linguistics of Chinese Academy of Social Sciences, led by Lu Shuxiang and Ding Shusheng, two renowned Chinese linguists, and finally published in 1978 by the Commercial Press. The dictionary is a milestone in Chinese lexicography because it is the first Chinese word dictionary arranged in Pinyin, and providing phonetic standard of the language. *The Contemporary Chinese Dictionary* is characterized by its clearly articulated criteria for selection of entries and for entry format as well as succinct definitions and illustrative examples. It is now into its 6th (2012) edition and its bilingual (Chinese-English) version was published in 2002. The sixth edition contains about 69,000 entries including characters, words and expressions and idioms. While pocket-sized *Xinhua Zidian* targets native learners at the primary to secondary levels, *Xiandai Hanyu Cidian* is a popular tool among students at tertiary levels and the general public. By 2002, 40 million copies had been sold. The compilers have attempted to add some new words in each edition, but admit they have been ‘open and cautious in the choice of new words and senses’ (Lu 2002:8). Only 1,200 new words and new senses of existing words were added from 1996 to 2000.

The above Chinese word dictionaries were mostly compiled with linguistic intuition, and expert judgement of the lexicographers involved. In recent revisions, language corpora were used, but probably just for citation or examples. There seems to be limited effort to take a corpus-driven approach in dictionary making in China.

Corpora and dictionaries

The field of dictionary making has long been influenced by empirical and corpus-based methods. However, the early text collections ‘did not mean to be representative of the language; rather, dictionary makers stressed the normative function of their work, aiming to describe the ’proper‘ use of words’ (McArthur 1996: 235). Corpus today refers to a much larger collection of authentic data which is machine readable and can be processed by a computer with different queries. Language corpora have been used to construct dictionaries since the release of the Collins-Birmingham University International Database COBUILD (Sinclair 1987). The consensus among lexicographers and computational linguists is that statistical word modeling and corpus support are indispensible to modern dictionary compilation.

Corpus linguistics benefits lexicography in three aspects: providing authentic texts, building lexical database and helping dictionary compilation. A number of Chinese mega-corpora have been compiled in the last three decades; some were sponsored by government, others were developed at institutional level. Compared to English corpora, constructions of Chinese corpora started late when better computer technology became available and corpus linguistic theories had been well developed. Unlike English corpora, few Chinese corpora have so far been constructed for the explicit goal of lexicography. The corpus by the Center for Chinese Linguistics (CCL) of Peking University is a corpus with more than 500 million characters. The data was collected with balanced genres of spoken language, fictions, popular magazines, newspapers and academic journals. Like many Chinese corpora, the main corpus was not segmented or tagged. The small portion (1 million words) which was tagged and annotated with different grammatical and semantic markers and used as the basis of the book *The Grammatical Knowledge-base of Contemporary Chinese — A complete specification,* has become a reference dictionary for Chinese language processing in many institutions worldwide.

The Sinica Corpus was constructed in Academia Sinica in the 1990s under the direction of Keh-jiann Chen and Chu-Ren Huang in Taiwan (Chen et al. 1996). It is the first fully POS-tagged balanced Chinese corpus as well as the first Chinese corpus to be available on the world wide web. Like many modern balanced corpora its content distribution largely follows the original design of Brown Corpus but is also influenced by the designs of COBUILD and BNC. It is unique among modern Chinese corpora to have the full corpus manually checked word by word for both its segmentation and POS-tagging after its initial automatic annotations. The Sinica Corpus is publicly available and freely searchable on the internet (http://app.sinica.edu.tw/kiwi/mkiwi/). Its latest version, the Sinica 5.0, has more than10 million words.

Another widely used corpus of Chinese is the one million word Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery and Xiao 2004). Although smaller and later than the above-mentioned two corpora, the LCMC adopts the Brown/LOB (Lancaster-Oslo-Bergen) Balance Corpus format with 500 texts of roughly 2,000 words from 15 different genres. This conventional set-up allows users of the LCMC to readily compare it to English using LOB or Brown corpus. However, its size and format constraints also means that is often inadequate for modern computational lexicographic studies, which typically requires at least 10 million words (i.e. BNC size) of natural and non-trancated texts.

