ИСТОЧНИКОВЕДЕНИЕ, ИСТОРИОГРАФИЯ, ПЕРСОНАЛИИ

Sergey Zinin

University of Massachusetts, Amherst
Warring States Workshop Project

Pre-Qin digital classics: study of text length variations

Abstract

The paper analyzes length variation of fifteen pre-Qin classical Chinese texts[1] *in the digital corpora context. Data on length of classical Chinese texts as well as numbers of type-tokens are critical for quantitative linguistics' analysis of character frequencies. However, there is considerable variation of these parameters in existing digital sources, and there is no study on its causes. This article presents the data on lengths (collected together for the first time), starting from the earliest available date. Besides, the study delineates evolution of digital resources of classical Chinese, provides an up-to-date review of major available online resources and research corpora for* Shisanjing, *and traces history of their compilation. The article demonstrates scope of variation in lengths and addresses issues of multiple text versions, fluidity, and inherent inaccuracy of digital texts, which could be called "digital content gap", i.e., discrepancy between printed and digital versions of texts. The content gap could affect a traditional philological study, but it may be not very significant for a quantitative analysis. Finally, the article presents a comparative break-down of length statistics. The article concludes that results of practically any frequency study are mostly applicable only to the specific corpus that was utilized, and suggests that online availability of digital corpora for all researchers to re-use and verify results will increase reliability of studies.*

1. Introduction

There are a few studies on most-frequent character lists for classical texts[2], as well as on character sets of a few specific texts[3]; however, quantitative linguistics of classical Chinese texts is still in its early development. There are practically no studies on comparative frequency distribution of characters in vocabularies of several texts. For this type of study, numerical characteristics, such as text length and vocabulary size (in tokens) are essential. However, not many online and digital corpora provide this type of information in a convenient form.

1.1 Character as token. Currently practically all researchers report lengths of classical Chinese texts in characters, i.e., they use character as the basic measurement unit for texts ("token"). This is not a typical corpus approach. While in corpus linguistics "token" could be any meaningful grouping of characters[4], for most languages, by default, token is a word. The Chinese language corpora could be different, due to the nature of Chinese writing system. First, there are no clear word boundaries in written texts, and it is often hard to automatically identify "orthographic words"[5]. Second, there is an ongoing discussion on the definition of word in modern, as well as in classical, written Chinese[6]. Human experts identify words in Chinese texts better than computers; however, researchers reported disagreement of even human experts on supervised "word segmentation", at varying, but significant rate[7]. Therefore, while automatic word segmentation of classical Chinese texts is possible, it could be ambiguous.

This ambiguity affects accuracy of measurement of text length in words. Currently, there are no available authoritative digital editions of most classical texts with marked-up word boundaries. For a text of certain length in characters, different word segmentation algorithms (or even human experts) would produce different lengths in words and vocabularies. While such difference could be negligible for huge modern corpora, it is important for smaller classical Chinese corpora, where problem of punctuation and word segmentation has always existed. To implement "word" as "token" for classical Chinese, corpora should be provided with a stable vocabulary of words, and be properly marked-up. Only a few classical texts with such mark-up exist[8].

Therefore, although it is possible to calculate length of classical Chinese text in words, in most current studies text length is still calculated in characters[9]. This article will continue utilize characters, and not words, as the main measurement unit (token).

1.2 Two modes of approach. Proliferation of digital (online) text repositories, concordances, and text statistical analysis, being a comparatively recent phenomena, have started affecting more traditional ("philological") approach, bringing in new perspectives to text studies.

It is possible to identify two approaches (or modes) to digital text statistics. In "philological approach", one would ask, "If this specific character (word, phrase) is used in this text? In which version it could be found? How many times? Combining with what other characters? What other texts contain this character?" In quantitative linguistics (corpora studies) one would also ask: "What are the most frequent characters in the text? How they are distributed? How many type-tokens there are? How many hapax legomena (hapaxes) there are and what is ratio of them and non-hapaxes? What is distribution of sentence and word lengths?"[10]

1.3 Digital corpora and quantitative linguistics. Most of "philological" questions could be answered by paper-based concordances, but only digital corpora could provide answers for quantitative approach. Not surprisingly, collecting information on text lengths and frequencies has been closely related to development of electronic corpora of classical Chinese. Text data, not tractable for non-machine corpora studies, could be easily processed by computers[11]. Creation of electronic corpora rendered retrieval of statistical information from paper sources almost obsolete.

1.4 Text variation and digital corpora. Pre-Qin texts are an imminent part of every corpus of classical Chinese, and they are increasingly available online. However, these texts often demonstrate considerable variation. Chinese classics are generally known for having multiple versions, many of which could be considered acceptable, in various degrees, for philological studies (researchers always reference printed sources on which version was used). Computational linguistics is more experimental, and researchers need to be able to have access to corpora and repeat experiments. Existence of standard and free digital versions of classical texts would accelerate progress in this direction. This study will review the current situation, based on most popular available online corpora, as well as Warring States Workshop (WSW Ctexts) research system.

1.5 Digital data accuracy: Digital content gap. There is a discrepancy, or a "digital content gap", between printed and digitized versions of texts, due to first, digitization issues (OCR errors, manual entry errors, code-page character limitations); and second, text modification at preparation stage. The digital versions must feature some information loss or modification comparing to printed versions.

Digital corpora are created either by manual data entry or through optical character recognition (OCR) process, followed by multiple reviews. The development of OCR software for Chinese language started in the 60s, but commercial technologies became available only in the 90s. The OCR technologies for Chinese (Cherniet–2007) did not provide good accuracy until the end of 90s; by this time many academic corpora have been created by manual data entry[12]. All earlier OCR-based databases that utilized low-accuracy OCR approach may contain a considerable number of errors, even after multiple reviews. However, manual entry also brings inaccuracy which could persist even after multiple reviews.

Data entry errors could be gradually corrected; however, it means that content of these sites may be in permanent change (while changes are not always announced)[13].

Another source of content gap that plagues digital sources, especially earlier ones, is limitations in presentation of Chinese characters in computer coding pages. For a printed edition, practically any character could be custom-made or cut. In computer versions, whatever entry method is chosen, data entry operators are limited by number of characters, represented by so called code-pages. This issue has not been resolved even by introduction of Unicode. Therefore, practically all academic groups that created digital versions of classics, introduced some modifications to printed versions during digitization process, so these versions, while based on well-known editions, represent versions by their own[14].

1.6 Online availability It is important to have digital versions of texts available online, as text versions in various projects tend to be slightly different, as well as statistical results based on them. It will make possible to verify results by all researchers.

The rest of article will be structured as follows. Section 2 reviews most important online digital corpora of classical Chinese, their origins, and what role they could play in quantitative studies. It will introduce development processes of the digital corpora, how text lengths are going to be collected and what problems are going to be there. The section 3 will investigate how lengths of classic texts in characters were measured, and what lengths are available, from digital corpora. The results are discussed in Conclusion.

2. Online Corpora of Classical Chinese

2.1 Literature review. Few articles describe general evolution of Chinese electronic corpora; most of them were published in the second half of 1990s and beginning of 2000s. The most recent available review is written by Winnie Cheng (Cheng, "Corpora: Chinese Language"), and gives a short description of the most important directions in development of Chinese electronic corpora[15]. The most comprehensive report, written in 2006, belongs to Feng Zhiwei (Feng, "Evolution and present situation"), and it describes development of Chinese corpora from the beginning of 20th century to the mid–100s[16].

However, these articles focus on electronic corpora of modern Chinese and only cursory mention classical Chinese corpora. Certain amount of information on classical corpora is contained in Wang Jianxin's article of 2001 (Wang, "Recent Progress") describing early stages of electronic corpora development in mainland China and Taiwan. He list includes *Siku quanshu* electronic database (about 800 million characters), Scripta Sinica (140 million characters), Shanghai Normal University corpus (100 million characters, containing a classical Chinese section)[17]. A similar short description could also be found in the introduction to McEnery and Xiao (McEnery, Xiao, "Lancaster Corpus of Mandarin Chinese")[18]. Some information on electronic corpora of classical Chinese is present in a few articles dedicated to research corpora, which are discussed below.

This paper will start with corpora and concordancers that are available online. Although there are many websites simply featuring classical texts, this article will deal only with those that provide some advanced corpus linguistics tools and feature, beside full-text character search[19]. Beside these online corpora, some off-line research corpora, providing information of classic texts length, will be described.

2.2 Electronic corpora of classical Chinese. The most important websites (and digital resources behind them) featuring advanced search and statistical tools for classical Chinese[20] (in chronological order) are 1) Scripta Sinica, 2) C.H.A.N.T. database, 3) Academia Sinica corpus, 4) Beijing University corpora (PKU), 5) Thesaurus Linguae Sericae (TLS), and 6) Donald Sturgeon's Ctext project[21]. Four out of five, not surprisingly, are hosted by mainland Chinese and Taiwanese academic institutions (and some are at least partly commercial resources), one (Ctext) is hosted by a private organization (also based in Hong Kong), and only one (TLS) is hosted by a European institution (the latter two are free)[22]. The corpora behind CHANT and Academia Sinica resources have been digitized starting from the second half of 1980s, and two last resources started in the second half of 2000s.

Online corpora for classical Chinese studies heavily depend on availability of digitized texts and quality of the texts. The texts became to be produced since mid–80s, as soon as electronic standards for Chinese characters coding were introduced. At the same time, mid-range and personal computers became increasingly available to researchers, and it led to proliferation of digital versions of Chinese classics. Most major classic collections were digitized in the 1990s (e.g., *Siku Quanshu*), often on commercial base[23].

Printed versions often were not considered to be perfect by projects' philologists, and practically all research groups behind main East Asian online resources modified ("improved") printed texts making digital corpora, supported by resources of their academic institutions. Unfortunately, there is not so much detailed information available on this process; therefore, this paper will present just a preliminary description of this process, the description hopefully to be expanded and improved later.

Digital corpora of classical Chinese could be online and offline resources. It would be safe to say that most corpora that originated as offline resources, sooner or later went online (e.g., *Siku quanshu*). However, it seems not many online versions of corpora are available offline (or, available for download as a text version). Sometimes institutions transfer their data to other institutions (e.g., Scripta Sinica project to Academia Sinica, or CHANT to TLS), but it is rare occasions.

Most popular online electronic corpora[24] could be roughly divided into three major groups: academic, independent, and commercial. Academic corpora are usually a product of a large body of researchers, which are supported by university or academia resources. They may require subscription, but the price is not very high and most members of research community have access to it (but cannot experiment with the source). Independent resources could be academically affiliated (e.g., CText and WSW CTexts), but they are not supported by large research resources[25]. Finally, commercial corpora could be produced by members of academia (e.g., Erudition database[26]), but the corpora belong to a for-profit corporation and the access is usually limited by a high subscription price.

2.2.1 Academic corpora.

Scripta Sinica Corpus. The corpus has been developed, starting from 1984, at the Institute of History and Philology (IHP) of Academia Sinica[27]. The researchers initially planned to create a digital version of the 25 histories for a study of Chinese economy. That was definitely a pioneering work (and arguably the oldest digital corpus of classical Chinese[28]). The texts were entered manually (OCR was not available at this time), and went through multi-pass verification process. Soon, *Shisanjing* was added to the 25 histories. These texts became the core of the future electronic database. Creation of digital corpora was enabled by advancement in computer science and electronics: the BIG5 coding was introduced in 1984, and computers became available to institutions. At this time, BIG5 contained not so many characters (13,051[29]), so there should have been substitutions for missing characters (Juan–2005). The database had continued to grow, and it was eventually taken to the web (in 1997, Liu, “Impact of Digital Archives”, 4), where it became known as Scripta Sinica corpus[30]. This resource does not provide information on lengths of specific texts, and word mark-up.

It should be noted that most online academic corpora do not provide classics’ lengths (with exclusion of CHANT). The reason could be in text variants and emendations. There should be a strategy to select one version for calculation, and it is not an easy decision from philological point of view.

Academia Sinica Corpus. Shortly after the beginning of Scripta Sinica project, the Computing Center of the Academia Sinica (the Institute of Information Sciences, IIS) also decided to create their own electronic database of Classical Chinese, as a part of their bigger corpus of Chinese[31]. The group managed to receive as an intra-academia transfer the core of IHP database (1.5 million characters) and then entered themselves another 1.5

million characters, also manually. This corpus later became the database of Academia Sinica. It is not clear if texts added by this group were modified in the digitization process. The IIS was probably the first group which provided an estimate for the whole scope of pre-Qin corpora as 3 million characters[32]. This resource also does not provide information on lengths of specific texts.

