Building a Collation Element Table for a Large Chinese Character Set in YES

Xiaoheng Zhang[1] and Xiaotong Li[2]

[1] Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University
ctxzhang@polyu.edu.hk
[2] College of International Exchange, Shenzhen University
sharklxt@aliyun.com

Abstract. YES is a simplified stroke-based method for sorting Chinese characters. It is free from stroke counting and grouping, and thus much faster and more accurate than the traditional method. This paper presents a collation element table built in YES for a large joint Chinese character set covering (a) all 20,902 characters of Unicode CJK Unified Ideographs, (b) all 11,408 characters in the Complete List of Chinese Characters Used by the Media in 2013, (c) all 13,000 plus characters in the latest versions of Xinhua Dictionary(v11) and Contemporary Chinese Dictionary（v6）. Of the 20,902 Chinese characters in Unicode, 97.23% have one-to-one relationship with their stroke order codes in YES, comparing with 90.69% of the traditional method. Enhanced with the secondary and tertiary sorting levels of stroke layout and Unicode value, there is a guarantee of one-to-one relationship between the characters and collation elements. The collation element table has been successfully applied to sorting CC-CEDICT, a Chinese-English dictionary of over 112,000 word entries.

Keywords: Chinese characters, collation, Unicode, YES

1 Introduction

Collation, or determining the sorting order of strings of characters, is a key function in computer systems and a basic need of the human society. Whenever a list of texts—such as the records of a database and the entries of a dictionary—is presented to users, they are likely to want it in a sorted order so that they can easily and reliably find their target words. Collation is so important to computational linguistics that a special standard, i.e. *Unicode Technical Standard #10: Unicode Collation Algorithm* [1], has been developed to support automatic sorting. At the center of the standard is a default collation element table, which usually needs to be customized for different natural languages before real life application.

There are tens of thousands of different characters in the Chinese writing system. Collation of Chinese has been a long time challenge for lexicography. Search for a word in a Chinese dictionary is notoriously time consuming and sometimes frustrating [14 p243, 5]. Even a small dictionary may be difficult to use, especially if one does not know the exact pronunciation of the target Chinese character and have to rely on the radical method, as evidenced by the following quotation from section How to Use the Dictionary in *Oxford Chinese Mini Dictionary* [18 px]. After a lengthy introduction of the radical collation method, the editors say:

> “... If you have trouble finding the character in the Character Index, first check above or below in the list under that radical, **in case your stroke count was incorrect**. It is best to write the character down as you count, being sure to write it using the proper strokes and stroke order. If you still cannot find it, the character is probably listed under a different radical, and **you will need to start again from the beginning of the process described above, looking under a different radical**. Beginners sometimes find this process **frustrating**, but if you keep trying, it will become easier.”

Obviously there is an urgent need for improvement of Chinese collation. The dominant sorting methods for Chinese are Radical (部首法), Pinyin (拼音法), Four-corners (四角号码法) and Stroke-based (笔画法) [14 p189-207, 13 p67-69). The stroke-based approach is employed by all the other methods: to sort characters of the same pronunciation (homophones) in the phonetic method, to sort the radical list and the characters belonging to a common radical in the radical method, and to sort characters sharing the same code in the four-corner method. That means the stroke-based method is the simplest and most fundamental among them. Improvement of Chinese collation should first consider stroke-based sorting.

YES (or 一二三 in Chinese) is a simplified stroke-based sorting method with better performance than the traditional method. In the following sections, we will introduce YES and its application to the design and implementation of a collation element table of a large Chinese character set.

2 Stroke-Based Collation

The early stroke-based arrangement of Chinese dictionaries merely relies on stroke numbers, and was used as an auxiliary method for the radical system. In the 20th century, it was developed into an independent method by adding stroke order as a second level of sorting [15 p107, 17 p357).

Strokes (笔画, bǐhuà) are the most basic unit of Chinese writing. A Chinese character is written stroke by stroke in a certain order. For example the standard stroke order of character 福 (good fortune, happiness) is “丶フ丨丶一丨㇕一丨㇕一丨一”. If we regard it as a sequence of letters, then it can be ordered alphabetically as an English word, provided we have a stroke “alphabet” defining the sequence of different strokes.