Table 1 is a description of some important Chinese corpora:

Title	**Compiler**	**Time**	**Size**	**Website**
Modern Chinese Corpus	The State Language Commission of China	1992-2002	100 million characters; 50 million characters segmented and tagged	http://www.cncorpus.org
Balanced Corpus of Modern Chinese	Academia Sinica, Taiwan	1996-2006	14 million characters fully segmented and tagged (= 10 million words)	http://asbc.iis.sinica.edu.tw/ OR http://app.sinica.edu.tw/kiwi/mkiwi/
CCL	Chinese Linguistics Research Center, Peking University		58 million characters	http://ccl.pku.edu.cn:8080/ccl_corpus/

Modern Chinese corpus	Xiamen University Language research Centre, China	2001-2005	2500 million characters	http://ncl.xmu.edu.cn
Chinese Internet Corpus	Leeds Universality, UK	2005	280 million (automatically segmented)	http://corpus.leeds.ac.uk/query-zh.html
The Lancaster Corpus of Mandarin Chinese	Lancaster University, UK	1991-1993	1 million words fully tagged (Brown/LOB format)	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
Tagged Chinese Gigaword Corpus	Lexical Data Consortium, University of Pennsylvania, and Academia Sinica	2002-2004	Ove1,200 million characters, fully segmented and tagged (= 831 million words)	https://catalog.ldc.upenn.edu/LDC2009T14

Table 1. A List of Chinese Corpora

A lexical database generated from a corpus is the starting point of a corpus-based dictionary. It is normally built up by lexicon matching and statistic modeling. Generating an English wordlist is straightforward: words are separated by spaces so there is one-to-one correspondence between orthographic and morpho-sysntactic word tokens. Chinese running texts are written without space, which means that words are not identified in the raw data. The first task in data processing is segmentation: to identify wordbreaks or segmented units which can then be used as processing units for other data (Huang and Xue 2012). Since both segmentation and POS-tagging in Chinese is non-trivial many widely available Chinese corpora are not tagged. In addition, high quality manually checked corpora tend to be smaller (usually a few million words, with

10 million words Sinica Corpus being the largest). Larger tagged corpora, such as the 831 million words tagged Gigaword Corpus (Huang 2009), are automatically tagged with only a small sample checked. The lack of sizeable Chinese corpora with high quality tagging may have contributed to the fact that a limited number of corpora were used in Chinese lexicography. However, the few examples of corpus-driven dictionaries in Chinese do provide very promising results for future developments.

Guoyu Ribao Liang Cidian (《国语日报量词典》*The Mandarin Daily News Dictionary of Classifiers*, Huang, Chen and Lai 1997) published in Taiwan is probably the first fully corpus-driven Chinese dictionary. The Academia Sinica team selected classifiers as the target for the first attempt to compile a corpus-driven dictionary not only because the classifier is a unique feature of Chinese but also because the uses of classifiers depend crucially on their collocation with nouns (Chang et al. 1996). With the fully tagged Sinica Corpus the selection of lexical entries of classifiers can be automated by selecting the POS and setting a frequency threshold. This also means that all attested usages of classifiers and classifier-noun collocations can be extracted and studied for generalization. The research team identified 537 types of measure words from the Sinica corpus and set up a lexical database of the relevant grammatical information for each classifier, which was then exported through a dictionary interface for the dictionary entries. To fully utilize and explicate the corpus-based information, the dictionary contains two parts: a classifier dictionary and a noun-classifier collocation dictionary. The noun-classifier collocation dictionary is organized by the head of noun because the head of the noun determines the semantic class of the noun and hence predicts the selection of classifiers. For instance, regardless of the length and nature of the modifier X all compound nouns X 牛 *niu* 'cow, bull' will take the classifier 头 *tou*. Chinese linguists and lexicographers are well aware of this characteristic of Chinese and dictionaries organized by the last (instead of first) character of words were occasionally compiled and referred as reverse-order dictionary (逆序辞典 *nixu cidian*). However, such dictionaries are tedious to compile manually. With the tagged Sinica Corpus the compilation of reverse-order noun-classifier collocation involves the same automatic extraction rules, generating comprehensive data much more than the manually compiled ones.

