QUANTITATIVE CHARACTERISTICS OF DIGITAL CHINESE CLASSICS: A PRELIMINARY STUDY

Sergey Zinin
Warring States Project
University of Massachusetts, Amherst

To the memory of Prof. Tatiana Grigoryeva

Abstract[1]

The study analyzes quantitative aspects of digital versions of fifteen pre-Qin classical Chinese texts[2], such as text length and number of character types. These parameters are critical for the analysis of character frequencies, but they vary in existing digital corpora. This article analyzes the available data (collected for the first time), beginning from the pre-digital period. It delineates the evolution of digital resources of classical Chinese and provides an up-to-date review of major available online resources as well as offline research corpora. The paper demonstrates the scope of variation and discusses the inherent inaccuracy of digital texts ("digital content gap," i.e., the discrepancy between printed and digital versions of the same text). The digital content gap could affect traditional philological studies, but it

[1] An earlier version of this article was published as Zinin, Sergey. "Pre-Qin Digital Classics: Study of Text Length Variations". This material is an extended version of the former material.

[2] 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 "Chinese Text Project (CTP)").

may not be very significant for a quantitative analysis. The article presents a comparative analysis of length statistics, and concludes that because of the existing variety in text characteristics, any quantitative study of Chinese classics will be corpus-specific. There is a need for the creation of standard, open-source online digital corpora of classical Chinese texts. While these corpora could be used as search tools for philological research, they should rather be targeting quantitative linguistic goals.

Content

1 Introduction

There are a few studies on most frequent character lists for classical texts[3], as well as on character sets for a number of specific texts[4]. However, quantitative linguistics of classical Chinese texts is still in its early development; for example, there are practically no studies on comparative frequency distribution of characters in vocabularies of multiple texts. For the latter type of study, quantitative characteristics such as text length and vocabulary size (in tokens) are essential. However, not many online or digital corpora offer this type of information in a convenient form.

1.1 **Character as token** Any frequency or length analysis of classical Chinese texts should choose either a "character-as-token" or a "word-as-token" approach. Practically all past and present researchers of classical Chinese corpora report lengths of texts in characters; i.e., they use characters as basic measurement unit for texts ("token"). This "character-as-token" approach is not a typical corpus linguistic approach. In corpus linguistics, "token" could be any meaningful grouping of characters; however, for most languages token is a word by default ("word-as-token" approach). Because of the nature of the Chinese writing system, Chinese language corpora often deviate from this rule, and choose a "character-as-token" approach[5].

---

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

[4] 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".

[5] On deeper philosophical foundation of the concept of type and tokens, see, e.g., Bromberger, *On What We Know*, 170-203, Wetzel,*Types and tokens: on abstract objects,* Hutton, Christopher. *Abstraction and instance: the type-token relation in linguistic theory.* While an "uncritical" approach to the choice of characters as tokens is still popular (characters are chosen as tokens simply because this is an established practice), the current author sees other reasons for this approach. Even though in (modern) Chinese language most words are bi-syllabic and written with two characters, their semantic construction depends on character semantics (see e.g., Packard, Jerome L. *The*

This is partly due to the unavailability of authoritative digital editions of most classical texts with marked-up word boundaries for research. An automatic word segmentation of classical Chinese texts is possible, but it could be ambiguous. Firstly, there are no clear word boundaries in written texts (classical or modern), and it is often hard to automatically identify "orthographic words"[6]. Secondly, there is an ongoing discussion on the definition of words in modern as well as in classical written Chinese[7]. Human experts, who mark-up corpora identify words in modern Chinese texts better than computers can; however, researchers have reported disagreements also between human experts on supervised "word segmentation," at varying but significant rates[8].

Morphology of Chinese: A linguistic and cognitive approach) Therefore, the character still could be a token in its own right, even though it is often (even in classics) part of a bigger word.

[6] Sproat et al., "Stochastic Finite-State Word-Segmentation", 378.

[7] That 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 to be an especially intuitive concept in Chinese language, 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 corpora, e.g., Sinica Corpus, are defined as "word-based" corpora 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, multiple character word-tokens practically do not apply. E.g., as late as in 2012, Lee ("Classical Chinese Corpus", 76) mentions that the status of "wordhood" in classical Chinese still needs consensus. In his own work, Lee generally states that "following the practice of the Academia Sinica Ancient Chinese Corpus, each character is initially presumed to be a monosyllabic word." (Lee, *ibid*, 78).

[8] Sproat reports that human judges disagree on many cases, and the agreement rate is 76% (Sproat, *ibid*, 394); i.e., every fourth word is arguable. 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.

This ambiguity will affect the accuracy of text length measurements in words. For a text of a certain length in characters, word segmentation algorithms (or even human experts) would produce different lengths in words and vocabularies. While such a difference could be negligible for very large modern corpora, it is important for smaller classical Chinese corpora, where problems of punctuation and word segmentation have always existed. To implement "word" as "token" for classical Chinese texts, corpora should be provided with a stable vocabulary of words and be properly marked up. Currently, only a few classical texts with such markup exist[9].

Therefore, although it is possible to calculate the length of classical Chinese texts in words, in most current studies text length is still calculated in characters[10]. After all, single characters satisfy most definitions of "word," as was especially analyzed by Linda Wetzel (Wetzel, "Types and Tokens: On Abstract Objects," 114-116). This article will continue to utilize characters, not words, as main measurement unit (token).

1.2 Two research modes The proliferation of digital (online) text repositories, concordances, and text statistical analysis, which is a comparatively recent phenomenon, has started to affect the more traditional ("philological") approach, thus introducing new perspectives to text studies.

[9] While word segmentation is important for syntax analysis, character approach may be 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 content in text documents. The computational evidence presented in this paper supports the argument that Chinese characters can be used as the basic semantic units in Chinese language modeling. (Zhao et al., *ibid*, 156).

[10] In the 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 for classical Chinese texts. They compared it with the data from 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).

It is possible to identify two approaches (modes) to digital text statistics. In a "philological approach," one would ask, "is this specific character (word, phrase) used in this text? In which version could it be found? How many times? Combined with what other characters? What other texts contain this character?" In a quantitative linguistics approach (corpora studies) one would also ask: "What are the most frequent characters in the text? How they are distributed? How many types (or type tokens) are there? How many hapax legomena (hapaxes) there are, and what is ratio of them and non-hapaxes? What is the distribution of sentence and word lengths?"[11]

It is not an exaggeration to state that the existing online classical Chinese corpora are mostly used by researchers for quick searches and comparative analysis of characters and expressions. However, as online resources are inherently inaccurate, studies verify results against paper sources. Meanwhile, the use of online (digital) corpora for quantitative linguistic research is still insignificant.

1.3 Digital corpora and quantitative linguistics. Most "philological" questions can be answered by paper-based concordances, but only digital corpora can provide answers in a quantitative approach. Not surprisingly, collecting information on text lengths and frequencies has been closely related to the development of electronic corpora of classical Chinese texts. Text data, not tractable for non-machine corpora studies, could be easily processed by computers[12]. The creation of electronic corpora rendered retrieval of statistical information from paper sources almost obsolete.

[11] The difference could be seen in two series of paper concordances for classical texts, HY and ICS. The ICS concordances (based on an electronic database, later to become the foundation of the CHANT online system) feature text lengths and type lists with frequencies. The same information could have been provided for HY, but it was not. One reason could be the difficulty to define a text without character variants and emendations.

[12] The total size of pre-Qin and Han classical texts could be roughly placed between three and eight million characters, according to the evaluation of the Academia Sinica project team (see below), which is not a huge amount.

It is also important to identify the form of electronic text that is used for length counting. First of all, the text should be cleared of punctuation symbols (similar to Western corpora). Second, the researcher has to decide whether such text elements as chapter titles should be included into the length count. This paper will consider text lengths where titles are removed. Keeping titles, especially for pre-Qin texts, will skew some hapax legomena statistics (as some singleton characters could be repeated in a title) and numerical statistics. It is probably useful to discern between concordances for "literary text" versions (including titles, etc.), and basic versions that are stripped of titles.

1.4 Text variation and digital corpora. Pre-Qin texts are an imminent part of every classical Chinese corpus, 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 to indicate which version was used). Computational linguistics is more experimental, and researchers need to be able to have access to corpora and repeat experiments. The 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 the most popular available online corpora, as well as on the Warring States Workshop (WSW Ctexts) research system.

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

Digital corpora have been created either through manual data entry or through an optical character recognition (OCR) process, followed by multiple reviews. The development of Chinese language OCR software started in the 60s, but commercial technologies became available only in the 90s. The OCR technologies for Chinese (Cherniet-2007) were not very accurate until the end of 90s; by this time, many academic corpora had been created through manual data entry[13]. All earlier OCR-based databases that utilized a low-accuracy OCR approach may contain a considerable number of errors, even after multiple reviews. However, manual entry also causes inaccuracy that could persist even after multiple reviews.

Data entry errors could be gradually corrected; however, it means that the contents of these sites may be in permanent change (while changes are not always announced)[14].

Another source of the content gap that plagues digital sources, especially earlier ones, is the limitation inherent in presenting 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 to a specific number of characters, represented by so-called code pages. This issue has not yet been resolved, not even by the introduction of Unicode. Therefore, practically all academic groups that have created digital versions of classics introduced some modifications to printed versions during the digitization process; hence these versions, while based on well-known editions, represent their own versions[15].

---

[13] 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 the 60s (Dai et al., "Chinese Character Recognition").

[14] 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 all changes are documented and available for reviews and corrections.

[15] For an introduction to code sets for Chinese characters and a description of the problem of missing rare characters, see Zhao and Zhang, "Totality of Chinese Characters". For a description of creation methods of those

While the article demonstrates variations in the length of digital versions of the same texts, it does not analyze the causes of the variety in specific texts, some of which were named above. In most cases (especially in regard to so-called "research corpora") this is simply impossible, because the texts are not available for analysis, and this raises again the issue of free online availability of texts for experiments.

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

The rest of the article will be structured as follows. Section 2, "Online Corpora of Classical Chinese," reviews the most important online classical Chinese digital corpora, their origins, and what role they could play in quantitative studies. It will introduce development processes for the digital corpora, namely, how text lengths are going to be collected and what problems are going to be encountered. Section 3, "Text Length as Indicator of Variety," will investigate how the lengths of classic texts in characters were measured, what lengths are available from digital corpora and from earlier editions, and what variety these lengths demonstrate. The results are discussed in the final section, "Conclusion."