CHinese ANcient Texts (C.H.A.N.T.). About the same time as the Academia Sinica project, a Hong Kong research group started creating their own electronic database of classical Chinese texts (at this time, researchers regularly used term "database" for what later became "corpus"). The initial goal of the project was continuation of Harvard-Yenching concordance project, under senior editors Professor D.C. Lau and Dr. F.C. Chen of the Institute of Chinese Studies at the Chinese University of Hong Kong (McLeod, *ibid*, 48). Eventually, this corpus also was put online and became CHANT database. The source was also modified (improved) during digitization. The data was entered manually. The first implementation of this electronic database was a (pre-web) series of printed concordances – it is the first time concordances to classical text were based on their electronic versions[33]. CHANT project provides information on text lengths and number of type-tokens.

Beijing University Corpus (PKU). This is the only well-known academic project on classical Chinese that has been developed in the mainland China. The project started in Beijing at the beginning of 2000s, and got abbreviation of "PKU" from the "Peking University" spelling[34]. It is not clear, if compromises were made when coding pages contained a limited set of characters were reworked, when UTF and more advanced Big5 and GB coding became available. PKU provides information on text lengths; however, it reports data on file lengths in kilobytes, not in characters[35]. Therefore, it was not possible to use this information in this article.

Thesaurus Linguae Sericae (TLS). Although TLS is claiming to be a "dictionary", or "interactive database", or "an historical and comparative encyclopedia of Chinese Conceptual Schemes", it is in reality an important digital corpus of classical Chinese texts, which is freely available, and is the only large academic collection of classical texts online, created outside China. Its development is unusual, because the input work is distributed among dedicated specialists, who curated their texts[36], and data entry is in Unicode[37]. It is not very large (e.g., some texts in *Shisanjing* are missing), but contains many important texts. This resource also does not provide information on lengths of specific texts.

2.2.2 Independent corpora.

Chinese Text Project (CText). This online corpora collection seems to be an individual enterprise of Donald Sturgeon (Sturgeon, "Zhuangzi"), who created it practically single-handedly, working out of Hong Kong. According to the site, text entry is based on OCR digital versions of old printed sources (that solves copyrights issue). Presumably, there are not many work resources, and accuracy of online texts may not be very high. However, it is a free resource, with a community formed around it, which constantly improves quality of texts (but not Wikimedia style, i.e., correctors cannot fix errors themselves[38]). According to personal observations of the present author, Ctext is the most popular source of informal references to classical Chinese texts among Western researchers[39]. This resource does not provide information on lengths of specific texts.

WSW Chinese Texts (WSW CTexts) is a research corpus, with a focus on *Shisanjing*, and it currently does not feature many texts. However, it provides the most extended set of tools for text research that is currently available online. The source of texts is Wikisource (see the resource for specific links)[40]. This resource provides information on lengths of specific texts and their vocabularies.

2.2.3 Commercial corpora. From personal observation of the author of this paper, despite all academic databases are still online, and some new texts are being added to them, it seems that their interface has not changed much from the beginning of 2000s. Meanwhile, starting from the mid–2000s considerable progress in development of online digital corpora of classical Chinese has been made by commercial companies. As early as in the 90s, there were several commercial projects, selling digital versions of classical texts, e.g., *Sikuquanshu*, but they could not compete with academic online corpora.

Since the mid–2000s, however, it seems that commercial projects take the lead[41]. There are two leading commercial projects in the area of classical Chinese: Unicode Inc., which produced two online databases ("Unihan" and Wenyuan Ge Siku quanshu) and "Erudition database", as well as "Hytong". Unihan presumably has good accuracy, as well as using Unicode coding from the beginning, despite using OCR technology. It is interesting that Erudition, which claimed reaching the level of precision of printed texts also started from OCR approach, but switched to manual entry – it probably means that even modern OCR precision was not satisfactorily[42]. It should also be noted that commercial companies tend to produce full-text search systems, not online corpora.

While it is possible that currently these commercial corpora are the most advanced source of classical Chinese corpora, it does not seem that Western research community is using these tools more than academic or independent ones[43]. It seems that the future could belong to independent or free corpora, as it is difficult to imagine that international academic community will be using a pay-walled resource, which is not available to everyone, for presentation of results.

3. Text lengths as indicator of variety

The commercial digital corpora probably reached a very high degree of accuracy but scholars still check their classic quotations by printed versions of texts. However, it is very unlikely that modern researchers will be calculating text lengths using a printed text. It was not possible, actually, to find any article, reporting text lengths, based on any of modern printed editions[44]. Moreover, as all computational linguistics experiments are going to be run on digital corpora, such basic characteristics as text length and type lists have to be calculated using these corpora, not printed versions.

The first half of this article investigated, how these corpora are built, and what problems should be expected, in comparison with printed text versions. This part of the article will present the lengths of *Shisanjing* texts, calculated on the basis of various digital corpora. To display variation of lengths, not only data of electronic resources, but all available data will be included, to delineate scope of the problem.

3.1 Text lengths in this study. This study started as collecting basic information on quantitative characteristics of *Shisanjing* texts, primarily, the length in characters of WSW Ctexts classics. However, the first attempt to compare results of this study with results of other studies, and, first of all, with data available on classical text lengths led the author to disappointing results, due to reasons, described above.

Eventually, the study concentrated on another question, which became its main subject, “what information is available on the lengths of *Shisanjing* texts and how WSW Ctexts data relates to it”? It is well-known that there are many versions of classics, and they often differ considerably. Text lengths vary for this and other reasons: they are affected by inclusion into the count such “secondary” text components, as text title, chapter titles[45], as well as punctuation and non-character symbols. These issues were approached differently by researchers; however, not all of them reported on what approach they used.

Currently, there is no available comparison of Shisanjing length data, what should be expected, and how it affects quantitative linguistics studies. This paper will try to fill this gap, as well as delineate the data framework and bring up the numbers for further evaluation.

3.2 Length measuring history. It is possible to identify a few periods in quantification of classical texts. Feng, “Evolution and Present Situation” divides the 20th century into three periods (from 20s to before 1979, 1979 to 1991, and modern), starting with first frequency lists[46], moving to 80s, when first digital versions became available[47], at that time, for frequency studies, and finishes with the modern period[48]. Similar description could be found in introduction to the paper of Zhang et al. (Zhang et al., *ibid*).

Expanding the time frame, it is possible to identify four chronological periods: traditional period, modern period, early digital period and mature digital period. In a way, all time from the beginning of literacy to appearance of first concordances could be called the "traditional" period. The modern period is the period of paper-based concordances. Early digital period began with digitalization of texts and computer analysis of electronic versions of text. Finally, the current period, in addition to electronic texts is defined by online concordances, and especially by wikifying of online editions.

3.2.1 Traditional period. Chinese bibliographical descriptions (especially, those in dynasty histories), starting from *Han shu*, and ending with *Siku quanshu*, describe book size by the number of *pian, juan and ce*[49]. The length in characters is not present in bibliographical descriptions, probably, because bibliographers perceived manuscripts as "books" or "works", not as abstract "texts".

It does not mean that the Chinese scholars did not try to calculate manuscripts' length in characters. Recent discoveries have shown that a Qin or Han scribe (or another person) could indicate text length in characters on the book cover[50]. However, these numbers were not entered into bibliographical descriptions (even if they were present on a copy used by a bibliographer, even though they could be more important for "version control" than "chapters"). None the less, these numbers were often known by scholars, but ignored by bibliographers[51].

Creation of "stone classics" (*shijing* 石經) played an important role in history of calculation of text length in characters[52]. Winkelman notes that "stone canons" functioned as publicly available authorative texts (Winkelman, *ibid*, 32), and, despite government moved later to wood-block-printed versions as standard texts of canons, most of known traditional records of numbers are based on calculations on these stone canons, not on manuscripts or wood-block-printed books. The character numbers, scribbled on manuscript covers, disappeared.

Song data. The earliest consistent measurements of length in characters of several classical texts from *Shisanjing* located by the author of this article are dated by the Song period[53]. It seems that recently created system of thirteen classics (twelve of which were displayed on *Kaicheng* stone classics, 833–837) and their role in state examinations prompted Song scholars to estimate time necessary to memorize the classics (e.g., memorizing by 300 characters a day). "Chayu kehua" (茶餘客話) compiled by Ruan Kuisheng (阮葵生)[54] contains a chapter ("Jiu jing zi shu" [Numbers of Characters of Nine Classics] (CYKN, 264), on lengths in characters of thirteen classics, quoting numbers, some of them presumably calculated by Song's time Zheng Genglao (鄭耕老) in his "Quan xue" (勸學)[55] (this and following data are provided in tables in the Appendix I)[56].

Qing data. More data is available from the Qing period (see Appendix I for numbers). This data seems to be of interest to their compilers as a property of texts, not for pedagogical reasons. The first data set is provided by Zhu Yizun (朱彝尊) (1629–1709)[57] in treatise "Jing Yi Kao" (經義考)[58]. Qian Taiji (錢泰吉) (1791–1863) in his treatise *Pushu zaji* (曝書雜記)[59] quotes numbers, calculated by Zheng Genglao, as well by Wu Yingdian (武英殿) who was using Qianlong stone classics.

3.2.2 In the 20–30s of 20th century, Chinese philology started creating Western-style concordances of classical texts. On early stages of this process, concordances were created manually and creators of concordances still did not perceive texts from the point of view of their length in characters (or words)[60]. These concordances did not feature text lengths in characters, as well as frequency lists. Moreover, there are no reports about number of characters of classical texts in printed editions. Meanwhile, the paper concordances were an ultimate answer to most questions that a classic philologist would like to ask.

3.2.3 In the 80–90s, early digital period, the situation has improved. The ICS concordances, based on electronic database, were published, where number of characters in text, and volume of vocabulary were indicated – probably, the first time since early traditional calculations[61]. These texts became foundation of later online concordances, such as CHANT or Scripta Sinica.

In the modern period, with ubiquitous internet presence, many independent online electronic editions and concordances started to appear, alongside older online concordances, which are extensions of earlier electronic databases. The most notable are Wikimedia and Ctext resource. As rule, they do not use existing electronic media[62], but re-scan similar or same editions. This means that there could be more mistakes in these concordances. In Wikimedia (and to some degree in Ctexts), texts are open to corrections. This means that they are improving with time, but also that their vocabulary is not permanent. However, to a lesser degree, same is applicable to "official" online versions and concordances. The list of these concordances is provided in Appendix II.

Finally, in mid–2000s, a slowdown could be observed in development of non-commercial electronic (and online) databases of classical Chinese texts. It coincidences with and is corroborated by the fact that most reviews of available systems are dated not later than 2010 (mostly mid–2000s). At the same time, commercial digital resources sprout up, suggesting that there are paying customers for their services. It is possible that private enterprises will intercept academic activity in this area, as roll-out of Erudition databases could signal. At the same time many emerging academic groups in China build their own classic Chinese corpora, instead of re-using existing resources, due to copyright complications. Neither type of resources is available for other researchers.

3.3 Modern research corpora. As development of academic corpora, and research activity in respective centers, starts to slowdown in mid–2000s, the activity in quantitative linguistics in classical Chinese is moving into smaller research groups, who, despite themselves being academic, cannot re-use existing academic corpora for their experiments, and build their own corpora for their experiments. As one of these researchers summarized recently, "there is no free database which can be used to get the statistical data of the Pre-Qin Chinese" (Li et al., "Corpus-Based Statistics", 145)[63]. Below described four of such efforts (in chronological order), and their data is reflected in tables of Appendix I.

Guo Xiaowu calculates highest frequency characters in classics, looking for the most frequent characters in classical Chinese. He provides data on length of texts with and without punctuation, as well as number of character (Guo, "Gudai Hanyu", 73, fig.2–2). Guo claims that his corpora are a selection of existing data[64], with no exact indication which text came from what source, and how they were processed.

Qin Qin claims that instead of using online resources, their group created their own corpus, based on Song engraved editions, digitized it (it is not reported, through manual data entry or OCR) and went through a few checks (Qin, "Xianqin guji", 112). The researchers encountered typicalissues: unencoded characters, etc., and they claim that it was resolved through some manual statistical approach (Qin, *ibid*).