Most dictionaries in China classify strokes into 5 categories or groups, each represented by a primary stroke. There are two popular sequences: “丶,一 ,丨,丿,㇖” in Hong Kong and Taiwan, and “一 ,丨,丿,丶,㇕” which is the official standard of the Mainland, as shown in Table 1.

Table 1. The standard 5-categories stroke list

Primary Stroke	Secondary strokes	Name of group
一	㇀	Héng 横 (horizontal)
丨	亅	Shù 竖 (vertical)
丿		Piě 撇 (left falling)
丶	㇏	Diǎn 点 (dot)
乛	㇕, ㇗, ㇛, ...	Zhé 折 (bending)

Contemporary Chinese Dictionary（现代汉语词典）[4], *Oxford Chinese Dictionary* [2] and many other dictionaries follow the standards by the National Language Commission of China [8, 9]. The basic rules are:

(1) Sort the characters by their number of strokes in ascending order, i.e. all the characters of one stroke are put before two-stroke characters, followed by three-stroke characters, and so on.

(2) Characters of the same number of strokes are arranged by their first stroke categories in the order of Table 1. If they belong to the same category, then check the second strokes, and so on.

3 The YES Collation Method

Briefly speaking, the YES Chinese character collation method is formed by eliminating stroke counting and grouping from the traditional stroke-based method. Arranging Chinese in YES order is similar to arranging English in alphabetic order, if we consider the stroke sequence of a Chinese character as the letter sequence of an English word.

Two Chinese characters are sorted by their first stroke positions in the YES stroke alphabet (Table 2). If the first strokes are the same, then compare the second strokes, and so on. For example, the different characters in "一二三排检法|一二三排檢法" (the YES Sorting Method) are sorted as:

一 (一)
二 (一一)
三 (一一一)
檢 (一丨丿丶丿㇏一丨㇕一丨㇕一丿丶丿丶)
检 (一丨丿丶丿㇏一丶丶丿一)
排 (一亅㇀丨一一一丨一一一)
法 (丶丶㇀一丨一㇜丶)

In the rare cases of more than one glyph or stroke order for a Chinese character, YES follows the standards of the National Language Commission of China [7, 8, 11].

Words of multiple characters are sorted by their first characters in YES order. If the first characters are the same, then check the second characters, and so on. Non-

Chinese characters appear after Chinese characters in alphabetical/Unicode order. For example,

覺
覺醒
觉
觉醒
觉悟
B超
T恤

Table 2. The YES Stroke Alphabet

Stroke	Stroke Name	Example Characters
㇐	横	十/七
㇕	横折竖	口 达 贯/敢 為
㇅	横折竖折横	凹 卍
㇎	横折竖折横折竖	凸 [illegible]
㇡	横折竖折横折竖钩	乃/杨
㇋	横折竖折横折撇	及 延
㇊	横折竖折提	计 颓 鸠
㇍	横折竖弯横	朵 投
㇈（乙）	横折竖弯横钩	几 九/艺 亿
㇆	横折竖钩	同 却 母 仓 羽/也
㇇（㇖）	横折撇	又 之/令 了/买 宝
㇌	横折撇折撇钩	阳 部
㇠	横折捺钩	飞 风 执
㇀	提	堆 打/江
㇑	竖	中/五
㇗	竖折横	山/母 乐/发 降/车
㇞	竖折横折竖	鼎 卐 亞 吳
㇉	竖折横折竖钩	马 与 钙/号 弓
ㄣ	竖折横折撇	专/[illegible]/矢
㇙	竖折提	长 鼠 以 瓦 收 岭
㇄	竖弯横	四 西 朮
㇟	竖弯横钩	己 匕 电 心 乱
㇚	竖钩	小 水 了
㇒	撇	千/人/月
㇜	撇折提	公 离 红 乡 亥
㇛	撇折点	女 巡
㇢（㇁）	撇钩	犹 家/㐅
㇔（㇀）	点	主 丸/火 刃 然
㇏（〇）	捺	八边/〇
㇂	捺钩	代 我

The YES stroke alphabet is based on the standards of the Standard Bending Strokes of GB13000.1 Character Set [10] and the Unicode CJK Strokes [16]. There are totally 30 strokes, sorted by the standard basic strokes sequence of “横 (一) 提 (㇀) 竖 (丨) 撇 (丿) 点 (丶) 捺 (㇏)” and bending points sequence of “折 弯 钩” [21].

