*Sergey Zinin**

Keyword analysis of Chinese classics: "Thirteen Canons" and Zhuangzi

ABSTRACT: This study applies keyword analysis to the Warring States Workshop (WSW) Ctexts corpus of classical Chinese texts. The WSW Ctexts corpus presents a good opportunity to extract a keyword character list of a limited number of linguistically and semantically related texts in a mid-size corpus. The initial goal of the author in this study is the production of the keyword lists, which could be further used for extraction of semantic information about texts. The contents of these lists depend on the type of scoring measure, "the keyness." The lists could be combined into a synoptic table, arranged by a special set of thematic categories. This table could serve as a kind of a "semantic map" of the texts. Analysis of this synoptic table should help in estimating which scoring methods (log-likelihood (LL), chi-square (CHI), effect size, or pre-h-point (PHP) content characters) are more productive for content investigation of Chinese classics. The article demonstrates that, although some methods extract characters that describe topics, other methods could be better for selecting characters, characterizing stylistics. In addition, the synoptic table allows conducting some preliminary genre clustering of the texts.

Content

* Zinin Sergey, Warring States Project, University of Massachusetts, Amherst, USA; E-mail: szinin@research.umass.edu

1. Introduction

1.1. Keyword analysis

Keyword analysis of text contents in corpus studies has been growing since the beginning of the 1990s. In this approach, keywords are defined as "word forms that occur in a text more frequently than expected by chance alone and are often closely connected to the overarching themes and genre of a text or set of texts" [Crvcek and Fidler, "Not all keywords are created equal," 55].[1] They could be identified by "comparing the frequencies of words in a corpus with frequencies of those words in a (usually larger) reference corpus" [Baron et al., "Word Frequency," 1]. It allows compilation of an ordered keyword list and discovering significant words. This list is "a useful tool for directing researchers to significant lexical differences between texts … keywords can reveal a great deal about frequencies in texts which is unlikely to be matched by researcher intuition" [Baker, "The question is," 8]. Keyword analysis allows studying such areas as thematic, stylistic, and linguistic analysis.[2]

Keyword analysis takes roots initially in theories of Firth and then Williams's research of culturally significant words.[3] Word frequency lists

[1] "A standard technique in corpus linguistics and corpus-assisted discourse studies consists in automatically extracting keywords from corpora, to identify the content words that stand out in terms of frequency and keyness" [Gaspari and Venuti, "A golden keyword," 131].

[2] "Keywords ought to identify words are phrases that may be specific to individual topics or questions asked, but they may also be revealing of authorial style, and if used by multiple authors in the corpus, could identify language features associated with a particular regional register" [Baker et al., "Triangulating Methodological Approaches", 40], following earlier Scott's "Key-words provide a useful way to characterize a text or a genre. Potential applications include: language teaching, forensic linguistics, stylistics, content analysis, text retrieval" [Scott, *Wordsmith Tools Version 3,* 70].

[3] Williams called these words "Keywords in two connected senses: they are significant, binding words in certain activities and their interpretation; they are significant, indicative words in certain forms of thought" [Williams, *A Vocabulary of*

have been used for content analysis for a long time, but using keyword lists is a comparatively new method, developed with advancement of personal computers at the beginning of the 1990s. The pioneers of this method often referred to the concept of "culture keywords" as a predecessor of their analysis [Scott, "PC Analysis of Key Words," 233], and it is still treated as such by Stubbs [Stubbs, *Words and Phrases*, 145–194]. The term "keyword" had already took on several meanings by the time of development of keyword analysis [Bondi and Scott (eds.), *Keyness in Text*], but in the latter method, a vaguely defined "cultural significance" had been replaced by statistical significance. Therefore, in this context, "keywords are those whose frequency (or infrequency) in a text or corpus is statistically significant, when compared to the standards set by a reference corpus" [Bondi, "Perspectives on keywords and keyness," 3].

In keyword analysis, the significance of a word as a potential keyword is measured by its "keyness score." There are various keyness measures to implement keyword extraction and ranking. There are a few software packages in the market, which will allow extracting keywords automatically.[4] In this article, the author will consider three popular measures: log-likelihood (LL), chi-square (CHI), and recently introduced %DIFF relative frequency measure.[5]

Depending on the measure of keyness, keyword scores may be positive and negative.[6] Positive keywords could be roughly defined as "comparatively overused" words, comparing with word use in the reference corpus, and negative keywords will be "comparatively underused." The specific measures of these properties will be described below.

There are more complex concepts built around keywords, most of which were developed by Mike Scott. One of them is the concept of key-keywords, which are keywords shared by a few texts of a corpus: a "key-

---

Culture and Society, 15]. As Richardson surmises, this concept is rooted in ideas of Valentin Voloshinov [Richardson, "Keywords Revisited", 101]; Richardson's interpretation has affected development of keyword analysis in the late 1990s. As Mike Scott notes in [Scott, *Wordsmith Tools Version 5*, 165] "The term 'key word,' though it is in common use, is not defined in Linguistics," as it has cultural theory roots.

[4] The most popular is Mike Scott's WordSmith, which has served as the main force in popularizing the concept of keyword [Scott, *WordSmith Tools Version 3*].

[5] They are explained below.

[6] By Scott's definition [Scott, *WordSmith Tools Version 3*, 71], "A word which is positively key occurs more often than would be expected by chance in comparison with the reference corpus. A word which is negatively key occurs less often than would be expected by chance in comparison with the reference corpus".

keyword" is one which is "key" in more than one of a number of related texts. The more texts it is "key" in, the more "key key" it is. This will depend a lot on the topic homogeneity of the corpus being investigated.[7] Other related concepts are "associates" and "clumps." An "associate" of a key-keyword X is "another keyword (Y) which co-occurs with X in a number of texts" [Scott, *Wordsmith Tools Version 5*, 158]. In addition, "clumps" is the name given to groups of keywords associated with a key-keyword [Scott, *Wordsmith Tools Version 5*, 161].[8]

The main reason for analyzing texts using keywords is the presumption that they express "aboutness,"[9] i.e., they allow understanding of text content, based on automatic extraction of frequent words. The key problem there is to select words that are relevant to the content of the text, and this is what the concept of "keyness" does. However, simply creating a list of these words is just the beginning. As Marina Bondi suggests [Bondi, "Perspectives on keywords," 3], "identifying elements that are repeated to a statistically significant extent does not in itself constitute an analysis or an interpretation of the text or corpus. It does however point to elements that may be profitably studied and need to be explained." Baker indicates one of the ways of further analysis—creating concordances

[7] "In a corpus of City news texts, items like bank, profit, companies are key key-words, while computer will not be, though computer might be a key word in a few City news stories about IBM or Microsoft share dealings" [Scott, *Wordsmith Tools Version 5*, 166].

[8] "The idea here is to refine associates by grouping together words which are found as key in the same sub-sets of text files. The example used to explain associates will help. Suppose the word wine is a key key-word in a set of texts, such as the weekend sections of newspaper articles. Some of these articles discuss different wines and their flavors, others concern cooking and refer to using wine in stews or sauces, and others discuss the prices of wine in a context of agriculture and diseases affecting vineyards. In this case, the associates of wine would be items like Chardonnay, Chile, sauce, fruit, infected, soil, etc." [Scott, *Wordsmith Tools Version 5*, 161].

[9] This concept was popularized in the 1970s in the information sciences and content analysis by William Hutchins, and Scott related it to keywords, earliest in his Wordsmith manual of 1998 [Scott, *Wordsmith Tools Version 3*] and in the article of 2000 [Scott, "Picturing the Key Words," 44]. Afterwards, it found popularity in corpus linguistics community. Hutchins discerned between document's topic, summarization, and aboutness, where "concept of 'aboutness' … associates the subject of a document not with some 'summary' of its total content but with the 'presupposed knowledge' of its text" [Hutchins, "The Concept of 'Aboutness'", 180].

based on keywords [Baker, "The question is," 3], which is a future work for this author.[10]

It should be noted that keyword analysis is not a unique method of identifying text content by creating a list of significant words. Computational linguistics has been developing many other methods, e.g., topic extraction and topic signature (see [Kao and Poteet, *Natural Language Processing*]). Some of the methods implemented there also use LL and reference corpora, similar to keyword analysis.[11]

This author also investigated the content character list of this corpus, using "pre-h-point" (PHP) list method [Zinin, "Analysis of character-frequency lists"], and its results will be compared with the results of the keyword analysis.[12]

1.2. Keywords and Chinese classics

There are many studies on frequency of characters in classical Chinese texts, and a few of them analyze text content and genre attribution based on lexis.[13] However, although keyword analysis methods have been around for more than 20 years, this author did not find any research implementing keyword analysis to classical Chinese texts; therefore, this article may be one of the first studies.

Any frequency research on lexis of classical Chinese texts should choose the lexical unit of study. Up to the end of the 20th century, in Chinese and Western studies, this unit had been character. Increasingly, though, researchers have attempted to introduce words (monosyllabic and polysyllabic) as the lexical units. This approach has been working comparatively well for modern texts, but it has been hindered for classical texts by lack of modern authoritative and publicly available classic corpora with word segmentation. The few teams that implement word segmentation usually do not share their corpora resources. Calculating frequency

[10] Examples of such analysis could be found in articles of Karen Donnelly ([Donnelly, "Risk, chance, hope,"] and [Donnelly, "Dr. Condescending"]).

[11] However, the researchers in the keyword analysis area rarely compare these approaches with their methods.

[12] H-point divides the frequency list into two parts: PHP and post-h-point. For a frequency list, the function f(r) is introduced, where f(r) is the frequency for rank r. The "h-point can be defined as that point at which the straight line between two (usually) neighboring ranked frequencies intersects the y = x line," i.e., where r = f(r). The autosemantic words above h-point (PHP) tend to be "content words" [Popescu et al., *Aspects*, 24].