---

rare or obsolete Chinese characters (almost four thousand) that are 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 the digitalization character substitution process. These problems are also addressed and 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.

2 Online Corpora of Classical Chinese

2.1 Literature review A small number of articles describe the general evolution of Chinese electronic corpora; most of these were published in the second half of 1990s and the first half 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 the development of Chinese electronic corpora[16]. The most comprehensive report (published in 2006) was written by Feng Zhiwei (Feng, "Evolution and present situation"), and describes the development of Chinese corpora from the beginning of the 20th century to the mid-2000s[17].

These articles focus on modern Chinese electronic corpora and only cursory mention classical Chinese corpora. A certain amount of information on classical corpora is contained in Wang Jianxin's 2001 article (Wang, "Recent Progress"), which describes the early stages of electronic corpora development in mainland China and Taiwan. His list includes the *Siku quanshu* electronic database (about 800 million characters), the Scripta Sinica (140 million characters), and the Shanghai Normal University corpus (100 million characters, containing a classical Chinese section)[18]. A similar short description can also be found in the introduction to McEnery and Xiao (McEnery and Xiao, "Lancaster Corpus of Mandarin Chinese")[19]. Some information on

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

[17] A concise (not up-to-date) list of corpora can be found in Yang Xiaojun's article (Yang, "Survey and Prospect").

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

[19] Very helpful (however concise) information is often featured on university libraries' websites. E.g., Berkeley has a very good 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 the Indiana university article by Liu Wenling (Liu, Commercial Databases").

electronic classical Chinese corpora is presented by a few articles dedicated to research corpora, which are discussed below.

2.2 Electronic corpora of classical Chinese

This study will start with corpora and concordancers that are available online. Although there are many websites featuring classical texts, this article will deal only with those that provide advanced corpus linguistics tools and features, as well as full-text character search options[20]. Besides these online corpora, some off-line research corpora that provide information on classic texts length will be described.

The most important online resources (and digital resources behind them) that feature advanced search and statistical tools for classical Chinese texts[21] are (in chronological order): 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 *Chinese Text Project (CTP)*[22]. Not surprisingly, four out of these five are hosted by mainland Chinese and Taiwanese academic institutions, one (CTP) is hosted by a private organization (also based in Hong Kong), and one (TLS) is hosted by a European institution[23]. The corpora behind CHANT and Academia Sinica resources have been digitized from the second half of the 1980s, and two last resources were started in the second half of 2000s.

---

[20] 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 the "Palace Museum Classical Chinese Database" will not be addressed in this article, and neither will Wikisource.

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

[22] The electronic corpora for modern Chinese texts, like Penn Corpus, etc., will not be reviewed in this paper, as they are unrelated to its subject.

[23] "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 only cursorily touch upon the subject of the many available online electronic texts (some of which having full-text search options).

Online corpora for studies of classical Chinese heavily depend on the availability of digitized texts and the quality of texts. Digital classical Chinese texts were first produced from around the mid-80s, as soon as the electronic standards for Chinese character coding were introduced. At this time, mid-range and personal computers became increasingly available to researchers, and this led to the proliferation of digital versions of Chinese classics. Most major classic collections were digitized in the 1990s (e.g., *Siku Quanshu*), often on a commercial basis[24].

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

Digital corpora of classical Chinese texts can be online or offline resources. It would be safe to say that most corpora that originated as offline resources sooner or later went online. However, it seems not many online versions of corpora are available offline (or, available for download as texts). Sometimes projects transfer their data to other projects (e.g., the Scripta Sinica project to Academia Sinica, or CHANT to TLS), but these occassions are rare.

Most popular online electronic corpora[25] can be roughly divided into three major groups: academic, independent, and commercial. Academic corpora are usually a product of a large body of researchers, who are supported by universities or academic resources. They may

[24] While texts themselves are freely available in block-prints, or xylographs, the digitalization of them (especially in pre-OCR times) is a time-consuming and expensive process, hence produced versions are expensive.

[25] Digital versions of printed collections, like SKQS, even though available online now, have been excluded from this list.

require subscription, but the price is not very high and most members of the research community have access to it (but cannot experiment with the source). Independent resources can be academically affiliated (e.g., CTP and WSW Ctexts), but they are not supported by large research resources[26]. Finally, commercial corpora can be produced by academics (e.g., the Erudition database[27]), 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 at the Institute of History and Philology (IHP) of Academia Sinica since 1984[28]. Researchers initially planned to create a digital version of the 25 dynastic histories for a study of Chinese economy. This was definitely pioneering work (and arguably the oldest classical Chinese digital corpus[29]). The texts were entered manually (OCR was not available at this time), and went through a multi-pass verification process. Soon, *Shisanjing* was added to the 25 histories. These texts became the core of the future electronic database. The creation of digital corpora was enabled by the advancement in computer science and electronics: BIG5

[26] However, they could be crowd-sourced (e.g., Wikisource, partly Sturgeon's CTP).

[27] 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 that 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).

[28] Probably, it was initially part of the joint project with the Library of Washington in 1984-1985 (see Wu, "Twenty-Five Dynastic Histories" about the library's participation).

[29] Paul Thompson mentions that as early as 1979-80 there were attempts to create classical digital 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).

coding was introduced in 1984, and computers became available to institutions. At this time, BIG5 did not contain many characters (13,051[30]), so there should have been substitutions for missing characters (Juan et al., "Resolving the Unencoded Character Problem"). The database continued to grow, and it was eventually released on the web (in 1997, Liu, "Impact of Digital Archives," 4), where it became known as the Scripta Sinica corpus[31]. This resource does not provide information on the length of specific texts or word markup.

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

Academia Sinica Corpus Shortly after the initiation of the 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 texts, as a part of their bigger corpus of modern Chinese texts[32]. The group managed to receive the core of IHP database (1.5 million characters) as an intra-academia transfer, and then entered another 1.5 million characters by themselves, also manually. This corpus later became the database of Academia Sinica. It is not clear whether the texts added by this group were modified in the digitization process. The IIS was probably the first group to provide an estimate for the entire scope of pre-Qin corpora as

[30] GB-2312 contained even less, 6,763 (see e.g., Juang et al., "Resolving the Unencoded Character Problem").

[31] 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").

[32] An interesting material on details of creating full-text search tools for Academia Sinica data (actually, 24 histories, 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)

3 million characters[33]. This resource does not provide information on the length of specific texts.

CHinese ANcient Texts (C.H.A.N.T.) At 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 the term "database" for what later became called "corpus"). The initial goal of the project was the continuation of the 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 released online and became the CHANT database. The source was also modified (improved) during digitization. The data were entered manually. The first implementation of this electronic database was a (pre-web) series of printed concordances–this was the first time concordances to classical text were based on their electronic versions[34], and can be considered as the end of the era of manually-created concordances. The CHANT project provides information on text length and the number of type-tokens (types).

Beijing University Corpus (PKU). This is the only well-known academic project on classical Chinese texts that has been developed in mainland China. The project was initiated in Beijing in the beginning of the 2000s, and was abbreviated as "PKU" in "Peking University" spelling[35]. It is not clear if compromises were made when coding pages

[33] "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 8 million characters.

[34] 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).

[35] See general description in Zhan et al., "Recent Developments". It was developed jointly by the "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). It was initiated in 2003 as a part of one of four

containing a limited set of characters were reworked, or when UTF and the more advanced Big5 and GB coding became available. PKU provides information on text length; however, it reports data on file length in kilobytes, not in characters[36]. Therefore, it was not possible to use this information in this article.

Thesaurus Linguae Sericae (TLS). Although TLS claims to be a "dictionary," an "interactive database," or a "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 it is the only large academic collection of classical texts online created outside China. Its development is unusual, because the input work was distributed among dedicated specialists, who curated their specialty texts[37], and data entry was done in Unicode[38]. 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 the length of specific texts.

2.2.2 Independent corpora

Chinese Text Project (CTP). 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, the text entry is based on OCR-

---

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).

[36] The data on frequency and text length 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.

[37] It could be that the Erudition database was built in same way, but there is not enough information.

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

processed digital versions of old printed sources (that solves copyright issues). Presumably, there were not many work resources, and the accuracy of the online texts may not be very high. However, it is a free resource, with a community formed around it that constantly improves the quality of the texts (but not in a "Wikisource style," i.e., correctors cannot fix errors themselves[39]). According to personal observations of the present author, CTP is the most popular source for informal references to classical Chinese texts among Western researchers[40]. This resource does not provide information on the length of specific texts.

WSW Chinese Texts (WSW Ctexts). This is a research corpus with focus on *Shisanjing*, and it currently does not feature many texts. However, it provides the most extended set of tools for linguistic research that is currently available online. The source of the texts is Wikisource (see the resource for specific links)[41]. This resource provides information on the length of specific texts and their vocabularies.

2.2.3 Commercial corpora

From personal observations of the author of this paper, despite all academic databases that are still online and some new texts that are being added to them, it seems that their interface has not changed much from the beginning of the 2000s. Meanwhile, from the mid-2000s onwards, considerable progress has been made in the development of online digital corpora of classical Chinese texts by commercial

[39] 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, these could be other online corpora, like Sinica. Some texts could be automatic conversions of GB codes into Unicode. Therefore, its accuracy may be no higher that of Sinica, etc. But its copyright policy allows it to be used for free, and texts, unlike for Sinica, etc., are gradually cleaned up by the community (similar to CTP, but correctors can do it themselves for Wikimedia, which simplifies the process.)

[40] Unfortunately, it does not feature information on text length.

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

companies. As early as the 90s, there were several commercial projects that sold digital versions of classical texts (e.g., *Sikuquanshu)*, but they could not compete with the academic online corpora at that time.

Since the mid-2000s, however, it seems that commercial projects have taken the lead[42]. There are two leading commercial projects in the area of classical Chinese texts: Unicode Inc., which produced two online databases ("Unihan" and "Wenyuan Ge Siku quanshu"), and Erudition, which produced the "Erudition Database," as well as "Hytong." "Unihan" presumably has good accuracy, and also has been using Unicode coding from its initiation. It is interesting that "Erudition," which claimed to have reached the level of precision of printed texts, started with an OCR approach similar to Unicode Inc., but switched to manual entry. This probably means that even modern OCR precision was not satisfactorily[43]. It should also be noted that commercial companies tend to produce full-text search systems rather than online corpora.