The Chinese name of the sorting method, i.e. “一二三”, are the first three characters in YES order. The English name “YES” is the abbreviation of the Chinese name’s Pinyin (**Y**ī **È**r **S**ān).

Table 3 presents the “code:characters” distribution of the 20,902 Chinese characters of GB13000.1 (the same character set as the Unicode CJK Unified Ideographs primary character set) in the traditional stroke-based method and in YES. YES performs much better than the traditional method. 97.23% of the characters (or 20,324 out of 20,902) have one-to-one relationship with their stroke order codes in YES, much higher than the 90.69% of the traditional method. The maximum number of characters sharing one code is 9 in the traditional method, and 4 in YES. In the traditional method the nine characters of “夕久夂夊么勺凡丸及” share the five-category stroke order of 354 [9 p8]. In YES, they are further classified into 6 groups: 及/凡丸/勺/夕/夊夂久/么.

Table 3 Code:characters distribution of GB13000.1characters in Traditional and YES sorting

	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8	1:9	Total
Codes (Traditional)	18956	756	85	25	9	3	1	0	1	19836
Characters (Traditional)	18956	1512	255	100	45	18	7	0	9	20902
Percentage (Traditional)	90.69%	7.23%	1.22%	0.48%	0.22%	0.09%	0.03%	0.00%	0.04%	100%
Codes (YES)	20324	258	14	5						20601
Characters (YES)	20324	516	42	20						20902
Percentage (YES)	97.23%	2.47%	0.20%	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%	100%

Details of the YES collation can be found in [20].

4 Further Improvement of YES

There still exists a handful of stroke orders shared by two or more Chinese characters. For example, 犬 (dog) and 太 (too (much)) have the same stroke order of “一 丿 ㇏

丶”. Among the 20,902 characters of the GB13000.1 character set, 578 or 2.77% do not have their unique stroke orders. The maximum number of characters sharing one stroke order is 4, totally 5 groups of them, as shown in Table 4.

Table 4 Five groups of 4 characters sharing a stroke order in GB13000.1

Characters	Stroke order
甲㆙叶申	丨㇕一一丨
另叧叨叻	丨㇕一㇆丿
叭叺央史	丨㇕一丿㇏
父从仌爻	丿丶丿㇏
八人乂入	丿㇏

Even if the four characters are randomly sorted in the group, we can still easily find a character in a one-tier search. However, an ideal collation method should be “total” (not partial), or be capable to arrange all characters into a reasonable order in which every character has its one and only one position.

When two characters have the same stroke order, their difference is in the appearance and layout of strokes in the 2-dimentional block area of a Chinese character, including each stroke’s position, size and orientation. For example, the difference between character 犬 and 太 is in the position of their last stroke 丶, the difference between 口 and 囗 are in their sizes, and the 丿 strokes in 千人 have difference in orientation. In other words, when the position, size and orientation of each stroke in the Chinese character are decided, the form of the whole character is decided. And these factors can be accurately defined by the positions of the starting point and ending point of a stroke. Though the absolute positions of the points can be detected by the computer easily, it is unlikely to be totally feasible for human’s eyes. Hence it seems preferable to focus on the character distinguishing differences easily recognizable by the reader. For example, the distinguishing difference between character 犬 and 太 is in the position of their last stroke 丶, higher in犬than in 太. According to the rule of “top-to-down”, 犬 is put before 太. If the two strokes are at the same height, than consider the difference in horizontal positions. For example, characters人入八乂share the stroke order of “丿㇏”. The starting point of “丿” in 入 is distinguishingly lower than the others, thus入 is put at the end. 人八乂are sorted according to the horizontal positions of the starting points of their first strokes (丿) in a left-to-right order, resulting in the reasonable order of 八人乂入.