[13] The author omits here a review of this literature, as it is not immediately relevant to the subject of keyword analysis. These works were reviewed in another article by the author [Zinin, "Vocabulary Richness"].

of characters, instead of words, remains the mainstream method in frequency research on Chinese classics. Most of the available articles in this area are character-frequency studies, not word frequency studies. This study belongs to the former category. It creates frequency lists using characters, not words. This author supports an opinion that, when it comes to the analysis of the contents and genre, using character as the unit of study should not considerably affect the results.

1.3. Research corpus

The corpus in this research is the Warring States Workshop (WSW) Ctexts open-source corpus, introduced in an earlier work of the author (see [Zinin, "Pre-Qin Digital Classics"])[14]. Despite the existence of a few available digital academic corpora, there is a lack of publicly available research open-source corpora in classical Chinese. Texts in digital academic corpora vary in textological decisions, which complicates comparisons of results. This is the reason why an open-source corpus, compiled on the basis of Creative Commons licensed digital texts, was used for this research.[15] It offers reproducibility of experiments, at the cost of a few textological flaws. The author believes that, because of the statistical nature of the research, potential text issues should not affect considerably its results.

This study applies keyword analysis methods to investigate the WSW corpus and analyzes the results. The WSW Ctexts corpus presents almost an ideal opportunity to analyze keyword lists of a limited number of linguistically and semantically close texts in a mid-size corpus. Each of the 14 texts, ranging in size from the Xiao Jing to the Zuo Zhuan, could be analyzed against the rest of the corpus[16].

The WSW corpus is very well suited for keyword analysis research. Researchers who studied the relationship between the number and size of texts in reference corpora and keywords state that "the answer to the question 'what is the ideal size of a reference corpus' is five," and that "a

---

[14] This author expresses gratitude to Bruce E. Brooks for continuing support of the project.

[15] This resource, created by the author, is based on Creative Commons digital versions of texts of Wikisource. It contains 12 texts of the Thirteen Classics (the Chun Qiu, the Gongyang Zhuan, the Guliang Zhuan, the Li Ji, the Lun Yu, the Mengzi, the Shi Jing, the Shu Jing, the Xiao Jing, the Yi Li, the Zhou Yi, and the Zhou Li). It does not include the Er Ya, but it includes the Zhuangzi (added as a balancing text). It also allows treating the Chun Qiu and the Zuo Zhuan as separate texts (instead of combined the Chun Qiu Zuo Zhuan), so there are 14 texts in total.

[16] The corpus includes also two small texts (the Guo Dian and the Mao Shi) which are not analyzed in this article.

reference corpus does not need to be more than five times larger than the study corpus" [Berber-Sardinha, "Comparing Corpora," 12]. The WSW corpus complies with all of these criteria.[17] Looking at the role of genre variety, Goh, who analyzed the relationship between keywords and genre composition of reference corpora (RC), states, "genre difference of spoken and written RCs is not an important factor in keyword calculation" [Goh, "Choosing a reference corpus," 249]. Therefore, even if WSW texts varied in genre, it should not affect the results. Considerable diachronic differences between texts could have affected the results [Goh, "Choosing a reference corpus," 254], but it is not the case for the WSW corpus.

Some of the texts are comparatively large (e.g., the Zuo Zhuan and the Li Ji), but according to researchers who analyzed the effects of corpus and text sizes on keyword analysis, these could be considered within the norm.

The author applied previous statistical methods for content analysis of the corpus. In a previous study, the concept of the h-point was applied to identify most significant characters, describing text content. In that study [Zinin, "Analysis of character-frequency lists"], the significant content words were identified and then distributed over categories. These categories have been designed to make historical genre texts prominent.

This study is mostly of explorative character. The main goal of the author in this study is to identify keyword lists that could be used to extract semantic information about texts. The content of these lists depends on methods of scoring "keyness." These lists will be combined into a synoptic table, arranged with a special set of thematic categories. This table could serve as a kind of a "semantic map" of the texts. Analysis of this synoptic table should help in identifying which methods are more successful. The article will not only identify LL and %DIFF keywords (and key-keywords) but also compare them with character lists from the h-point method. Better methods of keyness scoring could be applied to analyze other texts.

Many of the texts of the WSP Ctexts corpus are collections of smaller works, and, in reality, themselves are mini-corpora. But even smaller texts that have been assigned an "authorial figure" to them (like the Meng-zi or the Lun Yu), are not products of an "identifiable original author" [Boltz, "Why So Many Laozi-s?", 10], as was well-known to the Chinese philological tradition. The recent discovery of manuscripts of Warring States and Han period instigated new analysis of textological structure of received canons. Many classics should be viewed as dynamic, flexible collections of smaller fragments. Two concepts should be noted

[17] Only one text, the Zuo Zhuan, is about five times smaller than the entire corpus.

as indicating productive research directions. The first is the idea of a "growth text" by Bruce Brooks [Brooks, "Before and After Matthew", 1] and the second is the idea of "building blocks" by William Boltz [Boltz, "Why So Many Laozi-s?", 12]. Brooks utilizes stylistic analysis to trace trajectory of text building, and Boltz stresses the role of "thematically based selection" [Boltz, "Why So Many Laozi-s?", 12] in assembling texts from "building blocks". Keyword analysis may provide useful information for any of these approaches.

This study is striving to be an open-source, reproducible, and verifiable effort. All source data should be available for download; in addition, the accompanying site contains all extended original results, which would not fit into the article's limited space (see Appendices for links). The author is grateful to Bruce E. Brooks for continuing support of the project.

2. Keywords in Chinese classics

Keyword lists are created by assigning significance values ("keyness scores") to words or characters and then selecting the most significant ones. Mostly, keyness scores are statistically significant values obtained by calculating LL and CHI scores. Some researchers (Gabrielatos and Marchi, "Keyness") consider the statistical significance approach not the most effective measure and offer their own %DIFF measure, based on "effect size" approach. In this study, the author computed the values of all three measures (LL, CHI, and %DIFF) and compared the results.

2.1. Keyword character list compilation

LL keywords for the WSW corpus were calculated using formula[18]

$$-2\ln\lambda = 2\sum O_i$$

In case of a two-by-two matrix,
G2 = 2*((a*ln (a/E1)) + (b*ln (b/E2)))[19].

The list of keyword characters, produced by this method, is very large and needs filtering to obtain meaningful results. It is well known

[18] See [Rayson and Garside, "Comparing Corpora," 3], and UCREL http://ucrel.lancs.ac.uk/llwizard.html.

[19] Where "a" is the frequency of a character in Corpus 1 (Reference corpus), and "b" is the frequency of a character in Corpus 2 (corpus under testing); "c" is the number of characters in Corpus 1 and "d" is the number of characters in Corpus 2. "a+b" will be the total number of a character in both corpora, and "c+d" is the number of all characters in both corpora. In these terms, expected values E1 (for Corpus 1) and E2 (for Corpus 2) will be E1 = c*(a+b) /(c+d) and E2 = d*(a+b) / (c+d). See UCREL http://ucrel.lancs.ac.uk/llwizard.html.

that most rare words and characters go to the top of lists created using statistical significance. Most researchers recommend discarding of rare characters, often starting with minimum three counts.[20] In this study, owing to the size of the texts, only characters with frequency count four or higher were included in the resulting character lists. Another cutoff criterion is the p-value: it should be low enough. The value of $p = 0.05$ is standard, which corresponds to an LL critical value of 3.84. In this study, the minimum of the LL values is 11.0[21], so with one degree of freedom (d.f. = 1), p-value should be less than 0.001.[22] This is a more rigorous criterion than a p-value of 0.05. Even if these criteria are quite rigorous, for some large texts, the list of keyword characters could reach hundreds of entries. Finally, in keyword analysis, function words are usually removed from a keyword list [Archer, "Does frequency really matter?" 2–3]. Therefore, function (or "empty") characters were removed from the resulting keyword character lists, with some exceptions.[23] The full list of LL keywords could be downloaded from the project's Github site.[24]

CHI keywords were calculated using the contingency table and the formula from [Baron et al. Word Frequency, 44], see Table 1[25].

Table 1. Contingency table for the CHI test

	Corpus 1	Corpus 2	Total
Frequency of feature	a	b	a+b
Frequency of feature not occurring	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

[20] For example, Scott recommends three counts as default, but it depends on text's size: "The default setting of 3 mentions as a minimum helps reduce spurious hits here. In the case of short texts, less than 600 words long, a minimum of 2 will automatically be used" [Scott, *WordSmith Tools Version 5*, 177].

[21] The critical value for 0.001 is 10.83 (DOI: UCREL http://ucrel.lancs.ac.uk/llwizard.htm).

[22] "Any word with a LL greater than or equal to 6.63 ($p < 0.01$ for 1 d.f.) was considered key, and any word with a frequency less than 5 in either the Innsbruck Letter Corpus (before or after standardization) or the BNC sample was removed from the key word list," [Baron et al, "Word frequency," 55–56]. There was no intention in this study to look for very low p-values.

[23] Function words do not go into keyword lists, but could go if their usage is "function words can occur in a keyword list, if their usage is strikingly different from the norm established by the reference text" [Archer, "Does frequency really matter?" 3].

[24] See the Excel spreadsheet "chinese_classics_keywords_log_likelihood _short". See Appendix for the file structure's explanation.

[25] Usage of terms a, b, c and d is different there than in case of LL calculations.

The CHI statistics (X^2) will be calculated as follows:
$X^2 = N(ad-bc)^2/(a+b)(c+d)(a+c)(b+d)$

The criteria for frequency counts and p-value for this list were the same as for LL (11 and 4), function characters were also removed.[26]

A comparison of lists (see Table 2) shows that, at those criteria, there are more CHI keyword characters than LL keyword characters. However, the contents of the two lists are very close, with mostly the same characters but in a slightly different order. A more detailed analysis shows that, indeed, LL keyword list (with a few exclusions) is a subset of CHI keyword list. Therefore, CHI data will not be discussed in this article, as LL keyword list is representative enough well for statistical significance methods to be compared further with %DIFF keyword characters.