While it is possible that currently these commercial corpora are the most advanced sources for classical Chinese corpora, the Western research community does not seem to be using these tools more than academic or independent tools[44]. It seems that the future may belong to independent or free corpora, as it is difficult to imagine that the

[42] Some of them were initiated earlier, but they were not as successful, e.g., Guoxue baodian (see critique in Yang, "Chinese Classic Text Database").

[43] See Yang, *ibid*. The author of this paper did not have access to either commercial source, and there is no available public data on their statistics; therefore, these data are not featured in this article. One author mentioned the 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 the Song era some library owners valued manual copies over printed, because it allowed them to collate books in the process of copying (Winkelman, "Imperial Library", 28).

[44] 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, however, that the "Erudite" database is now officially a quotation source in China.

international academic community will be using pay-walled resources for research, which are not available to everyone.

3 Text length as indicator of variety

The commercial digital corpora probably reached a very high degree of accuracy, but scholars still check their quotations of classics in printed versions of the texts. However, it is highly unlikely that modern researchers will calculate text length by using a printed text. In fact, no article could be found reporting on text length based on any modern printed edition[45]. Moreover, because all computational linguistics experiments will be run on digital corpora, basic characteristics such as text length and type lists will have to be calculated using these corpora, not printed versions.

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

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

Eventually, the study concentrated on another question, which became its main subject: "what information is available on the length of *Shisanjing* texts and how do WSW Ctexts data relate to it"? It is well

[45] The opposite is the case, in fact; the latest printed concordances (the ICS series) are based on electronic versions of texts. It also seems that traditional calculations were made not by using printed editions, but "stone classics" *shijing*.

known that there are many versions of classics, and they often differ considerably. Text length varies for many reasons: it is affected by inclusion of "secondary" text components into the count, such as text title, chapter titles[46], as well as punctuation and non-character symbols. These issues have been approached differently by researchers; however, not all reported on what method they used.

Currently, there is no available comparison between the lengths of Shisanjing texts. What should be expected in terms of variety, and how does it affect quantitative linguistics studies? This paper will try to fill this gap, as well as to delineate the data framework and bring up the numbers for further evaluation.

3.2 History of text length measuring. It is possible to identify several periods in the quantification of classical texts. Feng, in "Evolution and Present Situation," divides the 20th century into three periods (from the 20s to before 1979, from 1979 to 1991, and the modern period), starting with the first frequency lists[47], then moving to the 80s, when digital versions first became available[48]for frequency studies, and finishing with the modern period[49]. A similar description can be found in the introduction to the paper of Zhang et al. (Zhang et al., *ibid*).

Expanding the time frame, it is possible to define four chronological periods: the traditional period, the modern period, the

[46] In manuscripts, book chapters are often untitled (e.g., Richter, "Textual Identity", 212), and text divisions are ambiguous, but it is not a rule. Definitely, in later times, chapter titles are found more often.

[47] 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).

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

[49] The year of 1991 marks the time when the National Chinese Corpus was initiated (Feng, "Evolution and Present Situation", 181).

early digital period, and the mature digital period. In a way, the entire time from the beginning of literacy to the appearance of the first concordances could be called the "traditional" period. The modern period constitutes the period of paper-based concordances. The early digital period began with the digitalization of texts and computer analyses of electronic versions of texts. Finally, the current period is defined, in addition to electronic texts, by online concordances, and especially by the wikifying of online editions.

3.2.1 Traditional period. Chinese bibliographical descriptions (especially those in dynasty histories), starting from *Han shu*, describe the size of books by the number of *pian*, *juan*, and *ce*[50]. The length in characters is not present in bibliographical descriptions, probably because bibliographers perceived manuscripts as "books" or "works," not as abstract "texts."

This does not mean that 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[51]. However, these numbers were not entered into bibliographical descriptions (even if they were present on a copy used by a bibliographer, and even though they could be more important for "version control" than "chapters"). Nonetheless, these

[50] Tsien addresses these units in a special section (Tsien, *Written*, 120-122). However, he does not pay special attention to character count, which is sometimes written on manuscripts. In the West, the 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.

[51] E.g., Richter, "Punctuation", 9, reports that in Mawangdui manuscripts text length in characters is sometimes found at the end of the texts. Interestingly, Tsien (Tsien, *Written*) does not mention it. Also, Loewe, "Early Chinese Texts", 8, mentions that for a 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 exactly sums up the number for chapters, "testifying that they were taken from a single source".

numbers were often known to scholars, but ignored by bibliographers[52].

The creation of "stone classics" (*shijing* 石經) played an important role in the history of calculating text length in characters[53]. Winkelman notes that "stone canons" functioned as publicly available authoritative texts (Winkelman, *ibid*, 32), and, despite the fact that the government later moved to wood-block-printed versions as standard canon texts, most of the known traditional records of numbers are based on calculations using these stone canons, not manuscripts or wood-block-printed books. The character numbers, written on manuscript covers, disappeared.

Song dynasty data The earliest consistent measurements of length in characters of several classical texts from *Shisanjing* that were located by the author of this article are dated to the Song period[54]. It seems that the recently created system of thirteen classics (twelve of which were displayed on *Kaicheng* stone classics, 833-837 CE) and their role in state examinations prompted Song scholars to estimate the 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)[55] contains a chapter ("Jiu jing zi shu" [Numbers of Characters of Nine Classics] (CYKN, 264), on the length in characters of thirteen

[52] 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.

[53] See, e.g., Tsien, *Written*, 78-83.

[54] 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.

[55] 阮葵生 (1727-1789), see Wang, "Kuan Kuishen Nianpu" for more details.

classics, quoting numbers, some of which presumably were calculated in the Song period by Zheng Genglao (鄭耕老) in his "Quan xue" (勸學)[56] (this and following data are provided in tables in Appendix I)[57].

Qing dynasty data. More data are available from the Qing period (see Appendix I for numbers). These data seem to have been 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) [58] in the treatise "Jing Yi Kao" (經義考)[59]. Qian Taiji (錢泰吉) (1791-1863), in his treatise *Pushu zaji* (曝書雜記)[60], quotes numbers, calculated by Zheng Genglao as well by Wu Yingdian (武英殿), who was using Qianlong stone classics.

3.2.2 Modern period. In the 1920-30s, Chinese philology began to create Western-style concordances of classical texts. In the early stages of this process, concordances were created manually and creators of concordances either did not perceive texts from the point of view of their length in characters (or words)[61], or deemed such calculations impossible due to text variations. These concordances (Harvard-Yenching series) did feature neither text length in characters, nor frequency lists. Moreover, there are no reports about the number of

[56] 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.*

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

[58] See e.g., Jiang, "Cheng Yue Chunqiu", 186.

[59] 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.

[60] PST, "Shisanjing zishu", juan 1, 2-4. These data are also referenced by Wang, who relates them to an edition of "Shisan jing zhushu", in the earliest of PRC publications on the length of classics (contains some discrepancies).

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

characters of classical texts in printed editions. Meanwhile, the paper concordances were an ultimate answer to most questions that a classic philologist would want to ask.

3.2.3 Early digital period. In the 1980-90s, the early digital period, the situation had improved. The ICS concordances were published, based on electronic databases, where the number of characters in texts and the volume of vocabulary were indicated–probably, the first time since the early, traditional calculations[62]. These texts became the foundation for later online concordances, such as CHANT and Scripta Sinica.

In the modern period, with ubiquitous internet presence, many independent online electronic editions and concordances began to appear alongside older online concordances, which are extensions of earlier electronic databases. The most notable are the Wikimedia and CTP resources. As a rule, they do not use existing electronic media[63], but re-scan similar or same editions. This means that there could be more mistakes in these systems. In Wikimedia (and to some degree in CTP), 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, the same is applicable to "official" online versions and concordances. The list of these concordances is provided in Appendix II.

Finally, in the mid-2000s, a slowdown has been observed in the development of non-commercial electronic (and online) databases of classical Chinese texts. It coincides 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.

[62] 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.

[63] It quite possible that texts in Wikisource are "borrowed" (at least partly), from Scripta Sinica or CHANT, but as far as they are featured on the site, they are covered by the Creative Commons license, and are available to all researchers.

It is possible that private enterprises will intercept academic activity in this area, as might be signaled by the roll-out of the Erudition databases. 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 restrictions. Neither type of resources are available for other researchers.

3.3 Modern digital research corpora. As the development of academic corpora and the research activity in respective centers starts to slow down by the mid-2000s, the activity in quantitative linguistics on classical Chinese texts is moving into smaller research groups. These groups, despite themselves being academic, cannot re-use existing academic corpora for their experiments, and therefore 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)[64]. Four of such efforts are described below (in chronological order), and their data are reflected in the tables of Appendix I.

Guo Xiaowu has calculated the highest frequency of characters in classics, looking for the most frequent characters in classical Chinese texts. He provides data on the length of texts with and without punctuation, as well as the number of characters (Guo, "Gudai Hanyu," 73, fig.2-2). Guo claims that his corpora are a selection of existing data[65], with no exact indication which texts came from what source, and how they were processed.

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

[65] They could be called "non-identified" as "most are 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.

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 whether through manual data entry or OCR) and went through several checks (Qin, "Xianqin guji," 112). The researchers encountered typical issues (unencoded characters, etc.), and they claim that it was resolved through a manual statistical approach (Qin, *ibid*).

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

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

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

3.4 Set-up of data sources. This article will use eleven sources of classics text length data[66]. 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. This approach is very close to the method that was used by the author of this article for the WSW Ctexts data. However, their text versions may

[66] This article considered data only as "source" when 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".

sometimes have been different from versions that were used for modern digital corpora.

There is a gap in the available data between the 19th century and the1990s, because no researcher reported data on printed books[67]. Even printed concordances, created in the 20s-30s and later, do not feature text length and vocabulary data. When Wang Genbao (Wang, "Shisanjing jing") reported on the text length of classics for SSJZS edition in the 80s, he referred to traditional sources (Qian Taiji, PST). The next available data are ICS, reporting numbers for printed versions of digital texts, i.e., from electronic sources[68]. Other digital online corpora, such as Scripta Sinica, do not provide this information.