On the other hand, the computer can be over accurate. For example, the computer can easily detect a difference in height of the 一 (horizontal) stroke in 犬太, but moving stoke 一 in 犬 up to the height of its counterpart in 太 does not change the character’s meaning. However, if we move the dot down to the position of its counterpart in 太, then 犬 becomes 太. Hence, the difference in the position of the dot strokes is character distinguishing between犬 and 太, while the difference in the positions of stroke 一 is not character distinguishing and not easily recognizable by human’s eyes.

Therefore, we add a second-level sorting rule to YES as follows:

> If two Chinese characters are of the same stroke order, then find their first (according to the stroke order) pair of character distinguishing strokes and put the character with a higher or lefter stroke before the other character. More accurately,
>
> If the starting points of the two corresponding strokes are in different height, then put the character with a higher stroke starting point before the other character,
>
> else if the ending points are in different height, then put the character with a higher stroke ending point before the other character,
>
> else if the starting points are in different horizontal positions, then put the character with a comparatively left stroke starting point before the other character,
>
> else put the character with a left stroke ending point before the other character.

For example, 土 and 士 have the same stroke order of “一丨一”. The first strokes (the upper 一) in both characters are of similar height, but horizontally the stroke starting point is comparatively more left in 士 than in 土, as shown in Figure. 1. Hence, 士is put before 土.

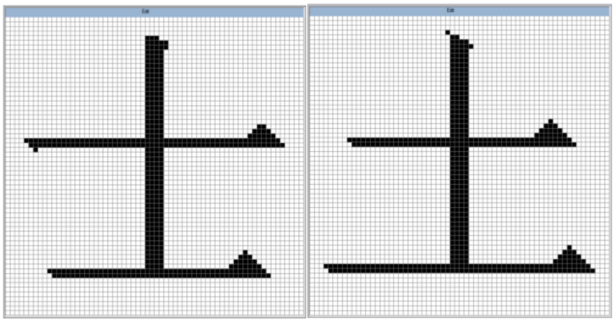

Fig. 1. Two characters in font SimSun

An advantage of considering the vertical position of the ending points before the horizontal positions of the starting points is two-fold: easier to operate and putting the character in up-down (⿱) structure before the character in left-right (⿰) structure. For example 炎炏（丶丿丿丶丶丿丿㇏）, the starting point of the first stroke 丶 is lower in 炏 than in 炎, hence 炏 is put after 炎. In characters 甼町（丨㇆一丨一一亅）, the starting points of the first stroke 丨are of similar height, if we compare the horizontal positions of the starting points first, it is more left in町than in 甼, and 町is put before 甼. The sorting result is

炎炏，町甼

In the first pair, the character in up-down (⿱) structure is before the character in left-right (⿰) structure. In the second pair, the character in up-down (⿱) structure is after the character in left-right (⿰) structure.

If we compare the vertical positions of the ending points right after the vertical positions of the starting points, the sorting is

炎炏，甼町.

In both cases, the character in up-down (⿱) structure is before the character in left-right (⿰) structure, which is a more consistent result.

Even when the top stroke is heng2 (一) in both characters, it is still reasonable to put the character in up-down (⿱) structure before the character in left-right (⿰) structure. Because the top point is the top of the triangle（**Fig. 1.**）at the end of stroke 一 (we use the standard font of SimSun 宋体), when the upper component of the character is stretched downward, the starting point of stroke héng 横（一）becomes proportionally lower. And there is space around the character, which will also make the left component lower than the upper component, for instance 泵砅 (一丿丨㇆一亅㇇丿㇏).