The next example of corpus-driven dictionary is also based on a part of speech (POS) tagged corpus but it is meant to be used both by a computer for natural language processing as well as human readers. The Institute of Computational Linguistics of Peking University compiled *Grammatical Knowledge-base of Contemporary Chinese*

(《现代汉语语法信息词典》*Xiandai Hanyu Yufa Xinxi Cidian*, Yu 2001), listing over 70,000 words in 18 different categories with their grammatical and statistical information. Based on both linguistic and statistical analysis each entry word is marked with POS as well as its syntactic/semantic context and frequency. Working closely with collaborators in the Chinese Department of Peking University the team set up a very detailed segmentation and POS-tagging system. Electronic version of this dictionary has been used as a resource for many applications in Chinese language technology.

Routledge's *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners* (Xiao, Rayson and McEnery 2009) is a recent example of corpus-driven Chinese dictionary published overseas for non-native speakers. The dictionary draws on the Lancaster Corpus of Mandarin Chinese (LCMC), a balanced 73-million-character Chinese corpus composed of spoken, fiction, non-fiction and news texts in current use. The data was processed with the ICTCLAS, a Chinese Lexical Analysis System developed by the Institute of Computing Technology of Chinese Academy of Science, an automatic tool widely used in Chinese language processing in China. Since it is automatically processed the dictionary cannot go beyond the original 80,000 words in the system even with mechanism to guess word meaning based on role tagging (Zhang et al. 2002). From this list Xiao et al. found similar distribution of mono- and disyllabic words as Huang et al. (2002) discovered: disyllabic word consist of most word types. The usually high token frequency of monosyllabic words at 54% is probably due to the fact that automatic segmentation and tagging typically fails to recognize many out-of-vocabulary words and leave parts of these words as monosyllabic words. Xiao, Rayson and McEnery (2009) only extracted 84,883 word types from a 73 million-word corpus. In contrast, Huang et al. (2002) extracted nearly 200,000 word types from the manually checked 5 million word Sinica Corpus, while Huang (2009) extracted nearly 3 million word types from the 831 million word Tagged Chinese Gigaword Corpus v.2.0.

With learners of Chinese in mind the dictionary by Xiao, Rayson and McEnery provides the user with a detailed frequency-based list as well as alphabetical and part-of-speech indexes. All entries in the frequency list feature the English equivalent and a sample sentence in Chinese character, Pinyin and English translation. The dictionary also contains thirty thematically organized lists of frequently used words on a variety of topics such as food, weather, travel and time expressions. The authors cherish the wish 'to enables students of all levels to maximize their study of Mandarin vocabulary in an efficient and engaging way.'

Kilgarriff summarises a number of aspects of dictionary creation supported by the corpus:

- Headword list development;
- For writing individual entries;
- Discovering the word senses and other lexical units (fixed phrases, compounds);
- Identifying the salient features of each of these lexical units;
- Their syntactic behavior;
- The collocations they participate in;
- Any preference they have in particular text-types or domains;
- Providing examples;
- Providing translations.

(Kilgarriff 2013:78)

Data-driven Chinese dictionaries are generally for computer to align with words or mark word boundaries in a dataset; they are not for human use. They normally have part of the features in the above list. To meet users' needs, more lexicographic information, such as collocations, usage notes and examples should also be provided.

4 Chinese Dictionaries for Foreign Learners

There are three types of learner's dictionaries: monolingual, bilingual and bilingualised, the last one combining information and translations from the previous two. The primary concern is the needs of target users. Some dictionaries are for native speakers to learn their mother tongue; others are for foreign learners to learn the language. The content and organisation can be very different because the learning needs, language problems and learning focus of users are different. Chinese dictionaries for foreign learners are bilingual dictionaries mostly with Chinese and English. With the increased economic and political influence of China the interest of learning Chinese is dramatically increased worldwide. The need for Chinese dictionaries for foreign learners is obvious. Although Chinese publishers have made effort to produce dictionaries for foreign learners research has revealed that few foreign learners are using dictionaries published in China (Xie 2010; Li 2013).

Features of learner's dictionary

The English learner's dictionary, which started in the 1930s in Japan by native English educators, has become a major branch of lexicography and a big business in the publishing world. After the Big Five learner's dictionaries (by Oxford, Longman, Cambridge, Collins and McMillan) were published in the UK, Merriam-Webster in the US finally published its first advanced learner's dictionary in 2008. Dictionaries of this type have striking features: they are user-cautious, supported by research findings in linguistics, cognitive science and behavioural science. The word lists are carefully selected and the definitions of the words are explained by a core vocabulary, such as the 3,000 defining words in the Oxford Advanced Learner's Dictionary (OALD). Most importantly, they are based not only on mega-corpora of native English language but also EFL writings so that the learning problems of users can also be addressed. The compilers make particular effort to help learners with traditional dictionary elements: grammar, pronunciation, definition and example, as well as with innovative ideas in the order of senses, illustrations, learning panels and appendixes. The success of English learner's dictionary in the world is due to the fact that the compilation focuses entirely on users' needs, although the importance of English is an obvious reason.