Finally, most data from research corpora was provided at the beginning of current century: Gou Xiaowu, Qin Qin, Li Xiang, Li Bin et al., and WSW Ctexts. If commercial corpora contain this data, they 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
12	**CTP**	2006

[67] Except estimates, like Tsien 2004, p.25 and Zhang Guogan's data on stone classics. Zhang's data (see Table 2) are very interesting, however, they are not considered in this article as a regular source.

[68] 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 include punctuation characters in one account.

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 and discussion. The numbers on text length in characters for classics, as well as on vocabulary volume for single characters, are presented in Appendix I. The data volume is rather small, and sources vary considerably, so it would be excessive to apply a real statistical approach. However, averages and standard deviations were calculated where available, to create a numeric framework.

In most cases, length variation reflects the differences in versions of texts and the choices of the edition's creators; however, there are many other factors affecting these characteristics (e.g., including in the count commentaries, punctuation, and titles of chapters, etc.) For example, ICS and CHANT often display the same number of types, but

different lengths[69]. Appendix I contains comparison tables of text length (and vocabulary size, where available) in characters for Shisanjing and Zhuang-zi texts, with a short discussion of these changes[70].

The range of the numbers' variation varies, sometimes considerably, sometimes a little, but it is clear that any quantitative linguistics features, as well as philological information, obtained from these corpora will be different[71].

Surprisingly, however, most text lengths of Confucian classics are falling within the standard deviation range (with exclusion of some obvious deviations from the populations), e.g., such texts as Gongyang and Guliang demonstrate good clustering. Meng-zi and Liji demonstrate closeness in variation. The least varied text is Lunyu. Shujing and Shijing demonstrate more variation.

The versions of some texts had to be excluded from the population, because they deviated too mch from other texts, e.g., Shijing in the version of Li 2009, and Shujing in versions of Guo 2001 and Qin 2005 (all new sources). Earlier versions of Xiaojing also had to be excluded, but it is an otherwise very consistent text.

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

That is why, for the WSW Ctexts site, Wikisource was chosen. Any researchers can check it out, copy the text, and use it in experiments.

---

[69] E.g., for Guliang, 42056 in the CHANT version and 40914 in the ICS version, with the same number of types, 1604. Most probably, the CHANT version counts punctuation.

[70] WSW Ctexts, which does not include repeated titles and punctuation, usually features the minimal length.

[71] This is especially applicable for Chunqiu and Zuozhuan. Despite Chunqiu and Zuozhuan being different texts, divided by a large time gap, they are usually 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.

However, the Wikisource texts could always be questioned for their reliability from a philological point of view, and downloaded texts should be updated from time to time, to be in synch with online versions.

4 Conclusions

It may seem that the transfer of Chinese classics from printed form to digital should be similar to such transfers of Western classics, and most discrepancies should be due to a typist's inaccuracy, which could be eventually corrected. The reality turns out to be different. While digital versions of classics could be very close to their printed sources, there are some differences that are still not corrected.

This study focuses on quantitative characteristics of digital texts, such as the length and vocabulary size of texts. It assumes the "character-as-token" approach, i.e., single Chinese characters, not words, are still valid quantification units for quantitative linguistic analysis. Therefore, text length in characters and the number of type-token characters are critical values for any consistent quantitative linguistics study. However, the conversion of Chinese texts into electronic form leads to ubiquitous errors and inaccuracies, and this, alongside modifications by researchers, creates a *digital content gap* between paper-based and electronic-based corpora.

This study did not analyze any specific discrepancies between concrete texts[72]. However, the analysis of the available data on the length of modern electronic corpora of Shisanjing shows that there is considerable variety in length. The scope of variation depends on the text. Some texts, like Lunyu and Xiaojing, show very little variation, while others, like Shujing and Shijing, display more variety. Sometimes,

[72] The comparison table of the top 100 most frequent characters for sources where these numbers are available (Appendix III) is revealing enough.

there is a historical tradition that includes commentaries (e.g., Yijing, and especially Chunqiu), so the numbers for the canon part itself and the "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 text Y" will get different answers for some characters, not only if one compares e.g. corpora of Li et al. and ICS or CHANT, but even for the same texts in ICS and CHANT. Appendix III presents a comparative table of the 100 most frequent characters in Ctexts, CHANT, and Li Xiang's dissertation[73]. The table demonstrates that even for the most frequent characters, the frequency numbers could be drastically different, as well as the frequent character order. 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 errors and inaccuracy, even more discrepancy is brought in by the differences in text versions and changes during the transformation process. Some databases, created in 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[74].

In spite of the digital content gap, any large modern quantitative linguistics study is 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 to

[73] Only Ctexts and CHANT readily provided statistics on all character frequencies in convenient form. Li Xiang's dissertation (Li, "Shisanjin jigao") is, arguably, the first study to present assembled statistics on Shisanjing in print.

[74] These electronic projects could be considered an editorial activity in progress, which has been applied only to electronic, not printed, media.

analyze texts. The quality of the electronic version of the text source plays a critical role in research accuracy. Electronic sources for reliable online corpora should be open to the academic community and, preferably, created by the academic community.

The ideal situation would be to have a standard and free digital corpus of classical Chinese texts. Definitely, any specific version of a canonical text could then be easily criticized in regard to various philological aspects. Therefore, two types of online corpora are desirable (or two versions of the same text). Philological research requires a multi-layered text, allowing for the display of various versions of parts of the text, as well as for character variants. For quantitative studies, a streamlined and simplified version could suffice. If such free standard versions come into existence, they should be supported by a body of experts through 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, an opposite trend has been observed towards the commercialization of digital resources. It is possible that commercial databases, like the Erudite database, provide more accurate digital corpora, but it does not seem that these databases are going to be available for independent examination and quantitative linguistic experiments any time soon. As a result, research groups in that area start building their own digital resources, and this leads to fragmentation of the field and the creation of many digital resources that differ from each other. Most available printed resources have been already digitized, but high-quality resources are mostly commercialized.

It can be stated with some certainty that online (digital) corpora of classical Chinese texts are being used by modern researchers mostly as convenient search tools. Information, obtained through such searches, is consequently being verified by traditional philological (printed) editions. Due to the digital content gap, digital corpora are not

very reliable sources of philological information, but their inaccuracy is not very significant for quantitative linguistic studies. However, they are not extensively used for this type of studies, and one of the reasons for this is their inconvenience and unavailability for this type of research.

This article tries to show that the results of quantitative linguistic study heavily depend on digital text versions. Creating a standard digital resource of classical Chinese texts that is open-sourced and available to the entire research community[75] will provide the proper level of reliability and repeatability.

Acknowledgements

The author would like to thank E. Bruce Brooks for his constant encouragement, inspiration, and support of the WSW Ctexts project from its very beginning. Lively discussions with Bruce and Chris Beckwith greatly accelerated the development of the project.

The project development put the author in contact with many people whose ideas and assistance had a great influence, especially, the ongoing dialogue with Rodo Pfister, who read an early draft of the text and suggested many changes. Wolfgang Behr and Matthias Richter provided valuable information and unpublished materials on manuscripts and character calculations. Donald Sturgeon also read the draft, and provided important information from his CTP resource.

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.

[75] 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).

While finishing the article, the author received the sad news of passing away of Prof. Tatiana Grigoryeva on December 22nd, 2014. Grigoryeva's personality and work have been an inspiration for the author for studying Chinese texts since the earliest stage. This article is dedicated to her memory.

APPENDIX 1 Text Lengths

1 Chunqiu

For each text, a table is created, displaying length numbers for various sources[76].
The data on length and vocabulary size of Chunqiu is only provided by WSW Ctexts corpus, other sources usually combine the text with Zuozhuan. The PKU offers number of bytes for 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	Comment
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	
12	**CTP**	**n/a**	n/a	

2 Zuozhuan

Only WSW Ctexts and LI_2013 treat it as a separate text. The PKU offers the number of bytes, 495584. 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	Comment
	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	

[76] Here and below in table headers, “source” means the source of text, “N” means the length of text in characters, and “V” is the size of vocabulary, i.e., the number of unique types.

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	
12	**CTP**	n/a	n/a	

3 Chunqiu and Zuozhuan

All sources, except Li_2013, provide numbers for this combination of texts.
Starting from this table, averages for text length and vocabulary size, as well as the standard deviation, will be offered. However, text length, taken into square brackets, will be taken out of population, due to various reasons (mostly, their versions deviate too much for other texts). Averages and standard deviation numbers are only given for comparisons; there is no significance tests for the data.
Qian's, CHANT's and Ctexts' versions are marked by sterisks, because their lengths are beyond standard deviation[77].
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
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[78]
9	**GUOXUE**	197294		Yin, "Guji shuzihua"
10	**LI_2013**	n/a	n/a	Li et al., ibid, 146

[77] Here and further such texts will be marked up with an asterisk sign.

[78] 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 Chuniu and Zuozhuan (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)

11	**[Ctexts]**	195354*	3251	
12	**CTP**[79]	197834	n/a	
AVG		196885	3271	
Stdev		1233	33	

4 Gongyang

Almost all data sources provide numbers for 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 following numbers: 27583 for Han stone classics (Zhang, *ibid*, 1:1a,b), 44748 and 44738 for Kaicheng and Shu stone classics (Zhang, *ibid*, 3:1a,b; 4:1a,b).

#	Source	N	V	Comment
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	n/a	Yin, "Guji shuzihua"
10	**LI_2013**	44366	1642	Li et al., ibid, 146
11	**Ctexts**	44224	1640	
12	**CTP***	45503*	n/a	
avg		44600	1645	
stdev		295	n/a	

5 Guliang

[79] The CTP data on length are not available directly, it is necessary to make a search call, e.g.,http://ctext.org/pre-qin-and-han?searchu=%E4%B8%AD&reqtype=stats. There is no data on vocabulary size.

