The Totality of Chinese Characters – A Digital Perspective

Shouhui Zhao
CRPP, NIE of Nanyang Technological University, Singapore 647616
Shouhui.zhao@nie.edu.sg

Dongbo Zhang
Department of Modern Language, Carnegie Mellon University, PA 15231, USA
dongboz@andrew.cmu.edu

Abstract

That the Chinese IT industry has been recurrently plagued by the deficiencies of the Chinese writing system has been well known. In this paper, we provided an examination of the complex factors involved in the process of streamlining the total number (TN) of Chinese characters (hanzi), and critically reviewed the existing and emerging theoretical frameworks advocated by language policy (LP) practitioners over the years in addressing the issue of fixing the TN. We first identified three referencing points when people from different fields talk about the TN of hanzi. Based on this review of the concept of TN, we discussed how the ever-expanding, indefinable and unpredictable TN poses problems to Chinese information processing. From a sociolinguistic perspective, we then pointed out the main external causes that lead to the dilemma of TN restriction. Various proposals, articulated since the early 20th century, were then reviewed and their failures were briefly discussed. Lastly, the most recent visionary model to deal with hanzi's TN was introduced and how it informs the planning of TN of Chinese characters was also discussed. It is hoped, this study will shed some light on the in-depth understanding towards one difficult aspect of Chinese script modernization and computerization.

Keywords

Chinese characters, total number, variant forms (VF), rarely used characters (RC), encoding set

1. Articulating the Case – Counting Chinese Characters

The Chinese writing system is notoriously well known for its large pool of characters. It is generally held that *hanzi* is an open system, so the total number grows over time, making it almost impossible to tell precisely how big the total is. This is indubitably truer when all non-Chinese and non-Mandarin characters are added, including: a) non-Chinese *hanzi*, mainly referring to what Lunde (1999) called JKV (Japanese, Korean and Vietnamese) characters and *hanzi*-derived characters (over 20 systems throughout history), created by Chinese ethnic minorities within China proper; b) Chinese regional/dialectal characters. Although only Cantonese characters are visible in modern publications, character specificity within various dialects did exist historically and they are still being circulated locally, with some having the possibility of playing a more active role in written communication (for more information see Chen, 1996; Jordan, 2002); c) obsolete characters in ancient scripts such as *Jiaguwen* (Turtle Bones Inscription) and *Jinwen* (Metal

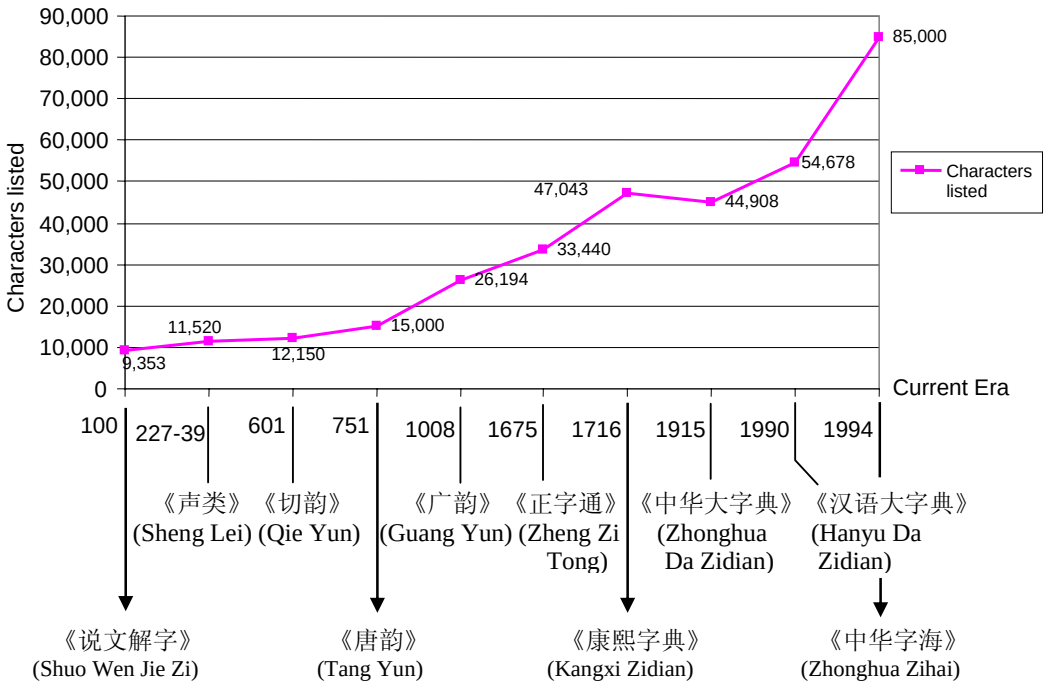
Inscription), which represent the embryonic forms of *hanzi* engraved on animal bones and metal utensils about three millenniums ago.

As described below, when people are talking about the totality of *hanzi* today, they have actually three streams of referencing points in their minds:

1.1. Milestone Dictionaries

Hanzi's traceable history began from pottery inscriptions, which were found to have existed 6,000-7,000 years ago, and have been archaeologically verified as the earliest signs for which the genetic link with *jiaguwen* can be established. Out of 4,672 character types found and identified on *jiaguwen*, only 1,723 can be deciphered without dispute among scholars. The good thing about Chinese characters is that numerous dictionaries were compiled throughout history. The number of characters (headwords) listed in a dictionary is an important parameter to measure the numerical increment of Chinese *hanzi* as they have evolved. Therefore, these dictionaries, either compiled by individual etymologists or mandated by imperial edicts in different dynasties, enable us to draw a clear picture about the *hanzi*'s developing trajectory in light of quantitative change. The following graph, which is marked by a token proportion of ten influential character dictionaries, roughly reflects the incremental tendency over a nearly 2000-year-history that started with *Shuowen Jiezi*.