The majority of Chinese-English dictionaries published so far target native Chinese users who may need dictionaries to translate texts from Chinese to English; grammatical and phonological information of Chinese is less relevant to this group of users. The dictionary for foreign learners is, by contrast, for users to render their ideas from English to Chinese. Much like a pedagogical grammar which receives sustenance and support from linguistic advances in language description the pedagogical dictionary is an educational aid, a major learning resource, whose form and function must be determined by its users. The best judgment on learners' dictionaries is probably what Dr Johnson noticed as long as 250 years ago: "In lexicography, as in other arts, naked science is too delicate for the purposes of life. The value of a work must be estimated by its use: It is not enough that a dictionary delights the critic, unless at the same time it instructs the learner" (Johnson 1747:6).

Chinese dictionaries for foreign learners

The history of Chinese bilingual lexicography reflects the development of China's relations with foreign countries (e.g. Chien and Creamer 1986). An interesting historical development can be observed based on the chronogially ordered list of early bilingual dicitonaries given in Table 2, compiled based on the survey of Chien and Creamer (1986: 41-43) and our own research. Since early 16th century, the interaction between China and the world can be attributed to European interest in culture, and religion via missionary contact. However, after the Opium War (1839-1842) which forced China to open trade opportunities, the role of England becomes more prominent and the emphasis of dictionaries helping foreigners to communicate with local people is underlined. This is attested by multiple dictionaries devoted to local vernacular languages in Southern China, where most direct contacts happen. Another important characteristic is the awareness of directionality of bilingual dictionary, as both E-C and C-E dictionaries were compiled and published equally frequently.

Time	Compiler	Dictionary
1583	Matteo Ricci	Dizionario Portoghese Chinese
1667	Michael Boym	Chinese-Latin Dictionary
1670	Michael Boym	Chinese-French Dictionary
1813	M. de Guignes	Dictionnaire Chinois Français et Latin
1815	Robert Morrison	Dictionary of the Chinese Language (Latin)
1828	Robert Morrison	Vocabulary of the Canton Dialect
1832	Walter Henry Medhurst	Dictionary of the Hok-këen Dialect of the Chinese Language
1848	Walter Henry Medhurst	English Chinese Dictionary
1853	Elihu Doty	Anglo Chinese Manual of the Amoy Dialect
1856	Samuel Wells Williams	A Tonic Dictionary of the Chinese Language in the Canton Dialect
1873	Castairs Douglas	Chinese-English Dictionary of the Vernacular Or Spoken Language of Amoy
1883	John Macgowan	English and Chinese Dictionary of the Amoy Dialect
1892	Herbert A. Giles	A Chinese-English Dictionary
1896	Samuel Wells Williams	A Syllabic Dictionary of the Chinese Language

Table 2 The first Chinese- foreign language bilingual dictionaries

The dictionaries in Table 2 aimed to help learners/users of Chinese to communicate with local Chinese people, for example, *A Syllabic Dictionary of the Chinese Language* was arranged according to the *WuFangYuanYin*, and records pronunciation of the characters as heard in Peking, Canton, Amoy and Shanghai. *WuFangYuanYin* 五方元音(Gong 1660?) was a Chinse word dictionary popular at that time after first published in the late 17th century. It is one of the firsr Chinese dictionaries arranged phonologically, by multiplying 20 initial consonants and 12 ryhme groups. It is interesting to notice that even today dictionaries used by foreign learners of Chinese are mostly compiled on foreign soil, probably because the compilers, some of whom are even non-native Chinese speakers, can understand the needs of foreign learners better. Many Chinese-English (C-E) dictionaries have been published in China in recent years, but they are mostly for native Chinese users to write in English or to translate Chinese texts into English, although some claim to be for foreign learners. The entry is normally structured as below:

Chinese headword

Pinyin

English translations of the word

Chinese example

English translation of the example

e.g. 曝光 [puguang] (喻) (将隐蔽的或不光彩的事公布于众) expose; reveal; make sth public;对于贪污腐败给予公开~ give public exposure to graft and corrupt (A New Contemporary Chinese-English Dictionary)

Entries like this are not very helpful to foreign learners. First, there is no grammatical information of the word; second, there is no Pinyin for the example. Usually beginner learners of Chinese concentrate on the spoken language rather than learning to read and write Chinese characters. If a reader cannot read characters the Chinese example cited is completely useless.