It should be noted that standard deviation is larger for Guliang, and more values are beyond it, i.e., Guliang's population of lengths is not 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	Comment
1	**[Zheng]**	n/a; 41512 ruan	n/a	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*	n/a	Yin, "Guji shuzihua"
10	**LI_2013**	40913*	1593	Li et al., ibid, 146
11	**Ctexts**	40835*	1594	
12	**CTP**	41995*	n/a	
Avg		41523	1597	
Stdev		563	n/a	

6 Liji

Li-ji 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	Comment
	Zhang	57111	n/a	
1	**Zheng**	99020*	n/a	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	

12	**CTP**	98089	n/a	
Avg		98271	n/a	
stdev		379	23	

7 Lunyu

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	Comment
1	**Zheng**	12700	n/a	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	
12	**CTP**	15962	n/a	
Avg		15719	1353	
stdev		964	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	Comment
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	
12	**CTP**	35426	n/a	
avg		35251	1896	
dev		286	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	Comment
1	**Zheng**	39124	n/a	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	
12	**CTP**	30497	n/a	
avg		34683	2878	
dev		5112	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	Comment
	Zhang	18650	n/a	Qian, *ibid*., 1:1:2-4;
1	**Zheng**	25700	n/a	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*	
12	**CTP**	25796	n/a	
avg		26418	1989	
dev		1387	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	Comment
1	**Zheng**	1903	n/a	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	
12	**CTP**	1840	n/a	
Avg		1844	373	
stdev		51	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	Comment
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	
12	**CTP**	53917	n/a	
Avg		55097	1524	
stdev		1722	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	Comment

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	
12	**CTP**	49420	n/a	
Avg		49497	2208	
stdev		395	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	Comment
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	
12	[CTP]	28745		
Avg		22229	1360	
stdev		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	Comment
1	**Zheng**	n/a	n/a	
2	**Zhu**	n/a	n/a	
3	**Qian**	n/a	n/a	
4	**ICS**	65406	2937	ICS, *Zhuangzi*
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	
12	**CTP**	65242	n/a	
Avg		65056	2923	
dev		367	32	

APPENDIX II Electronic Databases and Digital Corpora of Classical Chinese[80]

Resource	Type	URL	Start
Scripta Sinica *Hanji dianzi quanwen ziliaoku* 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/ Paid	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's Chinese Text Project (CTP)	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 (includes ww.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

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

Project Descriptions

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[81]. 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 1986 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 media. 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[82]. 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."[83] (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

[81] 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.

[82] 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.

[83] 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)

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[84]. Unfortunately, there is no available information on text lengths and vocabulary.

Sturgeon's Chinese Text Project (CTP) 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[85]. 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 the present 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 Wikisource. 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

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

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

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[86].

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

APPENDIX III Comparative Table of Top 100 Characters[87]

Chunqiu Zuozhuan	1		Gongyang zhuan	2		Guliang-zhuan	3		Liji	4	
CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT
之 7349	之 7342	之 7357	也 1496	也 1496	也 1496	也 2125	也 2124	也 2128	之 4119	之 4126	之 4129
子 4986	子 4986	子 4986	之 1285	公 1351	公 1339	之 1289	之 1289	之 1291	也 2619	也 2620	也 2623
曰 3735	曰 3731	曰 3735	何 1134	之 1280	之 1281	不 961	公 1145	公 1133	不 2437	不 2436	不 2438
不 3635	不 3627	不 3639	公 1097	何 1134	何 1134	公 891	不 961	不 963	子 2040	子 2042	子 2038
也 3606	也 3598	公 3615	子 930	子 930	子 928	月 779	月 779	月 779	而 2022	而 2024	而 2026
以 3432	以 3422	也 3607	不 839	不 839	不 840	其 760	其 760	其 760	以 1925	以 1922	以 1924
而 3121	公 3403	以 3439	月 738	月 739	月 739	人 758	人 759	人 759	其 1629	其 1632	其 1634
公 3114	而 3128	而 3129	人 706	人 708	人 707	子 722	子 722	子 723	者 1513	者 1520	者 1520
其 2663	其 2659	其 2669	者 689	者 690	者 690	於 611	以 607	以 607	於 1450	於 1322	曰 1308
人 2577	人 2577	人 2581	以 660	以 659	以 658	以 606	侯 598	侯 600	曰 1303	曰 1307	於 1228
於 2090	于	于 1910	於 654	為	為 634	侯 598	者	者 516	人 1140	人	有 1140
晉 1823	晉	晉 1823	為 633	其	其 626	者 514	齊	年 515	有 1132	有	人 1137
君 1771	於	君 1776	其 626	侯	侯 564	齊 507	于	齊 507	則 1086	則	則 1087
有 1763	君	於 1773	侯 563	于	于 547	有 445	有	于 456	君 977	君	君 979
侯 1736	有	有 1768	齊 512	齊	年 547	而 428	而	十 450	大 944	大	大 961
為 1661	侯	侯 1743	言 447	有	齊 511	曰 418	曰	有 443	為 932	為	為 930
于 1585	為	為 1663	而 391	言	有 466	何 413	何	而 429	夫 760	夫	夫 766
大 1438	大	大 1441	晉 380	而	言 447	晉 361	晉	曰 421	禮 731	禮	禮 738
月 1418	月	月 1417	大 366	晉	十 436	正 318	正	何 413	天 673	天	天 675
師 1378	師	師 1379	曷 347	大	而 390	十 310	言	晉 362	故 620	故	故 622
王 1358	王	王 1362	有 343	曷	晉 380	伐 310	伐	正 319	無 522	無	無 521
使 1329	使	使 1333	君 341	君	大 372	言 310	會	言 311	所 499	所	所 501
齊 1320	齊	齊 1322	曰 335	曰	曷 347	會 310	十	伐 310	三 476	三	三 477
楚 1269	楚	楚 1270	則 331	則	君 341	宋 308	宋	會 310	喪 467	喪	喪 467
無 1170	無	無 1171	宋 328	師	曰 336	伯 307	伯	宋 308	祭 450	祭	祭 450
鄭 1166	鄭	鄭 1166	師 327	宋	則 331	師 300	師	伯 306	事 438	事	事 438
將 1037	將	年 1165	伯 319	伯	師 329	鄭 299	大	大 302	後 425	是	是 408
伯 948	伯	將 1038	年 305	年	宋 329	大 292	鄭	師 301	是 406	諸	諸 401
國 938	國	十 1027	伐 300	伐	伯 319	為 282	為	鄭 299	諸 399	下	下 388
諸 926	諸	伯 950	夫 296	夫	伐 300	年 273	年	為 282	下 386	民	民 384
焉 892	焉	國 938	十 296	十	夫 297	夏 272	王	王 272	民 383	後	樂 381
如 886	如	諸 927	鄭 291	鄭	鄭 291	王 272	夏	夏 272	樂 379	樂	行 378
夫 877	夫	焉 893	王 288	王	王 288	春 270	春	春 270	行 376	行	服 374
與 859	與	如 888	此 285	此	此 286	衛 267	衛	衛 267	服 373	服	與 360
伐 840	伐	夫 879	衛 282	衛	衛 281	秋 264	秋	秋 265	與 360	與	可 356
是 838	是	二 867	春 280	春	春 279	夫 262	夫	二 264	可 357	可	如 356
矣 830	矣	與 861	會 274	秋	秋 274	卒 256	卒	夫 263	如 353	如	士 345
可 820	及	伐 840	秋 272	會	會 274	君 248	君	卒 256	國 342	國	國 344

[87] The table presents data statistics, collected on Ctexts and CHANT in December, 2014. Li Xiang data see in Li, "Shisanjin jigao". Li Xiang only presents frequency numbers for the top 10 characters. Li Xiang's dissertation does not contain data on Zhuang-zi.