Li Xiang also claims that created his own corpus for his dissertation, based on SSJZS (Li, "Shisanjing jigao"), with titles removed.

Li Bin's group (who was quoted above complaining on unavailability of academic resources for research), does not disclose data on their own corpus, and how it was built, but they demonstrate good coverage of classics. Also, it is claimed to feature multiple-character word and part-of-speech mark-ups (Li et al., "Corpus-Based Statistics").

All in all, recent experimental efforts in quantitative linguistics of classical Chinese are based, at least, officially, not on available online corpora (due to licensing issues and not encouraging interface), but on in-house corpora, whose accuracy and versions are unknown. These corpora are not available for other researchers for reproducing experiments.

3.4 Data sources set-up. This article will use eleven sources of classics text length data[65]. They are listed in the chart below. The first three sources belong to the traditional period, and their authors presumably retrieved their data from "stone canons". These texts are not punctuated, and sometimes chapter titles were not included into the account. This approach is very close to the approach that was used by the author of this article for the

WSW Ctexts data. However, their text versions could sometimes be different from versions that have been used for modern digital corpora.

There is a gap in available data between XIX century and 1990s, because no researcher reported data on printed books[66]. Even printed concordances, created in 20s–30s and later, do not feature text length and vocabulary data. When in the 80s Wang Genbao (Wang, "Shisanjing jing") reported text lengths of classics for SSJZS edition, he referred to traditional sources (Qian Taiji, PST). The next available data is ICS, reporting numbers for printed versions of digital texts, i.e., from electronic sources[67]. Other digital online corpora, such as Scripta Sinica, do not provide this information.

Finally, the beginning of this century provides most data from research corpora: Gou Xiaowu, Qin Qin, Li Xiang, Li Bin et al., and WSW Ctexts. If commercial corpora contain this data, it remained unavailable for this article.

#	Source	Date
1	Zheng	XII CE
2	Zhu	XVIII–XIX CE
3	Qian	XIX CE
4	ICS	1990s
5	CHANT	1990s
6	GUO	2001
7	QIN	2005
8	GUOXUE	2005
9	LI_2009	2009
10	LI_2013	2013
11	Ctexts	2008

Table 1. Data Sources in Chronological Order

Text	Han stone classics (Xiping 175–183) (Zhang, *ibid,* 1:1a,b)	Wei stone classics (Zhengshi 241) (Zhang, *ibid*, 2:1a,b)	Tang (Kaicheng) stone classics (Zhang, *ibid*, 3:1a,b)	Houshu 后蜀 (951–958) (Zhang, *ibid*, 4:1a,b)
Chunqiu	16572	16572	n/a	n/a
Chunqiu &Zuozhuan	n/a	-	198945	197265
Gongyang	27583	n/a	44748	44738
Guliang	n/a	n/a	42085	41890
Liji	n/a	n/a	98994	98545

Lunyu	15710	n/a	16509	15913
Mengzi	n/a	n/a	n/a	n/a
Shi	40848	n/a	40848	41021
Shu	18650	18650	27134	26286
Xiaojing	n/a	n/a	2003	1798
Yili	57111	n/a	57111	52802
Zhouli	n/a	n/a	49516	50508
Zhouyi	24437	n/a	24427	24052

Table 2. Zhang Guogan’s Data on Shijing Texts

3.5 Results discussion. The numbers of text lengths of classics in characters, as well as vocabulary volume, for single characters, are presented in the Appendix I. As the data volume is too small, and sources vary considerably, to apply a proper statistical approach to it would be excessive. However, to create some numeric framework, averages and standard deviation were calculated where available. Mostly, length variation reflects versions of texts, the editions creators choice, but there are many other factors, affecting these characteristics (e.g., counting in commentaries, punctuation or titles of chapters, etc.). The Appendix I contains comparison tables of text lengths (and vocabulary size, where available) in characters for Shisanjing and Zhuang-zi, with a short discussion of these changes.

The range of numbers’ variation varies[68], sometimes considerably, sometimes little, but it is clear that quantitative linguistics characteristics, as well as philological information, obtained from these corpora, will be different. Despite Chunqiu and Zuozhuan are different texts, divided by a large time gap, they are mostly treated as one text by most researchers. This makes separate quantitative studies of them difficult. Of all observed sources, only WSW Ctexts provides separate numbers.

Surprisingly, most text lengths of classics are falling within standard deviation range. WSW Ctexts, which does not include repeated titles and punctuation, usually features the minimal number. Gongyang and Guliang also demonstrate good clustering. Meng-zi and Liji demonstrate closeness. But the lest variative text is Lunyu.

Shujing and Shijing demonstrate more variation. Some versions had to be excluded from population, because they were too deviating from other texts, e.g. Shijing in version of Li_2009, and Shujing in versions of Guo_2001 and Qin_2005 (all new sources). Earlier versions of Xiajing also had to be excluded, but otherwise, it is very consistent text.

Variation demonstrates importance of availability of source texts for all researchers. However, not all of referenced sources provide this option, or it is not easy to obtain text (e.g., it has to be downloaded by paragraphs).

That is why for WSW Ctexts site, the Wikimedia source was chosen. Any researchers can copy the text and use it in experiments. However, the Wikimedia texts always could be questioned for its reliability, and downloaded texts should be updated from time to time, to be in synch with online version.

4. Conclusions

The study assumes that for quantitative linguistic analysis, Chinese characters, not words, are still valid quantification units. Therefore text length in characters and the number of type-token characters are critical values for any quantitative linguistics study. However, conversion of Chinese texts into electronic form leads to ubiquitous errors and inaccuracies, which, alongside with modifications by researchers, creates a *digital content gap* between paper-based and electronic-based corpora.

The analysis of available lengths of modern electronic corpora of Shisanjing shows that there are considerable discrepancies. The scope of variation depends on text. Some texts, like Lunyu and Xiaojing, show very little variation, while others, like Shujing and Shijing, display more variance. Sometimes, there is a historic tradition to include commentaries (e.g., Yijing, and especially, Chunqiu), so numbers for just canon part and "text" as it is perceived in philological tradition could be very different.

A typical philological question, e.g., "how many times the character X is found in the text Y" will get different answers for some characters, not only if one compares corpora of Li et al. with ICS or CHANT, but even for some texts in ICS and CHANT. Eliminating all errors and discrepancies for large corpora (however much effort applied to it) is very difficult; therefore, any results from electronic corpora will carry some inaccuracy (however small it could be). Some characters, present in paper-form text could be missing in an electronic resource.

Although digitization introduces some error and inaccuracy, even more discrepancy is brought in by differences in text versions and changes during transformation process. Some databases, created on the early digitization process stages, were modified by creators (e.g., in CHANT and Corpus Sinica projects), so they should not have direct counterparts in paper versions[69].

Despite the digital gap, any massive quantitative linguistics studies are only possible by using electronic corpora. Even though the size of pre-Qin corpora is limited by a few millions characters, it is rather problematic to return to paper concordances. The quality of the electronic version of the text source plays critical role in research accuracy. Electronic sources for reliable online corpora should be open to academic community and, preferably, created by academic community.

The ideal situation would be having a standard and free digital canon. Definitely, any specific version of a canonical text could be criticized from various textologic aspects. Therefore, if such free standard version comes to existence, it should be very well supported by philological analysis and discussion. One good prototype for this approach could be TLS, if it develops further and provides free downloads for entire texts.

However, since the mid–2000s, the opposite trend to commercialization of digital resources has been observed. It is possible that commercial databases, like Erudite database, provide more accurate digital corpora, but it does not seem that these databases are going to be available for independent examination and experiment any soon. As a result, research groups start building their own digital resources, and it leads to fragmentation of the field and creation of many digital resources that differ from each other. Most of available printed resources have been already digitized, but high-quality resources are mostly commercialized.

This article tried to show that results of quantitative linguistic study heavily depend on digital text version. Creating a digital resource of classical Chinese texts that is open-sourced and available to entire research community[70] will provide proper level of reliability and repeatability.

Acknowledgements

The author would like to thank E. Bruce Brooks for constant encouragement, inspiration, and support of WSW Ctexts project from its very beginning. Lively discussions with Bruce and Chris Beckwith greatly accelerated the development of the project. Another inspiration, on computer science side, was attending seminars of Gerald Penn, who attracted author's attention to the problem of relationship of Chinese characters and words, and multiple aspects of corpus building.

The project development placed the author in contact with many people whose ideas and assistance had a great influence, especially, the ongoing dialogue with Rodo Pfister. Wolfgang Behr and Matthias Richter provided valuable information and unpublished materials on manuscripts and character calculations.

Keeping the project alive on the web has been an unending work, and the author is grateful to University of Massachusetts's Yvette Mushenski and Mike Barnard, who were helpful in overcoming many a system hurdle, allowing the project to prosper on Amherst's servers. The author also would like to thank the staff of the Cheng Yu Tung East Asian Library of the University of Toronto.

References

Online resources (corpora)

1) CHANT (CHinese ANcient Texts) database (http://www.chant.org/)

2) CTEXT Chinese Texts Database http://ctext.org/

3) ERUDION database http://server.wenzibase.com/

4) GUOXUE (Guoxue baodian) http://www.gxbd.com/

5) HYTONG (Hytung Ancient Book Database) http://www.hytung.cn/Default.aspx

6) PKU (Peking University Corpus of Ancient Chinese) http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=gudai

7) SCRIPTA SINICA Scripta Sinica database http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm

8) SHEFFIELD (Sheffield Corpus of Chinese) http://www.hrionline.ac.uk/scc/

9) SINICA (Academia Sinica Tagged Corpus of Old Chinese) http://old_chinese.ling.sinica.edu.tw (http://app.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh)

10) TLS Thesaurus Linguae Sericae tls.uni-hd.de

11) WSW CTEXTS http://www.umass.edu/ctexts/index.php

Abbreviations of classical sources

CYKH: Ruan Kuisheng 阮葵生. Cha yu ke hua 茶餘客話 [Dialogues over a Cup of Tea] 商務印書館, Taibei Shi: Shang wu yin shu guan, Minguo 65 [1976].

JIK: Zhu Yizun 朱彝尊. Jingyi kao 經義考 (General Bibliography of the Classics[71]). (Sibu beiyao edition, 156.1a–8b, 157.1a–10b) 臺灣中華書局, Taibei Shi: Taiwan Zhonghua shu ju, Minguo 54 [1965].

PST Qian Taiji 錢泰吉. Pavilion for Airing My Books (Pushu Ting quan ji) 曝書亭全集. 台灣中華書局, Taibei Shi: Taiwan zhonghua shu ju, Minguo 54 [1965] Sibu beiyao, ji bu.

SSJZS Ruan Yuan 阮元. Shi san jing zhu shu: fu jiao kan ji 十三经注疏 (附校勘记) 中华书局: 新华书店北京发行所发行, Beijing: Zhonghua shuju : Xin hua shu dian Beijing fa xing suo fa xing, 1980.

SKPX Hongfu Zeng 曾宏父. Shi ke pu xu 石刻鋪叙, Taipei]: Taiwan shang wu yin shu guan, 1983.

Concordances

ICS: The Ancient Chinese Texts Concordance Series (Hsien-Ch'in Liang-Han ku-chi chu-tzu so-yin ts'ung-k'an 先秦兩漢古籍逐字索引叢刊), edited by D.C. Lau Lau Din Cheuk; Liu Tien-chueh 劉殿爵, Ho Che Wah 何志華 and Chen Fong Ching 陳方正, The Chinese University of Hong Kong, Institute of Chinese Studies,. (Hong Kong: Commercial Press, 1992–)

ICS LUNYU A Concordance to the Lunyu (論語逐字索引), 1995

ICS MENGZI A Concordance to the Mengzi (孟子逐字索引), 1995

ICS ZHUANGZI A Concordance to the Zhuangzi (莊子逐字索引), 2000

ICS YILI A Concordance to the Yili (儀禮逐字索引), 1994

ICS LIJI A Concordance to the Liji (禮記逐字索引), 1992

ICS ZUOZHUAN A concordance to the Chunqiu zuozhuan (春秋左傳逐字索引), 1995

ICS GONGYANG A concordance to the Gongyangzhuan (公羊傳逐字索引), 1995

ICS GULIANG A concordance to the Guliangzhuan (穀梁傳逐字索引), 1995

ICS SHIJING Aconcordance to the Maoshi (毛詩逐字索引), 1995

ICS SHUJING A Concordance to the Shangshu (尚書逐字索引), 1995

ICS YIJING A concordance to the Zhouyi (周易逐字索引), 1995

ICS ZHOULI A Concordance to the Zhouli (周禮逐字索引), 1993

ICS XIAOJING Erya, Xiaojing《爾雅、孝經逐字索引》, 1995

H-Y: The Harvard-Yenching Institute Sinological Index Series [Ha-fo Yen-ching hsueh-she yin-te 哈佛燕京學社引得] (Pei-p'ing: 1931–1947; rpt. Taipei: China Materials Center, 1965–69)

APPENDIX 1. Tables for single texts

1. Chunqiu data

The data on length and vocabulary size of Chunqiu is only provided by WSW Ctexts, other sources usually combine it with Zuozhuan. The PKU offers its number of bytes in a separate Chunqiu text file (78626 bytes), but it is not clear how many characters there are (and it most probably includes punctuation, spaces, etc.). Zhang Guogan provides calculated data for Han and Wei stone classics- 16572 characters (Zhang, *ibid*, 1:1a,b; 2:1a,b).