The organization of a C-E dictionary for foreign learners can also be a problem. ‘Chinese characters are without a doubt cumbersome to index for foreign learners’ (Li 2013:35). To use a C-E dictionary organized in Pinyin the user has to know how to read a Chinese word. To use a dictionary organized in radicals the user has to know how to write a Chinese character. Both are nearly impossible tasks for learners at elementary level. The complexity of Chinese makes the organization of C-E dictionaries very challenging.

The *Concise English-Chinese and Chinese-English Dictionary* by Martin Manser with translation by Wu et al. (1999) is a good attempt to help beginners of Chinese. The first edition was published in 1986 by Oxford University Press in Hong Kong and the Commercial Press in Beijing. The book was nicknamed ‘little red book’ by many foreign learners and received very good feedback. As the title indicates the dictionary has two parts. The A to Z English-Chinese section helps users to find both Pinyin and Chinese characters of an English word. The Chinese-English section is organised in Pinyin enabling users to understand Chinese words. Each entry has a POS marker, usage example of the Chinese word, its English translation and Pinyin of the whole example. The book has had four editions, with the latest in 2011. It has been among the best-selling Chinese-English books in the world.

ABC Chinese-English Dictionary was edited by John DeFrancis and published by Hawaii University Press in 1999. The ABC Dictionary has been welcomed by users in that the entries are organized in Pinyin rather than in characters. It innovative organization clears up the misconception that the Chinese language is made of monosyllabic characters. Most meaning units in Chinese are words with multiple syllables and hence with multiple characters. Because of this Chinese characters are without a doubt cumbersome to index. There will always be a few characters where it is difficult to figure out the exact number of strokes or the exact radical to find the character. With the ABC Dictionary, as long as one knows the pronunciation of the word then it is very easy to find its meaning. This dictionary is particularly useful for finding words that one hears spoken but is not sure of the meaning.

The ABC dictionary was compiled using lexical data from both China and Taiwan with simplified characters and their traditional forms. Where there are differences in usage between the two places, they are noted with PRC or TW in the entry. Another striking feature of the dictionary is that where homophones occur they are ranked in order of frequency. The dictionary does not have a radical index; therefore it is difficult to look up for a word if its pronunciation is unknown.

Chinese-English MRDs

As Li (2013) observed most foreign learners of Chinese find Chinese words hard to read, hard to recognize, hard to write and hard to remember (78). The lack of popularity of Chinese learner’s dictionaries published in China is probably because of the difficulties in finding a word and lack of information on how to use the word. To cater for different users' needs more space will be required in paper dictionary. Machine readable dictionaries (MRD) or electronic dictionaries are the only way to solve the problem.

A machine readable dictionary is a dictionary stored in an electronic form on a computer that can be linked to a database and can be queried in different formats via application software. MRDs can be loaded onto different electronic medium such as on the internet, on a PC, on a CD ROM, on a hand-held gadget, and on a mobile phone. Some major Chinese dictionaries have their online versions now, such as *Kangxi Zidian* (http://kangxizidian.com/), *Cihai* (http://tool.gaofen.com/cihai/), *Xinhua Zidian* (http://xh.5156edu.com/). Users can search a word by Pinyin or by radicals with number of strokes. There are also hyperlinks to related words, different word classes, synonyms, antonyms, grammar, history and literature.