及 820	可	是 838	夏 267	夏	夏 267	冬 243	冬	君 249	士 341	士	矣 337				
宋 798	宋	矣 832	書 256	書	二 256	楚 234	楚	冬 244	矣 337	矣	父 332				
衛 778	衛	可 823	乎 245	乎	書 256	來 233	來	楚 234	父 332	父	食 331				
叔 766	叔	及 822	冬 239	冬	乎 245	孫 225	孫	來 233	食 330	食	必 329				
孫 756	孫	宋 797	孫 236	孫	冬 241	國 218	國	孫 225	必 330	必	日 319				
何 755	何	衛 781	與 225	與	孫 237	如 213	如	國 218	焉 319	日	焉 318				
我 747	我	叔 767	卒 224	卒	卒 224	日 202	日	如 213	日 317	焉	命 317				
十 741	乎	孫 755	楚 214	楚	與 224	葬 196	葬	日 203	命 316	命	上 307				
乎 739	氏	何 754	國 213	國	楚 214	二 195	二	葬 196	上 308	上	公 306				
氏 731	十	我 747	來 213	來	來 213	陳 179	陳	三 183	公 305	公	乎 306				
二 726	二	乎 739	婁 190	婁	國 212	則 176	則	陳 179	乎 303	乎	于 305				
故 680	故	氏 731	如 188	如	三 196	盟 172	盟	則 177	道 301	十	道 302				
年 679	命	三 703	二 186	二	如 190	無 162	無	盟 171	主 288	道	主 287				
命 666	年	故 680	陳 184	陳	婁 190	自 159	自	無 162	言 276	主	此 276				
吾 645	吾	命 668	葬 182	葬	陳 184	非 157	非	自 159	此 273	此	言 275				
盟 640	盟	吾 648	邾 180	邾	葬 181	及 156	及	於 157	侯 272	言	侯 272				
三 622	三	盟 640	無 177	無	邾 180	邾 151	於	非 157	親 269	侯	親 271				
夏 621	夏	夏 621	至 168	使	無 177	諸 149	邾	及 156	王 263	親	王 265				
者 620	者	者 620	使 167	至	至 168	故 147	諸	邾 152	十 263	王	十 263				
歸 588	歸	歸 589	歸 166	歸	使 167	曹 145	故	諸 149	五 259	五	五 260				
從 583	從	從 583	曹 165	曹	歸 166	殺 145	殺	故 147	一 258	哭	先 256				
陳 574	陳	陳 576	爾 165	爾	曹 165	至 143	曹	殺 146	哭 258	先	哭 256				
會 560	會	會 560	稱 163	稱	爾 165	三 143	三	曹 145	至 255	至	至 255				
請 554	請	請 560	盟 163	盟	盟 163	歸 141	至	至 143	母 254	母	母 253				
能 554	能	能 555	正 161	正	稱 163	天 140	歸	歸 141	先 254	謂	後 250				
必 549	則	必 549	自 158	自	正 161	使 138	天	天 140	謂 247	一	謂 250				
則 549	必	則 548	三 157	三	自 158	事 136	使	四 140	成 241	成	成 243				
若 540	若	若 542	諸 156	諸	諸 155	乎 133	事	使 138	中 238	中	一 240				
謂 531	謂	謂 529	殺 149	殺	殺 149	帥 129	乎	事 135	敬 230	自	中 240				
來 527	來	來 528	矣 143	矣	矣 142	入 128	經	乎 133	自 230	敬	自 232				
禮 526	禮	禮 528	天 141	天	叔 142	與 126	帥	帥 129	在 229	在	敬 232				
殺 514	殺	殺 516	叔 141	叔	天 141	父 122	入	一 128	知 228	問	在 230				
臣 512	臣	臣 511	未 139	父	父 139	矣 117	與	入 127	能 226	知	知 229				
事 505	乃	事 506	父 139	未	未 139	四 116	父	五 127	問 224	能	能 226				
乃 503	出	成 504	及 138	及	及 138	辭 115	矣	與 126	死 221	出	問 224				
出 500	春	乃 504	我 135	我	我 134	叔 113	辭	父 122	出 221	死	出 222				
春 499	秋	出 500	弒 125	弒	四 126	一 112	四	七 121	然 219	然	死 221				
秋 496	事	春 499	然 123	季	弒 125	所 108	叔	矣 118	婦 215	入	然 220				
入 484	成	秋 495	季 123	然	七 124	可 107	一	辭 115	入 214	婦	入 216				
自 478	入	入 484	日 118	日	季 123	我 105	所	六 114	皆 211	于	婦 214				
冬 477	自	自 479	可 116	於	然 123	奔 103	可	叔 113	廟 210	皆	皆 209				
行 470	冬	冬 478	入 115	可	日 120	五 103	我	八 110	月 206	義	廟 208				
在 468	行	行 472	譏 114	入	一 119	是 101	奔	所 108	齊 204	廟	月 206				
卒 467	所	所 469	帥 114	譏	入 118	志 101	五	可 107	用 202	月	齊 205				
所 465	在	在 468	出 107	帥	五 117	未 100	是	我 105	義 201	齊	小 203				
成 461	卒	卒 468	是 107	出	可 117	此 99	志	奔 103	小 201	小	明 202				
先 453	死	先 454	莒 103	是	六 115	七 99	未	志 102	明 200	明	義 202				
死 452	遂	死 453	七 101	桓	帥 115	莒 97	此	是 101	臣 195	用	用 201				
遂 450	先	遂 451	取 101	莒	譏 114	弒 97	七	未 100	立 194	立	立 196				
至 443	至	至 442	四 101	七	於 110	焉 96	莒	此 99	門 193	臣	臣 195				
民 436	民	民 435	吾 101	取	八 108	蔡 95	焉	焉 97	弗 193	弗	見 193				
告 433	告	告 433	氏 100	四	是 107	出 94	弒	莒 97	見 193	見	門 193				
吳 422	言	言 422	一 100	吾	出 107	內 92	蔡	弒 96	衣 191	門	弗 192				
言 422	吳	吳 421	立 97	一	莒 104	六 91	出	蔡 95	敢 191	衣	后 192				
知 414	知	知 414	五 94	氏	桓 103	道 90	成	九 94	地 187	敢	敢 192				
奔 398	文	文 413	執 94	立	取 101	外 88	內	出 94	使 187	二	衣 190				
季 397	奔	五 408	焉 94	焉	氏 100	八 88	六	內 92	何 186	使	使 188				
弗 380	季	季 398	奔 93	奔	吾 100	惡 88	道	成 92	非 184	從	地 186				
對 377	弗	奔 398	六 92	五	立 97	朝 85	外	道 90	從 183	地	何 186				
文 377	對	一 390	辭 92	執	九 96	敗 85	八	外 88	孔 183	何	從 186				

許 375	許	弗 382	滅 90	辭	焉 94	取 84	惡	惡 88	既 182	非	非 185
又 375	又	對 378	許 90	六	奔 93	侵 83	桓	取 85	反 182	孔	孔 184

(Comparative Table of Top 100 Characters continued)

Lunyu	5		Mengzi	6		Shi-jing	7		Shu-jing	8	
CTEXTS	LI 2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT
子 971	子 975	子 972	之 1901	之 1899	之 1900	之 1018	之 1391	之 1353	惟 647	惟 647	惟 682
曰 757	曰 759	曰 758	也 1232	也 1232	也 1232	不 629	不 1038	不 774	於 585	于 574	于 634
之 611	之 613	之 613	不 1082	不 1084	不 1085	我 586	有 952	章 732	曰 475	曰 472	曰 500
不 583	不 583	不 584	曰 956	曰 958	曰 958	其 542	于 923	其 674	不 409	不 409	不 479
也 533	也 533	也 533	子 938	子 945	子 942	有 534	其 832	有 654	王 407	王 408	王 464
而 345	而 345	而 343	而 772	而 774	而 772	子 456	我 818	我 595	有 394	有 395	有 455
其 269	其 270	其 270	以 636	者 639	者 639	於 375	惟 648	子 554	乃 364	乃 363	之 410
者 220	人 219	人 222	者 635	以 635	以 637	兮 317	王 605	也 535	之 332	之 338	乃 384
人 219	者 219	者 219	人 613	人 611	人 611	彼 306	子 600	以 444	厥 318	厥 318	厥 336
以 209	以 211	以 211	其 585	其 587	其 586	以 304	曰 552	句 385	其 287	其 287	其 317
有 199	有	有 200	於 516	為	為 514	無 296	以	人 373	天 273	民	天 312
矣 182	矣	矣 181	為 515	於	於 512	人 267	人	于 371	民 273	天	民 299
於 179	為	於 176	有 464	有	有 464	維 258	無	無 369	人 244	人	命 275
為 170	於	為 170	則 423	則	則 425	如 248	天	王 343	命 233	命	人 266
君 159	君	君 159	王 319	王	王 321	既 226	爾	兮 330	爾 230	爾	以 256
可 156	乎	乎 158	孟 308	孟	孟 307	爾 211	乃	彼 312	以 230	我	我 245
乎 156	可	可 156	天 292	天	天 293	矣 207	民	四 303	我 229	以	德 243
如 154	如	如 154	無 271	無	無 271	君 199	厥	君 276	德 222	德	爾 242
與 142	與	與 144	可 258	可	可 258	王 193	命	維 266	予 220	予	予 238
無 132	無	言 130	是 256	是	是 257	言 184	兮	如 255	無 193	無	作 222
言 130	言	無 130	君 254	矣	矣 253	在 170	予	矣 251	汝 175	汝	無 217
則 124	則	則 124	矣 253	君	君 252	心 168	彼	既 234	若 171	若	若 192
問 121	問	問 120	下 236	下	下 237	天 166	既	公 227	克 159	克	汝 181
何 118	何	何 118	所 235	所	所 237	是 149	在	三 219	罔 155	大	大 170
知 117	知	知 118	與 231	與	與 226	何 139	德	爾 218	大 152	罔	克 164
吾 112	吾	吾 113	民 209	乎	民 209	止 122	言	言 214	用 148	用	用 163
仁 109	仁	仁 109	乎 208	民	乎 209	為 119	如	天 208	作 142	作	罔 155
夫 107	夫	夫 105	何 199	何	何 199	方 113	維	是 197	時 140	時	子 154
焉 88	道	道 90	然 182	然	然 185	載 112	君	而 194	子 129	子	在 146
道 88	焉	焉 88	得 177	得	夫 177	思 108	大	在 185	帝 127	帝	帝 146
行 82	行	行 82	夫 176	夫	得 177	女 101	心	大 184	在 125	在	時 146
謂 78	謂	謂 78	大 165	大	大 166	來 100	矣	心 179	哉 117	哉	三 128
禮 75	必	必 76	我 162	我	我 162	予 99	克	思 161	明 115	明	明 127
必 75	禮	孔 75	仁 160	仁	仁 158	匪 98	公	國 161	邦 109	邦	哉 125
孔 74	孔	禮 75	道 150	如	如 153	民 98	方	十 153	弗 108	弗	公 124
斯 71	三	斯 71	如 149	道	道 150	大 98	作	何 148	三 103	三	自 119
三 70	斯	三 70	非 147	非	非 147	公 97	四	刺 148	自 100	五	周 118
能 69	能	能 69	謂 144	謂	謂 145	斯 95	用	為 147	一 99	一	邦 113
見 68	見	見 67	焉 140	焉	焉 140	亦 94	時	六 145	五 99	自	弗 112
學 65	學	學 65	能 135	能	能 135	莫 93	自	女 144	後 98	今	則 112
事 61	事	哉 61	行 133	行	行 133	四 92	若	民 144	今 97	公	言 110
哉 61	哉	事 61	吾 127	吾	吾 128	歸 91	是	二 138	言 96	四	一 108
是 60	是	是 60	心 124	心	國 125	也 90	汝	止 128	四 96	言	五 107
聞 59	公	聞 59	一 124	公	心 124	且 89	亦	則 128	公 96	則	四 104
未 57	聞	公 57	國 124	國	一 123	行 88	哉	方 126	則 94	茲	茲 102
公 57	未	未 57	公 121	一	公 121	此 86	為	載 125	茲 92	小	殷 101