Finally, %DIFF value of keyness had been calculated by the formula (Gabrielatos and Marchi, Keyness: Appropriate metrics, 12):

((NF in SC – NF in RC) x 100)/ (NF in RC)[27],

where NF = normalized frequency, SC = study corpus, and RC = reference corpus

As with LL and Chi lists, only characters with an absolute frequency of more than three are included.[28] As to the %DIFF value cutoff, it was also taken as 11, simply for conformity with LL and Chi numbers.[29] Table 2 contains comparative numbers for numbers of keywords retrieved by these methods.

[26] The full list of CHI keywords is contained in "chinese_classics_keywords_CHI_short" Excel spreadsheet and could be downloaded from the Github site. See Appendix for the file structure's explanation.

[27] 0.0000001 was added to avoid division by zero.

[28] The full list of %DIFF keywords is contained in "chinese_classics_keywords_diff_short" Excel spreadsheet and could be downloaded from the Github site. See Appendix for the file structure's explanation.

[29] The meaning of %DIFF is different from that of statistical significance, but, according to the author's estimation, depending on text size, a similar LL value will be a bit higher than 10. There will be generally almost twice more %DIFF values than LL or Chi.

Table 2. Comparative list of keyword characters' numbers. V is the number of types in a text, N is the number of tokens, LL column contains the numbers of LL keyword characters, %DIFF is the number of %DIFF keyword characters, and CHI is the number of CHI keyword characters.

Text[30]	V	N	LL	%DIFF	CHI
XJ	374	1800	33	73	40
ZY	1035	13354	165	299	190
LY	1365	15928	97	256	108
CQ	941	16791	146	261	159
SHU	1911	24539	262	523	293
SHI	2833	29622	462	838	537
MZ	1895	35372	168	480	198
GL	1595	40836	163	387	179
GY	1641	44232	190	438	219
ZL	2215	49462	360	684	395
YL	1551	53929	320	493	341
ZHZ	2968	65251	352	771	400
LJ	3050	99007	325	918	358
ZZ	3236	178564	437	1042	447

Charts on Figure 1 and Figure 2 display the relationships between the numbers of keywords and lengths of texts and vocabulary size, respectively. Owing to the differences in the type of measures, it is not possible to compare the numbers of keywords directly, but these charts display a good visual correlation between ratios of keywords and text characteristics.

[30] See the list of abbreviations in the Resources section.

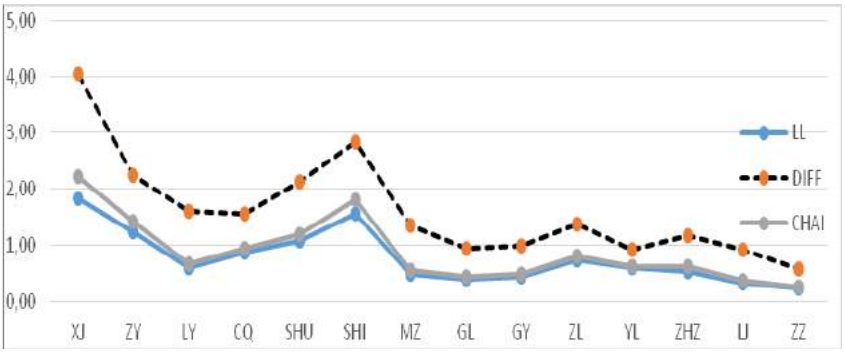

Figure 1. LL, %DIFF, and CHI keyword numbers/text length ratios, ordered by text length

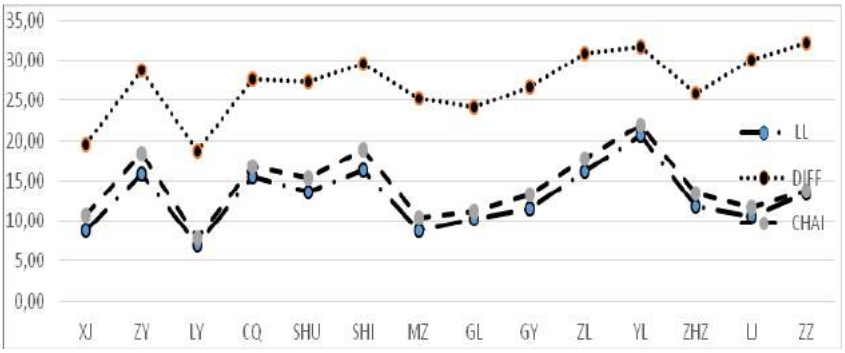

Figure 2. LL, %DIFF, and CHI keyword numbers/vocabulary ratios, ordered by text length

The ratios of all three scores correlate well. We can see that the ratio for keyword number/text length is generally falling with length increase, and this for keyword number/vocabulary probably depends on other text characteristics.

2.2. Analysis of the top list of keyword characters and comparison with PHP content word lists

In keyword analysis, it is often hard to handle all words/characters, even if strict criteria were imposed. For topic and genre analysis, most researchers recommend to select only the top part of keyword list, from

50 to 100 words/characters. The sizes of sample numbers are arbitrary and do not usually reflect text characteristics.[31]

This author earlier implemented another approach to the topic analysis, based on the concept of h-point, introduced by Popescu.[32] H-point should reflect texts' inner characteristics, and PHP lists of content word/characters should reflect texts' content. It may be beneficial to compare all three methods of selecting content characters based on their frequency: PHP content characters and top lists of LL and %DIFF scores. Although the h-point approach provides a "native" number of top characters, the LL and %DIFF methods lack such measure. The author decided to take 50 or 100 top characters from LL and %DIFF lists, depending on which number is closer to the corresponding number of PHP characters for this text. See Table 3 for the specific numbers.

Table 3. Matching LL and PHP keyword characters

#	1	2	3	4	5	7	8
Text	CQ	ZZ	GY	GL	SHI	SHU	ZY
Matching positive in all keywords	45	69	45	43	16	22	25
Total	49	125	68	63	37	37	37
% matching to all KW	91,84	55,2	66,18	68,25	43,24	59,46	67,57
Number of pre-h-point characters	60	182	94	91	66	67	51
Selected top KW	50	100	50	50	50	50	50
Matching positive in top keywords	36	48	35	35	9	14	14
Total	49	125	68	63	37	37	37
% matching to top KW	73,47	38,4	51,47	55,56	24,32	37,84	37,84

#	9	10	11	13	14	15	16
Text	ZL	YL	XJ	LJ	MZ	LY	ZHZ
Matching positive in all keywords	50	63	5	45	23	17	28
Total	74	81	7	91	43	29	56
% matching to all KW	67,57	77,78	71,43	49,45	53,49	58,62	50
Number of pre-h-point characters	96	114	20	136	78	50	101
Selected top KW	100	100	all	100	50	50	100
Matching positive in top keywords	44	54	5	30	16	16	23
Total	74	81	7	91	43	29	56
% matching to top KW	59,46	66,67	71,43	32,97	37,21	55,17	41,07

[31] "The vast majority of studies do not examine all keywords, but the top X (usually the top 100)" [Gabrielatos and Marchi, "Keyness," 3].

[32] See: Zinin, "Analysis of character frequency lists".

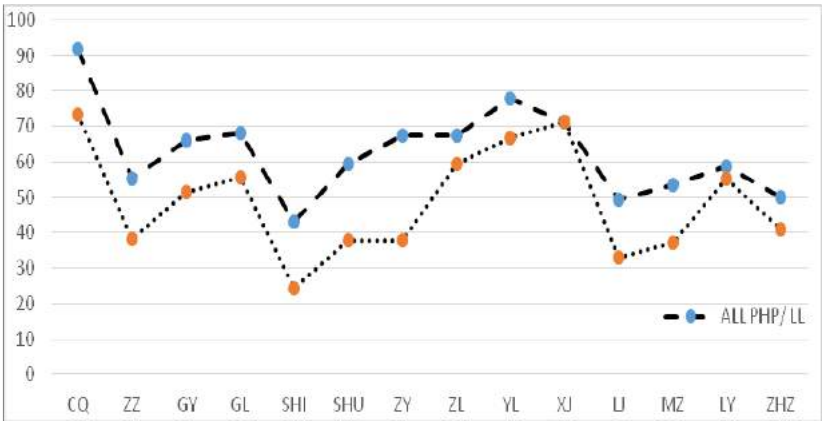

Figure 3. Ratio of PHP content characters, matching all and top LL keyword characters

Do PHP content characters match LL characters? Table 3 displays this information. For almost all texts, PHP content characters are on the top 100 most frequent list, as PHP characters are very frequent in text, and LL characters do not need to be very frequent. The study shows that, with the exception of the Shi Jing, more than 50% of the PHP characters belong to the LL lists. It means that the LL lists contain very frequent non-function characters, too. However, if the top 50 or 100 LL keyword characters are selected, and compared to the PHP characters, there is less significant overlap (from 30% to 50%) between the lists of PHP characters and LL characters. The top lists of LL keyword characters include not too many very frequent characters; otherwise, they would be overlapping more with the PHP lists.