The historical change in the number of hanzi



(Sources: the Research Team of Computer Information Processing, 1980: 70-71)

In this graph, four identifiable phases can be figured out to signpost the key milestones in the developing course of *hanzi's* number. The starting point was *Shuowen Jiezi*, or literally, 'script explanation and character analysis', which was the first Chinese dictionary and regarded as canon in ancient philology. It was compiled by Xu Shen (58-147 CE) during the East Han (25-220 CE) period, and 9,353 characters were etymologically analyzed in a systematic manner, based on their structural features, laying down the foundation for the study of Chinese script history. The figure shows that from CE 100 until up to CE 751, *hanzi's* number had witnessed a gradual and stable increase for six and half centuries. The *Tang Yun* was made during the Tang Dynasty. Although being formulated by an individual scholar (Sun Mian), it got blessed by the imperial court and was thus seen as setting the official standard for both pronunciation and semantic explanation for the characters circulating during a period when people enjoyed the most sophisticated life in Chinese history. The next historically attested dictionary, still somewhat used today, is *Kangxi Zidian* (1716), which was made under royal patronage during the Kangxi Reign in the Qing Dynasty (1664-1911), marshalling 47,035 characters. During this wide-stretched nearly period of one millennium, *hanzi's* number grew by over 30,000, producing an increase rate of 30 characters per year. Coming to modern times, *Zhonghua Zihai* (the *Ocean of Chinese Characters*), with its monstrous collection of 85,000 entries – of which a large number are variant forms and obsolete signs listed under the same entry and in different style fonts such as *Zhuanshu* (seal script) and *Lishu* (cleric script) – is by far the most inclusive dictionary in terms of amassing as many characters as possible. Notwithstanding that, a widespread estimate is that the ultimate number may be well beyond 100,000, if all Chinese characters and the derived forms that have ever existed were taken together, including variant forms, non-Chinese *hanzi* and dialect *hanzi*.

It is worth remembering that the quantities listed here do not necessarily give an accurate reflection of the period in question. On the one hand, it is impossible for any dictionary or author to collect all characters in use, as many must be missed, and in addition to that, the character inclusion is also influenced by the compiler's preferences, which often contributed to deviations from this general tendency, e.g. the decline from 47,043 in 1716 to 44,908 in 1905, rather than an increase; on the other hand, if we do not count the huge number of variant forms (see subsequent discussion), the actual number for each period would be a lot smaller.

1.2. Purpose-specific Character lists

Some well established pedagogic character lists and the high profile character tables, both official ones and those developed by individual scholars, show the TN required for different purposes. Two of the most frequently quoted tables are the *Table of Common Characters* (3,500 characters) and the *Table of General Characters* (7,000 characters), published by Chinese language administration authorities in 1988. Another widely quoted but has not been publicized till now is the *General List of Print Fonts of Chinese Characters* of 1965 (6,196 characters). It has been only 'in-house' circularized and has proved to basically meet the demand for the whole printing industry after forty-year in use.

Three education-oriented tables of Chinese characters are another strand of purpose-specific character lists. *The Table of Characters for Illiteracy Elimination* contains 2,000 characters, a number believed to cover 93.9936% of the characters in publication, targeting general readers; *The Table of Characters for Primary Students* has 3,071 characters; the third pedagogical character table is *Graded Outline of Chinese Vocabulary and Character Proficiency Criterion*. It was formulated by the National Office of Teaching Chinese as a Second Language and the Testing Centre of Chinese Language as a Second Language in

1991. It includes 2,905 characters arranged at four levels (A=800, B=804, C=601, D=700) for non-Chinese-speaking students.

1.3. Standard Character Encoding Set for Information Interchange

Since 1980, when the first character set, Chinese Character Code for Information Interchange (CCCII, 53,940 characters) was published in Taiwan, numberless character codes have been devised and published by respective governments or some major IT industries in *hanzi*-using polities. Most familiar to computer users' eyes are GB 2312-80 and Big 5. The former enumerates 6,763 characters[1] and was enforced by China's National Bureau of Standards in 1981; the latter, Big 5, is GB 2312-80's equivalent in Taiwan (*de facto* official standard), which includes 13,053 characters and was jointly endorsed by five big computer companies that collaborated in its development in 1984. These standard sets have established themselves as an important parameter for examining the quantity of *hanzi*. Some other influential systems include: the Government Chinese Character Set in Hong Kong (GCCS, 3,049 characters, 1994); the Japanese Industry Standard (JIS Code 6226, 6,349 characters); and the Korean Information Processing Standard (KIPS, 2,192 characters). ISO10646.1/GB 13000.1 was devised by ISO and the Unicode Consortium in 1993. 20,902 CJK (Chinese, Japanese and Korean) sinographs (a word coined by Mair (1991) to collectively refer to the *hanzi*-derived ideography of East-Asia) were included as a result of merging or unifying over 20 character sets and telegraphy codes (totaling 121,403 characters), introduced by the USA, Taiwan, Mainland China and Korea (Lunde, 1993: 49-53). The number of *hanzi* encoded in these national and international standard sets has grown year by year. On March 17, 2000, the Ministry of Information Industry and the former State Bureau of Technological Quality Supervision in China jointly issued GB 18030-2000, another national encoding standard for 27,484 *hanzi*. This is by far the most fundamental encoding set and will define the country's computer system for an infinite period into the future.

Having briefly described the *hanzi* profile in quantitative terms, both synchronically and diachronically, in what follows we will turn to what this means for Chinese character computerization.

2. From Oracle Bone to Computer – An Emerging Real-life Problem

In spite of the fact that it has served Chinese language generally well since its evolvement from *jiaguwen*, *hanzi* is a writing system that has been notorious for its three "technical hitches", i.e. enormous variations, a complex structure and an unstable number. The above brief introduction about the number of Chinese characters shows, on one hand, the massiveness of the TN; on the other hand, it demonstrates that there is a great disparity between the TN used for different registers and domains. Traditionally, the educational consideration is the main thrust of the argument supporting the need of having some tables prescribing the character quantity and usage. For the elimination of illiteracy, for example, it is important to know how many *hanzi* an illiterate needs to learn in order to operate successfully in a literate world. It is equally desirable to have some kind of restrictions on character use for school education in writing and reading. Because, so the saying goes, no one knows all the characters – mastery of the most frequently occurring *hanzi* is important, as they provide a relatively high usage and recognition rate in the written discourse.

With the ascent of the computer era, the entirety of *hanzi* has increasingly become troubled waters for both language planners and computer scientists – the need to regulate the use of characters appears rather immense. This is so because:

2.1. Why Is It a Problem

Firstly, in order to devise the structure-based *hanzi* input software, it is essential to know entire types of strokes/components that are the basic units of *hanzi* representation on the computer screen. An optimized stroke/component classification system that can best reconstruct all Chinese characters has been attempted, but yielded no successful result due to the unknown TN. Secondly, as mentioned already, the computer industry across the world employs at present numerous different encoding character sets for information exchange. The coexistence of standards, adopted in isolation, detrimentally affects each other and prevents information from communicating and interchanging stably and smoothly. Frequent failures to correctly decode and display files wrapped in *hanzi*, for different end-users or on different platform applications, have become a way of life for sinograph users. Despite the rapid rise of Unicode, with such a big and unstable number of characters there is no hope of unifying all *hanzi* code standards. The third reason is more evident and straightforward. Computers are assumed to have the capacity of processing all characters, but because of the three unique features of *hanzi*, i.e. big and unstable number, complex structure, and incapability of indicating the correct pronunciation, it is in fact impractical for users to be able to deal with the characters outside their daily purview, even if every computer was equipped with a large enough character database.