With literally no limit to space, MRDs can have many multimedia features embedded in the dictionary, for example, the sound files of entries and examples, static graphics- such as photos and colourful images, dynamic graphics-such as animations and video clips, hyperlinks to other resources, which have been proved to be able to positively stimulate learning from different cognitive channels (Mayer 2005:46). Mayer's cognitive theory of multimedia learning (CTML) is based on three principles: the human information processing system includes dual channels for visual and pictorial and auditory /verbal processing; each channel has limited capacity for processing; and active learning entails carrying out a coordinated set of cognitive processes during learning (ibid.:31). The most attractive advantage of MRD is its speed and accuracy in locating information. The wildcard function can save input time and correct user's typing errors. Xie (2010) surveyed his students who were learning Chinese as a foreign language and found that few students used paper dictionaries; majority used online dictionaries because they are fast and free. Also the high speed of dictionary consultation increased the amount of student reading because looking up a word in a paper Chinese dictionary was very time consuming. He believed that teachers cared more about the quality of e-dictionaries: whether the definitions are correct, the orders of senses are clear and examples are easy to follow, while the students paid more attention to convenience and speed. Other dictionary user studies are consistant with this finding: learners 'really want their dictionaries to be cheap, complete, portable, comprehensible and easy to use' (Nesi 1999:55). However there has been concern that with more reliance on e-dictionaries and online translation software the language acquisition process might not be as effective as in the traditional learning mode (Xie 2010:61).

As mobile devices have become part of our life, e-dictionaries have rapidly transferred to a new media, in connection with the ‘apps on the boom’ by Android, IOs and Windows. There are now dictionary apps on iPhone, iPod Touch, iPad and various Samsung products. Similar to the rapid growth of language learning apps, mobile dictionaries have changed the design and use of reference works even further. A free app Youdao Cidian (有道词典) by EasyNet, can translate words automatically in two directions between Chinese and English, Chinese and French, Chinese and Korean, or Chinese and Japanese. When a Chinese word is keyed in, its English equivalent will pop up. Further link can show Pinyin, grammar and examples. If a user wishes to know more about the word, there is a link to lead him/her to standard Chinese dictionaries *Xiandai Hanyu Da Cidian* (*A Comprehensive Modern Chinese Dictionary*) and *Xin Hanying Da Cidian* (*The New Comprehensive Chinese-English Dictionary*). It can further draw bilingual examples from the iCloud and provide audio pronunciation of the sentences both in English and in Chinese. In addition, it links the search word with specialised dictionaries. Looking up an English word can obtain abundant information at different levels of request. The app also has a built-in look-up function. The user can get a pop-up screen with the meaning and/or translation of a word highlighted in an example.

Modern technology can realize many lexicographic functions in MRDs which traditional lexicographers could never dream of. “Whatever the dictionary of the future will be like, there is still ample room for improvement, and the metalexicographer is in no danger of being unemployed: there is still much that has to be done in order to adapt the dictionary to its users and different uses” (Béjoint 1994:2).

5 Summary

Chinese dictionaries are the records of the language and learning tools for both native Chinese users and foreign learners of Chinese. The needs of users differ so the organization, components and examples of dictionaries should also vary. With the advancement of modern technology and corpus linguistics automatic processing of Chinese has started to benefit dictionary compilation from the most basic issues of word identification and entry selection to the more advanced issues of word meaning definition and induction, identifications of grammatical and pragmatic patterns, selecting examples and modeling the language.

Although the corpus approach has contributed greatly to the advancement of

English lexicography (e.g. Sinclair 1987; Kilgarriff 2013) the same cannot be said of Chinese dictionary making yet even though there were a few successful cases (e.g. Huang et al. 1997; Yu, et al. 1998). The fact shows that there are significant gaps between computational/corpus linguists and lexicographers. In particular, there has not been significant dialogue between them and the only examples of computational Chinese lexicographical works so far were done by computational and corpus linguists. In addition, in also seems that dictionary publishers in China have been more conservative and have not been involved in any significant research projects, such as COBUILD or subsequent projects by other publishers of English dictionaries. With an increasing demand from Chinese learners all over the world for high quality state-of-the-art references and dictionaries, the challenge is on now for Chinese lexicographers, corpus linguists, and dictionary publishers to work together to identify the user needs and resolve many outstanding issues in Chinese lexicography.

Bibliography

Dictionaries

Cihai Bianji Weiyuanhui (eds) (2009) *Cihai* (《辭海》*Sea of Words*) (6th edition), Shanghai: Shanghai Lexicographic Publishing House.

Ciyuan Xiuding Zu (1983) *Ciyuan* (《辭源》*Sources of Words*) (2nd edition), Beijing: The Commercial Press.

Fan, Tengfeng (1660?). 〈五方元音〉 *WuFangYuanYin*. Accesed at https://archive.org/details/02077328.cn

Hanyu Da Zidian Dictionary of Chinese Characters (2nd edition), Sichuan: Sichuan Lexicographic Publishing House.