Table 4. Matching %DIFF and PHP keyword characters

	1	2	3	4	5	7	8
	CQ	ZZ	GY	GL	SHI	SHU	ZY
DIFF							
matching	45	71	52	46	23	24	26
LL matching DIFF	91,84	56,8	76,47	73,02	62,16	64,86	70,27
REDUCED DIFF	12	1	5	4	1	3	8
	24,49	0,8	7,35	6,35	2,7	8,11	21,62

	9	10	11	13	14	15	16
	ZL	YL	XJ	LJ	MZ	LY	ZHZ
DIFF							
matching	37	64	6	52	25	20	32
LL matching DIFF	50	79,01	85,71	57,14	58,14	68,97	57,14
REDUCED DIFF	11	19	5	1	2	4	1
	14,86	23,46	71,43	1,1	4,65	13,79	1,79

Analysis of the overlapping of %DIFF list and PHP characters shows even a higher percentage of PHP characters on the complete %DIFF lists than on the complete LL lists. It should be caused by the size of the %DIFF lists. Analysis of the top part of the %DIFF lists shows that they do not practically overlap with the PHP characters, and the reason is the lower (than that for LL) frequencies of those top characters on the %DIFF lists.[33]

3. Key-keywords in Chinese classics

The notion of key-keyword is a derivation of the keyword concept. Key-keywords are (Scott, "PC Analysis of Key Words", 237) "words which are key in a large number of texts of a given type." There could be positive and negative key-keywords, as well. Table 5 presents the top part of the list of positive key-keywords for the WSW corpus, ordered by the number of texts to which they belong.[34]

Table 5. Most frequent key-keywords in the WSW Ctexts corpus

Character	Number of texts	Texts
天	7	shu, zhouyi, xiaojing, guodian, liji, mengzi, Zhuangzi
民	7	shi_mao, shu, zhouli, xiaojing, guodian, liji, mengzi
道	7	shi_mao, zhouyi, guodian, liji, mengzi, lunyu, Zhuangzi
子	6	chunqiu, zuozhuan, gongyang, liji, mengzi, lunyu
故	6	zuozhuan, shi_mao, xiaojing, guodian, liji, Zhuangzi
無	6	zuozhuan, shi, shu, zhouyi, mengzi, lunyu
萬	6	shi, shu, zhouli, guodian, mengzi, Zhuangzi
行	6	zhouyi, xiaojing, liji, mengzi, lunyu, Zhuangzi
三	5	chunqiu, zhouyi, zhouli, yili, liji
來	5	chunqiu, gongyang, guliang, shi, zhouyi
公	5	chunqiu, zuozhuan, gongyang, guliang, shi_mao
古	5	shi_mao, shu, liji, mengzi, Zhuangzi
國	5	zuozhuan, gongyang, guliang, shi_mao, zhouli

[33] For example, for the Chun Qiu, the h-point is 60; that is, all PHP characters have absolute frequencies more than 60. However, on the %DIFF list for the Chun Qiu, only 12 characters have a frequency equal or higher than 60. This situation is even worse for matching in other texts.

[34] It includes key-keywords found in at least seven texts (including the Mao Shi and the Guodian, which are otherwise not considered in this study). The full list of positive and negative keywords (appearing at least from three texts) could be found at the online reference Excel spreadsheet "chinese_classics_keywords_log_likelihood_short" (See Appendix for Resources).

圍	5	chunqiu, zuozhuan, gongyang, guliang, zhouli
地	5	zhouyi, zhouli, guodian, liji, Zhuangzi
姜	5	chunqiu, zuozhuan, gongyang, guliang, shi_mao
孔	5	shi, liji, mengzi, lunyu, Zhuangzi
孰	5	gongyang, guodian, mengzi, lunyu, Zhuangzi
己	5	chunqiu, gongyang, guliang, lunyu, Zhuangzi
思	5	shi, shi_mao, xiaojing, mengzi, lunyu
怨	5	zuozhuan, shi_mao, shu, mengzi, lunyu
政	5	zuozhuan, shi_mao, shu, zhouli, lunyu
敗	5	chunqiu, zuozhuan, gongyang, guliang, guodian
方	5	shi, shu, zhouli, liji, Zhuangzi
明	5	shu, zhouyi, xiaojing, liji, Zhuangzi
曰	5	zuozhuan, shu, zhouyi, mengzi, lunyu
正	5	chunqiu, gongyang, guliang, zhouyi, yili
歸	5	chunqiu, zuozhuan, gongyang, guliang, shi
父	5	chunqiu, gongyang, guliang, xiaojing, liji
王	5	chunqiu, zuozhuan, shi_mao, shu, mengzi
百	5	shi, shu, zhouli, liji, mengzi
義	5	shi_mao, xiaojing, liji, mengzi, Zhuangzi
詩	5	zuozhuan, shi_mao, xiaojing, liji, mengzi
變	5	guliang, shi_mao, zhouyi, liji, Zhuangzi
足	5	guodian, liji, mengzi, lunyu, Zhuangzi
身	5	xiaojing, guodian, liji, mengzi, Zhuangzi
邦	5	shi, shu, zhouli, guodian, lunyu
雨	5	chunqiu, gongyang, guliang, shi, zhouyi

If many characters are "overused" in several texts together, relative to the reference corpus, there could be some semantic affinity between these texts. Depending on the share of these texts in corpus, positive key-keyword, participating in many texts, may be "underused" in other texts, i.e., be a negative key-keyword for those texts. If some texts have affinities based on instances of positive and negative key-keywords, it could be that they fall together in semantic sense.

For example (see Table 6), characters *tian, min*, and *dao* are positive key-keywords in philosophical and fiction texts and negative in historical texts, and characters *zi* and *gong* are positive key-keywords in historical texts and negative key-keywords in philosophical and fiction texts.

Table 6. Text distribution of several key-keywords

Character	Positive	Negative
天	shu, zhouyi, xiaojing, liji, mengzi, Zhuangzi	chunqiu, zuozhuan, gongyang, guliang, zhouli, yili
民	shu, zhouli, xiaojing, liji, mengzi	zuozhuan, gongyang, guliang, Zhuangzi
道	zhouyi, liji, mengzi, lunyu, Zhuangzi	zuozhuan, gongyang, shi, zhouli, yili
子	chunqiu, zuozhuan, gongyang, liji, mengzi, lunyu	shi, shu, zhouyi, zhouli, yili, Zhuangzi
公	chunqiu, zuozhuan, gongyang, guliang	shi, shu, zhouyi, zhouli, yili, liji, mengzi, lunyu, Zhuangzi

It is possible to expand beyond just a list of key-keywords, and Scott introduces [Scott, "PC Analysis of Key Words", 238] "associates", as "words found to be key in the same texts as a given key key word"[35]. This notion is instrumental for the concept of (Scott, "PC Analysis of Key Words", 241) "clump of associates", which "is a set of associates formed by co-occurrence in the same texts which gave rise to associates." Clumps could be used for defining a semantic area for groups of texts.

There is another way to find out if key-keywords could be used as a feature to group texts together. A matrix of texts and key-keywords was prepared[36], where only the presence/absence of a character was entered and then clustered, using the free online tool ClustVis.[37] The resulting graphic representation of text/key-keyword clusters (so-called HeatMap) is presented in Figure 4.

[35] As Scott comments [Scott, PC, 239], "This notion of Associate is strikingly close to the early 1950s and 60s discussions of collocation". Therefore, this paper will not address neither associates (nor clumps) nor collocations.

[36] See the list of these key-keyword under "Top key-keyword character for clustering experiment" in Appendix.

[37] http://biit.cs.ut.ee/clustvis/

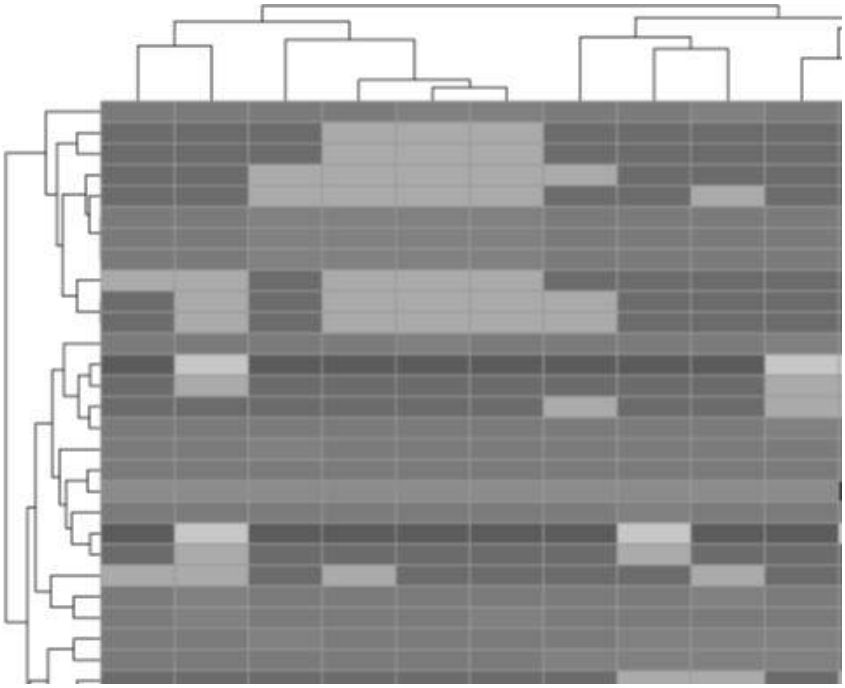

Figure 4. HeatMap of key-keyword clusters

There are two aspects of the ClustVis's HeatMap. On one side, it clusters texts according to their term (key-keywords) vectors; on the other side, it clusters terms according to their text vectors. Term clustering will not be analyzed in this article. Text clustering produces, most prominently, a historical group, and then, it groups together the Shi Jing and the Shu Jing (and the Zhou Li) and philosophical texts (the Lun Yu, the Meng-zi, the Xiao Jing, and the Zhuang-zi). Ritualistic texts are scattered, though. This opens another avenue for using key-keywords for genre and semantic analysis.

4. Comparison of outputs of keyword analysis methods

The main benefit of applying keyword analysis, as well as h-point methods to frequency lists of Chinese classics, is in compiling concise lists of characters that could be used for topic and genre analysis of corpus texts, which could be called a semantic map of the texts. Having three instances of lists allows us to estimate the effectiveness of the applied methods. In addition, it may happen that some methods could be better suited for presenting specific aspects of texts. The breakdown of lists by texts and by categories is presented in Appendices.