Broadly speaking, the conflict between the *hanzi* and the computer in terms of TN manifests in two aspects. On the one hand, the TN currently encoded in standard character sets is too small to process some big corpus of text in special areas. As noted earlier, at present the TN encoded in the largest IT-oriented Character Sets, issued by the government, are 20,902 in ISO 10646/GB13000.1 (1993), and 27,484 characters in GB18030 (2000). These numbers are obviously far from being sufficient to process all the orthographic forms that ever existed. For example, the paucity of the Chinese classical written heritage in cyberspace has been talked about for quite some time. To computerize the ancient texts is necessary for the analysis, categorization and encoding of every character used in the colossal body of ancient works, which inevitably involves an overhaul of the whole repertoire of *hanzi*. Given the fact that *hanzi* is the oldest writing system, surviving thousands of years in an uninterrupted civilization, it is inexorably difficult.

On the other hand, the TN used in the general texts in public domains has proven too big and unstable for common readers to deal with electronically. To overcome this difficulty, it is also necessary to control character use through restricting the TN for general purposes. For non-specialist computer users, the biggest trouble are the so-called Rarely Used Characters (henceforth RC). Despite the fact that their TN may be incredibly large, an average person, reading modern written material for good comprehension, needs to know relatively few. It is possible that various domains and corpora favor a certain type of character, but it is generally agreed that 2,500 represent the lower end and 3,500 the higher goal for a reader in mainland China who wishes to gain an over 99 per cent understanding of modern texts (it might be a bit higher for readers in traditional character-using polities). In real terms, we can safely say that 3,000 can be considered as the watershed; mastering the knowledge of additional characters beyond this baseline does not give the reader a great advantage or net gain[2].

4.2. Small Number, Big Trouble

However, regardless of their very small number, RC can make endless trouble in social use and in machine processing. As noted by Ao Xiaoping (2000: 74), although 3,000 characters cover more than 99 percent of text, no one can guarantee the 3,001st character does not appear. “Out of seven or eight thousand characters in current circulation, more than half are

non-common characters", Ao continued. This is a well-attested phenomenon of what Zhou (1992: 156) called the Rule of Decreasing Percentage Coverage (*Hanzi Xiaoyong Dijian Lv*), i.e. a relatively small number of high frequency *hanzi* typically make up a very high percentage of modern texts, whereas a large number of lower-frequency characters occur only a few times. One ramification of this frequency distribution is that the last few percentage points of coverage are made up of a great number of RC.

The instability of TN has been causing great confusion in the IT industry. In 1986, Xinhua News Agency, the biggest news provider in China, has investigated its 90,672 items of news reports, and 395 characters were found beyond the GB 2312-80, whose 6,763 characters are based on some of the best-known household dictionaries, such as the *Dictionary of Modern Chinese* (Yang, 2000: 195-6). This means that the agency had to create them by using a software package that can generate new *hanzi* upon demand. However, that small number of characters accounts for over 15 percent of the total number that the agency used in 1986. It is often reported that some customers were refused banking services just because some characters used in names cannot be found in the national standard code sets for information exchange. The Chinese banking system requires that Chinese names must be precisely identified by Chinese characters (Wang, 2002: 576). The examples given here are only the tip of the iceberg. As a token of the problem, Zhao and Baldauf' (2007) study exemplified the endless problem caused by the disparity of between small number of commonly used characters and disproportionate huge number of rarely used character needed in naming purpose. According to a media report, during the ID card updating process for the 9.8 million people living in Beijing in 2006, 231 cards could not be renewed because characters on their I D cards were not available in the updated and specially expanded character database. Even if these characters could be created temporarily by using special software, the uncoded characters would not be displayable and transmittable on other computers or over the Internet. With computer use spreading into more areas such as residential management, human resources, banking, insurance, security and transportation, it is not too far-fetched to imagine that these problematic characters cause processing break-downs and entail unnecessary work when they cannot be automatically processed by computer.

To facilitate the computability of the Chinese writing system through the optimization of *hanzi's* repertoire, *hanzi* standardization was singled out as one of the core working agendas of Chinese language planning at the hallmark National Conference on Language and Script Work in 1986. The major undertaking of the standardization have been the so-called 'Four Fixations' (see subsequent discussion). Each of these four fixations has encountered difficulties since then, and *hanzi* continue to remain the bottleneck in Chinese language computerization. This again made the issue of obtaining a fixed and frozen TN a paramount concern among Chinese LP decision makers and researchers over the last two decades.

3. Sociolinguistic Investigation into the Instability of *Hanzi*

A natural question arising from the above discussion would be: As the characters used in modern written texts are quite limited, what prevents us putting an upper limit on their growing number?

Chinese character use is a heavily culture-charged writing system, where individualism in the use of written words, even any deviation from the norm, has been extensively tolerated. Limiting people's character use, in most cases, has long been seen as constituting a form of behavior control. Early research about the quantity of *hanzi* was confined to the internal factors of *hanzi* per se; nowadays, language planners have come to realize that no linguistic event takes place in a theoretical vacuum outside of a specific socio-cultural

context, the solution to *hanzi's* misleadingly large number should be looked at from a broader perspective. Two forces, roughly speaking, prevent putting a straitjacket on the expansion of the TN: people's inclination for language novelty and cultural obsession. Each of these two reasons, which are associated with variant forms and rarely used characters respectively, will be discussed in turn in the following two sections.

It is generally agreed that, in addition to the aforementioned RC, obsolete characters and variant forms (henceforth VF) are two other major sources for making the *hanzi* TN uncontrollable. Obsolete characters, also known as historical characters, are those that have always existed for recording historical events. After developing for over three millennia without undergoing any thorough streamlining, their number in the voluminous classics, preserved and handed down from remote antiquity, has accumulated and they were deposited in dictionaries. These characters, although most have disappeared forever, were one important factor accounting for the explosion of the quantity of *hanzi* in some dictionaries. Since they no longer pose a severe impediment in modern use, the following analysis will focus on VF and RC.