Huang, Chu-Ren Chen, Keh-jiann, and Lai, Ching-hsiung. (eds) (1997) *Mandarin Daily Dictionary of Chinese Classifiers* (《國語日報量詞典》), Taipei: Mandarin Daily Press.

Kangxi Zidian (《康熙字典》*Kangxi Dictionary*) (1716) Reprinted 1993, Beijing: China International Culture Press.

Li, X. J. et al. (eds) (2004) *Xiandai Hanyu Guifan Cidian* (《现代汉语规范词典》*Modern Standard Chinese Dictionary*) Beijing: Foreign Language Teaching and Research Press.

Lu, Shuxiang (2002) *Xiandai Hanyu Cidian* (Han-Yin shuangyu ben) (《现代汉语词典》(汉英双语本）*The Contemporary Chinese Dictionary* (Chinese-English Edition)), Beijing: Foreign Language Teaching and Research Press.

Lu, Shuxiang (2012) *Xiandai Hanyu Cidian* (《现代汉语词典》*The Contemporary Chinese Dictionary*) (6th edition), Beijing: The Commercial Press.

Manser, Martin (1999) *Concise English-Chinese and Chinese-English Dictionary* (translated by Wu Jingrong, Mei, Ping, Zhu, Yuan and Liangbi Wang), Beijing: The Commercial Press and Oxford University Press.

The Chinese Academy of Social Science (eds) (2011) *Xinhua Zidian* (《新华字典》) (11th edition), Beijing: The Commercial Press.

Wang Yuwu (1928) *The Wang Yun-wu Da Cidian* (《王雲五大辭典》). Shanghai: Commercial Press.

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

Xu, Yuanhao and Fucun Ouyang (1978) *Zhonghua Da Zidian* (《中华大字典》*The Great Chinese Dictionary*), Beijing: Chinese Publishing House.

Yao, Naichiang (2003) *Xinhua Dictionary with English Translation.* Hong Kong: Commercial Press.

Yu, Shiwen (1998) *Xiandai Hanyu Yufa Xinxi Cidian* (《现代汉语语法信息词典》*Grammatical Knowledge-base of Contemporary Chinese*), Beijing: Tsinghua University Press.

Zhang, Jing (2002) *Dangdai Xinbian Hanying Cidian* (《当代新编汉英词典》*A New Contemporary Chinese-English Dictionary*), Shanghai: The World Book Publishing House.

References

Béjoint, Henry (1994) *Tradition and Innovation in Modern English Dictionaries*, Oxford: Clarendon Press.

Bloomfield, Leonard (1926) 'A Set of Postulates for the Science of Language', *Language* 2: 153-164. (Reprinted in Hockett, 1970:128-138).

Chang, Li-Li, Chen Keh-jiann and Chu-Ren Huang (1996) 'The Use of Corpus in the Compilation of Dictionaries', in *Proceedings of the Ninth R.O.C. Computational Linguistics Conference*, 255-279. Also in ACL Anthology.

Chen, Keh-jiann, Huang Chu-Ren, Chang Li-Png, and Hsu Hui-Li (1996) ‘Sinica Corpus: Design Methodology for Balanced Corpora’ in B. S. Park and J.B. Kim (eds) *Proceeding of the 11th Asia Pacific Conference on Language, Information and Computation*, Seoul: Kyung Hee University, 167-176.

Chien, David and Thomas Creamer (1986) ‘A Brief History of Chinese Bilingual Lexicography’, in Reinhard Hartmann (ed) *The History of Lexicography*, Amsterdam and Philadelphia: John Benjamins Publishing Company, 36-45.

Francopoulo, Gil (2013) (ed) *LMF – Lexical Markup Framework*, ISTE / Wiley.

Francopoulo, Gil and Chu-Ren Huang (2014) ‘Lexical Markup Framework: An ISO Standard for Electronic Lexicons and its Implications for Asian Languages’, *Lexicography* 1(1): 37-51.

Hartmann, Reinhard (2003) *Lexicography: Reference Works across Time, Space, and Languages*, London: Taylor and Francis.

Hsieh, Ching-chun 谢清俊 (1996) 〈电子古籍中的缺字问题〉(The Issue of Missing Characters in Electronic Classics), in〈第一届中国文字学会学术讨论会〉(*Proceedings of the First Conference of the Association of the Chinese Characters*), 25-30 August, Tianjin.

