*С.В. Зинин**

Частотный иероглифический словарь классических китайских текстов и его использование в тематическом и жанровом анализе

АННОТАЦИЯ: В данной статье анализируются иероглифические словари корпуса WSW Ctexts с использованием методологии, разработанной Попеску и Альтманном, с целью выявления жанровых и тематических сходств между текстами корпуса. Методология Попеску-Альтманна вводит понятие h-point для разделения частотных списков иероглифов текста на так называемые синсемантическую и автосемантическую части. С этой целью автор разработал специальный список синсемантических иероглифов классического китайского языка. Хотя не все методы Попеску-Альтманна оказываются плодотворными, анализ автосемантических иероглифов, оказавшихся в верхней части списка (так называемых «тематических иероглифов») позволяет получить интересные результаты, в особенности в применении к историческим текстам. Предложенный метод может оказаться полезным для жанрового и тематического статистического анализа древнекитайских текстов.

КЛЮЧЕВЫЕ СЛОВА: статистические методы в лингвистике; корпус древнекитайских текстов; «Тринадцать канонов»; «Чжуанцзы»; частотный спектр; насыщенность словаря; словарное покрытие; синсемантические иероглифы; автосемантические иероглифы; тематические иероглифы; точка h; точка k; тематическая концентрация; макроуровневое единство текста; теория органического роста текста.

* Зинин Сергей Васильевич, к.ф.н., Торонто, Канада; Исследовательский проект «Сражающиеся царства», Массачусетский университет (Амхерст); E-mail: szinin@research.umass.edu

Content

1. Introduction

This article analyzes the character frequency in the WSP Ctexts corpus[1] following the methodology developed by Popescu and Altmann (PA)

[1] This resource, created by the author, is based on Creative Commons digital versions of texts of Wikisource. It contains twelve 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, the Zhou Li). It is excluding the Er Ya, but is including the Zhuang Zi (added as a balancing text). It also treats combined the Chun Qiu and the Zuo Zhuan as a separate text, the Chun Qiu Zuo Zhuan.

(see Popescu et al., *Aspects)*, which uses such concepts as h-point, k-point, cumulative spectra, and thematic concentration (described below). The character vocabulary of these texts may be considered the most important and representative character set of the pre-Han period (even extending to the Han period). The study uses the open source digital resource of the WSP Ctexts[2] and is reproducible. Furthermore, all results are available on GitHub[3].

This is the third study in a series of articles describing the character vocabularies of texts in the WSP Ctexts corpus[4]. It concentrates on the character frequency lists of the texts in corpus to lay a foundation for further studies of individual text[5]. It investigates aspects of character frequencies of the vocabularies of the Thirteen classics, starting from character-frequency lists ("wordlists").

The wordlists, based on word frequencies, have well-known issues that limit their use for study of text topicality. One such limitation is that the highest-frequency positions on such lists are usually occupied by words

[2] "Character vocabulary" is a text vocabulary, having characters, not "words" as its entries.

[3] DOI: https://github.com/wsw-ctexts/vocabulary_richness

[4] The first article (Zinin, "Pre-Qin Digital Classics") delineated the area of study and the state of the art of digital resources in classical Chinese texts, including history and problems of their digitization. To expose discrepancies among existing digital resources, this earlier study concentrated on numerical parameters of texts, such as text lengths (in characters) and numbers of type characters, avoiding addressing specific textology issues. These parameters vary considerably, not only due to variation in text versions, but also due to difficulties in the digitization process. In addition, no open-source academically verified digital resources are yet available for all researchers, to serve as a "gold standard"; therefore, most studies are not reproducible because their resources are not available to other researchers. The author offered the Ctext platform as free source of reproducible research. The second article (Zinin, "Vocabulary Richness") provided a general description of character vocabularies of texts in the Ctexts corpus. It concentrated on the analysis of vocabulary richness and variety in Ctexts. As most of existing measures ("indicators") have not yet been tested on classical Chinese texts, the author applied to these texts the methods of Tweedy and Baayen (Tweedy and Baayen, "How Variable May a Constant be?"). The author analyzed the vocabulary variety by using a dynamic Text-to-Token-Ratio (TTR) indicator. This approach revealed some interesting text groupings and found that some indicators work better than others. The Shi Jing and the Yi Li were identified as two "extreme poles" of vocabulary growth of the Thirteen Classics.

[5] "[E]xamining the wordlist from a text or a specialized corpus can be a starting point for examining the lexis of a particular text or text type" (Kytö and Lüdeling, *Corpus linguistics: an international handbook*, 726).

that serve mostly grammatical functions and do not provide much content information. Separation of these words from "regular" words leads to splitting all words (or characters) into two groups, known under various terms[6]. Among these terms are pairs such as grammatical-content, functional-content, form-structure words, etc. (Klammer et al., *Instructor's Manual*)[7]. In computational linguistics, another category exists that is close (but not identical) to function words; namely, stop words[8]. In the frequency range, function words tend to be at the top of the list. However, some content words, which are important for understanding the text's topic or genre, could be also close to the top of the frequency list. The problem is to find a method to separate these groups and select a group of relevant words.

Various methods are designed to use vocabulary for analyzing text topicality[9]; for frequency lists, it is necessary to decide how many words

[6] "The most frequent words in a language tend to be grammatical or "closed class" words. When the user wants to look beyond the most frequent grammatical words to see which are the "content" words, which are used most often, a "stop list" may be used. Looking at frequently occurring words may also tell you something about the themes or topics in the texts in a corpus" (Kytö and Lüdeling. *Corpus linguistics: an international handbook*, 729).

[7] In China, function words are often named "empty words" (*xu ci*). See, e.g., Nakagawa et al. "Chinese term extraction".

[8] "This is a list of words that a program omits from its searches. Such lists typically contain grammatical or "closed class" words, and/or simply the most common words in a corpus. If the user is interested in the grammatical words, then it may be necessary to make sure that a stop list is not being used by the program by default. Stop lists are sometimes used by software to prevent users from searching for the most common words, as this would be too big a task for software and would produce too many results to analyze" (Kytö and Lüdeling. *Corpus linguistics: an international handbook*, 729).

[9] For example, the concept of "keywords": "[t]he keywords of a text, in the sense intended here, are words which can be shown to occur in the text with a frequency greater than the expected frequency (using some relevant measure), to an extent which is statistically significant" (Kytö and Lüdeling. *Corpus linguistics: an international handbook*, 730). "Keywords are an attempt to characterize the topic, themes or style of a text or corpus. As such, compared to some other forms of analysis using a corpus, keywords analysis tends to focus on the ways in which texts function, rather than on overall characterizations of a corpus, or focusing on isolated linguistic elements in the corpus" (Kytö and Lüdeling. *Corpus linguistics: an international handbook*, 733). "A final problem relates to frequency and salience. The words, which are perceived by the reader as the most significant in a text, are not necessarily only those, which occur more frequently than the reader would expect… Keywords can suggest ways to start to understand the topics or style of a text, and provide statistical evidence for certain textual phenomena, but cannot pro-

from the top are required or where to put divider between important and "not-so-important" (topically) parts of the list. Various approaches exist to select a few high-frequency content words, which are representative of the text topic. For example, it could be the top 10% or top 100 words. The key is to identify where to stop to select words that are important from the point of view of characterizing topicality.

Recently, PA offered a new solution to this problem, based on the application of the "h point" concept to linguistics[10]. They also developed other interesting indicators for handling frequency spectra of texts, such as "a-index," "b-index," etc. The h-point method is a modification of Hirsch's "h-index" method, applied for word frequencies[11]. Upon creating a frequency ranking of words, 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" (Popescu et al., *Aspects*, 24), i.e., where r = f(r). For example, in the ranking

r	1, 2, 3, 4, 5
f(r)	4, 2, 1, 1, 1

the h-point is 2 (Popescu et al., *Aspects*, 24).

Popescu and Altmann use specific terms "synsemantic" and "autosemantic" for the dichotomy, which correspond to the idea of "grammatical-content" division. "The h-point seems to be an important indicator in rank-frequency phenomena. [….] The number of synsemantic tokens in texts is always greater than that of autosemantics, and usually they occupy the first ranks by frequency. The h-point forms a fuzzy threshold between these two kinds of words. […] autosemantics may occur more frequently than f(h) and their occurrence in the pre-h domain signalizes their association to the theme of the text. […] The more autosemantics are in this domain and the more frequent they are, the greater the thematic concentration of the text" (Popescu et al, *Aspects*, 24).