路 53	路	好 53	言 118	故	故 119	命 85	則	行 124	小 91	上	小 100
我 53	好	我 53	故 118	言	言 117	樂 85	罔	焉 118	殷 87	下	下 99
好 53	我	路 53	見 115	此	此 115	自 83	帝	文 113	庶 86	殷	今 98
在 51	在	已 51	事 114	見	見 114	可 82	明	樂 112	百 86	百	百 97
得 50	得	在 51	善 114	善	知 114	則 81	何	一 110	上 84	庶	既 97
所 50	所	得 51	後 113	事	善 114	憂 81	邦	歸 109	呼 83	呼	上 95
已 50	已	所 50	知 113	知	事 114	曰 79	百	南 108	下 83	敢	朕 94
小 49	民	小 49	此 113	後	後 113	日 78	三	命 107	嗚 82	事	事 93
民 49	天	天 49	亦 110	亦	亦 111	靡 77	下	周 106	敢 82	嗚	庶 92
天 49	小	民 49	食 108	食	食 108	山 75	日	於 105	事 82	周	成 91
大 48	邦	大 48	義 108	義	義 108	將 75	周	斯 105	周 81	先	方 89
樂 48	樂	邦 48	今 106	今	今 107	與 73	弗	予 105	亦 81	亦	先 89
邦 48	使	樂 48	若 106	若	若 106	月 72	上	下 104	先 81	朕	非 89
使 47	亦	亦 48	問 105	使	使 105	南 71	小	來 104	朕 81	非	亦 89
亦 47	大	使 47	使 105	問	問 105	豈 71	今	自 103	非 78	方	敢 89
下 45	下	下 45	諸 103	諸	諸 104	德 71	載	能 103	方 73	文	呼 85
貢 44	貢	貢 44	必 101	必	必 102	百 70	文	山 102	文 72	既	嗚 85
從 43	欲	從 44	哉 101	哉	哉 101	孔 68	一	百 102	既 71	至	后 82
諸 43	從	欲 43	舜 100	皆	皆 100	國 68	來	作 101	至 70	刑	文 80
欲 43	政	諸 43	皆 98	舜	舜 100	中 67	止	者 101	越 69	越	心 74
善 42	諸	政 43	欲 97	欲	欲 96	實 66	思	將 101	刑 68	服	日 74
文 42	善	文 42	士 94	士	士 93	見 65	事	匪 101	服 67	敬	刑 74
政 42	食	食 42	樂 91	樂	樂 91	采 65	行	德 100	敬 66	日	至 74
食 42	文	善 42	齊 90	未	未 90	或 63	中	采 98	心 65	心	君 73
德 41	德	德 40	已 88	已	已 88	酒 63	庶	八 97	受 64	受	師 73
對 40	後	求 39	百 87	百	百 88	謂 61	殷	夫 97	日 64	保	為 71
惡 39	對	後 39	未 87	齊	聞 87	車 61	敢	莫 97	保 63	成	越 71
後 39	惡	惡 39	聞 86	聞	齊 87	胡 60	可	亦 97	正 62	正	敬 71
求 39	求	對 39	孔 82	父	父 83	者 60	東	憂 95	成 62	后	受 68
死 38	然	死 38	將 82	將	孔 82	而 57	月	月 93	師 60	多	服 68
然 37	死	信 38	父 82	孔	將 82	下 57	五	可 92	從 58	師	正 67
信 37	信	然 38	惡 80	惡	惡 80	哉 56	山	車 92	多 58	從	保 67
仲 35	由	由 35	自 75	自	自 76	弟 56	女	曰 91	降 57	降	夏 67
由 35	仲	仲 35	小 74	賢	小 74	周 56	士	東 90	士 57	君	從 64
非 33	一	非 33	賢 73	小	賢 74	侯 56	多	風 90	二 56	士	功 63
出 33	非	一 32	受 72	受	受 72	文 55	先	日 88	臣 55	二	臣 63
過 32	出	出 32	足 71	足	足 71	東 55	成	此 87	功 55	臣	二 62
一 32	過	過 32	至 70	至	至 70	乎 54	歸	中 87	夏 54	功	告 61
足 31	十	足 31	臣 69	臣	三 68	福 54	受	且 86	君 54	夏	降 61
雖 31	雖	雖 31	禮 68	居	臣 68	生 53	茲	時 86	土 53	罰	罪 61
居 30	足	居 30	居 68	禮	居 68	食 52	斯	美 85	為 53	土	多 60
夏 29	居	及 29	三 68	三	禮 68	父 52	匪	與 85	罰 53	為	東 60
父 29	父	父 29	侯 66	侯	侯 66	所 52	從	見 84	丕 52	中	土 59
及 29	及	夏 29	在 65	在	在 65	士 52	至	小 81	罪 51	丕	士 59
己 28	己	己 28	養 65	親	養 64	式 52	國	侯 80	告 51	永	而 59
予 28	夏	爾 28	親 64	去	親 64	皇 51	南	靡 77	東 50	告	萬 57
爾 28	爾	予 28	去 64	養	去 64	上 51	將	武 76	萬 50	罪	罰 57
色 27	予	友 27	五 63	五	五 63	誰 51	服	所 75	永 50	東	武 56
成 27	成	成 27	相 63	相	相 63	多 49	與	五 74	亂 49	萬	亂 56
友 27	季	色 27	中 62	中	中 62	俾 49	降	道 73	中 49	亂	咸 55
上 26	色	上 26	日 61	日	日 61	馬 49	萬	成 72	休 47	後	丕 54
遠 26	友	今 26	堯 60	生	生 60	明 49	莫	故 72	訓 47	訓	訓 54
齊 26	張	立 26	生 60	堯	堯 60	作 49	保	豈 71	辟 46	辟	永 53
中 26	臣	臣 26	猶 60	上	猶 59	邦 49	而	上 69	迪 45	休	中 52

(Comparative Table of Top 100 Characters continued)

Xiao-jing	9		Yili	10		Zhouli	11		Zhou-yi	12		Zhuang-zi	13
CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	LI-2009	CHANT	CTEXTS	CHANT
之 92	之 92	之 92	於 1391	于 1607	于 1470	之 2511	之 2515	之 2519	**也 624**	**也 960**	也 961	之 3087	之 3090
不 62	不 62	不 62	人 1264	人 1265	人 1288	人 1694	人 1694	人 1695	曰 515	之 612	之 614	而 2150	而 2158
以 52	以 52	以 52	拜 1106	拜 1106	之 1121	其 1399	其 1401	其 1404	象 446	曰 587	曰 587	不 2001	不 2011
也 46	也 46	也 46	主 1063	主 1063	拜 1106	以 1343	以 1342	以 1345	不 279	象 483	象 484	也 1684	也 1691
其 43	而 43	而 43	之 1004	賓 1037	主 1069	二 760	二 761	二 763	吉 252	以 474	以 474	其 1259	以 1260
而 43	其 43	其 43	西 917	之 1002	賓 1037	而 680	而 681	而 681	有 249	不 424	不 424	以 1258	其 1259
子 41	子 42	子 41	面 842	西 921	西 924	則 628	則 628	則 629	以 241	其 386	其 386	者 1165	者 1166
於 40	於 40	於 40	爵 754	面 841	面 845	凡 610	凡 609	凡 610	之 234	而 370	而 370	曰 1022	曰 1019
孝 34	孝 37	孝 34	者 707	爵 748	者 801	四 603	四 604	四 605	其 223	有 345	有 343	為 1012	為 1013
事 28	人 29	人 28	上 651	者 708	爵 751	大 580	大 579	大 583	六 216	无 300	□ 298	人 1006	人 1004
人 28	事	事 28	賓 610	上	以 714	掌 571	掌	掌 572	九 206	吉	吉 288	子 979	子 978
則 27	則	則 27	以 602	東	上 658	有 535	有	有 537	無 206	六	六 230	於 849	乎 837
者 27	者	者 27	東 602	以	東 602	士 522	士	士 521	利 179	大	大 224	乎 829	於 830
天 25	天	天 25	北 591	北	北 593	曰 474	曰	曰 473	大 167	九	九 219	有 808	有 817
故 24	故	民 24	升 560	升	不 587	十 457	十	十 455	貞 167	利	利 215	无 748	□ 784
民 24	民	故 24	階 547	階	升 568	者 440	者	者 445	而 163	天	天 214	天 674	天 675
親 23	親	親 23	坐 523	坐	階 546	事 391	事	事 396	上 159	人	人 209	知 616	知 619
敬 23	敬	敬 23	降 512	降	降 534	國 381	國	國 381	咎 143	子	子 192	所 581	所 584
父 20	父	父 20	不 503	不	坐 523	為 377	為	為 379	人 130	上	為 183	則 493	則 495
無 20	君	無 20	如 497	如	如 507	三 355	三	三 358	中 126	為	上 181	是 467	是 466
曰 19	無	曰 19	受 496	受	受 505	一 349	一	一 350	子 123	者	者 180	矣 454	矣 456
君 19	曰	君 19	執 465	執	大 470	六 338	六	六 334	行 113	貞	貞 176	與 451	與 451
下 17	第	下 17	夫 438	夫	夫 468	王 326	王	王 329	君 110	咎	咎 159	夫 441	夫 447
教 15	章	教 15	賓 428	南	執 465	下 317	下	下 318	三 98	故	行 153	吾 419	吾 423
德 14	下	有 14	南 428	大	南 426	中 312	中	中 313	用 98	行	中 152	下 412	下 412
行 14	德	行 14	大 422	皆	皆 401	五 296	五	五 296	天 98	中	故 152	然 410	然 410
有 14	教	德 14	皆 392	奠	奠 378	徒 290	徒	徒 290	于 98	君	君 151	大 381	大 367
可 13	行	可 13	奠 373	祭	祭 374	八 275	八	八 278	往 96	下	下 139	道 367	道 360
愛 13	有	愛 13	祭 369	尸	其 373	令 274	令	令 274	亨 96	乎	乎 131	謂 349	可 358
道 12	愛	矣 12	屍 363	射	尸 365	祭 268	祭	祭 269	凶 87	用	用 122	若 349	若 352
矣 12	道	道 12	司 343	司	也 350	史 260	史	史 259	得 84	于	于 120	可 348	謂 350
順 11	可	上 11	射 342	公	公 345	府 250	府	府 250	二 81	凶	凶 118	物 346	物 344
莫 11	大	大 11	公 336	其	司 343	祀 247	祀	祀 248	初 79	三	三 115	故 330	故 331
爭 11	矣	云 11	其 325	命	射 342	邦 241	邦	邦 241	四 75	可	可 110	非 318	非 319
上 11	敢	先 11	卒 321	卒	曰 339	於 234	于	民 225	五 73	往	往 109	得 309	得 313
先 11	莫	爭 11	命 321	興	命 329	民 224	民	于 218	剛 73	道	道 106	生 298	生 300
敢 11	上	莫 11	興 310	一	一 324	如 215	如	如 216	小 69	得	得 104	何 286	何 286
大 11	順	順 11	出 301	出	卒 323	及 213	及	及 212	下 69	地	地 103	言 265	能 267
云 11	先	敢 11	一 301	洗	子 316	夫 207	夫	夫 207	終 68	亨	亨 101	能 264	言 266
生 10	云	王 10	位 300	位	興 309	不 205	不	不 204	志 66	物	物 100	此 256	此 256
乎 10	爭	生 10	洗 300	答	出 304	禮 200	禮	禮 200	孚 65	四	終 97	至 253	至 253
為 10	治	言 10	左 284	觶	位 301	共 198	共	共 199	彖 64	終	四 96	見 239	見 236
王 10	乎	所 10	觶 283	左	洗 301	治 197	司	治 197	正 64	二	剛 95	一 233	一 234
所 10	明	是 10	入 280	入	荅 295	司 195	治	司 195	可 62	剛	二 93	行 222	行 222
言 10	三	為 10	婦 274	三	三 294	物 193	物	物 193	道 59	小	小 93	自 209	自 211
是 10	生	乎 10	三 273	婦	婦 292	用 190	用	用 191	位 59	易	則 90	德 205	德 205