3.1. Variant Forms

VF is defined as several characters having the same meaning and pronunciation but different forms. The type of VF is categorized by the relationships between characters in the same group. Supposing a range of different forms of a lexical entry (*hanzi*) were found to exist in various sources, the prospective standard character is A and the remaining ones are B, then the relationships between them are as follows (Zhao and Baldauf, forthcoming):

Type I – Absolute VF: B is/are completely the same as A, both semantically and phonetically; B is/are called VF of A. E.g.:

A – 窗 (chuang, window) = B –窓, 窗, 牕, 牕.

Type II – Containing Relationship: A is more inclusive than B in meaning and all the meanings of B are included in A. E.g.:

A – 布 (Bu): ① Clothing; ② To declare or to issue; ③ To spread or to distribute; ④ To arrange or to plan; ⑤ A kind of ancient currency; ⑥ Surname.

B – 佈 (Bu): ① To declare or to issue; ② To spread or to distribute; ③ To arrange or to plan.

Type III – Overlapping Relationship: A and B are overlapping semantically or phonetically. In the following example, 媮 can be pronounced in two ways. When A is pronounced as 'yu', its meaning is not included in B – 偷 (tou).

A – 偷 (tou): ① To steal; ② Stealthily; ③ To spare time; ④ Perfunctory, being content with temporary comfort.

B1 – 媮 (tou): ① Perfunctory, being content with temporary comfort.

B2 – 媮 (yu): ① Delightfulness; ② To look down upon, to despise.

Absolute VF are purely duplicates without any functional role in semantic and/or phonetic differentiation from the standard counterparts. In the 1950s and 60s, removing a large

number of such VF was a parallel effort with reducing the structural complexity and lowering the number of characters in use. But a snag developed where lexical factors were involved, not only because of the technical difficulty to identify the different types of VF, but also because of people's perceptions about the use of *hanzi*.

Hanzi users have idiosyncratic ways of expressing subtleties. This makes a large number of such characters a necessity, constraining the effectiveness of character restriction. As Coulmas (1989: 242) points out, "character standardization is hence first and foremost a lexicographic task". Doing away with VF is in effectiveness a matter of striking a balance between the distinctness in meaning and the cutback in number. A big number of VF were coined to signify minuscule differences that the creators considered important enough, but most often because they just want to show off their erudite knowledge about the character and their skill in discerning the subtle semantic dissimilarities.

There are two kinds of increase in *hanzi's* total number with relation to VF; one is positive and the other negative. The positive increase stems from the requirement of meeting accuracy in expression, and this kind of increase is necessary to make the script function well in order to survive social development. The negative increase is a kind of ineffective or counterproductive increase, mainly caused by the accumulated variant forms of the same character. There is general agreement that these obsolete VF, resulting from a long historical development, are the sediment in *hanzi*, meaningless and superfluous, an absolutely unnecessary burden on users' memories. However, once created, they generate a niche for themselves because of what Wang (1989: 573) has termed a "Backward Compatible Principle":

> the authorities in all dynasties were found to be tolerant of the existing forms in the previous
>
> texts, but imposed stringent restrictions on the newly created characters that were in current
>
> circulation. Thereby a large number of variant forms were shielded under this policy of 'stress
>
> the past and suppress the present', which inevitably led to a sharp increase of *hanzi*.

To optimize Chinese characters, the Committee of Language Reform and the Ministry of Culture jointly promulgated the *First Table of Variant Forms* in 1956, and 1,053 VF were eliminated through careful selection (some twenty-six have been resumed since then). This is the number confined to commonly used characters in modern times. If the increased quantity is to be counted, then the number of VF will be many times greater. For example, out of 47,035 characters in *Kangxi Zidian*, over 20,000 are VF, accounting for 40 per cent of the total (Gao, 2002: 276). In the *Hanyu Da Cidian* (Great Dictionary of Modern Chinese, 1990), which lists over 54,000 characters, approximately 20,000 are VF. Therefore, eliminating or merging VF has been an effective way of reducing the total number. But if too many are being seen as VF and are given the death sentence (eliminated from the writing system), this may occur at the expense of the ability to make distinctions and increase the ambiguity in meaning, imperiling the expressive power of the language. The most discussed issue is the handling of a number of variant forms that have their own phonetic value, thus discharging the less prestigious and complex variants, giving inevitably rise to homophonous substitutions. It is, in essence, a matter of testing to what extent the general population would tolerate these ambiguities.

3.2. Rarely Used Characters

Rarely Used Characters or infrequently used characters are in a special register and domain. Some of them have been created for special purposes, e.g. denoting newly found chemical elements. The RC number twenty times the common characters in the TN. These characters

differ greatly in frequency of usage, with most appearing only occasionally but anytime and anywhere, acting like submerged rocks in an ocean of writing and reading.

RC can be discussed under two rubrics: specialty characters (*Zhuan Yong Zi*) and literary characters (*Wenxue Secai Zi*).

Specialty characters are necessary for special topics and purposes. It is generally agreed that specialty characters are sourced from six to nine areas. These areas can be fitted into two broad headings, omitting the second category due to space limitation:

– Characters for proper names, including characters for foreign proper name translations, and for ethnic minorities and religious purposes; In broader terms, there are also peripheral proper names, i.e. names of buildings, commercial brands and shops, etc.;

– Characters for science/technology, medicines, animals and plants.

Apart from personal names, characters for place names are an important part of the category of proper names. One successful attempt in delimiting specialty *hanzi* has been the replacement of place names under governmental fiats from March 30, 1955 to August 29, 1964. Changing these characters, be it their physical shape or their pronunciation, is an extremely emotional and controversial issue, particularly for names with historical implications or those used by ethnic minority groups (see, Zhao & Baldauf, 2007). Characters for personal names are another area that promises a radical reduction in the total number. The proposed list that is being studied, designated especially for naming, ought to include 12,000 characters (including their traditional and variant forms). Most accommodate the minute number of rare naming characters. Research (Su, 2004) shows that 2,500 *hanzi* can cover 98 percent of modern Chinese names in Mainland China. It is debatable whether it is practically possible to standardize every character for everybody's name in such an enormous country with such a long recorded civilization[3]. A very heated nation-wide debate was triggered off, revolving around 'Shall We Have a Restriction on Name Giving Rights?', when the *Table of Standardized Characters for Naming* was included in a national language research program (www.shyywz.com/page/jsp/showdetail.jsp?id=1080, 30/8/2003).