Following the PA methodology, the terms synsemantic and autosemantic are used in this article. These concepts allow us to set up the area of frequency list where the content words are important for defining the text topic. This article uses this methodology to look at frequency lists of

---

vide a list of all the interesting words, reveal all stylistic devices, and cannot explain a text" (Kytö and Lüdeling. *Corpus linguistics: an international handbook*, 733).

[10] See Hirsh, "An Index to Quantify".

[11] Hirsh's concept of h-point is not new; it is an "extension of the mathematical fixed point to the actual discrete rank-frequency distribution" (Mačutek et al., "Confidence intervals", 45).

the vocabularies of WSP Ctext corpus and investigates how these lists could be used to analyze text topicality and genre classification.

One problem is how to define exactly the set of synsemantic characters for classical Chinese. Even for English, no clear-cut definition of this term exists in general linguistics, and it is even more difficult to define such a character list for the classical Chinese[12]. Computational linguistics has a slightly different concept of "stop-words," i.e., the list of words that are "not important" for text content, but these lists are also not available for classical Chinese. A synsemantic character list is thus developed in this work by combining a few available resources, such as existing stop-word lists for modern Chinese (created for Baidu search engine[13]) and a list of grammatical characters from classical Chinese grammar books[14].

The article analyzes "h-points" of character lists of texts, and "pre-h" and "post-h" domains in the view of synsemantic and autosemantic character sets. The primary goal is to investigate whether character frequency lists provide topical information that could be used for text classification and genre attribution, and how the PA methodology could be used for this[15]. Based on the application of the h-point concept, the autosemantic lists of characters that provide topical information for each text in the corpus is developed. The comparison of these sets helps to group similar texts and their genre attribution[16].

Previous work

General works on frequency statistics of characters are not lacking, especially in modern texts (see, e.g., the review in Da, "A corpus-based

[12] For example, Hao states that "no commonly acceptable stop word list has been constructed for Chinese language" (Hao, 2008, 718). This does not mean that such work has not been done. Many individual stop word lists were produced, but most are not available publicly.

[13] DOI: http://www.baiduguide.com/baidu-stopwords/

[14] There are many grammar references on function words and characters in classical Chinese, e.g., Dobson, *A Dictionary of the Chinese Particles*; Chi, *Gudai hanyu*; Bai and Chi, *Gudai hanyu*; Wang, *Gui hanyu*; Yang Bojun, *Gui hanyu*. For this study, a filtered list was created that combined Baidu, GitHub, and Wang's lists (see Appendix 2).

[15] Topical analysis of texts by their vocabulary is not directly connected to genre attribution. However, in some cases genre attribution may use vocabulary analysis.

[16] Whereas it is obvious, that multiple character words existed in this period there is no recognized and freely available digital resource of the texts with marked word boundaries. Meanwhile, not only were most words at this period single-character words, but character vocabulary by itself was also the main staple of all studies on text vocabularies of this period.

study of character and bigram frequencies," 1–2). However, for classical Chinese, most statistics started to be collected only recently, after digital versions of texts became available (i.e., from the 1980s). Initially, researchers presented statistics on text length, but soon statistics on characters became available. Da is one of pioneers (Da, "A corpus-based study of character and bigram frequencies") and published data on general frequency of characters in Classical Chinese[17].

In 2001 Guo published a pioneering article on Chinese classics frequencies (Guo, "Gudai hanyu"). The study concentrates on three classes of the most frequent characters and breaks them down by genres, as well as by stroke distribution, etc. Guo, as is popular in Chinese linguistics, implements a "frequency-zone" approach to identify meaningful areas of frequency lists. Guo selects the first hundred most-frequent characters, divides them into three zones: A (1–10), B (11–30), and C (31–100) (Guo, "Gudai hanyu", 70), and investigates their properties. Guo provides these data on individual texts (Shi jing, Lunyu, Zuozhuan) (Guo, "Gudai hanyu", 71) and on most pre-Qin and Han texts, breaking them down by genres (historical, philosophical, poetical) and by part-of-speech (POS) categories (Guo, "Gudai hanyu", 73). He also analyzes semantics and compares with frequencies of modern Chinese characters. Guo also compares character frequencies with word frequencies and looks at stroke distribution and phonology aspects.

Qin provides general statistics for characters in classics (Qin, "Xianqin guji") and frequencies of singletons and related most-frequent characters. She divides characters by frequency into five zones and analyzes their characteristics.

Li Bo (Li, ***Shiji zipin yanjiu***) published a monograph on character frequency in the Shi Ji. It provides character statistics for all parts of the text (Li, ***Shiji zipin***, 20–38). Li divides the frequency list into five zones: most-frequent core, high, medium, low, and singletons, and compares frequencies with other texts. Similarly to Guo, Li analyzes the part-of-speech (POS) breakdown, the distribution of personal and geographical names, etc.[18]

Li Xiang (Li, ***Shisanjing jigao***) provides the first study devoted to the Thirteen classics. Li Xiang breaks down the hundred most-frequent characters into four groups (1–10, 10–30, 30–50, 50–100 or A, B, C, D) and analyzes them (Li, ***Shisanjing jigao***, 22). Li Xiang also implements a

[17] Da is the author of the popular online table of frequencies of classical Chinese characters, which is still the most cited; DOI: http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=CL

[18] There is also frequency comparisons with other historical texts and the Thirteen classics.

genre-oriented approach (philosophical, historical, language, ritual, fiction, ethical) and produces a distribution by POS.

In a monograph on frequencies of the Thirteen classics, Hai Liuwen (Hai, *Shisanjing zipin*) implements the schema of five frequency zones (Hai, *Shisanjing zipin*, 19–20), analyzing them for each classic (Hai, *Shisanjing zipin*, 30–37, 39–42) and comparing texts by the number of characters in each layer. He also introduces modern dictionary rank frequencies (Hai, *Shisanjing zipin*, 45–85).

In a recent book on information management of pre-Qin cultural monuments, Chen Xiaohe analyses twenty-five pre-Qin texts (Chen, *Xian Qin wenxian*). Chen also applies genre classification to Pre-Qin classics, although it is a rather traditional "Siku-style" classification. Chen provides frequency statistics not only for characters, but also for words.

Thanks to this research, the statistics of character frequencies of the Thirteen classics is well developed. However, the majority of research has concentrated on creating lists of most-frequent characters for various classics. The importance of meaningful separation of frequency lists is often understood by dividing the lists into frequency zones and discussing the resulting stylistic characteristics. However, the size of zones and their partitioning were identified with no relation to text sizes, e.g., by selecting the "top ten characters," "characters with ranks from 30 to 50," etc. Such research, as a rule (excepting Chen, *Xian Qin wenxian*), did not separate function and content characters, and uses most-frequent characters for stylistic analysis, not content analysis. Therefore, these works have no significant impact on the research presented herein. To the best of this author's knowledge, the PA methods have yet to be applied to classical Chinese texts (although they have been applied to later medieval texts[19]).

This study consists of the following parts: First, we analyze character rank-frequency charts of texts in the corpus. Second, we study frequency spectrum charts and h-points for these texts. Finally, we compare autosemantic characters contained in pre-h parts of frequency lists, thereby building the foundation for thematic analysis (although we do not present complete content analyses[20]). We identify and use some features in these lists to prove that they may be used for attribution of historical narratives in the corpus.

---

[19] Liu Haitao et al. (Yu and Liu, "Comparison of vocabulary richness", Pan, Hui and Liu, "Golden section", etc.) used this index to investigate genre in classical Chinese novel. Liu also applies it to translations of the novel "Hongloumeng".

[20] This study assigns the terms "thematic analysis" and "content analysis" a similar meaning, resulting from the preference for the terms "thematic analysis", "thematic concentration" in the PA methodology, whereas "content analysis" seems to be a wider category, covering more types of qualitative research. At-

The author is grateful to E. Bruce Brooks, who has supported the WSP Ctexts project from its beginning, and has encouraged his research.