#	Source	N	V	Comments
1	Zheng	n/a	n/a	
2	Zhu	n/a	n/a	
3	Qian	n/a	n/a	
4	ICS	n/a	n/a	
5	CHANT	n/a	n/a	
6	GUO_2001	n/a	n/a	
7	QIN_2005	n/a	n/a	
8	Guoxhue	n/a	n/a	
9	LI_2009	n/a	n/a	
10	LI_2013	n/a	n/a	
11	Ctexts	16791	941	The source here and below is taken from Wikisource, with chapter titles and punctuation removed.

2. Zuozhuan

Similar to Chunqiu, only WSW Ctexts and LI_2013 treat it as a separate text. The PKU offers the number of 495584 bytes. Difference between WSW Ctext and LI_2013 is less than 1000 characters, and could be explained by possible presence of chapter titles in LI_2013.

#	Source	N	V	
	Zhang	n/a	n/a	
1	Zheng	n/a	n/a	
2	Zhu	n/a	n/a	
3	Qian	n/a	n/a	
4	ICS	n/a	n/a	
5	CHANT	n/a	n/a	
6	GUO_2001	n/a	n/a	
7	QIN_2005	n/a	n/a	
8	LI_2009	n/a	n/a	
9	GUOXUE	n/a	n/a	
10	LI_2013	179814	3312	Li et al., *ibid*, 146
11	Ctexts	178563	3235	

3. Chunqiu and Zuozhuan combined

Only Li_2013 does not provide numbers for this combination of texts. Starting from this, averages for text length and vocabulary size, as well as the standard deviation, are offered. The lengths of versions of Qian, CHANT and Ctexts are beyond standard deviation. There is a difference between ICS and CHANT numbers, probably, reflecting changes made over years of editing. Zhang Guogan cites 198945 and 197265 characters for Tang stone classics and Shu stone classics, respectively, including Chunqiu (Zhang, ibid, 3:1a,b; 4:1a,b).

#	Source	N	V	Comment/Reference
1	Zheng	196845	n/a	Qian, *ibid.*, 1:1:2–4; Ruan (Ruan, *ibid*, 64) indicates two numbers: 201350 and 196845
2	Zhu	197265	n/a	Zhu, *ibid*, 289, 3–4
3	Qian	198945	n/a	Qian, *ibid.*, 1:1:2–4; Wang, *ibid*, provides the number of Zheng: 196845
4	ICS	195792	3290	ICS, *Zuozhuan*, 2205
5	CHANT	198699	3320	CHANT website
6	GUO_2001	196043	3238	Guo, “Gudai hanyu”, 73
7	QIN_2005	195879	3257	Qin, “Zianqin guji”, 113

8	LI_2009	195792	n/a	Li, “Shisanjin jigao”, 11; here and further is not the number in the original paper, but re-calculated by the author of the present paper[72]
9	GUOXUE	197294		Yin, “Guji shuzihua”
10	LI_2013	n/a	n/a	Li et al., ibid, 146
11	Ctexts	195354	3251	
AVG		196791	3271	
dev		1257	33	

4. Gongyang

Almost every data source provides numbers on Gongyang (except GUO_2001), and they are very close. Only early Zheng version, and Guoxue lie outside of standard deviation. Numbers for text length vary for ICS and CHANT versions (unlike vocabulary). Standard deviation is not shown for vocabularies, as they are very close. Zhang Guogan provides numbers 27583 for Han stone classics (Zhang, *ibid*, 1:1a,b), and 44748 and 44738 for Kaicheng and Shu stone classics (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	n/a; 44015 Ruan	n/a	Ruan, *ibid*
2	Zhu	44738	n/a	Zhu, *ibid*, 289, 3–4 and Yin, “Guji shuzihua”
3	Qian	44748	n/a	Qian, *ibid.*, 1:1:2–4;
4	ICS	44379	1648	ICS, *Gongyang*, 551
5	CHANT	44521	1648	CHANT website
6	GUO_2001	n/a	n/a	Guo, “Gudai hanyu”, 73
7	QIN_2005	44338	1645	Qin, “Zianqin guji”, 113
8	LI_2009	44841	n/a	Li, “Shisanjin jigao”, 9
9	GUOXUE	44922		Yin, “Guji shuzihua”
10	LI_2013	44366	1642	Li et al., ibid, 146
11	Ctexts	44224	1640	
avg		44509	1645	
dev		295		

5. Guliang

It should be noted that standard deviation is larger for Guliang, and more values are beyond it, i.e., the Guliang population of lengths is no so close as Gongyang. However, vocabulary sizes are close. Zhang Guogan provides data for Guliang for Tang and Shu stone classics, 42085 and 41890 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	n/a; 41512 ruan		Qian, *ibid*., 1:1:2–4
2	Zhu	41890	n/a	Yin, "Guji shuzihua"; Zhu, *ibid*, 289, 3–4
3	Qian	42089	n/a	Qian, *ibid*., 1:1:2–4
4	ICS	40914	1604	ICS, *Guliang*, 517
5	CHANT	42056	1604	CHANT website
6	GUO_2001	n/a	n/a	Guo, "Gudai hanyu", 73
7	QIN_2005	40828	1590	Qin, "Zianqin guji", 113
8	LI_2009	41484	n/a	Li, "Shisanjin jigao", 10
9	GUOXUE	42242		Yin, "Guji shuzihua"
10	LI_2013	40913	1593	Li et al., ibid, 146
11	Ctexts	40835	1594	
Avg		41476	1597	
dev		571		

6. <u>Liji</u>

Liji is also one of the closest populations, with unexpectedly high Song period numbers. Zhang Guogan provides data for Liji for Tang and Shu stone classics, 42085 and 41890 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
	Zhang	57111		
1	Zheng	99020		Qian, *ibid*., 1:1:2–4; Ouyang Gong provides the number: 99010
2	Zhu	98545	n/a	Yin, "Guji shuzihua"; Zhu, *ibid*, 289, 3–4
3	Qian	98994	n/a	Qian, *ibid*., 1:1:2–4;
4	ICS	97973	3028	ICS, *Liji*, 943
5	CHANT	98123	3037	CHANT website
6	GUO_2001	98202	2973	Guo, "Gudai hanyu", 73
7	QIN_2005	98081	3016	Qin, "Zianqin guji", 113
8	LI_2009	98250	n/a	Li, "Shisanjin jigao", 13
9	GUOXUE	97985		Yin, "Guji shuzihua"
10	LI_2013	97994	2999	Li et al., ibid, 146
11	Ctexts	97994	3014	
Avg		98287		
dev		393	23	

7. <u>Lunyu</u>

Lunyu numbers is the only population with no numbers beyond standard deviation. They are practically identical, with exception of Song and

Qin. Zhang Guogan provides data for Lunyu for Han stone classics 15710 (Zhang, *ibid*, 1:1a,b), and for Tang and Shu stone classics, 16509 and 15913 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	12700		Qian, *ibid*., 1:1:2–4
2	Zhu	15913	n/a	Zhu, *ibid*, 289, 3–4
3	Qian	16509	n/a	Qian, *ibid*., 1:1:2–4;
4	ICS	15935	1355	ICS, *Lunyu*, 197
5	CHANT	15935	1355	CHANT website
6	GUO_2001	15962	1345	Guo, "Gudai hanyu", 73
7	QIN_2005	15920	1351	Qin, "Zianqin guji", 113
8	LI_2009	16013	n/a	Li, "Shisanjin jigao", 8
9	GUOXUE	15917		Yin, "Guji shuzihua"
10	LI_2013	15935	1349	Li et al., ibid, 147
11	Ctexts	15923	1361	
		15697	1353	
		1009	6	

According to Huang Kan, Song's scholar Ouyang Gong gives the number 11705. Huang also reports that Zheng Gengla's number could be 13700, as a version. See Huang_2006.

8. Mengzi

Again, numbers are pretty close, except Song's ones. However, vocabulary numbers show more deviation.

#	Source	N	V	
1	Zheng	34685		Qian, *ibid*., 1:1:2–4;
2	Zhu	n/a	n/a	Zhu, *ibid*, 289, 3–4
3	Qian	34685	n/a	Qian, *ibid*., 1:1:2–4;
4	ICS	35417	1913	ICS, *Mengzi*, 373
5	CHANT	35417	1912	CHANT website
6	GUO_2001	35289	1876	Guo, "Gudai hanyu", 73
7	QIN_2005	35258	1886	Qin, "Zianqin guji", 113
8	LI_2009	35454	n/a	Li, "Shisanjin jigao", 14;
9	GUOXUE	35385		Yin, "Guji shuzihua"
10	LI_2013	35389	1897	Li et al., ibid, 146
11	Ctexts	35354	1892	
avg		35233	1896	
dev		295	15	

9. Shijing

Shijing numbers depend on whether Preface and other comments are included. WSW Ctexts, which does not include punctuation and song titles, features the minimum number LI_2009 seems to be a huge deviation, and it was excluded from the population. Zhang Guogan provides data for Shijing for Han stone classics 40848 (Zhang, *ibid*, 1:1a,b), and for Tang and Shu stone classics, 40848 and 41021 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	39124		Qian, *ibid.*, 1:1:2–4; Wang, *ibid*, 39224 Ouyang Gong: 39234
2	Zhu	41021	n/a	Zhu, *ibid*, 289, 3–4
3	Qian	40848	n/a	Qian, *ibid.*, 1:1:2–4;
4	ICS	37438	2989	ICS, *Maoshi*, 467
5	CHANT	41077	2993	CHANT website
6	GUO_2001	30798	2810	Guo, “Gudai hanyu”, 73
7	QIN_2005	29752	2837	Qin, “Zianqin guji”, 113
8	LI_2009	55102	n/a	Li, “Shisanjin jigao”, 17;
9	GUOXUE	30387		Yin, “Guji shuzihua”
10	LI_2013	30954	2806	Li et al., ibid, 146
11	Ctexts	29622	2833	
avg		35102	2878	
dev		5185	88	

10. Shujing

GUO–2001 and QIN–2005 probably used shorter versions, and therefore were excluded from population. Zhang Guogan provides data for Shujing for Han stone classics 18650 (Zhang, *ibid*, 1:1a,b), for Wei stone classics 18650 (Zhang, *ibid*, 2:1a,b), and for Tang and Shu stone classics, 27134and 26286 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
	Zhang	18650		Qian, *ibid.*, 1:1:2–4;
1	Zheng	25700		Wang, *ibid.*, 25800 Qian, *ibid.*, 1:1:2–4: 25800
2	Zhu	26286	n/a	YIN_2007, Zhu, 289, 3–4
3	Qian	27134	n/a	Qian, *ibid.*, 1:1:2–4;
4	ICS	28073	2026	ICS, *Shujing*, 307
5	CHANT	28153	2025	CHANT website
6	GUO_2001	16357	1597	Guo, “Gudai hanyu”, 73
7	QIN_2005	17062	1623	Qin, “Zianqin guji”, 113

8	LI_2009	24657	n/a	Li, “Shisanjin jigao”, 15;
9	GUOXUE	25700		Yin, “Guji shuzihua”
10	LI_2013	28146	1995	Li et al., ibid, 146
11	Ctexts	24539	1911	
		26488	1989	
		1453	54	