Characters will be broken down by a set of categories, which were earlier utilized by the author for PHP content words. The category sets were not derived through topic discovery or modeling; they have been designed by the author and defined by the author's analysis of the PHP content characters of the Thirteen Classics.[38] The author decided to select categories that expose semantic dimensions of the classics (CQ, ZZ, GY, and GL, as well as SHU, LJ, YL, and ZL), as historical texts and, especially, chronicles. These categories have been informed mostly by the Chun Qiu PHP content characters' list. The main categories are Numeric, Calendrical, Social, and Politico-Geographical. The remaining characters were distributed across Part-Of-Speech (POS) categories: Nouns, Verbs, Adjectives/ Adverbs, and finally Miscellanea. Practically all PHP characters of the Chun Qiu could be assigned to non-POS categories, with the exclusion of some verbs.[39] The order of text also reflects this idea — on the top, there are properly historical texts, and other texts follow in their internal id number order. The most important element is the lists of keyword and content characters. They could be reorganized into any other set of categories.

In addition, the texts below are grouped according to genre categories, presumed by the author.[40] One of the possible usages of character-

[38] This approach is similar to what Karen Donnelly implements [Donnelly, "Dr. Condescending," 91] "Initial analysis was carried out using Wordsmith Tools to elicit the top 100 lexical keywords from each corpus, which were then grouped thematically in order to allow comparison across the 3 corpora and guide selection for further study using collocations and concordance lines."

[39] Many Chinese characters (words) could play the role of various POS – noun, verb, adjective, etc. This phenomenon is well known as "POS variation or categorial ambiguity (i.e., ambiguity in grammatical category)" [Tsou and Kwong, "Some basic and Salient Linguistic Features," 7]. There are various solutions for this situation. Character (or word) could be assigned to all possible categories, or only one, dominating, category could be assigned to it. Tsou and Kwong offer an interesting log-ratio measure for the latter solution [Tsou and Kwong, "Some basic and Salient Linguistic Features," 8], but it requires a POS-marked corpus. In this study, only one category was assigned uniformly for all texts. It could be not optimal, as it may depend on text, which syntax category the character belongs. The author considers it a temporary provisional solution. The characters could be re-shuffled in other studies, if necessary.

[40] These texts were classified in their own categories, starting from the Qi Lüe catalogue (except the Zhuangzi). Most of them are collections of various texts, sometimes from different genres. As Lisa Raphals defines, "The Chinese Classics are a group of texts of divination, history, philosophy, poetry, ritual and lexicography that have, to a significant extent, defined the orthodox Ruhist (Confucian) tradition of China. Since the Song dynasty (960–1279), they have consisted of the

frequency lists is attempting to find similarities between lists and relating them to the genre categories. A pure numeric approach (described above), consisting in clustering vectors, created based on the number of characters in topic/lexical categories, is not conclusive. It only showed that "historical" texts could be grouped together, but the category system was pre-built to favor historical terms. Some preliminary suggestions will be made below, but this should be the subject of a special investigation.

5. Discussion of results

This study will not analyze topics of specific texts in detail because of the lack of space. However, it will discuss briefly keyword characters (and content characters) extracted by all three methods and how these sets of characters differ in relation to established categories, for each of the texts. The following section briefly discusses relevant parts of the synoptic table[41].

Historical texts

1. Chun Qiu

Table 7. Comparative keyword chart for the Chun Qiu

Method	Numeric	Calendrical	Social	Politico-Geographical	Nouns	Verbs	Adjective/ Adverb	Misc
PHP	十二三七四八六五九	月年夏春冬秋	公人侯子師孫王夫伯叔曹	齊晉宋鄭衛楚陳邾		有伐來葬盟奔莒歸蔡殺杞帥侵	大正	
LL	十七二八	月冬春秋夏年癸辛亥酉巳壬午庚丑乙丁	公侯伯孫卒杞叔曹	齊宋衛晉鄭邾莒楚蔡薛陳		伐帥師來葬盟奔侵		

following thirteen texts" (Raphals. "Chinese Classics," 135). In the current article, the lexicographic text (the Er Ya) was omitted, the divinatory text (the Yi Jing) was classified as philosophical, and the Shu Jing and the Shi Jing were classified as fiction, instead of singling out just the Shi Jing as "poetry."

[41] The complete table could be restored from these parts, or found in GitHub resources, see Appendix.

%DIFF		酉癸冬 巳卯亥 丑月壬 秋春乙 辛 丙戊午 丁夏庚	杞曹帥 郎	鄆薛邾滕 邸鄫宋洮 衛鄰婼	鼷螽孛 蠆霄頓 款雩讙 彊鬍貜 羯	莒蔑匄		

In the Numeric section, only the PHP and LL methods bring numeric characters on top lists (see Table 7), and there are more characters in the PHP section than in the other measures.[42] In the Calendrical section, all methods bring up season names, but LL and %DIFF bring many cyclical signs on the top, unlike PHP, which could be explained by the lower frequency of cyclical signs. Therefore, LL and %DIFF methods seem to be considering dating by cyclical signs more prominent than by numbers. In the Social category, PHP and LL are close.[43] %DIFF only includes a few family terms on the top of its list. In the Politico-geography section, all methods bring in names of states or kingdoms, with some variations. The major difference between methods could be observed in Nouns and Verbs categories. The PHP and LL conceptual framework is consisting mostly of Social and Politico-geographical terms.[44] %DIFF output is practically missing any topical verbs.

Interpreted from the point of view of PHP content characters, the Chun Qiu could be perceived as a text, where there are social actors, related to some territories and states, who conduct some political actions, referred to seasons, month, and years. A similar interpretation could be given on the basis of LL. The type of interpretation, allowed by PHP and LL output, places the Chun Qiu into a genre category of historical chronicle. However, %DIFF only tells that the text is about some dates and states, but misses on actors and their actions.

[42] %DIFF method does bring numbers to the keyword character list, but they do not get on the top list, e.g., 十 gets into position 77 on CQ %DIFF, below cut-off of 50.

[43] LL missing such important categories as 人 and 王.

[44] However, it includes many high-frequency verbs. A few verbs overlap between PHP and LL (e.g., 伐, 來, and 葬).

2. Zuo Zhuan

Table 8. Comparative keyword chart for the Zuo Zhuan

Meth-od	Nu-meric	Calen-drical	Social	Politico-Geo-graph-ical	Nouns	Verbs	Adjec-tive/ Adverb	Misc
PHP	二三十一五四六	月年日秋春冬夏	子公人君侯王師夫伯氏叔孫歸臣民季父司軍孟仲士帥尹蔡	晉楚齊國鄭衛宋陳吳周城秦趙魯邑	禮事文書天德朝罪門馬亂女謀	曰為有可伐盟謂死殺行出成告言知入敢來聞欲立見命奔敗用亡取生執辭求政信令主懼獻問宣還獲乘棄服	大寡今武小難	是吾未下上中右
LL			子叔氏師公君尹諸軍武侯季臣伯孫貳族嬖	晉楚鄭吳趙秦韓邑國衛魯荀虢蘧	產罪禍書文圖鮑亂范郤賂駟	曰請懼盟宣寵討敗謀告棄叛奔歸許亡召洩殺怒臧遂伐逐囚死鬥賦襄獲逃昭免敢寘逞偪	寡欒敝穆難平魏俘絳彊戎崔	吾
%DIFF			尉蟜媯姚甥嬖褚頡姞潘	潁鄖蘧溱虢郇鄂邯邶巴	范帑孟裔胙旆胄躒魋衷蚡茷祚紇諺嚴龙篳櫝豭賂產儕鯈殿鮑黶廚繻董	洩紓悛逞縊謗寵犒闕狃殪懼迂蒸寘庇縛蒐鬥囚儆轘祓劫掠匄	俘絳暱徼抶訴齬鞏嗌罌鑾瑰錮尪綿楯	余

The Zuo Zhuan contains numeric and calendrical characters only in the PHP list, and even these numbers are lower than those in the Chun Qiu (see Table 8). All methods deliver more characters in Verbs and Social area

than in the Chun Qiu. %DIFF considerably expands its number of nouns. PHP and LL bring up nouns, too. In addition, there are characters in the Adjectives/Adverbs section. It is clear that nouns, produced by the PHP and LL methods, are important historical narrative terms,[45] which characterizes the Zuo Zhuan as a historical text. Although there is some overlapping with the %DIFF output, its nouns are less significant. The same could be said about Verbs. LL and PHP brought up characters, describing important historical actions, whereas %DIFF does not. All this allows, based on keyword and content characters, qualifying the Zuo Zhuan as a historical narrative, but may not be specifically a chronicle.

3. Gongyang Zhuan

Table 9. Comparative keyword chart for the Gongyang Zhuan

Method	Numeric	Calendrical	Social	Politico-Geographical	Nouns	Verbs	Adjective/Adverb	Misc
PHP	二三十一五四七	月年日秋春冬夏	子公人君侯王師夫伯氏叔孫歸季父帥曹婁	晉楚齊國鄭衛宋陳邾	書天	曰為有可伐盟殺出言入來立取執葬莒譏弒稱	大正然	未是吾
LL	十	月春秋冬年夏朔	公婁侯曹伯孫卒師季帥歸蔡杞	齊宋邾衛晉鄭陳桓	書災	譏言稱貶弒伐葬記諱莒來滅盟侵殺	隱	曷奈
%DIFF	什頃		婁袁侄鱄郎盱繆郜	閨濤盾系酈甗	鷁獟漸蝝翬鞍孛書貜	譏貶錄記篡托諱剽逡丐運弒褒稱諼	昧煬晅斐隱憊瞶	曷遫奈

The Gongyang Zhuan is closer to the Chun Qiu than to the Zuo Zhuan, even though in the LL list it has very reduced set of Numeric terms (see Table 9). Both PHP and LL produce rich sets of Social and Political terms, and both lists are practically missing Nouns. In Nouns, the characters that are available are similar to what the Zuo Zhuan list produces. The verbs also have a rich set of terms. Therefore, on the basis of these lists, the text could be characterized as a historical text with considerable time indications.