2. Rank-frequency charts

Popescu observes that word frequency is not an "intrinsic property because it cannot be measured directly on the word [by] using some operational definitions" (Popescu, *Word Frequencies*, 1). Only empirical frequencies may be obtained by studying text corpora, and these frequencies are text dependent. However, with the proper choice of corpus, the researcher can still obtain some important information about large areas of literature. In classical Chinese, corpus sizes are generally limited. Although the WSP Ctexts corpus is not the largest collection of texts, it is representative in terms of character vocabulary. Therefore, a study of character frequencies in the WSP Ctexts should provide valuable information on area of classical Chinese studies.

As noted above, character-frequency lists of digital versions of Chinese classics depend strongly not just on the given text, but also on the specific digital version of the texts and its digitalization procedure. For example, for three top-frequency lists of the same classics, which are publicly available, the top 100 characters are not identical (see Zinin, "*Pre-Qin Digital Classics*"). Identifying the reason for such discrepancies is difficult because source texts are most often not available, which is why the present study is based on open-source and replicable technology. The texts and frequency lists are available online for replication and verification.

Table 1 presents the 36 most frequent characters for the Chun Qiu in WSP Ctexts corpus.

Table 1. The top part of frequency list for the Chun Qiu. The graph column contains characters ordered according to their frequency ("freq" column) and assigned a rank ("rank" column). If a character belongs to synsemantics group, it has 1 or 2 (class of synsemantic) in the "function" column

rank	function	graph	freq
1		月	713
2	2	公	654
3	2	人	463
4		侯	417
5		齊	397

tempts have recently been made to establish distinctions between the two concepts (see, e.g., Vaismoradi M. et al., "Content analysis and thematic analysis"), but these distinctions are not relevant for this article.

6		子	396
7	1	于	363
8		晉	310
9		十	305
10		宋	279
11	2	有	263
12		年	261
13		夏	255
14		鄭	251
15		春	244
16		伯	239
17		冬	239
18		秋	239
19		衛	239
20		伐	233
21		師	229
22	1	會	214
23		孫	198
24		王	195
25		二	180
26	1	卒	179
27		楚	163
28	2	來	162
29		陳	150
30	1	如	148
31		邾	141
32		曹	128
33		葬	124
34		帥	121
35	1	自	118
36	2	大	113

Figure 1 presents a rank-frequency chart of the first hundred rank entries of character frequencies in the WSW Ctexts corpus. This type of chart shows character ranks on the x-axis and their frequencies on the y-axis. For example, in Table 1, the character *yue* has the first rank with an absolute frequency 713 tokens, the character *gong* has the second rank with an absolute frequency 654, etc. Rank-frequency curves ignore individual characters in frequency lists and can only provide information on potential stylistic differences in character use and text coverage.

The rank-frequency curves of texts in Fig. 1 seem to be very similar in form if we exclude the middle segment (between numbers 5 and 25, which is actually the most interesting part).

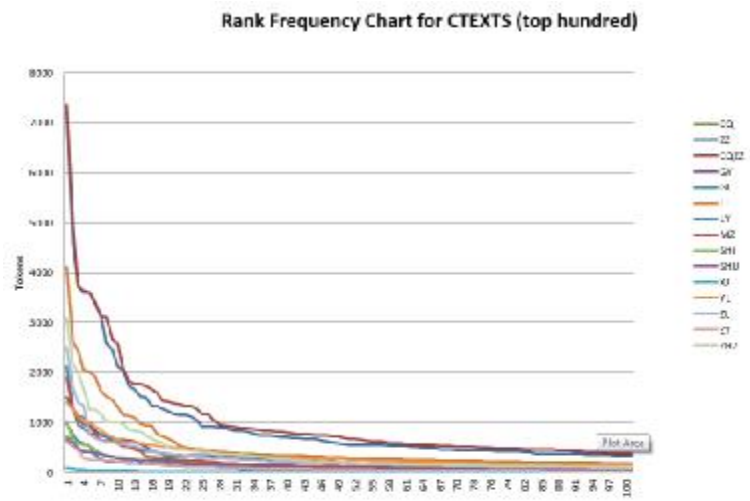

Figure 1. Rank Frequency chart for the WSP Ctexts (top hundred). The x-axis gives the position in the frequency list (rank) and the y-axis gives the number of entries for the character with this rank

Some understanding of these curves is gained by fitting them with curves formed from known parameters. Such an approach allows for numerical comparison of curves and indirect evaluation of parameters of unknown original distributions (i.e., stylistic differences). For this study, the fitting (for the first hundred positions) was done with the help of free MyCurveFit software[21]. Generally, rank-frequency curves should be best fit by power laws (according to Zipf's Law). However, in this case, the best fit is obtained with a fourth-order polynomial (4PL). Tables 2 and 3 give the fitting parameters[22].

Table 2. Fourth-order polynomial fittings of WSP Ctext corpus frequency curves, sorted by parameter B (decreasing). The first column is the

[21] DOI: https://www.mycurvefit.com/

[22] The parameters "a" and "b" control points at the beginning and the end of curve (the upper and lower asymptotes of the curve), and the parameters "b" and "c" control curve slope: the parameter "b" controls the steepness of curvature at point c (Hill's slope), and "c" controls the inflection point (i.e., the point on the S-shaped curve halfway between a and d). The greater the parameter "b", the steeper is the slope.

abbreviation of text titles (see abbreviation list at the end of the article), and columns a, b, c, d contain the respective parameters of the fitting curves

Texts sorted by parameter b of 4pl curves				
	a	b	c	d
LY	1201.805	1.269255	3.101328	18.46599
LJ	5460.579	0.94665	2.613665	-20.0199
GY	1839.398	0.904307	5.315445	-27.3502
MZ	2920.835	0.883385	1.837437	-32.1025
SHI	1694.393	0.794405	1.496774	-16.5317
ZHZ	5143.033	0.78716	1.658353	-123.241
ZL	8642.383	0.747191	0.309148	-19.3749
ZZ	13986.09	0.738874	1.025051	-157.811
CQZZ	17561.42	0.626911	0.585614	-343.63
ZY	1629.075	0.583592	0.543049	-58.8629
GL	5.85E+09	0.497013	1.13E-13	-131.875
XJ	398.9686	0.454172	0.087503	-13.1246
CQ	1717.798	0.420444	0.852837	-206.54
YL	-34.4658	-0.82093	6.802611	1713.88
SHU	-8.9651	-0.92738	6.095543	763.9737

Table 3. Fourth-order polynomial fittings of WSP Ctext corpus frequency curves, sorted by parameter c (increasing). The first column is the abbreviation of text titles (see abbreviation list at the end of the article), and columns a, b, c, d contain the respective parameters of the fitting curves

Texts sorted by parameter c of 4pl curves				
	a	b	c	d
GL	5.85E+09	0.497013	1.13E-13	-131.875
XJ	398.9686	0.454172	0.087503	-13.1246
ZL	8642.383	0.747191	0.309148	-19.3749
ZY	1629.075	0.583592	0.543049	-58.8629
CQZZ	17561.42	0.626911	0.585614	-343.63
CQ	1717.798	0.420444	0.852837	-206.54
ZZ	13986.09	0.738874	1.025051	-157.811
SHI	1694.393	0.794405	1.496774	-16.5317
ZHZ	5143.033	0.78716	1.658353	-123.241
MZ	2920.835	0.883385	1.837437	-32.1025
LJ	5460.579	0.94665	2.613665	-20.0199
LY	1201.805	1.269255	3.101328	18.46599
GY	1839.398	0.904307	5.315445	-27.3502
SHU	-8.9651	-0.92738	6.095543	763.9737
YL	-34.4658	-0.82093	6.802611	1713.88

The parameters of fitting curves vary considerably, which allows the texts to be classified. For example, in Lun Yu, the rank frequency falls quickly, whereas the character distribution in Shu Jing seems to be more even. Evaluating by the inflection point in the fitting curve identifies three distinctive groups of texts: $c < 1$, $c \sim 1$, and $c > 1$.

This behavior needs explanation. The steep fall of the curve means a shorter distance to the inflection point, which suggests a smaller group of high-frequency characters, but fewer characters exist between high- and low-frequency groups. A flatter curve means that character frequencies are distributed more evenly across the rank spectrum. Text vocabularies may be compared from this point of view, but this analysis is beyond the scope of this article. Upon sorting the texts of the WSP Ctexts corpus by parameters b and c of their curve fits, they do not fit the usual genre grouping of such texts[23].