11. Xiaojing

No deviations, except in Song and Qing versions, which were excluded from the population. Zhang Guogan provides data for Xiaojing for Tang and Shu stone classics, 2003 and 1798 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	1903		Qian, *ibid*., 1:1:2–4; Ouyang Gong: 1903
2	Zhu	1798	n/a	Yin, “Guji shuzihua” Zhu, *ibid*, 289, 3–4
3	Qian	2113	n/a	Qian, *ibid*., 1:1:2–4;
4	ICS	1800	373	ICS, *Xiaojing*, 27
5	CHANT	1800	373	CHANT website
6	GUO_2001	n/a	n/a	Guo, “Gudai hanyu”, 73
7	QIN_2005	n/a	n/a	Qin, “Zianqin guji”, 113
8	LI_2009	1906	n/a	Li, “Shisanjin jigao”, 18
9	GUOXUE	1903		Yin, “Guji shuzihua”
10	LI_2013	1801	373	Li et al., ibid, 146
11	Ctexts	1800	374	
		1845	373	
		55	1	

12. Yili

LI_2013 is exceedingly high, and excluded from the population. Otherwise, only Qian version of Song period length lies beyond deviation. Zhang Guogan provides data for Yili for Han stone classics 57111 (Zhang, *ibid*, 1:1a,b) and for Tang and Shu stone classics, 57111 and 52802 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	n/a		Qian, *ibid*., 1:1:2–4” 56624
2	Zhu	52802	n/a	Yin, “Guji shuzihua”; Zhu, 289, 3–4
3	Qian	57111	n/a	Qian, *ibid*., 1:1:2–4;
4	ICS	56809	1529	ICS, *Yili*, 467
5	CHANT	56809	1529	CHANT website
6	GUO_2001	n/a	n/a	Guo, “Gudai hanyu”, 73
7	QIN_2005	56758	1522	Qin, “Zianqin guji”, 113

8	LI_2009	53917	n/a	Li, “Shisanjin jigao”, 19
9	guoxue	53867		Yin, “Guji shuzihua”
10	LI_2013	71342	1507	Li et al., ibid, 146
11	Ctexts	53882	1536	
		55244	1524	
		1779	11	

13. Zhouli

There is little variation, with exception of Zhu’s version. Zheng’s version was excluded from population, as it is too short. Zhang Guogan provides data for Zhouli for Tang and Shu stone classics, 49516 and 50508 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	
1	Zheng	45806	n/a	Qian, *ibid.*, 1:1:2–4;
2	Zhu	50508	n/a	Yin, “Guji shuzihua”; Zhu, *ibid*, 289, 3–4
3	Qian	49156	n/a	Qian, *ibid.*, 1:1:2–4;
4	ICS	n/a	n/a	n/a
5	CHANT	49540	2236	CHANT website
6	GUO_2001	n/a	n/a	GUO_2001, p.73
7	QIN_2005	49417	2219	Qin, “Zianqin guji”, 113
8	LI_2009	49375	n/a	Li, “Shisanjin jigao”, 20;
9	guoxue	49413		Yin, “Guji shuzihua”
10	LI_2013	49238	2167	Li et al., ibid, 146
11	Ctexts	49410	2212	
		49507	2208	
		421	29	

14. Zhouyi

There is much variation, with two major groups: Song and Qing’ ones are 24K, while modern ones are around 21K. WSW Ctexts is considrebly lower, as does not include commentaries, so it was excluded from population. Zhang Guogan provides data for Shujing for Han stone classics 24437 (Zhang, *ibid*, 1:1a,b), and for Tang and Shu stone classics, 24427 and 24052 (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	

1	Zheng	24207	n/a	Qian, *ibid.*, 1:1:2–4;
				Wang, i*Ibid*, 24270;
				Ouyang Gong: 24107

2	Zhu	24052	n/a	Yin, “Guji shuzihua”; Zhu, *ibid*, 289, 3–4
3	Qian	24437	n/a	Qian, *ibid.*, 1:1:2–4;

4	ICS	21055	1363	ICS, *Zhouyi*, 275
5	CHANT	21055	1363	CHANT website
6	GUO_2001	21847	1357	Guo, “Gudai hanyu”, 73
7	QIN_2005	21083	1358	Qin, “Zianqin guji”, 113
8	LI_2009	21703	n/a	Li, “Shisanjin jigao”, 11
9	guoxue	21696		Yin, “Guji shuzihua”
10	LI_2013	21152	1363	Li et al., ibid, 146
11	Ctexts	13348	1030	
		22229	1360	
		1416	3	

15. Zhuangzi

As it is not a part of Shisanjing, there is no Song and Qing period numbers. All modern numbers, when available, show not much variation.

#	Source	N	V	
1	Zheng	n/a	n/a	
2	Zhu	n/a	n/a	
3	Qian	n/a	n/a	
4	ICS	65406	2937	ICS, *Zhunagzi*, pp
5	CHANT	65406	2937	CHANT website
6	GUO_2001	64464	2898	Guo, “Gudai hanyu”, 73
7	QIN_2005	65231	2924	Qin, “Zianqin guji”, 113
8	LI_2009	n/a	n/a	
9	guoxue	n/a		
10	LI_2013	64744	2888	Li et al., ibid, 146
11	Ctexts	65251	2968	
		65019	2923	
		398	32	

APPENDIX II. Electronic Databases and Digital Corpora of Classical Chinese[73]

Resource	Type	URL	Start
Scripta Sinica Han ji dian zi quan wen zi liao ku 漢籍電子全文資料庫 Institute of History and Philology, Academia Sinica, Taiwan Institute of History and Philology, Academia Sinica, Taiwan	Academic	http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm (punctuation,no word)	1984

CHANT (Chinese Ancient Texts) 漢達文庫	Academic/ Semi-commercial	http://www.chant.org/ (punctuation, no word)	1986
(Academia) Sinica Corpus 中央研究院[現代]漢語語料庫 *Academia Sinica Ancient Chinese Corpus*	Academic	http://hanji.sinica.edu.tw/ (punctuation, no word)	1986
PKU Peking University Corpus CCL语料库	Academic	http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=gudai simplified, punctuation, no words.	2003
Thesaurus Linguae Sericae (TLS)	Academic	http://tls.uni-hd.de/projectDescription/features/firsts.lasso	1989
D.Sturgeon’ Ctext	Independent/ Research	http://ctext.org/	2006
Warring States Workshop Ctexts	Research	http://www.umass.edu/ctexts/index.php	2009
Unihan (Unihan Digital Technology Co., Ltd. 北京书同文数字化技术有限公司)	Commercial	http://www.unihan.com.cn/	2009
Erudition Database (爱如生數據庫) Database of Chinese Classic Ancient Books (中國基本古籍庫)	Commercial	http://server.wenzibase.com/dblist.jsp	
OTHER RESOURCES			
Palace Museum Classical Chinese Database 故宮【寒泉】古典文獻全文檢索資料庫 (Palace Museum, Taiwan)		http://210.69.170.100/s25/	1999
古今圖書集成 East View Information Services United Data Banks (formerly Greatman) Taiwan		http://greatman.eastview.com/Chinesebookweb/home/index.asp The Complete Classics Collection of Ancient China 标点古今图书集成	1997
The Sheffield Corpus of Chinese		http://www.hrionline.ac.uk/scc/db/scc/manual.html (source – includes www.shuku.net, www.guoxue.com and www.chinapage.com/china.html)	2005

Guoxue baodian Corpus 国学宝典网络版正式发布		http://www.gxbd.com/	2005
Hytong		http://www.hytung.cn/Default.aspx	2003

Scripta Sinica (漢籍電子文獻) is arguably the oldest, and one of the largest classical Chinese electronic database projects that began in 1984 at the Institute of History and Philology (IHP), Academia Sinica (中央研究院, http://hanji.sinica.edu.tw/), with initial goal, as stated at its website (http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm), "to digitize all documents essential to research in traditional Sinology". It grew up into a full-text database for academic research, which eventually was deployed online. By 2013 the database contained 688 titles and 445,950,000 characters. Most notably, there are twenty-five histories and thirteen classics, as well as other classic texts. The online version has an elaborated search interface, however, it is not designed as a concordancer or annotated corpus[74]. During the data entry process, researchers encountered the problem of coding page limitations. It was partly resolved with a sophisticated "character replacement" method (see Wang and Hsie, *Chinese Classics*). Wu Yeen-Mai (Wu, "Twenty-Five Dynastic Histories", 21) also mentions about 135 textual changes, made to the original edition of dynasty histories. It does not contain statistical data on text lengths.

CHinese ANcient Text (CHANT) database is, like Scripta Sinica, one of the earliest and most comprehensive collections of classical Chinese texts in electronic form. It started in 1988 at the Institute of Chinese Studies (ICS) at Chinese University of Hong Kong, under the lead of by D.C. Lau (Lau Din Cheuk), as an electronic database of all classical texts pre–6th century A.D., with the original mandate to continue Harvard-Yenching series of paper concordances on the new basis. The texts were entered manually (bases mostly on Sibucongkan), and passed through multiple verification stages, that made it one of the most reliable electronic sources[75]. Eventually, a series of ICS paper concordances was published based on these electronic texts, as well as CD-ROMs (a separate study is needed to understand how characters not represented by coding pages were handled). Finally, at the beginning of 2000s, the project was taken online. The online version, as well as ICS paper concordances, features lengths of texts and number of type-tokens (which most often, but not always, are same).

Academia Sinica Ancient Chinese Corpus was developed by Chinese Knowledge Information Processing Group, Institute of Information Science (IIS) at the Academia Sinica (and Academia Sinica Computer Center (ASCC))

The group was founded by Hsie Ching-chun in 1986 (soon after Scripta Sinica group) (Huang and Chen, “A Chinese Corpus”, 1214) as a sub-project of CKIP. Hsieh Ching-Chun (Hsieh, “Full Text Processing”, 126) indicates that even earlier, in 1985, there was the Chinese Text Processor (CTP) project group at ASCC, which focused on creation of electronic version of 24 dynasties for a workstation for studies in humanities. Wu Yeen-Mai (Wu, ibid, 21) indicates that the project started about same time as Scripta Sinica, and was partly funded by East Asia Library of the University of Washington, “The Academia Sinica Computer Center began this project in 1984 with a trial data base of the economic chapters of the first eight dynastic histories. The East Asia Library of the University of Washington (EALUW) participated in this pilot project.”[76] (Huang and Chen, “A Chinese Corpus”, 1214), mention that the group estimated the size of whole pre-Qin corpus as three million characters, of which they managed to receive texts of 1.5 million characters as an intra-Academia transfer from IHP, and the rest they were going to entry manually by the end of 1992. The fact of sharing of the data is confirmed by reference to “IHPAS prepared the text and CCAS was responsible for input, quality control, etc.” (Wu, ibid, 21). Therefore, we might consider Scripta Serica and Academia Sinica corpora as one corpus. As other such groups, although, this group made some modifications to original printed texts, “IHPAS has carefully reviewed this edition and made 135 textual revisions based on information from other authoritative editions.” I.e., these changes were not simple, like-OCR input (Wu, ibid, 21). The Sinica Treebank of classical texts was based on Academia Sinica corpus (Huang et al., “Sinica Treebank”). It does not contain statistical data on text lengths.

Peking University Corpus (PKU) The project started about 2003 at the Center of Chinese Linguistics (CCL) of Department of Chinese Language and Literature. By January 2006, “the texts written in traditional Chinese in PKU-CCL-CORPUS have contained approximately 101 million Chinese characters (486 documents, 54 folders, 202,305,825 bytes), and the texts written in modern Chinese have contained 115 million Chinese characters (157 documents, 23 folders, 229,700,435 bytes)” (Zhan et al, “Recent Developments”). The PKU documentation provides text length for classics, but only lengths of files in bytes, which probably includes punctuation and extra-textual characters, which makes this data unusable for our goals.

Thesaurus Linguae Sericae (TLS) has been developed since 1989 by an international group of scholar, under editorship of Christoph Harbsmeier – as a part of the Cluster of Excellence “Asia and Europe in a Global Context” (Mueller et al, “Geschite Ostasiens”). TLS is defined as “the first synonym dictionary of classical Chinese in any Western language.” corpus (see its presentation at http://tls.uni-hd.de/projectDescription/features/firsts.lasso) It puts stress semantic analysis of Chinese texts, but it has a considerable value as a classical Chinese online. Each text was curated and reviewed (often entered) by a specialist. This approach has had probably some drawbacks (e.g., some classical texts could be missing, because there was no person who could be involved into editing), but it allowed to create digital copies of highest quality[77]. Unfortunately, there is no available information on text lengths and vocabulary.