[45] There are overlapping frequent terms, such as *luan*, *zui* and *wen*.

4. Guliang Zhuan

Table 10. Comparative keyword chart for the Guliang Zhuan

Meth-od	Nu-meric	Calen-drical	Social	Politico-Geo-graphical	Nouns	Verbs	Adjec-tive/ Adverb	Misc
PHP	二三十一五四七	月年日秋春冬夏	子公人君侯王師夫伯叔孫歸父帥蔡曹	晉楚齊國鄭衛宋陳邾	事天	曰為有可伐盟殺出言入來辭葬莒弒志	大正內	是未
LL	十	月春秋冬夏年日朔辛癸	公侯卒曹孫帥師蔡姬杞伯	齊宋衛鄭邾晉陳楚狄滕		正伐葬來弒言盟莒志侵殺奔戰敗稱繒		
%DIFF		酉朔	伉謚郜曹侄	邿鄟鄑鄆桷郰酈薛滕邾	鴞佾雩頰鼷累賵鞏螽貜翬彊	繒丐嫌借挈搜逾斥弒糴唁崩錄壅斲嫁葬髡	謹卑婼	

The Guliang Zhuan is similar to the Gongyang Zhuan; their Calendrical, Social, and Politico-geographical sections are very close (see Table 10). It also should be considered a historical narrative.

Ritualistic texts

5. Li Ji

Table 11. Comparative keyword chart for the Li Ji

Meth-od	Numer-ic	Calen-drical	Social	Politico-Geo-graphical	Nouns	Verbs	Adjec-tive/ Adverb	Misc
PHP	二三十一五四	月年日夏	子公人君侯王夫臣民士母婦孔	楚國	禮事文天德朝門世賓喪東樂方道廟義衣宗位	曰為有可謂死行出成言知入敢立見命用生執主問食祭教學哭拜居服	大小正內貴外然明反長尊	是未上中

LL		日	親君婦父夫孔妻舅士祖 櫛	鄉	禮喪故廟樂杖衣冠天氣麻尸孝宗義音棺肉節情聲齒事身本功器佩鬼味堂墓臭體繭粥緦禘絰踴衾裼	哭祭服衰學斂食練製修哀教養容問衿順浴飲發行附居沐除誌殯吊寢帶	親敬尊貴幼誠玄大素賤明袒疏	
%DIFF			婿紳婢	賁剡	繭禫綍胎緦緦粥綴雛蠟笏洞旒尸杖縞紛帨觿庠翌衧綬腎騮閣訝縰竽蔥鷹闚臭衾菜痛麻犢雁蚤棺酏祫禘禘簟紼幬疫辟墠麑箭苫咳絰裼	誌製襢詘僎練鋪漱減衿涂浣溜措紐堊綴伸殯掃煎吊浴哭	膻暖恒蓑糜褻菲腥峻暗揄歠涀	毋

The Li Ji has content character lists that are very different from the lists of historical texts (see Table 11). The PHP method extracts some Numeric and Calendrical terms, but any politico-geographical terms are practically missing. The Li Ji has a Social section, even though not so big as it is in historical texts. The main difference between those texts is that all methods bring in for the Li Ji many Nouns, Verbs, and Adjectives/Adverbs. There is a difference between how both the PHP and LL methods work and how %DIFF method works. The first two methods extract, on the top, meaningful terms, describing the conceptual frame of the text. The %DIFF

method also extracts important terms for the Li Ji, but these terms are not "conceptual." They rather characterize the text stylistically.

6. Yi Li

Table 12. Comparative keyword chart for the Yi Li

Meth-od	Nu-meric	Calen-drical	Social	Politi-co-Geo-graph-ical	Nouns	Verbs	Adjec-tive/ Ad-verb	Misc
PHP	二三一		子公人君夫爵賓婦賓		禮門東北西面馬南階屍位觶席阼酒堂首幣弓醢奠	曰為有出告入立見命取執主獻拜升降受祭射興洗揖祝送授答荅進退酌稽佐篚辭俎	大外長	下上右左反従
LL			司婦卒爵賓		西北東階觶屍南興席阼位耦筵醢房堂脯門豆幣酒弓矢首拾鼎衆面巾束羞戶肺楹薦奠脊鉶饌醴匕	拜主升坐降洗俎受揖執射設荅祝答篚祭初酌介送授取盥佐立稽加酬進辭徹出反遂退獻縮適臘薦侑擯釋嚌踊複箕	實祖	左右

%DIFF			糗爵		觶枋楅柳 肵鉶匕陔 塾觚笋階 胃饌飪拾 匝驸面鏃 冪髀肺腸 奠肫筵阼 脊槃壁楣 甒楹耦脡 茵茁屍房 脯梌西巾 韭羸骼箱 醓醢脅席 麷鞀繶笴 縓膞	荅洗啐嚀 侑篚擩箄 俎縮複搢 升答揖拜 坐扱擯醮 盥酬臘酌 撲降 概酳	膞纚謢 耍綪蕡 胖妥	毋第

The Yi Li's lists are consisting mostly of Noun and Verb characters, practically missing characters in Numeric, Calendrical, and Politico-Geographical categories (see Table 12). They could be characterized similarly to what was said about the Li Ji.

7. Zhou Li

Table 13. Comparative keyword chart for the Zhou Li

Method	Numeric	Calendrical	Social	Politico-Geographical	Nouns	Verbs	Adjective/ Adverb	Misc
PHP	二三一十五八九百	日歲	子人侯王夫氏民司士師史賓帥客官徒	國府邦	禮事馬寸掌方物車喪法器田刑	曰為有謂行入命用政令食祭祀治禁辨鼓受教服啐	大小正內長共胥	下上中
LL	二四八六十五三百九	歲	士徒史客職國野官賓軍群兵司仆吏	邦市地方田府	掌祀法寸事車尺物鼓刑旅器牲財喪圭節旗材獄金馬膳輪獸弓筋數角罰弊鐘畿	令禁治詔祭屬訟舞分役贊政參任縣兇守裸教建待巡奏頒馭敘蹕圜積戒辨	共胥均大隸	

%DIFF		昨	仆史嬪徒吏	府域邦酭	掌帟弩畿軓鉦壝琮弊鐃駑楬唇蜃瘍輻版鎛輈筋虡筍陁部廩轂澹鑮牷柝諭甕柮毳寸勛鞀案醫法尺藪珍鱗殳皋鼙輪橈鐘甬臡繢鎮枚	馭頒蹕療蔟裸覜迤斫圜穸辨搏揉詔禁煮敘染訟弇兕判舂齌芟擾	侔液均胥瑞倨植隸	

The Zhou Li's lists are similar to those of the Yi Li and the Li Ji, but they contain many Numeric characters, unlike the latter two texts (see Table 13). However, it practically does not contain calendrical terms, so the origin of the numbers should be different. Similarly to the Yi Li, it contains many Social terms, which is not surprising for a ritual text.

Philosophical texts

8. Lun Yu

Table 14. Comparative keyword chart for the Lun Yu

Meth-od	**Numer-ic**	**Calen-drical**	**So-cial**	**Politico-Geo-graphical**	**Nouns**	**Verbs**	**Adjec-tive/ Adverb**	**Misc**
PHP	三		子公人君夫孔		禮事仁道路	曰為有可謂行言知聞見問學	好	是吾未
LL			子孔友君	邦	仁路貢道張顏色樊善勇恥忠蔽淵鯉	曰問學知冉言聞回見信政求謂行無怨沽敏	直佞遠怡巍偲騫遲枉	吾
%DIFF			臾友	顓	鯉樊貢蔽仁路顏畔張牆淵勇恥禱色躬	沽冉諒切譬學敏誨回欺問惑恭	偲怡巍騫希枉便佞倦遲愚泰篤貧狂驕儉直	斯

In the Lun Yu lists, calendrical and numeric characters disappear in all lists (see Table 14). There are comparatively less keyword characters for Social actors than in many Ritualistic texts. Politico-geographical area is also non-existent. The main keywords there are defined by catch-all sections of Nouns and Verbs, similarly to the Yi Li. However, the LL and %DIFF methods also produce many characters in a rich Adjective/Adverb area. The character Nouns lists of the PHP and LL methods capture well the main conceptual frame of the text, whereas the characters produced by the %DIFF method rather characterize it stylistically. All lists contain just one, but very important term — *ren.* Finally, verbs are very basic and reflect the discursive side of the text.

9. Mengzi

Table 15. Comparative keyword chart for the Mengzi

Method	Numeric	Calendrical	Social	Politico-Geographical	Nouns	Verbs	Adjective/ Adverb	Misc
PHP	一百		子公人君王夫民父士孟孔舜	齊國	事天仁道樂心義	曰為有可謂行言知聞欲見問食	大今然	是吾未下
LL	百		孟堯子民王奚妻舜禹孔霸	鄒	仁智善義天心耕飢道性足里惡獸	養云問仕欲放充無愛居章餽	賢湯慕誠汙瞍悅	
%DIFF			氓舜叟孟堯嫂霸	鄒邠庾	智界眸弈謳庳囂飢烹驩畎沼供耕簞鴞屑濱	餽摟孳饜慕充覺	詘煖瞍悅傑紾惻潔壑褐餓巍沛餒汙	沓

The Mengzi is a philosophical text, like the Lun Yu (see Table 15). It also has no keywords in the Numeric, Calendrical, and Politico-geographical sections. However, it has a wider list of Social terms.[46] Characters in the Nouns section are also close to the Lun Yu's lists, which means that they discuss the same subject area. The Verbs list is shorter and reflects a discourse structure.