3. Frequency-spectrum charts

Another way to look at character coverage of text through frequencies involves the frequency spectrum. The characters should be ordered according to how many times they appear in text. For example, 327 unique characters appear in the Chun Qiu (singletons), 151 characters appear twice in the text, etc. Therefore, the value 327 is assigned to spectrum-rank 1, etc. The frequency-spectrum distribution charts for the WSP Ctexts corpus are presented in Fig. 2.

[23] The Thirteen classics may be classified as "historical texts" (the Chun Qiu, the Zuo Zhuan, the Gongyang Zhuan, the Guliang Zhuan), "ritualistic texts" (the Yi Li, the Li Ji, the Zhou Li), "philosophical texts" (the Lun Yu, the Mengzi, the Zhou Yi (and added the Zhuangzi)), and "fiction/poetry" (the Shi Jing and the Shu Jing). Naturally, this classification is not rigid. For example, the Shu Jing could be also classified with historical texts (stylistically, however, its vocabulary-richness parameters put it closer to the Shi Jing (see Zinin, "Vocabulary Richness")). On the other side, vocabulary-richness indicators do not group texts of the same genre together, so they are not decisive in genre attribution. For example, the developmental profiles of Type-Token-Ratio (TTR) suggests breaking the corpus into four groups of texts: (1) The Shi Jing; (2) The Shu Jing, the Zhuangzi, the Li Ji, the Zuo Zhuan, the Chun Qiu Zuo Zhuan, the Lun Yu, the Mengzi, and the Zhou Li; (3) The Zhou Yi, the Chun Qiu, the Guliang Zhuan, and the Gongyang Zhuan; and (4) The Yi Li. (Zinin, "Vocabulary Richness"). This arrangement does not group together ritualistic texts, and only vaguely groups together historical texts.

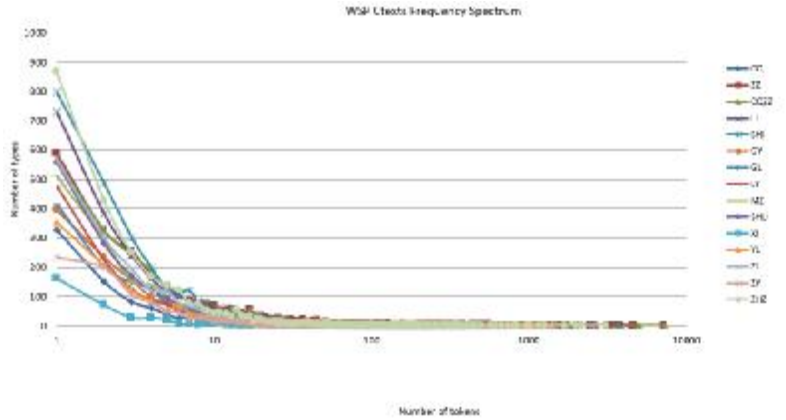

Figure 2. Frequency-spectrum chart for the Ctexts. The x-axis gives the position in the frequency spectrum list (rank) and the y-axis gives the number of characters for this spectrum rank

The most important part of the distribution is located within the first twenty positions, as presented in Fig. 3.

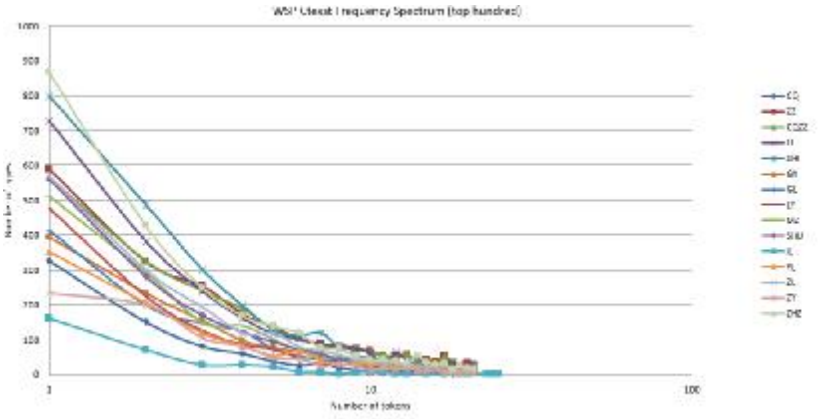

Figure 3. Frequency-spectrum chart for the Ctexts; first twenty positions. The x-axis gives the position in the frequency spectrum list (rank) and the y-axis gives the number of characters for this spectrum rank

The curve form is different from rank-frequency charts and are better fit by a power law, with the two parameters "a" and "b," displayed in Table 4[24]. The greater is parameter a, the steeper the slope descends. The steep fall of a curve means that many singletons and characters exist (i.e.,

[24] Therefore, these curves are closer to those suggested by Zipf's Law.

that appear only two or three times) and many characters have higher frequencies, with fewer of other types of characters in between.

Table 4. Power-law fits to frequency-spectrum curves for WSW Ctexts corpus, sorted by parameter "a" in decreasing order

Texts sorted by parameter a (power curves)		
	a	b
ZHZ	883.1699	-1.19849
SHI	846.4789	-1.12589
LJ	746.3928	-1.1124
ZZ	610.7547	-0.99076
CQZZ	590.8199	-0.97178
ZL	585.4565	-1.12037
SHU	573.9839	-1.17326
MZ	531.6745	-1.13227
LY	485.9311	-1.28618
GL	424.4473	-1.12391
GY	414.5473	-1.08534
YL	365.0004	-1.08784
CQ	332.6843	-1.31413
ZY	264.7787	-1.01289
XJ	162.8639	-1.42313

Text ordering is based on the parameters obtained from fitting the frequency-spectrum curves, deserves a separate study and is beyond the scope of this work.

4. H-point statistics

4.1. Stop-words

The third approach offered by the PA methodology to studying character frequencies is the analysis of pre- and post-h-point distributions. This method separates the general areas of synsemantic and autosemantic vocabulary components.

This approach requires a synsemantic list to be established. What are the synsemantic characters in classical Chinese? There is no strict definition of synsemantics in Popescu's book, which implies that the main property of a synsemantic word is its high frequency: "[it] is well known, the auxiliaries, the synsemantics etc., are usually more frequent than autosemantics" (Popescu et al., *Aspects*, 18) and "the *h-point* divides the vocabulary into two parts; namely, in a class of magnitude h of frequent synsemantics or auxiliaries (prepositions, conjunctions, pronouns, articles, particles, etc.) and a much greater class (V-h) of autosemantics which are not so frequent but build the very vocabulary of the text"

(Popescu et al., *Aspects*, 18–19). This reference to frequency rates, having no clear criteria, does not facilitate building a synsemantic list[25].

Such a list may be created based on the similarity between synsemantic and autosemantic and divisions such as content-grammar, structure-function, or structure-form. The resulting list may be merged with the input from a stop word list. However, while any grammar textbook offers examples of function words, no full ready-made list of them exists that could be used for building a synsemantic list, especially for classical Chinese. The situation is similar to that of the "stop-words" lists in computational linguistics, and they are often developed "ad hoc." Mahalakshmi and Sivasankar (Mahalakshmi and Sivasankar, "Cross Domain Sentiment Analysis", 83) note that "There is no single universal list of stop words used by all natural language processing tools. For a given purpose, any group of words can be chosen as stop words". Moreover, computational linguistics stop lists need more filtering to extract synsemantics than grammar- or function-word lists. Stop-word lists could include some of the most common words, including lexical words, such as "want," because researchers are mostly interested in the most efficient way to classify documents. Therefore, in general, there exist no clear criteria or "written-in-stone" samples of synsemantic lists.

Chinese stop words are usually words such as adjectives, adverbs, repositions, interjections, and auxiliaries (i.e., non-content words). Most stop-word lists developed for recent studies are not available publicly. Two stop-word lists exist in the public domain for modern Chinese: one was developed by Baidu[26], and the other is available from a Github repository[27].