5 Design of the Collation Element Table

5.1 The Unicode default collation element table

The Unicode default collation element table is Latin letters oriented [1]. A typical collation element consists of 3 weight values in hexadecimal, in the format of "[.base character .accent .case/variant]" representing 3 levels of comparison for sorting. The characters are first sorted by the weight 1 value of base character, if the base characters are the same, then sort by the weight 2 value of accent, if the accent are the same, then sort by the weight 3 value of case/variant.

Most of the mappings in a collation element table are simple: they consist of the mapping of a single character to a single collation element, as shown in Table 4.

Table 4 A collation element table with simple mapping

Character	Collation Element	Name
0300 "`"	[.0000.0021.0002]	COMBINING GRAVE ACCENT
0061 "a"	[.06D9.0020.0002]	LATIN SMALL LETTER A
0062 "b"	[.06EE.0020.0002]	LATIN SMALL LETTER B
0063 "c"	[.0706.0020.0002]	LATIN SMALL LETTER C
0043 "C"	[.0706.0020.0008]	LATIN CAPITAL LETTER C

The default collation element values for Chinese characters are generated from the Unicode code point. The Unicode code point order is realized by dividing the CJK characters into several charts, each roughly sorted by the Kangxi Radical order. Hence, very cumbersome for human usage.

Collation is not code point (binary) order. The only way to get the linguistically-correct order is to use a language-sensitive collation, not a binary ordering [1] (section 1.8). Hence we need a collation element table tailored-made for the Chinese language.

5.2 Design of the Chinese collation element table

Our design of a Chinese collation element table is in similar format of the Unicode default table, as shown in Table 5.

Table 5 A Chinese collation element table in YES

Character	Collation Element	Stroke Order
臧	[.177E.1.81E7]	一丿㇗一丿一丨㇕一丨㇗㇂丿丶
鴄	[.177F.1.9D04]	一丿㇄㇗丿丨㇕一一一㇆丶丶丶丶
兀	[.1780.1.5140]	一丿乚
兀	[.1780.1.FA0C]	一丿乚
尢	[.1780.2.5C22]	一丿乚

Where a collation element consists of a primary weight of stroke order, followed by a second weight of stroke layout, followed by a third weight of Unicode.

The characters are first sorted by the primary weight of stroke orders according to the YES stroke alphabet, putting 臧 before 鴄 before 兀兀尢. Characters 兀兀尢 share the stoke order of 一丿乚, and are arranged by the secondary weight of stroke layout. The first stroke 一 in 兀兀 is higher than in 尢, hence兀兀 come before 尢.

There are some duplicate characters in Unicode [19]. To make sure that every Unicode character can be sorted properly on the computer, we have added Unicode code point as the third-level weight. For example, 兀 (5140) and 兀 (FA0C) are in the same form, and are sorted by their Unicode values, resulting in 兀 (5140) before 兀 (FA0C). Such a design guarantees a strict one-to-one relationship between characters and collation elements.

6 Experiment Results and Analysis

We have built a collation element table for a large joint Chinese character set covering:

- All of the 20,902 characters in Unicode CJK Unified Ideographs (same as the national standard GB13000.1 Character Set);
- All 11,408 characters in the Complete List of Chinese Characters Used by the Media in 2013 (2013 年度媒体用字总表) [12];

- All 13,000-plus characters in the latest version of the Xinhua Chinese Characters Dictionary [3];
- All 13,000-plus characters in the latest version of the Contemporary Chinese Dictionary [4].

There are totally 21,976 Chinese characters in the joint character set, among which

- 21,335 characters have distinctive stroke orders from others
- 612 characters have similar stroke orders as one or more other characters but in different shapes
- 29 characters are of the same shapes (or forms) as one or more other characters but have different Unicode code points.

Table 6 presents the first 21 characters with similar stroke orders in the collection elements table.