Having seen the complexity of the specialty character reduction, let us turn to the literary characters – a more hazardous issue in the TN reduction venture. The purpose of literary characters, as the name suggests, is to make literary writing stylish and attractive to readers. The archaic *hanzi* and dialectal *hanzi* are a big constituent body of it. Archaic characters are the characters carried over into modern texts from classical Chinese in the form of archaic words and expressions, found predominantly in proverbs and idiomatic phrases. Well-known for the richness of its vocabulary, Chinese has created a myriad of works in its dynastic history since the turtle bone inscriptions. It is believed that at least 8,000 titles of ancient classical Chinese texts have survived into modern times[4]. The influence of this ancient heritage is considered the major source of RC found in modern general texts.

Akin to classical literature and traditional characters, many characters/words that are inherently associated with classical works but not semantically needed in modern texts, enjoy a high prestige and authentic status as they symbolize a time-honored heritage. Literary characters are, in essence, an exemplification of manifest archaism and cultural obsession rather than linguistic necessity. The Chinese spoken and written languages are markedly dissimilar. The degree of disparity in lexicon and syntax ranges so widely, that they are almost unintelligible to each other. After the "Vernacularization Movement" had been completed for a century, a large number of characters, exemplifying classical texts, still survived, posing as a major destructive factor in conforming written language to a uniform standard of using only a fixed number of prescribed characters. To complicate matters still further, there is a deep-seated literary superiority complex in employing millennia-old expressions to add an aura of elitism to contemporary texts. As DeFrancis

(1984: 286) notes, “attachment to characters which boast a vast body of literature, a system so deeply embedded in Chinese society, is naturally far more resistant to change.” After the 1990s, with the rise of a healthier attitude towards the traditional heritage that suffered serious destruction during the Great Cultural Revolution (1966-76), there came a return to traditional things. A ‘Back to the Ancients’ sentiment has been a new trend in popular culture and engendered a large number of classical works to be reintroduced into school education, which in turn boosted a re-emergence of archaic usage in graphic life.

This indicates that, rather than treating characters as a means of written communication, people, and some scholars in particular, use them as a vehicle to display scholarship and intellectual superiority. Although literary characters are an alternative and not strictly necessary for writing, they are the most uncontainable and irrepressible for restraining the abundance of characters. This kind of intellectual exercise reminds us of a similar problem that Japan has confronted. As Twine (1991: 215) notes, “It was a favorite ploy of scholars wishing to display their erudition to pad out the text of their discourse with unnecessarily complex characters”. However, Japan was quite successful in regulating the legitimate number of characters for modern use to an upper limit of 1,850 in 1946 (*Table of Contemporary Characters*, the number was increased to 1,945 in 1981), plus an official list containing extra characters for giving names to children born after May 25, 1951 (Watanabe, 2007).

From the point of controllability, these RC differ a great deal in terms of their dynamism and visibility. As the preceding discussion shows, except for VF and literary characters, all other types of RC are found to be relatively brought under control, and the authorities have actually never stopped trying to do this. Some experiences have been gained and progress is being made on how to manage RC. Wang Tiekun (2004), the Vice-Director of the Language and Information Management Department of the Education Ministry, says, “The work to standardize characters for personal names, geographical names and technological terms has never been so important and urgent”.

To sum up, in contrast to the alphabetic letter system, the Chinese character system is open to public creativity and productivity. Therefore, although 3,500 characters at the most are sufficient for the lexical representation of modern Chinese for general purposes, the real impediment for restricting the TN rests in the users’ attitudes towards *hanzi* rather than in the linguistic rationale. Wu (1995: 85-90) has investigated how, and under what circumstances, new characters were created, and has shown that from antiquity to the modern era, new characters may be produced at any time by anyone. In one sense, every Chinese person can be a *hanzi* creator, which consequently makes the number of the character shapes literally too large to describe. Probably, because writing characters is apt to be a very idiosyncratic process, parallel forms have learned to co-exist, and people have always been accustomed to using a wide range of diverse forms of characters. The psychological and attitudinal reasons for and types of character complication are an interesting topic worth exploring further.

4. Striving for a Solution

4.1. A Recap of the Past Experience – Great Efforts, Little Gains

Hanzi’s TN streamlining has been a vital part of one century of script reform movements. However, despite vigorous efforts over a long historical time span, little has been achieved in comparison with the collective efforts by generations of language reform pioneers. Lu Feikui was perhaps the first scholar who saw the importance of delimiting the commonly used characters while undertaking a structural simplification. In his suggestions published in 1922, he considered 2,000 characters were sufficient to satisfy the basic needs of daily

use for ordinary people at that time. But Lu did not elaborate on how to achieve this goal. Following Lu's example was Hong Shen, a famous playwright, who was also actively advocating confining the modern *hanzi* to an even smaller number. Probably inspired by the Ogden and Richard's (Carter & McCarthy, 1991) work in devising the so-called Basic English during the early 1930s to provide a basic minimum vocabulary for English learners, his method was based on the coinage of multiple-syllable words by using prescribed characters to replace the ones that are structurally complex or rarely used. Hong attempted to delimit 1,000 characters for general purpose, with 250 characters for special use. Although the schemes described above will not embrace full orthography, they will at least not be un-Chinese. In 1939, an even more radical advocate, Zhai Jianxiong, wanted to employ only 454 characters as syllables, to transliterate Chinese writing.

All these proposals wandered too far from script reform as they invariably affected the lexical system and resulted in limiting the expressive power of the language, thus leading to a wordy and artificially dumbed-down style for less educated readers. At the basis of these experimental schemes is the notion of a communicative adequacy. The writing system is an instrument to serve the language. When a large, new set of vocabulary is created to accommodate delimited characters, it reverses the functions of language and script. It came as no surprise that these efforts failed to get support from the population, for which these schemes were designed.

Nevertheless, the previous failures did not prevent other scholars, after the establishment of the P.R. of China in 1949, from continuing the ambitious course of putting a limit on character use. Because the radical change of the political climate ruled out an environment that was conducive to experimenting with individual schemes, scholars' efforts focused on theoretical exploration. In 1953, the Commission of Chinese Script Reform, the national official body of LP, had launched an experiment to test if the *List of 1,469 Characters* would be sufficient to deal with the wide variety of texts (twelve areas in all) of modern life, and the result was found to generate more problems than it resolved. In 1964, in a paper published in the public media, Zhou Youguang, the most prolific Chinese LP researcher, advocated the wider use of *pinyin* to supersede those 'difficult characters', thus reducing *hanzi* within the limit of 3,500. Unfortunately, his suggestions were made just before the Great Cultural Revolution started, and they did not draw much attention from either the authorities or the public. During the Cultural Revolution, in spite of the negative evidence accumulated in the past, under the thrust of the ultra-Leftist tenet of the so-called 'Mass Line', another set of standard *hanzi* was made by publishing house workers and tested in a variety of samples. The outcome showed that, owing to the extensive homophonous replacements, the misunderstandings found in the sample texts, printed with the prescribed 3,260 characters, were beyond an acceptable level.