This study uses existing lists of function words and stop-word lists to create a synsemantic single-character list of combined stop words[28] and classical Chinese function words, or "empty words" (Wang, *Gu Hanyu*). In fact, two lists were created: a core synsemantic character list, which is essentially the book's list (class 1 synsemantics), and an extended list, which includes those characters on lists that could be interpreted as content words (class 2 synsemantics, see Appendix 1). It is hoped that these

[25] Note that the term "synsemantic" in this approach could have two meanings. On the one hand, it is basically everything located in the "synsemantic area" of a given text. On the other hand, there is a list of vocabulary synsemantics. As part of thematic analysis, PA understand these as vocabulary autosemantics, which got into statistical synsemantics area are special.

[26] DOI: http://www.baiduguide.com/baidu-stopwords/

[27] DOI: https://github.com/ghpaetzold/questplusplus/blob/master/lang_resources/chinese/chinese.stopwords.txt

[28] Filtered Baidu stop words DOI: http://www.baiduguide.com/baidu-stopwords/

lists will be refined and improved. In this study, only class 1 synsemantics were considered synsemantics.

While function word distribution could be one of important tools for stylistic forensic analysis, these words are not so important when it comes to topical analysis. As Lescovec et al. note "in fact, the several hundred most common words in English (called stop words) are often removed from documents before any attempt to classify them. In fact, the indicators of the topic are relatively rare words" (Lescovec et al., *Mining of Massive Datasets*, 8).

4.2. H-point

This study presents most of indicators developed by the PA methodology for all texts in the WSP Ctexts corpus[29]. Table 5 presents all h-points for texts in the WSP Ctexts corpus (not ordered by h-point value, because h-point depends on text length).

Table 5. Frequency indicators for Ctexts corpus. The columns N and V contain data on the number of characters in a text (tokens) and the number of unique characters in the text (types). H is the h-point, and h/V is the result of dividing h-point value by V. See the list of abbreviations for abbreviations of text names

Text	N	V	h-point	h/V
CQ	16791	941	60	0.064
ZZ	178563	3235	182	0.056
CQZZ	195354	3251	191	0.059
GY	44224	1640	94	0.057
GL	40835	1594	91	0.057
LJ	97994	3041	136	0.045
LY	15923	1361	50	0.037
MZ	35354	1892	78	0.041
SHI	29622	2833	66	0.023
SHU	24539	1911	67	0.035
XJ	1800	374	20	0.053
YL	53882	1536	114	0.074
ZL	49410	2212	96	0.043
ZY	13348	1030	51	0.050
ZHZ	65251	2968	101	0.034

4.3. A-value

Many other indicators in the PA methodology are based on the h-point. One such indicator is the "a-value." The value of h (or h-1) is con-

[29] Also available online as Excel spreadsheets (see Reference section for Github links; and there are lists of characters for each text).

sidered by Popescu to be the natural "pace" of the text (Popescu, *Word Frequency*, 19). This "pace" depends on the text's length. However, a more length-independent indicator for texts is desirable, so Popescu, following Hirsh, offers the "a-value" indicator to describe an h-paced text. The influence of the text length on the indicator thus diminishes (Popescu, *Word Frequency*, 19). The expression is a = N/h^2, where N is the number of words in the text.

Unlike the h-point, which grows almost linearly with N, the a-point is a stable indicator and does not (in the case of the WSP Ctexts corpus) correlate with N in the corpus selection of texts. Fig. 4 compares h-point and a-value curves, and Fig. 5 presents a scaled a-value curve. The a-values of the WSP Ctexts corpus are presented in Table 6. They can be sorted, because the text length less affects the a-value.

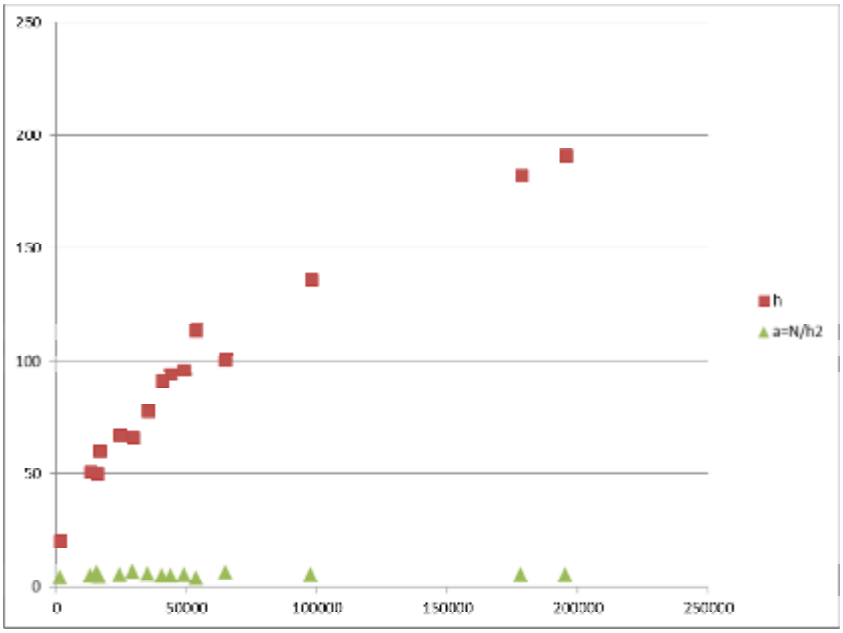

Figure 4. H-point chart depends on the number of characters in the text, whereas the a-value chart is less dependent on the number of characters

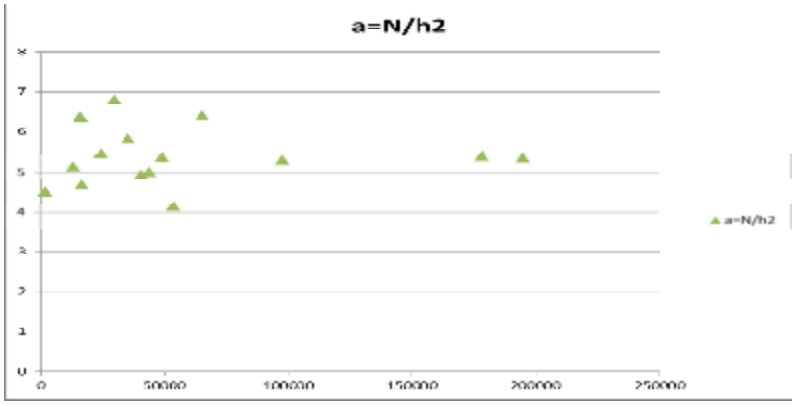

Figure 5. Scaled-up view of a-value chart from Fig. 4

Table 6. A-value sorted in descending order. The columns N and V contain data on the number of characters in a text (tokens) and the number of unique characters in the text (types). See the list of abbreviations for abbreviations of text names

Text	N	V	h-point	a-index
SHI	29622	2833	66	6.800
ZHZ	65251	2968	101	6.397
LY	15923	1361	50	6.369
MZ	35354	1892	78	5.811
SHU	24539	1911	67	5.466
ZZ	178563	3235	182	5.391
ZL	49410	2212	96	5.361
CQZZ	195354	3251	191	5.355
LJ	97994	3041	136	5.298
ZY	13348	1030	51	5.132
GY	44224	1640	94	5.005
GL	40835	1594	91	4.931
CQ	16791	941	60	4.664
XJ	1800	374	20	4.500
YL	53882	1536	114	4.146

This text ordering and ensuing group arrangements are closer to those created by vocabulary richness developmental profiles, unlike the case of rank frequency of frequency-spectrum curve parameters. As sorted by a-value, this indicator puts the Shi Jing at the top, and the Yi Li at the bottom.

Popescu remarks that, in the cross-linguistic perspective, from the point of view analyticity or syncretism, a lower average "a-value" means analyticity. In the WSP Ctexts corpus, a-value < 6.8 (between 4.1 and

6.8) for all texts and, placing it into Popescu's classification built on "a-value," classical Chinese could be placed among Polynesian languages (see Popescu, *Word Frequency*, 47–48).

5. Vocabulary richness

Popescu and Altmann also developed their own indicators of vocabulary-richness indicators: $\underline{F(h)}$ and G(k)[30].