Table 6 The first 21 characters with similar stroke orders

Character	Collation Element	Stroke Order
王	[.60.1.738B]	一一丨一
龶	[.60.2.9FB6]	一一丨一
玊	[.A4.1.738A]	一一丨一丶
玉	[.A4.2.7389]	一一丨一丶
珈	[.11E.1.73C8]	一一丨㇀㇆丿丨㇕一
玿	[.11E.2.73BF]	一一丨㇀㇆丿丨㇕一
玾	[.13A.1.73BE]	一一丨㇀丨㇕一一丨
珅	[.13A.2.73C5]	一一丨㇀丨㇕一一丨
末	[.217.1.672B]	一一丨丿㇏
未	[.217.2.672A]	一一丨丿㇏
亐	[.21C.1.4E90]	一一㇉
亏	[.21C.2.4E8F]	一一㇉
于	[.21F.1.4E8E]	一一亅
亍	[.21F.2.4E8D]	一一亅
开	[.228.1.5F00]	一一丿丨
亓	[.228.2.4E93]	一一丿丨
井	[.228.3.4E95]	一一丿丨
无	[.23B.1.65E0]	一一丿乚
元	[.23B.2.5143]	一一丿乚
天	[.255.1.5929]	一一丿㇏
夫	[.255.2.592B]	一一丿㇏

There is a one-to-one relationship between the characters and collation elements in the collation table, i.e., each character has one and only one distinctive collation element.

7 Conclusion and Further Development

Chinese collation has been a long-time challenge to lexicography and natural language processing. This paper introduces the YES stroke-based collation method and its application to the design and implementation of a user-friendly collation element table to support automatic Chinese sorting. The new method is significantly simpler and more accurate than the traditional approaches. According to our experiment on 20,902 Chinese characters in Unicode, 97.23% of the characters have one-to-one relationship with their stroke order codes in YES, comparing with 90.69% of the traditional method. The maximum number of characters sharing one code is 9 in the traditional method, and 4 in YES.

The collation element table is based on the Unicode Standard and covers a large number of 21,976 Chinese characters. With the three sorting levels of stroke order, stroke layout and Unicode value, there is a guarantee of one-to-one relationship between the characters and collation elements.

The collation table has been successfully employed to sort all 112,178 word entries in CC-CEDICT, a large Chinese-English dictionary downloadable from the Web [6]. A trial version YES-CEDICT Chinese Dictionary is also on the Web for free download [22, 23]. And it is our intention to further develop the collation element table to cover all the 70,000 plus Chinese characters in Unicode.

However there are a number of tricky issues. For example, 靼靻, 坦坥 and 旦且. A close look will find the heights of first stroke 丨shu5 in the components of 旦 and 且 are not consistent: higher in 靼 than in 靻, lower in 坦 than 坥, and similar in 旦 and 且. Another example, the starting point of the left stroke in 八 is slightly lower than the right stroke. But the difference is usually ignored linguistically. Moving the left stroke up to the level of the right stroke does not change the character into another one. A more interesting case. The first stroke 丨in 吅吕 (NSimSun 12p) are of similar height. When we down size the characters to吅吕(NSimSun 10p), the first stroke is clearly lower in 吅. The situation in 旱旰 and 旱旰 is similar. These inconsistencies have brought inconvenience to language processing, however YES can consistently put the top-down structure character before the left-right structure character in all cases.

On the whole, the glyphs need to be more consistent. And there is a need for an optimal balance between the glyph, the computer and people. Fortunately, the above-mentioned tricky issues only involve the secondary sorting level of stroke layout, hence unlikely to bring serious inconvenience to the application of YES.

Acknowledgements The project has been partially supported by a University research fund (Account Code: 4-ZZEW). The authors are also very grateful to the three anonymous reviewers, whose valuable comments played an important role in the revision of the paper.