Sturgeon's Ctext Project Donald Sturgeon started the project single-handedly in 2006, but gradually it grew a real community. Sturgeon does not state what were the origin and mission of the project (http://ctext.org/introduction), but his project is immensely popular due to texts' layout, accessibility and search tools. The lack of resources, having OCR as main method of digitization, affected accuracy of texts, though[78]. However, errors are being gradually corrected by members of community; although the process is not as easy as at Wikimedia. It does not contain statistical data on text lengths.

Warring States Workshop Ctexts The project started as online dimension of research database, created by its author. It contains the less number of texts, comparing to other online resources, but it provides sophisticated search, statistical and other research tools, which are more proficient than any other available resource. The source of digital resources is Wikimedia. It definitely contains some inaccuracies, but the main text bodies are most probably "loaned" from Academia Sinica or similar resources, so it is most probably accurate enough for a research tool, e.g., for calculation text lengths and vocabularies.

Of other full-text search resources *Guoxue baodian* and *Sheffield Corpus of Chinese* should be mentioned. Guoxue baodian database 国学宝典网络版正式发布 (see Liu, "Commercial databases") is a commercial resource, featuring more than 3800 texts, 800 million characters (simplified characters). It is important for this study, as it reports text length data. Sheffield Corpus of Chinese (SCC) is a small, but very important academic corpus of Chinese historical texts (see HU–2005). Its importance is particularly based on its being grammatically marked-up. Unfortunately, since the mid–2000s, this corpus is not growing, and is too small to be used for this study[79].

Notes

[1] All texts, except Zhuangzi, are from the «Thirteen Classics» (*Shisanjing*), and are available through a web-based concordancer Warring States Workshop Ctexts (thereafter, «WSW Ctexts», to discern it from another project with similar name, Donald Sturgeon's «Ctext Project»).

[2] E.g., see Qin, *Xianqin guji* and Liu, «Xizhou jin» (inscriptions on bronze), Guo, «Gudai hanyu», Lee, «Classical Chinese Corpus» (semantic frequencies), Li, «Shisanjing Jigao», Li et al., «Corpus-Based Statistics», Da, «Corpus-Based Study» (character frequencies).

[3] E.g., Che, «Han fei zi», see description of first dictionaries in Feng, «Evolution and present situation», also Liang, «State of Art», or syllable-to-character statistics in Li et al., «Corpus-Based Statistics».

[4] On deeper philosophical foundation of the concept of tokens, see, e.g., Bromberger, *On What We Know*, 170–203.

[5] Sproat et al., «Stochastic Finite-State Word-Segmentation», 378.

[6] This discussion is far beyond the scope of this article (see its historical review at Packard, *Morphology of Chinese*), but it is important at least to delineate a few points here. The extreme negative position was summarized by Richard Sproat (who does not necessarily support it) as follows: «Chinese simply lacks orthographic words … Partly as a result of this, the notion "word" has never played a role in Chinese philological tradition, and the idea that Chinese lacks anything analogous to words in European languages has been prevalent among Western sinologists" (Sproat, *ibid*, 378). Packard admits (Packard, *ibid*, 17) that «word» does not appear especially intuitive concept, as «in Chinese culture, the clear and intuitive notion of word is zi. For most speakers, zi as morpheme and zi as written characters are same. Word for word is ci» (Packard, *ibid*, 15). In the initial period of Chinese corpora linguistics, even the size of modern language corpora was indicated mostly in characters. However, while in general, e.g., Sinica Corpus is defined as «word-based» corpus, with POS-tagging, both measures are applied: «Version 2.0 of the Academia Sinica Balanced Corpus (Sinica Corpus) contains 5,345,871 characters, equivalent to 3.5 million words.» (Chen et al., «Sinica Corpus», 167). And in classical studies word-token practically does not apply. E.g., as late as in 2012, Lee, «Classical Chinese Corpus», 76, mentions that status of «wordhood» in classical Chinese still needs consensus. In his own work Lee generally «following the practice of the Academia Sinica Ancient Chinese Corpus, each character is initially presumed to be a monosyllabic word.» (Lee, *ibid*, 78).

[7] Sproat reports that human judges disagree in many cases, and the agreement rate is 76% (Sproat, *ibid*, 394). However, Nianwen Xue et al. report a higher degree of expert agreement, « Following (Sproat et al., 1996), we calculate the arithmetic mean of the precision and the recall as one measure of agreement between each output pair, which produces an average agreement of 87.6 percent, much higher than the 76 percent reported in (Sproat et al., 1996)» (Xue, «The Penn Chinese TreeBank», 6)). It is still lower than for most other languages.

[8] While word segmentation is important for syntax analysis, it could be that character approach is as good as words for topic analysis. One example is Zhao et al., What is the Basic Semantic Unit». This research suggests that the topic model with Chinese characters can also effectively capture the semantic contents in text documents. The computational evidence presented in this paper supports an argument that the Chinese characters can be used as the basic semantic units in Chinese language modeling. (Zhao et al., *ibid*, 156).

[9] In future, with improvement of classical Chinese word segmenting algorithms, length could be counted in words (Xue, «Chinese Word Segmentation»). Liang

Shehui reviews word segmentation attempts for various texts (Liang, «State of Art», 58), and Li Bin et al. provided new statistics on words classical Chinese texts, and compared with the modern corpora, stating that «the multiple-character words dominate the vocabulary as early as Pre-Qin period. … there are … 17,505 multiple character word types, which account for more than half of the total word types» (Li et al., «Corpus-Based Statistics», 150).

[10] The difference could be seen in two series of paper concordances for classical texts, HY and ICS. The ICS concordances (based on electronic database, later to become the foundation of CHANT online system) feature text lengths and type-token lists with frequencies. The same information could have been provided for HY, but it was not provided.

[11] The total size of pre-Qin and Han classical texts could be roughly placed between three and eight million characters, which is not a huge amount.

[12] For a general guide to OCR for Chinese characters see Cheriet et al., *Character Recognition Systems*. Dai Ruwei offers a historical review of OCR for Chinese characters, starting from 60s (Dai et al., «Chinese Character Recognition»).

[13] All online academic and commercial sites, similar to crowd-sourced sites like Wikisource, could be in permanent change. One strong side of Wikisource is that changes are documented and available for review.

[14] For introduction to code sets for Chinese characters, and description of the problem of missing rare characters see Zhao and Zhang, «Totality of Chinese Characters». For description of methods of creation of those rare or obsolete Chinese characters (almost four thousand), not found in existing computer writing programs (which were prepared mainly for business use) see McLeod, «Sinological Indexes», 48. See also Wang and Hsieh,»Chinese Classics Full-Text Database», 2011 on OCR and digitalization character substitution process. These problems are also addressed by described in Wittern, «Digital Editions». Yang Jidong and Yin Xiaolin (Yang, «Approaching Pre-modern China», 7 and Yin, «Guji shuzihua») address issues of text versions.

[15] In the reference section of Cheng's article, only one article is written after 2010, most other articles were published before 2007. Coincidentally, this is the time when commercial corpora started dominating the online market.

[16] A concise (not up-to-date) list of corpora could be found in Yang Xiaojun's article (Yang, «Survey and Prospect»).

[17] However, not only this article does not mention Hong Kong's CHANT/ICS database, but it also lacks description of Western corpora of classical Chinese.

[18] Very helpful (however, concise), information is often featured on university libraries' websites. E.g., Berkeley's resource list («Chinese Studies Electronic Databases», University of California, Berkeley, last modified September 15, 2013, accessed June 15, 2014, http://www.lib.berkeley.edu/EAL/resources/chinese_databaseA-Z.html), or Indiana university article by Liu Wenling (Liu, Commercial Databases»).

[19] Therefore, such important electronic collections of classical texts, as *Sibu quanshu*, *Sibu congkan*, and *Sibubeiyao* will not be reviewed here. Other similar and otherwise important resources like «Palace Museum Classical Chinese Database» will not be addressed in this article, as well as Wikisource.

[20] It should be noted that due to the Internet fluidity, some of these sites are non-functional or could be non-functional soon; on others, functionality could be damaged and not updated; still, most of these sites have played significant role in evolution of classical Chinese corpus linguistics. It will be noted below, how situation is changing in this area with advance of commercial corpora.

[21] The electronic corpora for Modern Chinese, like Penn Corpus, etc., will not be reviewed in this paper, as unrelated to its subject.

[22] «A dramatic growth of large-scale digitization efforts has taken place in Chinese studies. A few electronic-resources providers in China and Taiwan have produced the most influential electronic resources in the field.» (Liu, «Commercial Databases», 14) This paper will be only cursory touching on subject of many available online electronic texts (some of them having full-text search), e.g., Guoxue, Wikisource, etc.

[23] While texts themselves are freely available in block-prints, etc., digitalization of them, especially before OCR process, is time-consuming and expensive process, so produced versions were expensive.

[24] Again digital versions of printed collections, like SKQS, even available online now, are excluded from this list.

[25] However, they could be crowd-sourced (Wikisource, partly Sturgeon's Ctext).

[26] The other name is «Database of Chinese Classic Ancient Books» 中國基本古籍庫. It claims to contain «more than 10,000 titles of most important classical Chinese works in various subjects covering the period from Pre-Qin to the Republic of China. The size of the contents is at least three times of the well-known "Imperial Collection of the Four Libraries" (四庫全書) (see list of resources «Social History of the Chinese Silk Road», Yale University Library, last accessed June 15, 2014, http://guides.library.yale.edu/silkroad).

[27] Probably, it was initially a part of joint project with the Library of Washington 1984–1985 (see Wu, «Twenty-Five Dynastic Histories» about the library's participation)

[28] Paul Thompson mentions that as early as in 1979–80 there were attempts to create classical corpora (Lunyu, Mengzi, Liji) in Japan at the Institute of Asian and African Languages and Cultures at the Foreign Studies University in Tokyo, but they did not succeed (Thompson, «Chinese Text Input», 123).

[29] GB–2312 contained even less, 6,763 (see e.g. Juang et al., «Resolving the Unencoded Character Problem»).

[30] It was integrated in 2008 into TELDAP («Taiwan e-Learning and Digital Archives Program (TELDAP) initiative (see Liu–2009). The history of development is described by Mao Jianjun (Mao, «Zhongguo jiben guijiku»).

[31] An interesting material on details of creating full-text search tools for Academia Sinica data (actually 24 stories, probably, borrowed from IHS) could be found in Hsie, Full Text Processing». Wei Peichuan et al. mention word segmentation (Wei et al., «Historical Corpora», 132)

[32] «Farther in the future may be ICS in-house CD-ROM production. The body of extant Han and pre-Han texts totals about eight million characters» (McLeod, «Sinological Indexes», 50). This is why in this article the scope of pre-Qin and Han texts is evaluated from 3 to 5 million characters.

[33] In 1992, the Institute began publication of the ICS Ancient Chinese Text Concordance Series of some ninety-three planned volumes covering all 103 extant Chinese writings from antiquity to the end of the Eastern Han in a.d. 220. McLeod, *ibid*, 48).

[34] See general description in Zhan et al., «Recent Developments». It was developed jointly by «Center for Chinese Linguistics (CCL) of Department of Chinese Language & Literature, which is engaged in Chinese language research and teaching, the other is the Institute of Computational Linguistics (ICL), which is engaged in Chinese information processing» (Zhan et al., «Recent Developments», 3). Started in 2003 as a part of one of four corpora – «a very large scale of wide time-span Chinese corpus, which is processed with sentence segmentation (denoted as PKU-CCL-CORPUS).» (Zhan et al., «Recent Developments», 4), subcorpora – «Xiandai» (modern) and «Gudai» (classical).

[35] The data of frequency and text lengths are provided in lists of statistics, published by Beijing university, e.g., «Classical Chinese Character Frequencies», Beijing University, last accessed June 15, 2014, http://ccl.pku.edu.cn:8080/ccl_corpus/CCL_CC_Sta_Gudai.pdf, http://ccl.pku.edu.cn:8080/ccl_corpus/CCL_Gudai.pdf.

[36] It could be that Erudition database is built in same way, but there is no enough information.

[37] Text sources are probably digitized versions of printed books; some texts came from CHANT, etc.