5.1. $\underline{F(h)}$ indicator

F(h) is a cumulative function of ranked frequencies (see Table 7)[31]. The "relative frequency up to h-point, i.e., F(h) represents h-coverage of the text" (Popescu et al., *Aspects*, 30), it will be the "area covered by auxiliaries". Thus, Popescu offers a modified value $\underline{F(h)}$, which is given by

$$\underline{F(h)} = F(h) - h^2/2*N$$

This indicator offers better coverage by auxiliaries. Therefore, 1-$\underline{F(h)}$ is an aspect of vocabulary richness, reflecting the distribution of hapax legomena and other autosemantics which are, important for thematic aspects. Popescu considers $\underline{F(h)}$ to be a vocabulary-richness indicator that is independent of text length.

Table 7. F(h) sorted in descending order. The columns N and V contain data on the number of characters in a text (tokens) and the number of unique characters in the text (types). h/V is the result of dividing H by V. See the list of abbreviations for abbreviations of text names

Text	N	V	h-point	h/V	F(h)
SHI	29622	2833	66	0.023	0.396
XJ	1800	374	20	0.053	0.401
SHU	24539	1911	67	0.035	0.464
ZY	13348	1030	51	0.050	0.548
LY	15923	1361	50	0.037	0.550
ZL	49410	2212	96	0.043	0.569
MZ	35354	1892	78	0.041	0.584
LJ	97994	3041	136	0.045	0.590
ZHZ	65251	2968	101	0.034	0.595
GL	40835	1594	91	0.057	0.647
GY	44224	1640	94	0.057	0.648
ZZ	178563	3235	182	0.056	0.659

[30] Different from those covered in the previous article (Zinin, "*Vocabulary richness of early Chinese texts*").

[31] F(h), i.e., the sum of all frequencies up to rank h, or, rather, its relative value, divided by N.

YL	53882	1536	114	0.074	0.673
CQZZ	195354	3251	191	0.059	0.676
CQ	16791	941	60	0.064	0.676

Arranged by F(h), Table 7 groups together Shi, Shu, Zhouyi (including XJ), then philosophical texts (including the Li Ji), and finally the group of historical texts, including the Yi Li. This text grouping is close to the grouping obtained by dynamic TTR profiles mentioned above.

5.2. G(K) indicator

Another aspect of vocabulary richness is represented by the G(k) function. This function introduces the K-point, which is a counterpart of the h-point, but for the frequency-spectrum distribution (represented in Figs. 2 and 3). The K-point is defined in the same way as the h-point in relation to the frequency spectrum[32]. Characters from 1 to k–1 may be considered autosemantics, and characters above K may be considered synsemantics. The data on the k-point and G(k) are presented in Table 8.

Table 8. G(k) sorted in descending order. The columns N and V contain data on the number of characters in a text (tokens) and the number of unique characters in the text (types). k/V is the result of dividing k by V. See the list of abbreviations for abbreviations of text names

Text	N	V	k-point	k/V	G(k)
SHI	29622	2833	16	0.006	0.881
LY	15923	1361	14	0.010	0.868
SHU	24539	1911	16	0.008	0.854
ZHZ	65251	2968	20	0.007	0.852
ZY	13348	1030	14	0.014	0.842
XJ	1800	374	6	0.016	0.837
MZ	35354	1892	16	0.008	0.823
CQ	16791	941	12	0.013	0.821
ZL	49410	2212	18	0.008	0.807
GL	40835	1594	17	0.011	0.797
GY	44224	1640	16	0.010	0.787
LJ	97994	3041	19	0.006	0.767
CQZZ	195354	3251	26	0.008	0.737
ZZ	178563	3235	23	0.007	0.726
YL	53882	1536	14	0.009	0.723

Like F(h), G(k) groups texts in a way that is closer to the known vocabulary-richness groupings, with the Shi Jing at one end of the spectrum and the Yi Li at the other end.

[32] Intermediate points are not calculated.

5.3. Indicators k/V and b

Similar to the h-point, the k-point introduces indicators such as k/V and b = V/k^2; see Table 9.

Table 9. G(k) and b, sorted in descending order. The columns N and V contain data on the number of characters in a text (tokens) and the number of unique characters in the text (types). k/V is the result of dividing k by V. See the list of abbreviations for abbreviations of text names

Text	N	V	k-point	k/V	G(k)	b
SHI	29622	2833	16	0.006	0.881	11.066
XJ	1800	374	6	0.016	0.837	10.389
LJ	97994	3041	19	0.006	0.767	8.424
YL	53882	1536	14	0.009	0.723	7.837
SHU	24539	1911	16	0.008	0.854	7.465
ZHZ	65251	2968	20	0.007	0.852	7.420
MZ	35354	1892	16	0.008	0.823	7.391
LY	15923	1361	14	0.010	0.868	6.944
ZL	49410	2212	18	0.008	0.807	6.827
CQ	16791	941	12	0.013	0.821	6.535
GY	44224	1640	16	0.010	0.787	6.406
ZZ	178563	3235	23	0.007	0.726	6.115
GL	40835	1594	17	0.011	0.797	5.516
ZY	13348	1030	14	0.014	0.842	5.255
CQZZ	195354	3251	26	0.008	0.737	4.809

The graphic distribution of b-points is displayed in Fig. 6.

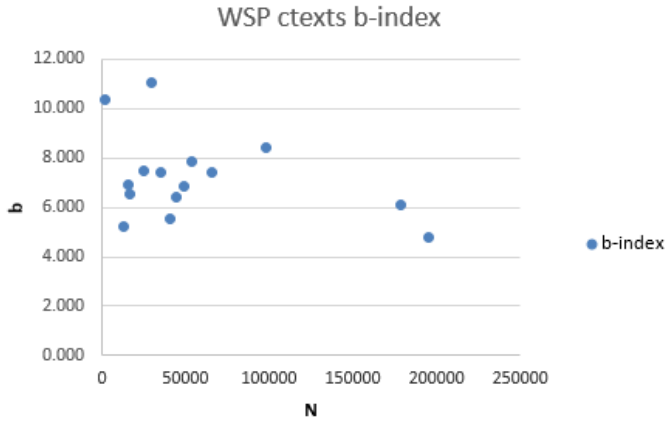

Figure 6 . b-points (y-axis) over the number of characters in the text (x-axis); the b value is less dependent on the number of characters

This picture is similar to the A-value distribution; no immediate correlation exists between text size and indicator value, in this text sample. For Popescu, the b-point has a linguistic meaning similar to that of the a-point. A greater value of b, as for a, means a more synthetic language.

5.4. Indicator of vocabulary exploitation A

In terms of the PA methodology, vocabulary richness means "more autosemantic, less synsemantic" text features (…, p. 76). Therefore, A = Ah/Amax is yet one more indicator of vocabulary richness offered by the PA methodology (see Table 10).

Table 10. Indicator A sorted in ascending order. The columns N and V contain data on the number of characters in a text (tokens) and the number of unique characters in the text (types). See the list of abbreviations for abbreviations of text names

Text	N	V	a-index	F(h)	A
XJ	1800	374	4.500	0.290	0.74027
YL	53882	1536	4.146	0.553	0.845089
CQ	16791	941	4.664	0.569	0.854369
SHU	24539	1911	5.466	0.372	0.863278
ZY	13348	1030	5.132	0.451	0.871152
GY	44224	1640	5.005	0.548	0.881051
GL	40835	1594	4.931	0.546	0.90113
SHI	29622	2833	6.800	0.323	0.913135
LY	15923	1361	6.369	0.471	0.913455
CQZZ	195354	3251	5.355	0.582	0.915681
MZ	35354	1892	5.811	0.498	0.918754
ZZ	178563	3235	5.391	0.566	0.919115
ZL	49410	2212	5.361	0.476	0.919184
LJ	97994	3041	5.298	0.495	0.922809
ZHZ	65251	2968	6.397	0.517	0.933892

Fig. 7 plots the indicator A. Whereas "the index A does not depend on text length N" (Popescu, *Word frequency*, 79), it does not group texts into arrangements discovered by the dynamic TTR profiles.