詩 9	是	臣 9	俎 272	俎	觶 286	車 180	車	也 181	未 55	則	五 86	焉 203	焉 203
治 9	所	身 9	再 267	再	左 285	也 180	也	車 181	來 54	五	易 85	我 202	我 201
身 9	王	明 9	取 265	取	入 281	上 175	上	上 175	明 51	初	初 84	相 196	相 198
臣 9	為	治 9	門 260	門	俎 271	喪 174	喪	喪 173	地 51	位	位 82	已 190	已 193
母 9	言	非 9	子 254	子	再 268	百 162	百	百 162	在 51	見	見 79	君 190	君 191
非 9	臣	詩 9	立 250	立	取 266	賓 162	賓	賓 162	攸 48	德	在 78	心 187	心 185
明 9	非	母 9	設 249	設	門 260	胥 159	胥	胥 159	如 47	在	德 78	形 182	聞 181
夫 8	夫	夫 8	席 238	席	設 251	與 154	地	與 158	柔 46	正	正 75	聞 181	未 180
義 8	母	義 8	曰 237	曰	立 250	亦 154	亦	亦 154	王 44	是	來 72	未 180	死 180
嚴 7	身	三 7	在 231	揖	為 247	地 152	與	地 152	元 43	受	受 72	死 180	形 180
此 7	詩	地 7	揖 231	在	則 246	皆 152	皆	皆 152	悔 43	明	明 71	王 175	王 178
三 7	十	此 7	初 226	初	席 241	法 151	濿	法 151	順 42	來	是 70	今 173	今 174
能 7	義	致 7	下 226	下	在 239	方 149	方	方 149	乎 42	彖	矣 69	問 173	問 172
地 7	此	能 7	右 219	右	下 236	客 147	客	師 147	无 42	言	言 69	邪 163	邪 165
致 7	地	嚴 7	于 215	遂	揖 231	師 145	師	客 147	見 42	矣	彖 69	使 162	使 163
思 6	后	後 6	遂 214	辭	初 227	服 141	服	服 141	時 41	孚	孚 68	將 161	將 163
聖 6	能	思 6	祝 214	祝	右 223	諸 141	諸	諸 141	厲 41	志	志 68	聖 149	地 149
后 6	聖	然 6	辭 213	則	有 218	政 137	九	九 137	日 39	柔	柔 66	惡 148	惡 149
然 6	嚴	聖 6	則 213	自	遂 216	九 137	政	禁 136	當 38	所	所 64	地 148	聖 149
禮 6	致	樂 6	自 203	有	祝 215	禁 135	禁	政 136	自 37	未	未 63	亦 147	亦 148
樂 6	樂	雖 6	有 199	乃	辭 213	小 132	小	侯 133	復 36	知	知 61	始 147	始 147
雖 6	禮	禮 6	乃 198	反	自 207	時 131	侯	小 132	勿 36	動	日 60	萬 147	萬 147
心 5	然	心 5	反 196	禮	從 199	侯 131	時	時 131	則 35	日	動 60	足 143	哉 144
哀 5	思	日 5	為 190	食	乃 199	辨 130	辨	辨 129	乃 35	悔	時 59	哉 142	足 143
自 5	四	失 5	食 187	爲	君 196	馬 127	馬	正 127	失 34	乾	悔 58	彼 142	彼 142
日 5	雖	在 5	某 180	從	反 195	正 126	正	馬 127	觀 33	時	乾 57	成 139	成 141
長 5	五	自 5	禮 178	某	食 190	會 123	會	會 123	出 31	與	與 56	日 138	且 139
失 5	至	至 5	中 177	介	中 187	入 119	入	入 119	亡 31	成	成 55	且 138	日 137
悅 5	宗	何 5	介 177	中	某 180	若 115	官	若 115	事 30	於	於 53	名 133	名 133
何 5	哀	居 5	薦 174	薦	外 178	歲 114	若	刑 114	德 29	如	如 52	世 131	世 131
宗 5	安	法 5	外 172	外	而 178	刑 113	歲	歲 114	征 29	順	順 51	樂 123	莫 126
祭 5	長	長 5	君 164	士	禮 178	官 111	刑	氏 111	過 28	元	元 50	莫 123	樂 124
曾 5	得	哀 5	正 164	酒	介 177	氏 111	氏	謂 111	矣 28	變	萬 50	明 123	明 123
終 5	法	悌 5	阼 164	君	薦 175	謂 110	謂	官 111	說 28	後	後 49	中 122	事 122
國 5	國	悅 5	士 163	正	與 173	行 107	日	日 107	命 28	萬	變 49	事 122	上 121
安 5	何	得 5	酒 162	阼	服 171	日 107	行	行 107	光 27	觀	觀 49	上 121	義 121
謂 5	祭	祭 5	實 162	實	弟 168	命 106	命	命 106	應 27	事	謂 48	方 119	中 121
在 5	居	終 5	送 161	送	阼 167	無 105	無	無 104	吝 26	謂	攸 48	無 118	方 118
法 5	名	善 5	而 160	而	正 165	鼓 103	鼓	長 102	義 26	攸	事 48	義 117	又 116
善 5	日	曾 5	獻 159	獻	士 164	長 102	長	鼓 102	於 25	王	王 47	又 117	三 116
得 5	善	謂 5	與 159	與	酒 163	食 101	寸	寸 100	巽 25	出	失 45	身 115	仁 115
悌 5	失	安 5	授 155	授	實 162	寸 100	食	田 99	遇 24	失	復 44	三 115	身 115
至 5	悌	宗 5	答 152	弟	送 160	帥 98	帥	食 99	或 24	當	當 44	治 113	治 114
養 4	謂	國 5	弟 150	告	獻 159	教 97	田	帥 98	所 24	復	出 44	神 112	神 112
刑 4	心	令 4	告 149	衆	無 158	受 97	教	內 97	物 24	夫	夫 43	仁 112	出 112
四 4	刑	兄 4	若 147	若	授 153	內 97	內	教 97	食 24	生	生 43	出 111	民 111
滿 4	一	用 4	及 143	及	眾 152	田 97	受	子 96	動 24	相	自 43	用 111	用 111
名 4	悅	刑 4	荅 143	進	若 151	器 96	器	受 96	履 23	自	相 43	民 110	後 109
忠 4	在	名 4	進 140	二	告 150	使 95	使	守 96	言 23	過	至 42	必 108	必 108
蓋 4	曾	如 4	從 139	堂	於 147	子 95	子	使 95	益 23	厲	過 42	後 107	雖 104
如 4	忠	成 4	堂 138	首	及 146	守 94	守	器 95	井 23	至	厲 42	雖 104	同 103
守 4	終	和 4	首 137	幣	首 143	相 94	相	於 94	我 23	民	民 41	國 103	汝 103
家 4	自	忠 4	二 137	長	二 140	各 93	各	相 94	窮 22	說	或 40	同 103	國 103
焉 4	成	昔 4	幣 133	弓	進 140	職 93	職	各 93	女 22	或	聖 40	欲 101	欲 101

REFERENCES

ONLINE RESOURCES (CORPORA)

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

2) **CTP** Chinese Texts Database (中國哲學書電子化計劃) 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[88]). (Sibu beiyao edition, 156.1a-8b, 157.1a-10b) 臺灣中華書局, Taibei Shi : Taiwan Zhonghua shu ju, Minguo 54 [1965]

[88] 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).

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)

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, Shuyan 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, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. "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, Cheng-Lin Liu, and 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, Chenglin Liu, and Baihua Xiao. "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*

Herdan, Gustav. *Type-token mathematics; a textbook of mathematical linguistics*. 'S-Gravenhage: Mouton 1960.

Herdan, Gustav. *The advanced theory of language as choice and chance*. Vol. 4 of Kommunikation und Kybernetik in Einzeldarstellungen. Heidelberg: Springer-Verlag, 1966.

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

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

Huang, Chu-Ren and Keh-jiann Chen. “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, Keh-Jiann Chen, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao, and Kuang-Yu Chen. “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 黃侃. *Shoupi Baiwen Shisanjing 手批白文十三經 [Hand annotated edition of Thirteen Classics].* Beijing: Zhonghua shuju, 2006.

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

Hutton, Christopher. *Abstraction and instance: the type-token relation in linguistic theory.* Language and communication library, v. 11. Oxford [etc.]: Pergamon Press, 1990.

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, Jenq-Haur Wang, Chen-Yu Lai, Ching-Chun Hsieh, Lee-Feng Chien, and Jan-Ming Ho. “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, and Zhichang Xu. *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, Ning Xi, Minxuan Feng, and Xiaohe Chen. “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. and 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 and 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, Chilin Shih, William Gale and Nancy Chang, "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 and Richard Xiao. 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. W*ritten 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, Genbao 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 and Hsiaolin Hsieh. *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, P. M. Thompson, Cheng-hui Liu, Chu-Ren Huang, and Chaofen Sun. "Historical Corpora for Synchronic and Diachronic Linguistics Studies." *Computational Linguistics and Chinese Language Processing* 2 no.1 (1997):131—145.

Wetzel , Linda. *Types and tokens: on abstract objects.* Cambridge: MIT Press, 2009.

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, Fei Xia, Fu-dong Chiou, and Martha Palmer. "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 and 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, Bo Xu, and Taiyi Huang. “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, Zengchang Qin, and Tao Wan. “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 and Zhang Dongbo. “The Totality of Chinese Characters - A Digital Perspective.” *Journal of Chinese Language and Computing* 17 no.2 (2007): 107-125.

Zinin, Sergey. *“Pre-Qin Digital Classics: Study of Text Length Variations”*.- Ученые записки отдела Китая, выпуск 15, 44 научная конференция Общество и государство в Китае, том XLIV, ч.2, М., Институт Востоковедения РАН (Scholarly Reports of the Department of China of the Institute of Oriental Studies, Russian Academy of Sciences, issue 15, The 44th Conference “Society and State in China”, vol. XLIV, pt.2, Moscow) (2014): 270-311.