[38] About Wikimedia, which is not included Wikimedia is a communal resource of classical Chinese texts. The sources of the texts are unknown, but, judging from some replacements for rare characters, it could be other online corpora, like Sinica. Some texts could be automatic conversions of GB codes into Unicode. Therefore, its accuracy may be not higher that Sinica, etc. But its copyright policy allows it to be used for free, and texts, unlike Sinica, etc., are gradually cleaned up by the community (similar to Ctext, but correctors could do it themselves, which simplifies process.)

[39] Unfortunately, it does not feature text lengths.

[40] It is possible that Wikisource incorporates some legacy corpora, but it provides Creative Commons copyrights.

[41] Some of them started earlier, but were not as successful, e.g., Guoxue baodian (see critique in Yang, «Chinese Classic Text Database»).

[42] See Yang, *ibid*. The author of this paper did not have access to either commercial source, and there is no available publicly data on their statistics; therefore, this data is not featured in this article. One author mentioned difference between *Sibucongkan* and *Sibubeiyao* – manual entry helps to correct errors in xylograph, but brings new ones. This discussion is very old: according to John Winkelman, in Song time some library owners valued manual copies over printed, because it allowed to collate book in the process of copying (Winkelman, «Imperial Library», 28).

[43] It is hard to evaluate the real use of online corpora through published materials. Whenever a researcher quotes Chinese texts, they mostly use printed versions. Therefore, to evaluate research access to these resources, one needs to have statistics of their usage, based on university IPs, which is not readily available. In Yang, *ibid*, it is stated, though, that Erudite database is now officially a quotation source in China.

[44] Even opposite, the latest printed concordances (the ICS series) are based on electronic versions of texts. It also seems that traditional calculations also were made not by using printed editions, but «stone classics» *shijing*.

[45] In manuscripts, book chapters are often untitled (e.g., Richter, «Textual Identity», 212), as well as text delineations at all are ambiguous, but it is not a rule, and definitely, in later epoch, chapter titles are found more often.

[46] It started in 1928, as Chen Heqin's «*The Applied Glossary of Modern Chinese* (语体文应用字汇)» was published by the Commercial Press in 1928» (Feng, «Evolution and Present Situation», 175)

[47] The first Chinese Modern Literature Work Corpus (in 1979), 5.27 million words, by Wu Han University (Feng, «Evolution and Present Situation», 176).

[48] The year of 1991 marks time when National Chinese Corpus has started (Feng, «Evolution and Present Situation», 181).

[49] Tsien addresses these units in a special section (Tsien, *Written*, 120–122). However, he does not pay special attention to character count, sometimes written on manuscripts. In the West, number of words or letters is also not usually entered in bibliographical catalogues, while it is sometimes mentioned in the printing data on the book itself.

[50] E.g., Richter, «Punctuation», 9, reports that in Mawangdui manuscripts text lengths in characters are often found at the end of the texts. Interestingly, Tsien (Tsien, *Written*) does not mention it. Also, Loewe, «Early Chinese Texts», 8, mentions that for chapter 15–19 of Mawangdui version, «there is a note at the end of each item giving the number of characters therein, and at the end of the group the total number is given as 2870» (which sums up exactly to numbers for chapters, «testifying that they were taken from a single source».

[51] Loewe indicates that lengths of texts in characters are often recorded in dynasty histories. E.g., for Zhuangzi, Shi-ji «refers to a text of some 100 000 words» (Loewe, «Early Chinese Texts», 57). As Winkelman, *ibid*, informs, at Song times, there was a quota of 2000 characters a day for copyists and collators. There was a process of accounting volume of work. E.g., the Imperial library reports that hired contractors recopied 50,000,000 characters in update process (Winkelman, *ibid*, 33), which means that lengths for specific texts in characters were accounted for and most probably well-known to librarians and whoever was related to libraries.

[52] See, e.g., Tsien, *Written*, 78–83.

[53] Zhang Guogan (Zhang, *Lidai Shijing Kao*) reports estimates of lengths for most early stone classics, starting from the II century CE, and they will be also cited in Appendix I.

[54]阮葵生 (1727–1789), see Wang, «Kuan Kuishen Nianpu» for more details.

[55] Zheng Genglao 郑耕老 (1108–1172) himself only counted numbers for «nine classics», as follows from the chapter's title, but his numbers were amended by the compiler. See Yin, «Guji Shuizihua», as well as Huang, *Shoupi Baiwen.*

[56] Zheng was not the only one interested in these numbers. Another Song's scholar, Ouyang Gong, in «Dushufa» (歐陽公 «讀書法»), provides some data on classics lengths, as well as probably others. But this subject should be a subject for a special research.

[57] See e.g., Jiang, «Cheng Yue Chunqiu», 186.

[58] The numbers (contained in chapter 289) were most probably added later. The original text contains numbers for both canon and commentary, and numbers for canon text itself are provided in commentary.

[59] PST, «Shisanjing zishu», juan 1, 2–4. This data is also referenced by Wang, who relates them to an edition of «Shisan jing zhushu», in the earliest of PRC publications on classics' lengths (contains some discrepancies).

[60] E.g., McLeod, *ibid*, 48 describes manual process of creation of *Shisan jing suoyin* in 1929.

[61] Reproduced in 1982 Wang, *ibid*, publication. Until the 80s, whenever scholars needed estimates of classics' vocabulary, they had to rely on their own calculations, e.g., Tsien, *Written on Bamboo and Silk*, 25.

[62] It quite possible that texts in Wikimedia are «borrowed» (at least partly), from Scripta Sinica or CHANT.

[63] It means, «legally», i.e., providing a link to work's sources. At the beginning of 2000s, it was still acceptable to publish data on «scraped online» sources without explicit permissions, like, e.g., Guo, «Gudai hanyu», 81.

[64] They could be called «non-identified» as «most downloaded from the Internet, a small part of the acceptance of the gift of friendship main sources of material used in the Web» (p.81) scraped from Sinica Corpus, PKU, and (currently unavailable) «bookbig» resource at http://www.bookbig.com/culture1.html.

[65] This article considered as «source» only data, where authors claimed they personally calculated numbers from an available source, or reported such data. Therefore, Zhang Guogan's very interesting data will be presented, but it is not listed as a «source».

[66] Except estimates, like Tsien_2004, p.25 and Zhang Guogan's data on stone classics. Zhang's data (see Table 2) is very interesting, however, it is not considered in this article as a regular data source.

[67] The CHANT website, based on the same digital texts, also reports these numbers, but they sometimes differ from ICS numbers. It may reflect some changes in digital sources, made over twenty years, or including punctuation characters in one account.

[68] While it is easy to calculate standard statistical characteristics, e.g., average, deviation, etc., this article will not be including this data, because its goal is to expose variation, rather than to discuss specific cases and its causes. A fruitful discussion would be only possible if most corpora are available for inspection, and this is not the case for our texts.

[69] These electronic projects could be considered a progressive editorial activity, which has been applied only to electronic media, not printed.

[70] The capabilities for automatic data retrieval during qualitative corpus analysis enable the scholarly community to replicate searches, with the purpose of reproducing and verifying outcomes of linguistic investigations, when corpora are publicly available and corpus markup, annotation, and problem-oriented tagging schemes are made available along with the published corpus. (Hasko, «Qualitative Corpus Analysis», 4)

[71] This is a popular translation. Elman (Elman offers translation that seems more accurate: «Critique of classical studies». See Elman, «Collecting and Classifying» and Elman, *On their Own Terms*, XX).

[72] Li–2009 does not provide absolute numbers, only relative percentage. Numbers for LI–2009 have been calculated, based on his character percentage numbers. E.g., for CQZZ, for the most frequent character, zhi, Li lists 7342 tokens, at 3.7499% (p. 11), which translates, rounded, into 195792 (incidentally, the same number as ICS).

[73] See a more complete list at Liu, *Impact of Digital Archives*.

[74] There is practically no research material describing Scripta Sinica; however it is possible to state that classics were entered from the 1970 edition of SSJZS and still need some post-entry editing.

[75] The CHANT group, as pioneers of digitizing classical texts with very complicated character vocabulary, went through immense difficulties, and brought some positive change into the area.

[76] Library of the University of Washington (EALUW) participated in this pilot project. In 1986, EALUW and CCAS signed an agreement to initiate a joint project to 1) develop a prototype of a Chinese full text processing system, and 2) design an integrated library system. Ibid In the future, this system may also be used to store Chinese texts created by Academia Sinica as, for example, Shih son ching (The Thirteen Chinese Classics), Chuang-tzu, Kuan-tzu, and Taiwan gazetteers. (Wu, ibid, 24)

[77] Some of texts were loaned from other corpora, e.g., CHANT.

[78] Sturgeon recommends always double check quotations. However, due to digital content gap, it is recommended for practically all other online resources.

[79] There have been other interesting attempts to mark-up classical Chinese texts grammatically, e.g., Academia Sinica, and Huang et al., «Statistical Part-ofSpeech Tagging», based on their own small corpus, but they are not available readily.

Literature

Bromberger, Sylvain. On What We Know We Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them. Chicago: University of Chicago Press, 1992.

Che Shuya 车淑娅. “"Han fei zi" cihui yanjiu《韩非子》词汇研究 [Vocabulary Study of Hanfeizi].” Chengdu: Ba Shu shushe, 2008.

Chen Keh-Jiann, Huang Chu-Ren, Chang Li-Ping, Hsu Hui-Li. “Sinica Corpus: Design Methodology for Balanced Corpora.” In Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC 11), 167–176. Seoul, 1996.

Cheng Winnie. “Corpora: Chinese Language.” In Encyclopedia of Applied Linguistics, edited by C.A. Chapelle. Chicester: Wiley-Blackwell, 2013.

Cheriet Mohamed, Nawwaf Kharma, Liu Cheng-Lin, Ching Suen. Character Recognition Systems: A Guide for Students and Practitioner. Hoboken: Wiley-Interscience, 2007.

Da Jun. “A Corpus-Based Study of Character and Bigram Frequencies in Chinese E-texts and its Implications for Chinese Language Instruction.” In The studies on the theory and methodology of the digitalized Chinese teaching to foreigners: Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese, edited by Zhang, Pu, Tianwei Xie and Juan Xu, 501–511. Beijing: Tsinghua University Press, 2004.

Dai Ruwei, Liu Chenglin, Xiao Baihua. "Chinese Character Recognition: History, Status and Prospects." Frontiers of Computer Science in China 1 no. 2 (2007): 126–136.

Elman Benjamin A. "Collecting and Classifying: Ming Dynasty Compendia and Encyclopedias (Leishu)." Extrême-Orient, Extrême-Occident no. 1 (2007): 131–157.

Elman Benjamin A. On Their Own Terms: Science in China, 1550–1900. Cambridge: Harvard University Press, 2009.

Feng Zhiwei. "Evolution and Present Situation of Corpus Research in China." International Journal of Corpus Linguistics 11 no. 2 (2006): 73–207.

Guo Xiaowu 郭小武. "Gudai hanyu jigao pinzi tansuo 古代汉语极高频字探索 [Exploration of most-frequent characters in classical Chinese]." Yuyan yanjiu 44 no. 3 (2001): 69–84.

Harbsmeier Christoph. "Thesaurus Linguae Sericae: an historical and comparative encyclopedia of Chinese conceptual systems." University of Heidelberg, posted May 26, 2007, accessed 25.12.2013, http://tls.uni-hd.de/projectDescription/acknowledgements/acknowledgements.lasso

Hasko Victoria. "Qualitative Corpus Analysis." In The Encyclopedia of Applied Linguistics, edited by Carol A.Chapelle et al. Malden, MA: Wiley-Blackwell, 2013.

Ho Che Wah. "CHANT (CHinese ANcient Texts): a Comprehensive Database of All Ancient Chinese Texts up to 600 AD." Journal of Digital Information 3 no.2 (2002): article 119.

He Jianye. "Acquiring High Quality Chinese Research Materials: A Case Study of Irregularities in Current Chinese Publishing." Journal of East Asian Libraries, no. 141 (2007): 11–18, https://ojs.lib.byu.edu/spc/index.php/JEAL/article/download/8826/8475

Hsieh Ching-Chun. "Full Text Processing of Chinese Language." Journal of library and information science 11 no. 2 (1985): 125–142.

Hu Xiaoling, Williamson Nigel, McLaughlin Jamie. "Sheffield Corpus of Chinese for Diachronic Linguistic Study." Literary and Linguistic Computing 20 no. 3 (2005): 281–293.