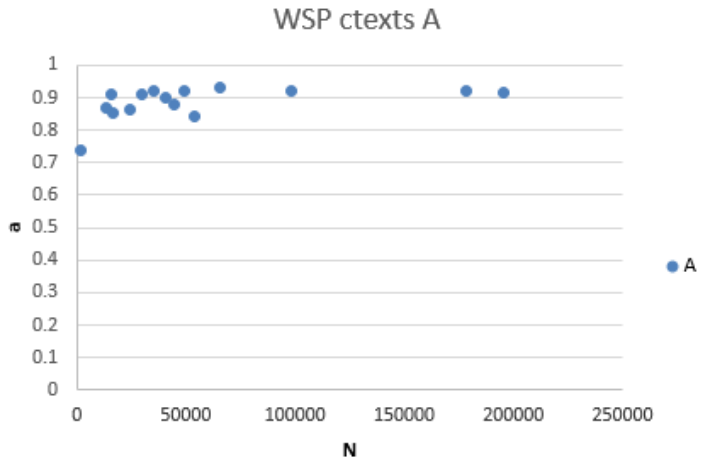

Figure 7. Indicator A (y-axis) over the number of characters in the text (x-axis); the A is less dependent on the number of characters

6. Thematic concentration

In the present work, the main interest of the h-point is its ability to open the perspective of analyzing thematic concentration and topicality of texts (Popescu, *Word frequency*, 95–96).

Even the simple indicator, 1-SYN/AUTO allows to evaluate the degree of thematic concentration[33].

Table 11. Numbers of Class 1 synsemantics (SYN), autosemantics (AUTO), Class 2 synsemantics (SYN_STO) and related indicators 1-SYN/AUTO and 1-SYN_STOP/AUTO. See the list of abbreviations for abbreviations of text names

		SYN	AUTO	1-SYN/AUTO	SYN_STOP	1-SYN_STOP/AUTO
1	CQ	11	60	0,8167	6	0,9
2	ZZ	56	182	0,6923	45	0,752747
3	CQZZ	57	191	0,7016	45	0,764398
4	GY	26	94	0,7234	26	0,723404
5	GL	28	91	0,6923	23	0,747253
6	LJ	45	136	0,6691	37	0,725926
7	LY	22	50	0,56	27	0,46
8	MZ	35	78	0,5513	32	0,589744

[33] The PA methodology offers a more complex indicator of thematic concentration (Popescu, Word frequency studies 95), which is not featured in this study.

9	SHI	29	66	0,5606	25	0,621212
10	SHU	30	67	0,5522	27	0,597015
11	XJ	13	20	0,35	10	0,5
12	YL	33	114	0,7105	24	0,789474
13	ZL	22	96	0,7708	32	0,663158
14	ZY	14	51	0,7255	11	0,784314
15	ZHZ	45	101	0,5545	38	0,595745

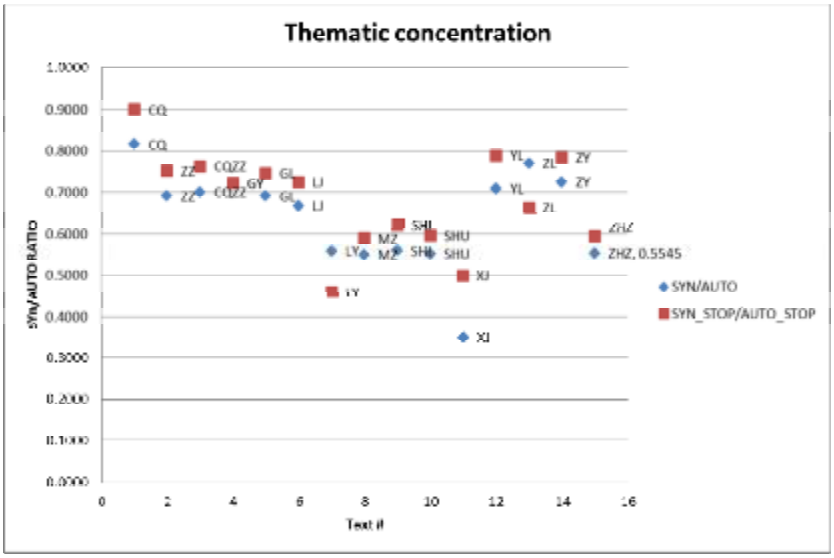

Figure 8. Values of indicators SYN/AUTO and SYN_STOP/AUTO (y-axis) over the texts (x-axis); thematic concentration is less dependent on the number of characters

The indicator consisting of the ratio of synsemantic to autosemantic in the pre-h list allows to group historic and ritual texts, grouping philosophic texts, the Shi Jing, and the Shu Jing together, whereas the Chun Qiu and the Xiao Jing are special cases.

7. Pre-h list autosemantics, thematic analysis, and genre attribution

The h-point allows resolving the problem of borders between important and unimportant (from the point of view of topicality) sections of the list of most frequent characters. Removing synsemantic characters from the pre-h-point part (or simply pre-h) of the list allows selecting the most

important content ("thematic") characters[34]. According to the PA methodology, autosemantic nouns represent the text theme, and verbs–"first-order predicates expressing the properties or actions of the central words" (Popescu, *Word frequency*, 95). Table 12 lists the pre-h-point autosemantics for all texts of WSW Ctexts corpus.

Table 12. Pre-h-point autosemantics in the WSP Ctexts corpus

CQ	月公人侯齊子晉十宋有年夏鄭春伯冬秋衛伐師孫王二楚來陳邾曹葬帥大盟正叔三七夫四八六奔五莒歸蔡九殺侵杞
ZZ	子曰公人君為晉有大侯王師楚齊國鄭是可夫伯月氏叔命吾伐孫二衛盟謂禮三宋歸事臣死殺行出十成民告入陳言年知吳敢文夏來聞季書天德欲見周城立未朝奔父一寡罪今司武秦敗用亡取五日軍趙生下門秋春馬孟仲執亂上冬辭中懼士小求政四魯獻信尹蔡令邑難女宣還獲六謀乘問右帥主棄服
CQZZ	子曰公人晉君有侯為大月師王齊楚鄭伯國夫伐是可宋衛叔孫十氏二年命吾盟三夏歸陳謂來禮殺臣事出春秋入冬行成死民告吳言知奔季文敢父城天朝聞五未一書秦周德見欲立敗四取帥日武寡罪今仲司用蔡趙六亡邾執生軍小門下孟馬曹葬正士亂上中七辭懼求獻政魯女尹還信申宣令邑難甲八圍獲食侵莒戰謀乘九
GY	公子月人為侯齊言晉大有君曰宋師伯年伐十夫鄭王衛春秋夏書冬孫楚來國婁二陳葬邾歸曹盟稱正三殺叔天未父弒季然日可入帥譏出是莒七取吾四一氏立五執
GL	公月人子侯齊有曰晉正伐十言宋伯師鄭大為年夏王春衛秋夫君冬楚來孫國日葬二陳盟邾曹殺三歸天事帥入父四辭叔一可五奔志是未七弒莒蔡出內
LJ	子曰人有君大為夫禮天三喪祭事是下民樂行服可國士父食日命上公道主言侯十王五一哭母謂成中知問出死然婦入廟月齊用小義明臣立見門敢衣孔反德長二外尊拜方百宗四位居執生朝內正學文未貴教東賓夏年世
LY	子曰人有為君可言問知吾仁夫道行謂禮孔三見學事是聞公未好路
MZ	曰子人為有王孟天可是君下民然夫大仁道謂行吾一國心公言見事知義食今問舜欲士樂齊未百聞孔父
SHI	有子人維君王言心天是為方載思女來予大民公四歸行命樂可憂曰日山月南德百國孔中
SHU	惟曰王有厥天民人命德予汝大用作子帝明邦三一五今公四言小殷百上下事敢周方文刑

[34] Popescu classify as "thematic" only nouns, verbs, and adjectives (Popescu, *Word frequency*, 95).

XJ	子孝事人天民父
YL	人拜主西面爵上賓東北升階降受執夫南賓大奠祭屍司射公命興一出位洗左觶入婦三俎取門子立席曰揖下右祝辭有反為食禮中外君正阼士酒送獻授答告荅進從堂二首幣弓長退酌馬稽佐醢篚
ZL	人二四大掌有士曰十事國為三一六王下中五徒八令祭史府祀邦民夫禮共治司物用車上喪百賓胥法方客師服九政禁小侯辨馬正入歲刑官氏謂日行命鼓長食寸帥內受教田器子
ZY	曰象吉有六九利大貞上咎人中子行君三天用亨凶二四五下小終志孚彖正可位道未來明
ZHZ	曰為人子有天知是夫吾下然大道謂可物生言見一行德君心形聞未死王今問聖足成日名世明樂中事上方義三身治仁神出用民同國欲

7.1. Thematic and genre analysis with pre-h list

Characters on the pre-h list may be used to attain various goals, such as thematic analysis of texts, or as features for genre classification. As an example of the application of autosemantic characters on pre-h lists for genre classification, this study will apply categorical analysis to demonstrate how historical narratives may be identified with the help of pre-h lists.