To sum up, we find that, methodologically, what has been previously trialed and tried can be generalized into four typical schemes as follows (also see the table), which are, we hope, to be representative of efforts by other scholars at large.

- Homophonous substitution. To substitute those infrequently used characters with the ones that have same/similar pronunciation included in the designated character list. Central to this kind of core or nuclear character is the idea that many notions can be expressed by using a more basic language, thus the learning burden of these characters is kept to a minimum. This has been the by far most widely explored method.
- Superseded with alphabetic spelling (*pinyin*). To use alphabetic symbols to supersede the rarely used characters in the text. This has been a long-debated scheme in previous efforts, as the biggest problem of this method is that the

systematic use of alien elements in a logographic script will inevitably result in a style of script that smacks of cross-hybridization, proving to be unacceptable to most users. However, this is the only method that has achieved success to some extent, and can be seen in today's graphic life, occasionally in publications and frequently in private text/handwriting.

- Semantic Substitution. To use explanatory words (oral expression) to annotate the rarely used characters (written language) through word paraphrasing, e.g. to use 'son's wife' to substitute 'xi' (媳). This is to use language means to compensate for the inefficiency of the writing system, where an adult's fundamental linguistic needs can be communicated periphrastically.
- Phonetic *Hanzi*. To use structurally simple characters as the basic phonetic syllable or symbol to 'spell' Chinese characters. This is the most radical of the four methods, obviously inspired by the traditional *fanqie*, ('cut and splice' – a method to describe the reading of an unknown character by means of a dual set of characters with known reading). The drawback of this scheme is salient. The users get frustrated trying to guess the supposed meaning from the sound element characters without tone and morphemic differentiation. Thus it was criticized of little practical value.

Advocate	**Advocate**	**Number**	**Work**	**Publication**	**Methodology**
Lu Feikui	1922	2000	*My suggestions on collecting and collating Chinese characters*	National Language Monthly, 1/1	?
Hong Shen	1935	1100+250	*Teaching methodology of 1100 basic Chinese characters*	By Shenghuo Press	Semantic Substitution
Zhai Jianxiong	1939	454	*Issues on a tool for eliminating illiteracy*	Fortnightly Journal of National Review, 222/12	Phonetic *Hanzi*
Commission of Script Reform	1953	1469	*Table 7685 Characters by categories*	by Commission of Script Reform	Homophonous Substitution
Zhou Youguang	1964	3500	*Delimiting and reducing*	Guangming Daily (22.7.1964)	Superseded with alphabetic

			characters in modern Chinese		spelling
Staff of Renmin Daily	1976	3260	*The design of "Standard Characters"*	Renmin Daily (18.7.1976)	Homophonous Substitution

What has been described above is a far from complete survey of the past attempts in keeping characters within the prescribed range. This succinct reflection just typically shows tortuous attainments over a long span of time in pursuance of a writing system with a fixed number of characters. Despite the previous failures or unproductive efforts, in an era dominated by mounting activities of frequent information exchange over the Internet, obtaining a relatively small number of characters appears even more imperative than it was in the past, and cutting the numbers will still be one of the most important tasks in LP for quite a long period into the future.

4.2. The Ongoing Exploration – Toward a Framework for the Future

The ongoing national research project of formulating the Comprehensive Table of Standardized Characters (CTSC) – which aims at settling four unstable attributes of *hanzi's* features, namely, the total number, the shape, meaning and the pronunciation – has been set up as the foremost task for LP practitioners in 2002 (for details, see Zhao, 2005). It is the ultimate summary of all the *hanzi* standards and tables promulgated by the national language authorities, and to arrive at a fixed number and to delineate what characters should be included in the table, are the first steps and the foundation for the other three fixations. Important as it is, the quantity determination has proved to be a hard nut to crack. Discrepancies are found to exist among the key protagonists of the CTSC project (e.g., 2004: Li, 149; Wang, 195; Zhang, 246). To be brief, there are two competing views on the TN that should be included in the table. A number of scholars are of the view that the chief purpose of the CTSC is to serve the modern *hanzi* users; there is no need to standardize the RC, so the TN should be restricted to a fixed number ranging between 8,000-12,000. The other side argues that it at least will be able to cover the extant standard codes for information exchange.

In view of the CTSC becoming the most important national standard, governing character use for both human and machine for decades to come, a relatively small number is convenient for general *hanzi* users. Computers, however, are expected to be able to process all real world characters, introduced by the borderless information inflow. This development will, therefore, inevitably lead to an apparent paradox in either case: from an educational point of view, an operationally small number, say about 5,200, can cover more than 99.99 per cent of modern publications. But this number will be obviously error prone if the written text extends to a broader domain. For example, computer applications would find 10,000 characters far too insufficient for the reproduction of ancient texts, and this number would be rather restrictive in dealing with the *hanzi*-wrapped information circulation through international communication networks. On the other hand, a larger number, in which most *hanzi* are only necessary for very special purposes, would, technologically speaking, only cause a kind of waste cyber resources. Furthermore, the unimaginably huge exertion aside, to include more characters in the national standard

would, from the perspective of LP, result in not only bringing more RC, variant forms and non-Mandarin characters into circulation for common use, but also entailing the revision of a large number of obsolete characters that are no longer relevant in the realm of modern life. One aspect, which was not mentioned by our previous discussion, is that once a government-mandated standard is put into force, it tends to trigger a 'the more, the better' competition among dictionary compilers, input program devisers and software venders, all contending for the largest-size product. Although the inclusiveness of dictionaries and the data processing ability of a larger number of *hanzi* is practically valueless, they do offer a selling point in the marketplace.