Huang Chu-Ren, Chen Keh-jiann. "A Chinese Corpus for Linguistics Research." In The Proceedings of the 1992 International Conference on Computational Linguistics (COLING–92), 1214–1217. Nantes, France, 1992.

Huang Chu-Ren, Chen Keh-Jiann, Chen Feng-Yi, Zhao-Ming Gao, Chen Kuang-Yu. "Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface." In Proceedings of 2nd Chinese Language Processing Workshop (Held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL–2000), 29–37. Hong Kong, 2000.

Huang Kan 黄侃. *Shoupi Baiwen Shisanjing* 手批白文十三經 [Hand annotated edition of Thirteen Classics]. Beijing: Zhonghua shuju, 2006.

Huang Liang Huang, Peng Yinan, Wang Huan, Wu Zhengyu. "Statistical Part-of-Speech Tagging for Classical Chinese." Lecture Notes in Computer Science no. 2448 (2002): 296–311.

Jiang Youyu 江右瑜. "Cheng Yue Chunqiu sixiang zhelun 陳岳《春秋》思想析論 [Analysis of Chen Yue's thought on Chunqiu]." Guo wenxue zhi 19 no. 12(1) (2009): 181–225.

Juang Derming Juang, Wang Jenq-Haur, Lai Chen-Yu, Hsieh Ching-Chun, Chien Lee-Feng, Ho Jan-Ming. “Resolving the Unencoded Character Problem for Chinese Digital Libraries.” In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05), 311–319, 2005.

Kirkpatrick Andy, Xu Zhichang. Chinese Rhetoric and Writing: An Introduction for Language Teachers. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press, 2012.

Lee John. “A Classical Chinese Corpus with Nested Part-of-Speech Tags.” In Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Avignon, France, 24 April 2012. Association for Computational Linguistics, 75–84, 2012.

Li Xiang 李想. “Shisanjing jigao pinzi fenbu ji zuci yanjiu 十三经极高频字分布及组词研究 [Study of Character Frequency Distribution and Word Formation in *Shisanjing*.]” MA Diss., University of Heilongjiang, 2009.

Li Bin, Xi Ning, Feng Minxuan, Chen Xiaohe. “Corpus-Based Statistics of Pre-Qin Chinese.” In Chinese Lexical Semantics – 13th Workshop, CLSW 2012, Wuhan, China, July 6–8, 2012, ed. by Donghong Ji and Guozheng Xiao 145–153, Berlin-Heidelberg: Springer-Verlag, 2013.

Liang Shehui. “State of Art of Pre-Qin Chinese Information Processing – Case Studies with Mencius and its Annotations and Commentaries.” International Journal of Knowledge and Language Processing 3 no.1 (2012): 54–63.

Liu Ts’ui-jung. “Impact of Digital Archives on Humanities.” Topic presentation at Pacific Neighborhood Consortium (PNC) Annual Conference and Joint Meetings, 6–8 October 2009, Academia Sinica, Taipei, 2009.

Liu Zhiji 刘志基. “Xizhou jin wenzi pin tedian cheng yin chutan 西周金文字频特点成因初探 [Preliminary Study of the Causes of the Character Frequency on Bronze Inscriptions of the Western Zhou Dynasty].” Yuyan kexue 1 no. 9 (2010): 80–90.

Liu Wen-ling “Commercial Databases in East Asian Studies.” Journal of East Asian Libraries, no. 151 (2010): 13–27.

Loewe Michael (Ed.) Early Chinese Texts: a Bibliographical Guide. Berkeley: The Society for the Study of Early China and the Institute of East Asian Studies, University of California, 1993.

*Mao Jian-jun*毛建军. “Zhongguo jiben gujiku” de tese yu qishi – jian tan guji quanwen shujuku de biaozhun yu guifan《中国基本古籍库》的特色与启示 — 兼谈古籍全文数据库的标准与规范 [Characteristics and Inspirations on Chinese Classic Ancient Books Database: On standards and norms on ancient books full-text database].” Guanli xuekan 22 no. 5 (2009): 104–106.

McEnery Anthony M., Xiao Zhonghua. “The Lancaster corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study.” In Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004). Lisbon: European Language Resources Association, 1175–1178, 2004.

McLeod Russell. “Sinological Indexes in the Computer Age: The ICS Ancient Chinese Text Concordance Series.” China Review International 1 no. 1 (1994): 48–53.

Müller Gotelind, Wolfgang Seifert, Joachim Kurtz. “Geschichte Ostasiens: Heidelberger Forschungsbeiträge.” In Arbeitsgemeinschaft historischer Forschungseinrichtungen in der Bundesrepublik Deutschland (Hrsg.): Jahrbuch der historischen Forschung, 61–72, München: Oldenbourg 2012.

Packard Jerome L. The Morphology of Chinese: A linguistic and cognitive approach. Cambridge: Cambridge University Press, 2000.

Qin Qin 覃勤. “Xianqin guji zi pin fenxi yuyan yanjiu 先秦古籍字频分析语言研究 [A Statistic Study on Character Frequency of Pre-Qin Literature].” Studies in Language and Linguistics 25 no. 4 (2005): 112–116.

Richter Matthias L. “Textual Identity and the Role of Literacy in the Transmission of Early Chinese Literature.” In Writing and Literacy in Early China: Studies from the Columbia Early China Seminar, Edited by Li Feng and David Prager Branner, *206–238,* Seattle: University of Washington Press, 2011.

Richter Matthias L. “Punctuation” and “Scribal Hands”. In Reading Early Chinese Manuscripts: Texts, Contexts, Methods, ed. Wolfgang Behr, Martin Kern, Dirk Meyer (Handbook of Oriental Studies) Leiden: Brill, forthcoming.

San Duanmu. “Word-length preferences in Chinese: a corpus study.” Journal of East Asian Linguistics 21 no.1 (2012): 89–114.

Sproat Richard, Shih Chilin, Gale William, Chang Nancy. “A Stochastic Finite-State Word-Segmentation Algorithm for Chinese.” Computational Linguistics 22 no. 3 (1996):377–404.

Sturgeon Donald. “Zhuangzi, Perspectives, and Greater Knowledge.” Philosophy East and West 65 no. 3 (Forthcoming).

Sun Qin 孙琴. “Liang da zhongwen guji shujuku bijiao yanjiu 两大中文古籍数据库比较研究 [A Comparative Study of Two Databases of Chinese Rare Books].” Xinshiji tushuguan no.1 (2007).

Tao Hongyin, Xiao Richard. UCLA corpus of Modern Chinese The UCLA Chinese Corpus (2nd edition). UCREL, Lancaster, 2012 http://www.lancaster.ac.uk/fass/projects/corpus/UCLA/

Tsien Tsuen-Hsuin. Written on Bamboo and Silk: The Beginnings of Chinese Books and Inscriptions. 2nd ed. Chicago: University of Chicago Press, 2004.

Thompson Paul M. “Chinese Text Input and Corpus Linguistics.” In Characters and Computers, edited by Victor H. Mair and Yongquan Liu, 122–130, Amsterdam: IOS Press, 1991.

Wang Fengyang 王凤阳. “Hanzi pinlǜ yu hanzi jianhua 《汉字频率与汉字简化》[Frequencies and simplification of Chinese Characters].” Yuwen Xiandaihua, no 3 (1980): 83–103.

Wang Enbao 王恩保. ““Shisan jing zhushu” de juan shu he zishu《十三经注疏》的卷数和字数 [Number of characters an length of Shisanjing texts.]” Wenxian no. 2 (1982): 82.

Wang Jianxin. “Recent Progress in Corpus Linguistics in China.” International Journal of Corpus Linguistics 6 no. 2 (2001): 281–304.

Wang Ya-Ping, Hsieh Hsiaolin. Chinese Classics Full-Text Database Digitization Procedures Guideline. Taibei: Taiwan e-Learning and Digital Archives Program, Taiwan Digital Archives Expansion Project, 2011.

Wang Zeqiang 王泽强. “Kuan Kuisheng Nianpu 阮葵生年谱 [Chronicle of Life of Kuan Kuisheng].《淮阴师范学院学报：哲学社会科学版》.” Huaiyin Teachers College Journal: Philosophy and Social Sciences Edition no. 1 (2006): 14–18.

Wei Pei-chuan, Thompson P.M., Liu Cheng-hui, Huang Chu-Ren, Sun Chaofen. "Historical Corpora for Synchronic and Diachronic Linguistics Studies." Computational Linguistics and Chinese Language Processing 2 no. 1 (1997): 131–145.

Winkelman John H. "The Imperial Library In Southern Sung China, 1127–1279: A Study Of The Organization And Operation Of The Scholarly Agencies Of The Central Government." Transactions of American Philosophical Society 64 no. 8 (1974).

Wittern Christian. "Digital Editions of Premodern Chinese Texts: Methods and Problems – Exemplified Using the Daozang Jiyao." Chung-Hwa Buddhist Journal, no. 25 (2012): 167–194.

Wu Yeen-Mei. "Twenty-Five Dynastic Histories Full Text Retrieval Database at the University of Washington." Journal of East Asian Libraries no. 94 (1991): 21–24.

Xue Nianwen. "Chinese Word Segmentation as Character Tagging." Computational Linguistics and Chinese Language Processing 8 no. 1 (2003): 29–48.

Xue Nianwen, Xia Fei, Chiou Fu-dong, Palmer Martha. "The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus." Journal of Natural Language Engineering 11 no. 2 (2005): 207–238.

Yang Haikun 杨海昆. ""Guoxue baodian" wangluo ban kaitong zishu neirong chao siku quanshu"《国学宝典》网络版开通字数内容超四库全书 ["Guoxue baodian" online text collection deployed; surpasses by size Sikuquanshu collection]." http://book.sina.com.cn/news/c/2005-03-08/3/172295.shtml (posted 2005-03-08, accessed 2013–12–24).

Yang Jidong. "Approaching pre-modern China through the computer: the benefits and risks of using electronic resources in sinological research." Panel presentation at Annual Meeting of the Association for Asian Studies, Boston, 2007.

Yang Jidong. "Chinese Classic Text Database by Erudition 中国基本古籍库及分库." Presentation at 2012 CEAL Conference Committee on Chinese Materials Annual Meeting, Toronto, 2012, www.eastasianlib.org/ccm/annual_meeting/2012/powerpoints/erudition.pptx

Yang Xiao-jun. "Survey and Prospect of China's Corpus-Based Research." In Corpus Linguistics Around the World, Language and Computers Series, edited by Andrew Wilson, Dawn Archer, Paul Rayson, 219–233, Amsterdam: Rodopi B.V., 2006.

Yin Xiaolin 尹小林. "Guji shuzihua de shiming yu qianjing [古籍数字化的使命与前景] The Mission and Perspectives of Digitizing Ancient Texts (http://www.guoxue.com/zt/gjszh/yjwz_026.htm)." Presentation at the conference: The first symposium on digitizing Chinese ancient texts, 第一届中国古籍数字化国际学术研讨会 (http://www.guoxue.com/zt/gjszh/gjszh.htm) Beijing, 2007.

Zhan Weidong, Chang Baobao, Duan Huiming, Zhang Huarui. "Recent Developments in Chinese Corpus Research." In Proceedings of The 13th NIJL (The National Institute for Japanese Language and Linguistics) International Symposium, Language Corpora: Their Compilation and Application. Tokyo, Japan. 2006.

Zhang Guogan 张國淦. Lidai shijing kao 歷代石經考 [Study of Stone Classics]. Beijing: Yanjing daxue guoxue yanjiuso, 1930.

Zhang Hong, Xu Bo, Huang Taiyi. "Statistical Analysis of Chinese Language and Language Modeling Based on Huge Text Corpora", in Proceedings of Third

International Conference, edited by Tan, Tan, Yuanchun Shi, and Wen Gao, Beijing, 279–286, 2000.

Zhang Shuangdi 张双棣. “Lüshi chunqiu” cihui yanjiu.《吕氏春秋》词汇研究 Vocabulary Study of Lv Shi Chun Qiu. Beijing: Shangwu yinshuguan, 2008.

Zhao Qi, Qin Zengchang, Wan Tao. “What Is the Basic Semantic Unit of Chinese Language? A Computational Approach Based on Topic Models.” In Proceeding of the Mathematics of Language – 12th Biennial Conference (MOL 12), Nara, Japan, September 6–8, 2011.

Zhao Shouhui, Zhang Dongbo. “The Totality of Chinese Characters – A Digital Perspective.” Journal of Chinese Language and Computing 17 no.2 (2007): 107–125.