[46] Including practically all of the Lun Yu's terms, their lists are close.

It should be noted that the absence of characters in the Numeric, Calendrical, and Politico-Geographical sections does not define the text's genre automatically. It only means that such text does not belong to historical literature. Its specific genre should be analyzed by content characters in the Nouns and Verbs sections; the Nouns play the most important role.

10. Xiao Jing

Table 16. Comparative keyword chart for the Xiao Jing

Method	Numeric	Calendrical	Social	Politico-Geographical	Nouns	Verbs	Adjective/ Adverb	Misc
PHP			子人民父		事天孝			
LL			親民父母		孝事故詩天身德思義	云爭愛教順移治敢生行	敬嚴悌滿聖明先	
%DIFF			親父母民兄家		孝蓋忠詩身法聖思事德宗刑廟天地善義	云移爭愛順教哀治養敢故守生失道	悌嚴滿悅敬昔安貴明終先	莫

XJ is a typical philosophical text. There are no characters in the Calendrical, Numeric, and Geographical sections, and a few characters in the Social and Nouns sections (see Table 16). The LL and %DIFF methods add characters to the Verbs section, and %DIFF has a large Adjective/Adverbs section.

11. Zhou Yi

Table 17. Comparative keyword chart for the Zhou Yi

Method	Numeric	Calendrical	Social	Politico-Geographical	Nouns	Verbs	Adjective/ Adverb	Misc
PHP	二三五四六九		子人君		天象位道咎彖利	曰有可行用貞志孚咎	大小正吉凶明亨終	未下上中
LL	九六				象利志頤雷井夬蠱巽兌坎	貞咎孚曰悔無終厲謙艮行蹇順剝征涉履震 應噬蒙遯渙	吉亨凶剛柔元險光壯需當攸吝	

%DIFF					央巽象拇棟頤頻兌蠱坎眚雷牝利	艮嗑嗃咝渙孚貞謙噬遯咎剝蹇悔窺婚姤	吝亨媾剛凶睽需吉贏拯汔攸妄漸惕壯柔牀	

The Zhou Yi lists characterize it as not a historical text; however, the PHP method extracts several numbers, because of their frequent usage in the text (see Table 17). However, the LL and %DIFF methods ignore numbers. The text content is best represented by PHP nouns.

12. Zhuangzi

Table 18. Comparative keyword chart for the Zhuangzi

Meth-od	Nu-meric	Ca-lendrical	Social	Politico-Geo-graphical	Nouns	Verbs	Adjec-tive/ Adverb	Misc
PHP	三一	日	子人君王夫民	國	事天道物神方德心形聖足名世樂義身仁坤	曰為有可謂死行出成言知聞欲見用生問治	大今同明然	是吾未下上中
LL	萬		尼堯奚莊墨倪	地	形物天道性足精衆意心俗身名神劍機情盜處故惡果德目海顏世累仁枝死耳跖聃崖竅聖	知生遊化謂始忘辯見回問聞惑似靜應理通論解況語觀存治譽殉齧避離調	邪然真虛冥滿老巧古愚窮全悲恬瞀默	吾嘗孰
%DIFF			倪	崑崙	崖跖筴扁影毫軀竅涯梱纆畦形蟻湖係狀機圓跂樞聃屠齋机脩杯尻精骸蜩暮籠斛竿符仞脛鑑垢疵戕鼈	攖憐闚避齧悟娛擢捐遊撓剖掊掇臥波攫響哼釣捶諛支調彫	真恬瞀僻徨彷腐漫謬默邪淡豪怪俄盲澹全蓬怵逍悲俱忮蹴	蒐嚮偏

The Zhuangzi is a regular non-historical text. Its Noun section contains many terms that qualify it as a philosophical text, especially in the PHP list (see Table 18).

Fiction

13. Shi Jing

Table 19. Comparative keyword chart for the Shi Jing

Meth-od	Nu-meric	Ca-lendri cal	Social	Politico-Geo-graphical	Nouns	Verbs	Adjec-tive/ Adverb	Misc
PHP	四百	月日	子公人君王歸民孔	國	天德女心山方思樂南方山	曰為有可行言來命維	大憂	是中予載
LL			牡		心胡思福方桑隰女	載止采俾憂式瞻飛徂寐逝烝陟懷	昊赫碩肅烈淑皇醉	予載
%DIFF				狉淇	玁駾駿鷺幡劬零荏黽坷鱮卉騏遑盬苣靁樛鎬	蹻維遹覲緝究儦咽啄泮	駉粲芾湑雝悠奕菀慘昊孑汎嚊緜	兮噦屆猗

The PHP method identifies a few Numerics and Calendricals in the Shi Jing (see Table 19). It has many Social terms in the PHP list, but almost nothing in this category in the LL and %DIFF lists. It is placed here in the "fiction" genre category, but this qualification needs to be elaborated later on the basis of keyword and content character analysis.

14. Shu Jing

Table 20. Comparative keyword chart for the Shu Jing

Meth-od	Nu-meric	Ca-lendri cal	Social	Politico-Geo-graphical	Nouns	Verbs	Adjec-tive/ Adverb	Misc
PHP	三一五四百		公人王民帝殷	周邦	事文天德	曰有言敢命用作刑	大今小明	下上惟予汝厥
LL			帝 王民禹	殷邦	德命天艱刑典厎	克訓宅越罰保休允格念逸祇恭嗣威率迪敷猷	丕永明咸誕欽庶	予惟汝厥

%DIFF				岳洛岱	忱眷敕厎彰豬洪艱棐疇	迪遜敷猷罔導沖佑肇格祇俞迓誥胤賚宅訓逸	丕亮欽惇毖訖俊誕灼彥奭	惟矧朕厥汝茲

In the PHP list of the Shu Jing, there are no characters in Calendrical and in Politico-geographical, but there are some Numeric characters (see Table 20). There are characters in the Social section in the PHP and LL lists, but the Noun section is rather small in the PHP and LL sections, and Verbs are abundant. This could be interpreted as a text about social actors.

6. Conclusion

This paper has investigated the character-frequency lists of the WSW Ctexts corpus to identify significant keyword characters. Three methods were utilized to calculate keyness score: LL significance test, CHI test, and %DIFF. Comparison of the results of the first two methods showed that LL test is very close to the CHI test, and the latter was omitted in this study. The results of the %DIFF test were different, especially for the top parts of the lists.

Beside regular keyword character lists, this study created lists of positive and negative key-keywords for the corpus. Analysis of those lists, as well as clustering texts across the top key-keywords, has shown that key-keyword groups could be used to identify genre groups of texts. It may be even more beneficial to analyze clusters of key-keyword associates and "clumps," which could be the subject of a future study.

The study further concentrated on the top parts of lists. Depending on the keyness score method, keyword filtering parameters, and text size, there could still be hundreds of keywords on lists for some texts. Although all of them could be relevant in understanding texts' content, researchers in keyword analysis usually concentrate on the top parts of these lists.[47] There is no "native" measure to identify how many characters to select for the top list, but usually a number from 50 to 100 is recommended.

Previously, the author has conducted a study on selecting most significant characters on character-frequency lists, employing "PHP" method to identify content characters. It allowed, depending on text size, to identify from 30 to 100 content characters for each text and can be considered a native measure.

It is logical to compare results from the PHP content character list and keyword analysis methods. Taking as guide the number of PHP content

[47] The full keyword lists are still available, for this study, for analysis on the accompanying GitHub website, see Appendix.

characters for a text, the number of characters from the top of the keyword lists was defined, provisionally, either 50 or 100 characters. Therefore, for each text, three content character sets were created, and for each set, characters were broken down into a set of categories, which were earlier used in PHP content study, losing the frequency order. This latter set of categories was created, based on specific characteristics of the Chun Qiu; therefore, it allows tracing how well other texts fit the Chun Qiu topic model – the model of a historical chronicle. Such set of categories is biased toward the genre of the chronicle, but in any case, it allows getting some basic genre specifics of texts.

Production of a synoptic table with keyword character lists for corpus texts has been one of main goals of this study. This table allows comparing various methods of extracting content characters. It could be called a semantic map for the texts. Owing to the scope limitation of this article, the comparison of the specific lists was not conducted thoroughly, and it is a matter of further research. The main point in the current comparison was to identify how well texts fitted the category model of the Chun Qiu and to try to group texts according to characteristics of their content lists.

Comparison of sets of extracted significant content characters demonstrates that the PHP and LL keyword character lists are most beneficial for understanding the corpus text content, and seemingly, the PHP list gives a slightly better picture of the text content. The %DIFF method, though, could be useful for analyzing the stylistic characteristics of texts.

Suggested set of categories, designed for the identification of historical texts (chronicles), demonstrated to be effective for content lists developed using the PHP and LL methods. The %DIFF method was not that effective in selecting characters, characteristics for chronicles. The developed synoptic table of content characters will be used in further studies of individual text characteristics.