Faced with this kind of paradox, particularly frequent criticism was heard from time to time over the past few years for the failures of various official standards to deal with situations that are more diverse. After extensive reflection and review of the previous practices of managing character use, a new model of what can be called function-specific multilayered standards has become the mainstream thesis among LP decision makers and was received very positively by many researchers. This model also deals with possible problems arising from the more complex circumstances of a new historical context. First proposed by Wang Tiekun (2004), it was further elaborated in detail by other predominant scholars such as Wang Ning (2004) and Fei and Xu (2005). Multilevel approaches to provide theoretic explanations for understanding the limitation of previously published tables, as well as envisaging a paradigm to define the characters that are going to appear in the CTSC, are to be formulated. According to Wang (2004), there are three key criteria to choose the characters to be included in any standard table: time, region and domain or purposes. Based on this principle, Wang proposes that instead of trying to work out a one-fits-all standard, the characters included in the CTSC should be arranged at three levels for respective purposes: characters in Layer 1 (approx. 3,500) are the commonest, in Layer 2 (approx. 4,500) are less frequently used characters and Layer 3 (approx. 4, 000) is designated to serve for special purposes. The total number was stemmed from a linguistic corpus of 70 millions *hanzi* tokens texts stretched over the 20th century. Wang's proposal is characterized by its limitation to a) common modern users b) within mainland China c) for general communication. While concurring with Wang's three determinate criteria, we made necessary modification and present it in a figure as follows.

Layer	Time	Region	Domain
L-I 3,500	…….. current	nationwide	basic education and daily life
L-II 4,500	……. current + modern	nationwide + regional	general publications and classical texts
L-III 8,000	……. current + modern + historical	local + regional + international	personal/placenames, chemical, medical, and religious, etc.

Our modification expends Wang's suggestion to include more character (at Layer III) for dealing with a more complex situation in a wider range of physical areas and longer historical period, characters are clustered around three hierarchic-layer stratification so as to show a concentric outward expansion from basic core characters at the top layer to the less used characters at lower layer in terms of time, region and function. The best feature of

such a domain-appropriate and function-specific multilayered standards model is that it discerns the previously intermingled relationships between the four relationships, namely, past vs. the present, the inside vs. the outside (of Mainland China), the majority vs. the minority and the human vs. the machine. Take the *General Table of Simplified Characters* (2,235) as an example. The table, which was made in the 1950s and officially promulgated in 1964 (revised in 1986), was originally to target on the common users, who were struggling to get a basic grasp of reading and writing the characters encountered in daily life. It by no means intended to embrace all characters used in all domains for all functions at all time. Failure to understand the functional areas and the sociolinguistic underpinnings of the table leads to over-or -unlimited generalization. Nowadays, the rising fever of the so-called Culture Renaissance makes it a profitable to publish classical canons in simplified character, but to reprint classical texts, which are normally read by a minority of the so-called elite intellectuals, require many times characters more than 2,235, publishers often find the needs to simplify characters by analogizing principles of simplified radicals/components. As a result, a number of 'simplified characters', which were neither included in the official simplification tables nor existed in history, were coined because of overgeneralization. Similarly, users are frequently annoyed by the inability of machine applications to match simplified characters one-to-one with their traditional counterparts in automatic conversions. This occurs, according to the model of function-specific multilayered standards, because character simplification was designed for the convenience of human written communication; the computer application is obviously beyond the functional areas it intends to serve. Fei and Xu (2005) emphasize the notion of level-specific utilization, and argue that most of the previously formulated standards about character use are in essence aimed at the level of the majority public; they have neither responsibility nor ability to regulate the exceptional instances which were usually only appreciated by a minority of privileged social groups. As this notion has never been explicitly articulated by the standard setters within the framework of this kind of relative and multidimensional model, many usages that were considered unofficial or inappropriate in the past can now be treated in an individual way and thus the discrepancies between the majority and minority are readily resolved.

The fact that the new model has been cogently expressed in a number of belated articles shows that, in order to address the discrepancy between the prescribed standards and the variations in real life, the problems born out of the transformation need to be looked at from a new angle. Although a fair number of issues are far from being settled, some are in fact rather fundamental questions. For instance, at a time characterized by the free information flow in a borderless virtual world, characters appear to be indefinable and elusive, users are more likely to read or write rare characters that are largely impervious to deliberate delineation. In other words, modern day users in the digital society are obviously exposed to an ever-growing body of written communication that cannot be defined or constrained by a fixed number of characters. A fixed number of characters, however, is of concern only in a binding domain and intends to restrict variation in a rather tight way. The specific interest of the multi-layered theory is largely due to the fact that it defines the role of the standard for more flexible development, giving more appropriate consideration to individual rights (e.g., naming rights) in character use, and that it recognizes the need of striking a balance between four sets of relationship. The author believes the model has become popular because the notion of functions and levels has been understood as offering a particularly useful solution for an as yet unsatisfactory outcome of standardization practices. Given the prolific nature of *hanzi's* functional distribution it may not be an ultimate solution. Nevertheless, new paradigms for new concepts could lead the way to new thinking about the demarcation in an increasingly diverse and complex society. It is quite legitimate to proceed in this way and to operate temporarily with a relatively defined concept, for this

enables all concerned to reflect upon more profound insights in their effort to standardize the TN.

5. Conclusion

In this paper, we first surveyed the historical growth of *hanzi's* inventory and made an inquest into the areas where *hanzi's* total number issue has become a concern for both educators and software developers/IT professionals, thus underpinning the urgency of overhauling *hanzi* invoked and driven by the increasing globalization and accelerated use of the technologies of modernity. In the second and third parts we highlighted the difficulties confronting both IT and language planning professionals in dealing with *hanzi* and its TN, pointing out the psychological and socio-cultural-political dimensions. The fourth part documents the previous attempts and currently ongoing undertakings in taming the unbridled numerical expansion of *hanzi's* TN, demonstrating the interplay between scholarly aspirations and pragmatic limitations.

In looking toward the future, technological development has taken over the educational requirements to become the *raison d'etre* in invoking the necessity of maintaining a relatively stable number of *hanzi*. In an increasingly digitalized society, *hanzi s*erves as the interface between humans and machines. As long as *hanzi* computerization and its cyberspace survival remain unresolved, issues concerning the amount of *hanzi* and their stability are bound to be brought to the fore, exerting pressure on the needs of *hanzi's* size regulation. To achieve the future success of language management and planning, a new set of mindsets and new strategies will be required to address the language's emerging function. Zhao (2005) notes that the recent development indicates that technology has increasingly become a major dynamo for linguistic growth, and that IT-oriented LP activities will remain a fundamental feature of most of the ensuing language reform. If this is the tendency, the visionary framework succinctly outlined in the conclusion should suffice to confirm the latest efforts by Chinese language planners. It is, therefore, hoped that this study will provide some preliminary illustration of the complexity and intricacies of one aspect of language planning, thus enhancing our overall understanding of a number of social and contextual factors that seem to affect the technological treatment of language attributes in a complex way.