Huang, Chu-Ren (1998) ‘Criteria for Computational Chinese Lexicography: A Study based on a Standard Reference Lexicon for Chinese NLP’, in *Proceedings of ROCLING XI*, 87-108, also in ACL Anthology.

Huang, Chu-Ren (2009) *Tagged Chinese Gigaword Version 2.0*. Philadelphia: Lexical Data Consortium, University of Pennsylvania, Available at: https://catalog.ldc.upenn.edu/LDC2009T14

Huang, Chu-Ren and Nianwen Xue (2012) ‘Words without Boundaries: Computational Approaches to Chinese Word Segmentation’, *Language and Linguistics Compass* 6(8): 494-505.

Huang, Chu-Ren, Chao-Ran Chen, and Claude C.C. Shen (2002) ‘The Nature of Categorical Ambiguity and Its Implications for Language Processing: A Corpus-based Study of Mandarin Chinese’, in Mineharu Nakayama (ed) *Sentence Processing in East Asian Languages*, Stanford: CSLI Publications, 53-83.

Jackson, Howard (ed) (2013) *The Bloomsbury Companion to Lexicography*, London: Bloomsbury.

Johnson, Samuel (1747) The Plan of an English Dictionary. In *The Works of Samuel Johnson, LL.D.: With An essay on his life and genius.* London: Luke Hensard & Sons (printed in 1820).

Jhuang, Der-Ming, Jenq-Haur Wang, Chen-Yu Lai, Ching-Chun Hsieh, Lee-Feng Chien, and Jan-Ming Ho (2005) 'Resolving the Unencoded Character Problem for Chinese Digital Libraries', Paper to the Joint Conference on Digital Libraries 2005, Denver, Colorado, 7-11 June 2005.

Kennedy, George (1953) *ZH Guide: An Introduction to Sinology*, New Haven: Yale University Press.

Kilgarriff, Adam (2013) 'Using Corpora as Data Sources for Dictionaries', in Jackson, H. (ed) *The Bloomsbury Companion to Lexicography*, London: Bloomsbury, 77-96.

Li, Yan (2013) 'On the Compilation of General Purpose Chinese Dictionaries for Foreign Learners of Chinese', *Cishu Yanjiu* (*Lexicographical Studies*) 5: 34-39.

Mayer, Richard (ed) (2005) *The Cambridge Handbook of Multimedia Learning*, Cambridge: Cambridge University Press.

McArthur, Tom (1996) *The Oxford Companion to the English Language*, Oxford: Oxford University Press.

McEnery, A. and Xiao Z. (2004) 'The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study', Paper presented at LREC 2004, May 2004, Lisbon.

Nesi, Hilary (1999) 'A User's Guide to Electronic Dictionaries for Language Learners', *International Journal of Lexicography* 12(1): 55-66.

Sinclair, John (1987) *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, London: Collins COBUILD.

Su, Xinchun (ed) (2001) *A Statistical Analysis of the Chinese Lexicon*, Xiamen: Xiamen University Press.

Teng, Suyu and Knight Biggerstaff (1971) *An Annotated Bibliography of Selected Chinese Reference Works* (3rd ed.), Cambridge, Mass: Harvard University Press.

Xiao, Z, A. McEnery, P. Baker, and A. Hardie (2004) 'Developing Asian Language Corpora: Standards and Practice' in V. Sornlertlamvanich, T. Tokunaga, and C. Huang (eds) *Proceedings of the Fourth Workshop on Asian Language Resources*, 25 March, Sanya, 1-8.

Xie, Tianwei (2010) 'Using Online Dictionaries in Teaching Chinese', *Journal of Chinese Language Teachers Association* 45(3): 53-65.

Xue, Shiqi (2003) ‘Chinese Lexicography Past and Present’, in *Lexicography: Critical Concepts II*, London and New York: Routledge, 158-173.

Yu, Pingfang and Jiali Du (2010) *A Comparative Study on Learner’s Dictionaries in English and in Chinese*, Beijing: China Social Sciences Press.

Zhang, Huaping, Liu Qun and Cheng Xueqi (2002) ‘Automatic Recognition of Chinese Unknown Words Based on Role Tagging’, in *Proceedings of the 1st SIGHAN Workshop*, *COLING 2002*, Taipei, 71-77.