Appendix

Abbreviations for texts in the WSW Ctexts Corpus

Chun Qiu	CQ / chunqiu
Chun Qiu Zuo Zhuan	CQZZ / chunqiu zuozhuan
Gongyang Zhuan	GY / gongyang
Guliang Zhuan	GL / guliang
Li Ji	LJ / liji
Lun Yu	LY / lunyu
Mengzi	MZ /mengzi
Shi Jing	SHI / shi
Shu Jing	SHU /shu
Xiao Jing	XJ /xiaojing

Yi Li	YI / yili
Zhou Yi	ZY / zhouyi
Zhuangzi	ZHZ / zhuangzi
Zhou Li	ZL /zhouli
Zuo Zhuan	ZZ /zuozhuan

Top key-keyword characters for clustering experiment

天子無行民道萬三來圍孔己方明曰正歸父百雨故公古國地姜孰思怨政敗王詩變足身邦義

Internet resources

The corpus website:
http://www.umass.edu/ctexts/index.php
Username: ctexts, password: umass

GITHUB resources

Github DOI: https://github.com/wsw-ctexts/vocabulary_richness/

chinese_classics_keywords_log_likelihood_short.xlsx

This file contains lists of LL keyword characters for the corpus. Positive and negative keywords for each text are contained in a dedicated sheet (e.g., for the Chun Qiu in the sheet "LL_CQ"). It has three areas: are of positive keywords, area of negative keywords, and comparison with PHP content characters (the list of PHP synsets is contained in "PHP_SYNSETS" sheet). Summary data is contained in the "KK_LL_STATS" sheet. The "KW_Charts" sheet contains comparative charts of LL, CHI and %DIFF keywords. The "KW_Summary" sheet contains a detailed comparison of PHP and LL characters (with a chart). The sheet "LL_ALL" contains the synoptic list of all LL keyword characters of the corpus. The sheet "LL_ALL_KKW" contains the list of all key-keywords. The sheet "LL_KKW_SUMMARY" contains the list of key-keywords with lists of their texts attached. The sheet "LL_KKW_POS_SUMMARY" contains a similar list of only positive key-keywords. The sheet "LL_KKW"NEG_SUMMARY" contains a similar list of negative key-keywords.

chinese_classics_keywords_CHI_short.xlsx

The file contains data on CHI keyword characters of the corpus. The file has the same structure, as chinese_classics_keywords_log_likelihood_short.xlsx, but only has "KW_CHI_CHARTS" and "CHI_SUMMARY" sheets as summary data.

chinese_classics_keywords_diff_short.xlsx

The file contains data on CHI keyword characters of the corpus. The file has the same structure, as chinese_classics_keywords_log_likelihood_short.xlsx**,** but is missing comparison with PHP content characters.

chinese_classics_keykeywords_matrix.xlsx

The file contains data on top LL key-keywords, in the form, used for clustering algorithms.

Literature

Archer, Dawn. "Does Frequency Really Matter?" In *What's in a Word-List? Investigating Word Frequency and Keyword Extraction*, ed. by In Dawn Archer, n.p.: Ashgate Pub. Limited, 2009, 1–16.

Archer, Dawn (ed.) *What's in a Word-List? Investigating Word Frequency and Keyword Extraction*. Digital Research in the Arts and Humanities. n.p.: Ashgate Pub. Limited, 2009.

Baker, Paul. 'The question is, how cruel is it?' Keywords, Foxhunting and the House of Commons. In *Word Frequency and Keyword Extraction AHRCICT Methods Network Expert Seminar on Linguistics 8 September 2006*, Lancaster: UCREL, 2006, 1–8.

Baker, Paul, Jesse Egbert, Tony McEnery, Amanda Potts and Bethany Gray. "Triangulating Methodological Approaches in Corpus Linguistic Research." In *Corpus Linguistics 2015, Abstract Book, ed. by Formato, Federica and Andrew Hardie.* Lancaster: UCREL, 2015, 39–41.

Baron, Alistair, Paul Rayson, and Dawn Archer. "Word Frequency and Key Word Statistics in Corpus Linguistics." *Anglistik*: *International Journal of English Studies* 20, no. 1 (2009): 41–67.

Berber-Sardinha, Tony. "Comparing Corpora with WordSmith Tools: How Large Must the Reference Corpus Be?" In *Proceedings of the Workshop on Comparing Corpora. Association for Computational Linguistics,* 9 (2000): 7–13.

Boltz, William G. "Why So Many Laozi-s?" In *Studies in Chinese Manuscripts: From the Warring States Period to the 20th Century, edited by Imre Galambos,* Budapest: Institute of East Asian Studies, ELTE, 1–32.

Bondi, Marina. "Perspectives on Keywords and Keyness: An Introduction." In *Keyness in Texts*, *ed. by Bondi, Marina, and Mike Scott.* Studies in Corpus Linguistics, 41. Amsterdam; Philadelphia: John Benjamins Pub. Co, 2010, 1–18.

Bondi, Marina, and Mike Scott (eds.) *Keyness in Texts*. Studies in Corpus Linguistics, 41. Amsterdam ; Philadelphia: John Benjamins Pub. Co, 2010.

Brooks, E. Bruce. "Before and After Matthew." In *Studies in the Didache, ed. by Jefford and Draper* (forthcoming).

Cheng, Winnie. "Corpora: Chinese-Language." In *The Encyclopedia of Applied Linguistics ed. by Carol A. Chapelle*, Oxford: Blackwell Publishing Ltd, 2012.

Cvrcek, Václav and Masako Fidler. "Not all keywords are created equal: How can we measure keyness?" In *Corpus Linguistics 2013 : Abstract Book, ed. by Hardie, Andrew, and Robbie Love*. Lancaster: UCREL, 2013, 55–57.

Donnelly, Karen. "Risk, chance, hope — the lexis of possible outcomes and infertility." In *Corpus Linguistics 2013: Abstract Book, ed. by Hardie, Andrew, and Robbie Love*. Lancaster: UCREL, 2013, 67-69

Donnelly, Karen. "'Dr Condescending' and 'Nurse Flaky': The Representation of Medical Practitioners in an Infertility Corpus." In *Corpus Linguistics 2015, Abstract Book, ed. by Formato, Federica and Andrew Hardie.* Lancaster: UCREL, 2015, 91–93.

Gabrielatos, Costas. "Selecting Query Terms to Build a Specialized Corpus from a Restricted-Access Database." *ICAME Journal* 31 (2007): 5–44.

Gabrielatos, Costas and Anna Marchi. "Keyness: Appropriate metrics and practical issues." In *CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?*, University of Bologna, 2012

Gaspari, Federico and Marco Venuti. "A golden keyword can open any corpus: theoretical and methodological issues in keyword extraction." In *Corpus Linguistics 2015, Abstract Book, ed. by Formato, Federica and Andrew Hardie.* Lancaster: UCREL, 2015, 131–133.

Goh, Gwang-Yoon. "Choosing a Reference Corpus for Keyword Calculation." *Linguistic Research* 28: 1 (2011): 239–256.

Hutchins, William John. "The Concept of 'Aboutness' in Subject Indexing." In *Aslib Proceedings*, 30, n.p.: MCB UP Ltd, 1978, 172–181.

Kao, Anne and Poteet, Steve R. (eds.) *Natural Language Processing and Text Mining*. London: Springer-Verlag, 2007.

Kilgarriff, Adam. "Language Is Never, Ever, Ever, Random." *Corpus Linguistics and Linguistic Theory* 1, no. 2 (2005): 263–276.

Lin, Chin-Yew, and Eduard Hovy. "The Automated Acquisition of Topic Signatures for Text Summarization." In *Proceedings of the 18th Conference on Computational Linguistics,* Association for Computational Linguistics, vol. 1, 2000, 495–501.

Nenkova, Ani, and Kathleen McKeown. "A Survey of Text Summarization Techniques." In *Mining Text Data*, *ed. by Charu C. Aggarwal and Cheng Xiang Zhai*, Boston: Springer US, 2012, 43–76.

Popescu, Ioan-Iovițˌ and Gabriel Altmann. *Word Frequency Studies*. Berlin; New York: Mouton de Gruyter, 2009.

Popescu, Ioan-Iovitz, J. Mačutek, and Gabriel Altmann. *Aspects of Word Frequencies*. Lüdenscheid, 2009.

Rayson, Paul, and Roger Garside. "Comparing Corpora Using Frequency Profiling." In *Proceedings of the Workshop on Comparing Corpora*, Association for Computational Linguistics, 2000, 1–6.

Rayson, Paul. "From Key Words to Key Semantic Domains." *International Journal of Corpus Linguistics* 13, no. 4 (2008): 519–49.

Richardson, Kay. "Keywords Revisited: The Present as History." *Social Semiotics* 5, no. 1 (1995): 101–17.

Scott, Mike. "PC Analysis of Key Words—and Key Key Words." *System* 25, no. 2 (1997): 233–245.

Scott, Mike. *Wordsmith Tools Version 3*. Oxford: Oxford University Press, 1999.

Scott, Mike. 'Picturing the Key Words of a Very Large Corpus and their Lexical Upshots — or Getting at the Guardian's View of the World" in *Teaching and Learning by Doing Corpus Analysis, ed. by B. Kettemann and G. Marko*, Amsterdam: Rodopi, 2002, 43–50.

Scott, Mike. *WordSmith Tools Version 5*. Liverpool: Lexical Analysis Software, 2010.

Stubbs, Michael. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford, Malden, MA: Blackwell Publishing, 2001.

Tsou, Benjamin K., and Olivia Oi Yee Kwong. "Some Basic and Salient Linguistic Features Across Chinese Speech Communities from a Corpus Linguistics Perspective," In *The Oxford Handbook of Chinese Linguistics, ed. by William S-Y. Wang and Chaofen Sun*. Oxford: Oxford University Press, 2015, 1–23.

Williams, Raymond. *Keywords: A Vocabulary of Culture and Society*. Oxford University Press, 2014.

Zinin, Sergey. *"Pre-Qin Digital Classics: Study of Text Length Variations". In* 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, 270–311, 2014.

Zinin, Sergey. "*Vocabulary richness of early Chinese texts: macroanalysis of the Thirteen classics and the Zhuangzi*." In Scholarly Reports of the Department of China of the Institute of Oriental Studies, Russian Academy of Sciences, issue 20, The 46th Conference "Society and State in China", vol. XLVI, pt. 1, Moscow, 197–253, 2016

Zinin, Sergey. *"Analysis of character-frequency lists of Chinese classics and its application to content analysis and genre attribution.*" In Scholarly Reports of the Department of China of the Institute of Oriental Studies, Russian Academy of Sciences, issue 22, The 47th Conference "Society and State in China", vol. XLVII, pt. 1, Moscow, 254–299, 2017

CONTENTS

History and ethnography

Source studies and historiography

Economics, politics and law

In memoriam