NOTES

1. It also includes 682 other non-*hanzi* symbols and signs. A discharged variant form of 'rong', a character in the given name of Zhu Rongji, was added to the coding set in 1993. The enlargement was designated to deal with the frequent confusion in the public media, due to the absence of 'rong' in the *hanzi* database of the Chinese character input program. Zhu Rongji was Chinese PM from 1998 to 2003.
2. Based on the latest statistics (Wang, 2007) of 0.978 billion of *hanzi* tokens, 591, 958 and 2377 characters cover 80%,

   90% and 99% of the texts selected from public media, internet and educational materials used in 2006.
3. In Japan, the government enacted a *Supplementary Table of Characters for Name Giving* in 1951 with 92 characters, increased to 166 in October 1981. However, the law was challenged when questions concerning individual rights were raised and a court case developed (Neustupný, 1983). A similar proposal from the Taiwan IT industry failed to be passed in the legislature in the 1980s (Tse, 1983: 16).

4. To exemplify the difficulty of streamlining the characters for literary purposes and classical style, we can look at classical encyclopedias written during various Chinese dynasties. The *Complete Book by Four Categories* (*Si Ku Quan Shu*) was the last encyclopedia compiled under royal patronage in 1773. This 79,309-volume book contains 3,461 titles with nearly one billion characters (token). The total number of characters in the *History of Twenty-Five Dynasties* (*Er Shi Wu Shi*), a small part of this encyclopedia, has 13,966 types of characters recurring 31,409,450 times.

References:

Ao, X. P. (2000) Talking about Chinese language and Chinese characters, In P. C. Su, Y. M. Yan & Y. B .Yin (eds.) *Forum on Language Modernization (4)*.64-79. Beijing: Beijing University Press.

Carter, R. & McCarthy, M. (1991) Wordlist and learning words: Some foundations. In R. Carter & M. McCarthy (eds.) *Vocabulary and Language Teaching*. 1-17. London and New York: Longman.

Chen, P. (1996). Modern written Chinese: dialects and regional identity. *Language Problems and Language Planning* 20, 223-43.

Coulmas, F. (1989) *The Writing System of the World*. Oxford: Blackwell Publishers.

DeFrancis, J. (1984) *The Chinese Language – Facts and Fantasy*. Honolulu: University of Hawaii Press.

Fei, J. C. & Xu, L. L. (2005). Examining Chinese character standardization from a non-canonical perspective/alternate view. *Language Review* (HK) 80, 30-6.

Gao, G. S. (2002) *On Standardization of Modern Characters*. Beijing: Commercial Press.

Jordan, D. K. (2002) Language left behind: Keeping Taiwanese off the World Wide Web. *Language Problems and Language Planning* 26/2, 111-27.

Lunde, K. (1999) *Understanding Japanese Information Processing*. Sebastopol Calif: O'Reilly & Associates, Inc.

Mair, V. H. (1991) Forward Preface: Building the future of information processing in East Asia demands facing linguistic and technological reality. In V. H. Mair and Y. Q. Liu (eds.) *Characters and Computers*. pp. 1-9. Amsterdam: IOS Press.

Neustupný, J. V. (1983) "Language planning and human rights". *Philippine Journal of Linguistics* 14-15, 2/1, 66-74.

Research Team of Computer Information Processing (1980) The five principles of formulating the standard Chinese character codes. *Language Modernization* 4, 64-78.

Su, P. C. (2004) The pro and the contra of standardizing the characters for personal names. Retrieved: November 15, from http://www.China-language.gov.cn/webinfopub/list.asp?id=1042&columind=154&columnlayer=000500300154.

Twine, N. (1991) *Language and the Modern State: The Reform of Written Japanese*. London: Routledge.

Tse, J. K-p. (1983) The standardization process for Chinese languages. Paper presented at Conference of Linguistic Modernization and Language Planning in Chinese-Speaking Communities (Hawaii), 1983.

Wang, F. Y. (1989) *Chinese Hanziology*. Changchun: Jilin Wenshi Press.

Wang, N. (2004) On the social nature and scientific nature of *hanzi* standardization. In Y. M. Li & J. C. Fei (eds.) *Various Views on Hanzi Standardization*. 1-18. Beijing: Commercial Press.

_____, (2006) Revisit of the Scientific Nature and Societal Nature of Chinese Character Standardization. *Applied Linguistics* 4, 2-11.

Wang, T. K. (2004) Some issues regarding research of the Comprehensive Table of Standardized Chinese Characters (CTSCC). In Y. M. Li & J. C. Fei (eds.) *Various Views on Hanzi Standardization*. 179-203. Beijing: Commercial Press.

_____, (2007) The investigative study of the usage of natural language and language planning. Paper presented to Global Mandarin Forum: The Chinese, Mandarin, Pedagogy. Singapore, 16, November.

Wang, Y. W. (2002) Some issues of fixing the number of modern *hanzi*. *The Nanyang Technological University Journal of Language and Culture* 5/2, 55-76.

Watanabe, N. (2007) Politics of Japanese naming practice: Language policy and character use. *Current Issues of Language Planning* 8/3. 344-364.

Wu, C. A. (1995) *Cultural Reflection of Chinese Characters*. Changchun: Jilin Educational Press.

Yang, R. L. (2000) *General Introduction to Modern Hanzi*ology. Beijing: Great Wall Press.

Zhang, S. Y. (2004) The tentative scheme for the formulation of the Comprehensive Table of Standardized Characters. In Y. M. Li & J. C. Fei (eds.) *Various Views on Hanzi Standardization*. 229-48. Beijing: Commercial Press.

Zhao, S. H. (2005) Chinese character modernization in the digital era – A historical perspective. *Current Issue of Language Planning* 6/3, 315-78.

Zhao, S. H. & Baldauf, B. R. Jr (2007) Language planning, naming and character use in China. *Current Issues of Language Planning* 8/3. 283-304.

Zhao, S. H. & Baldauf, R. B. Jr. (forthcoming) *Planning Chinese Characters: Evolution, Revolution and Reaction*. Boston: Springer/Kluwer Academics Publisher

Zhou, Y. G. (1992) *The Whole Story of Chinese Script*. Beijing: People's Educational Press.