References

1. Davis, M. Whistler, K. and Scherer, M.: Unicode Technical Standard #10: Unicode Collation Algorithm, version 8.0 (2015), http://www.unicode.org/reports/tr10/
2. Kleeman, J. and Yu, H. (editors): The Oxford Chinese Dictionary. Oxford: Oxford University Press (2010)
3. Linguistic Institute of the Chinese Academy of Social Sciences. Xinhua Dictionary (Xinhua Zidian, 新华字典), 11th edition, Beijing: The Commercial Press (2011)
4. Linguistic Institute of the Chinese Academy of Social Sciences. Contemporary Chinese Dictionary (Xiandai Hanyu Cidian, 现代汉语词典), 6th edition, Beijing: The Commercial Press (2012)
5. Mair, V. H. "The Need for an Alphabetically Arranged General Usage Dictionary of Mandarin Chinese: A Review Article of Some Recent Dictionaries and Current Lexicographical Projects". Sino-Platonic Papers 1:1–31 (1986)
6. MDBG. CC-CEDICT Chinese to English Dictionary (2015), URL: http://www.mdbg.net/chindict/chindict.php?page=cedict . (Downloaded on January 31, 2015).
7. National Language Commission of China (国家语委). Standard Stroke Order of Commonly-Used Characters of Modern Chinese (现代汉语通用字笔顺规范). Beijing: Language & Culture Press (语文出版社) (1997)
8. National Language Commission of China (国家语委). The Standard Stroke Order of the GB13000.1 Character Set (GB13000.1 字符集汉字笔顺规范). Shanghai: Shanghai Education Press (上海教育出版社) (1999)
9. National Language Commission of China (国家语委). The Standard (Stroke-Based) Order of the GB13000.1 Character Set (GB13000.1字符集汉字字序（笔画序）规范). Shanghai: Shanghai Education Press (2000)
10. National Language Commission of China (国家语委). The Standard Bending Strokes of GB13000.1 Character Set (GB13000.1字符集汉字折笔规范). Beijing: Language & Culture Press (2001)
11. National Language Commission of China (国家语委). Standard List of Commonly-Used Chinese Characters (通用规范汉字表). Beijing: Language & Culture Press (2013)
12. National Language Commission of China (国家语委). Language Situation in China: 2014. (中国语言生活状况报告(2014)). Beijing: Commercial Press (2014)
13. Norman, J.: Chinese. Cambridge: Cambridge University Press (1988)
14. Su, P.: Essentials of Modern Chinese Characters (现代汉字学刚要), 3rd Edition. Beijing: Commercial Press (2014)
15. Sun, C.: Chinese: A Linguistic Introduction. Cambridge: Cambridge University Press. (Ch. 5 Chinese writing) (2006)
16. The Unicode Consortium. The Unicode Standard, Version 8.0. Mountain View, CA: The Unicode Consortium, (http://www.unicode.org/versions/Unicode8.0.0/) (2015)
17. Yong, H., Luo, Z. and Zhang, X.: Chinese Dictionaries: Three Millennia. Shanghai: Shanghai Foreign Language Education Press (2010)

18. Yuan, B. and Church, S. K.: Oxford Chinese Mini Dictionary, 2nd edition. New York: Oxford University Press (2008)
19. Zhang, X.: Duplicate Encoding of Chinese Characters (中文的同形异码字问题). Journal of Chinese Information Processing, No. 4, Vol. 29 (2015), pp. 233-240 (2015)
20. Zhang, X. and Li, X. Handbook of the YES Stroke-Based Sorting Method for Chinese Characters (一二三笔顺检字手册). Beijing: Language & Culture Press (2013)
21. Zhang, X. and Li, X. Integration and Optimization of Standard Chinese Stroke Lists (标准笔形表的整合与优化). In Li, X., Jia, Y. and Xu, J. eds, Digital Teaching of Chinese Language 2014 (数字化汉语教学 2014). Beijing: Tsinghua University Press, pp 200-208 (2014)
22. Zhang, X., Li, X. and Lun, C. (editors). The YES-CEDICT Chinese Dictionary (一二三汉英大词典, Trial Edition, Sorted by Simplified Chinese). The Journal of Modernization of Chinese Language Education (中文教学现代化学报), Vol.4, No.1, June, 2015. (http://xuebao.eblcu.com/) (2015)
23. Zhang, X., Li, X. and Lun, C. (editors). The YES-CEDICT Chinese Dictionary (一二三漢英大詞典, Trial Edition, Sorted by Traditional Chinese). The Journal of Modernization of Chinese Language Education (中文教学现代化学报), Vol.4, No.1, June, 2015. (http://xuebao.eblcu.com/) (2015)