General categorical analysis of frequency lists has been implemented in earlier studies of Chinese classics (e.g., Guo, "Gudai hanyu", 73; Li, *Shisanjin*, 27–29; Chen et al., *Xian Qin wenxian*). Unlike these approaches, this work presents a feature analysis of a reduced number of categories, targeting historical narratives.

Fang and Cao indicate that analysis of the relation between linguistic variations and genres is based on two points of view: "(1) linguistic features have been extracted and examined across different text types; and (2) observable linguistic variations have been used to identify and classify texts automatically" (Fang and Cao, *Text Genres*, 51). The present work is driven by observing features in pre-h lists, translated into characteristics of historical narratives. Characters belonging to these features are grouped under related categories and all other characters are grouped under POS categories (to avoid building a full ontology). The featured categories are numerical, calendrical, social, and politico-geographical; the POS categories are nouns, verbs, adjectives, adverbs, and miscellaneous characters[35]. This approach allows us to test whether these characteristics of historical texts are more prominent by comparing with other texts and demonstrates

[35] The systematic approach would include other terms, e.g., ritual, ideological, and mantic terms.

that autosemantic characters in a pre-h list may be used for genre analysis. A subsequent study will implement a more complete categorical analysis.

The thematic characters (nouns, verbs, and adjectives) in autosemantic pre-h lists may be used to analyze the content of texts. This option is mentioned in the following analysis, but only to demonstrate the potential offered by the PA methodology for content analysis of classic Chinese texts.

7.2. Numeric and calendar categories

Numerical and calendar categories may be reviewed together because, in historical texts, numerical and temporal characters are often parts of dates. The most prominent text in this relation is the Chun Qiu, which features all numbers from 2 to 9 (missing one and ten) in the pre-h list, and characters for month, year, and all four seasons (missing "day"). This text has the highest percent of autosemantic tokens in these categories[36]. It characterizes the Chun Qiu, the Zuo Zhuan, and their combination in the Chun Qiu Zuo Zhuan as texts, where dating is a very important part of discourse. The same situation (with minor omissions) is observed in accompanying texts of the Gongyang Zhuan and the Guliang Zhuan.

Only the Li Ji could be closer in this aspect to the historical texts, mostly with respect to the Gongyang Zhuan and the Guliang Zhuan. In all other texts, calendar terms, except the character for day, are practically absent from the h-list part of the frequency list, so these terms may be a genre-classification feature. Numericals, as well as calendricals, are practically absent in the pre-h list of "ideological" texts, such as the Meng Zi, the Lun Yu, the Xiao Jing, and the Zhuan Zi. They are also not frequent in the poetical Shi Jing[37]. However, the "ritualistic texts," such as the Zhou Li (introduces *sui*), the Yi Li, the Li Ji, and the Shu Jing, feature a few numerical characters.

The distribution of terms in these categories demonstrates benefits of implementing the PA methodology. In general, numerical and calendar characters belong to the most-frequent characters for all these texts[38]. In philosophical texts, they could be lower on the list than in historical texts, but the significance of this difference is hard to evaluate in frequency-zoning approach. The presence of these characters in the list of autose-

[36] The character for day (*ri*), as well as numericals for one and ten, have higher frequencies in the narrative of the Zuo Zhuan. Only the combined text, the Chun Qiu Zuo Zhuan, finally, in the pre-h zone has the full set of numericals up to ten, and all temporal characters for day, month, year, and the four seasons.

[37] The Zhou Yi is an exception, due to his mantic formulas.

[38] For example, even in the Lun Yu many numbers and calendricals are among top 200 most-frequent characters.

mantics in the pre-h area allows them to be treated as a feature for genre classification of historical texts.

7.3. Social category

The "social" terms in this article include official titles, family-relation terms, and such social categories as *min* (people). The most terms (in absolute numbers), unsurprisingly, are observed in historical texts; the Chun Qiu Zuo Zhuan has the most (26 characters)[39]. Unlike numeric and calendrical terms, all texts include social terms in the autosemantics of the pre-h list. However, different types of social terms exist as groups. The historical texts include many specific terms, such as *shi*, *bo*, or other feudal titles. The philosophical texts, such as the Mengzi and the Lun Yu, include mostly general terms and appellations, such as *gong*, *jun*, *min*, *wang*. The Zhou Yi only contains *zi*, *ren*, and *jun*. This is the smallest set in the corpus[40].

7.4. Politico-geographical category

These terms are also prominent mostly in historical texts. Other texts, such as the Lun Yu, the Yi Li, the Zhou Yi, and the Xiao Jing, often do not contain these terms in pre-h lists. The most generic character for those that do contain them is *guo*. Pre-h lists for historical texts contain the most geographical and administrative terms, and the names of the most important kingdoms. In absolute terms, the Zuo Zhuan is the champion (in relative terms, it is the Chun Qiu).

7.5. Nouns

Although characters that do not belong to featured categories and that are lumped together under POS categories are not of special interest for the present work, they still merit a few words. The noun category contains conceptual skeletons of texts; they are thematic words that provide the main information on the topic of the text. Table 13 groups nouns from historical texts into politico-geographical and social categories. However, in philosophical and ritual texts, nouns are the main part of their autosemantics. In the Lun Yu and the Mengzi, the thematic nouns contain terms such as *ren*, *dao*, *xue* and *xin*, whereas in historical narratives, beside numerical and temporal terms that are not used in philosophical texts, thematic nouns include *zui*, *chao*, *luan*, etc.

The largest list in relative terms is that of the Xiao Jing, and the next largest is the Zhuangzi. In absolute terms, the list of the Yi Lin is largest.

---

[39] However, relatively, it is most high in the Xiao Jing (due to short text) and then in the Mengzi with its twelve terms.

[40] The next one, the Xiao Jing, lists *zi*, *ren*, *min*, and, not surprisingly, *fu*.

7.6. Verbs

Verbs create action for texts. All texts, with the exception of the Xiao Jing, which contains only social terms and nouns, contain verbs in their preh lists. These thematic verbs reflect the topic of the text, as well as nouns.

The Yi Li contains the most verbs in relative terms, followed by the Lun Yu. In absolute terms, the Chun Qiu Zuo Zhuan contains the most verbs.

7.7. Adjectives and adverbs

The Zhou Yi, which contains mantic formulas, contain the most terms in relative terms, whereas the Li Ji contains most absolutely. The Chun Qiu and Xiao Jing contain no adjectives or adverbs at all. Shi Jing contains surprisingly few adjectives in its h list.

7.8. Miscellaneous

Other terms included into autosemantics essentially border with synsemantics. The Shu Jing contains most of them, relatively, and the Chun Qiu Zuo Zhuan contains the most in absolute terms. They are mostly pronouns, such as ***shi***, and directionals.

Table 13. Thematic characters by category. Nouns, verbs and adjectives that fit Numeric, Calendar, Social or Political categories, are not duplicated in POS category column.

	Numeric	Calendar	Social	Political	Nouns	Verbs	Adjective/Adverb	Misc.
CQ	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	
ZZ	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服	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	大寡今武小難	是吾未下上中右

CQZZ	二三十一五四六七八九	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曹	晉楚齊國鄭衛宋陳吳周城秦邑趙魯邑邾	禮事文書天德朝罪門馬亂女謀服甲	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葬圍侵莒申食戰	大寡今武小難**正**	是吾未下上中右
GY	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
GL	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
LJ	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	禮事文天德朝門服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	大小**正**內貴外然明反長尊	是未上中
LY	三		子公人君夫孔		禮事仁道路學	曰為有可謂行言知聞見問	好	是吾未
MZ	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
SHI	四百	月日	子公人君王歸民孔	國	天德女心山方思樂南	曰為有可行言來命維	大憂	是**中**予載
SHU	三**一五四**百		公人王民帝殷	**周**邦	事**文天**德方	曰有言**敢**命**用**作刑	大**今小**明	**下上**惟予汝厥

XJ			子人民**父**		事**天**孝			
YL	二**三** 一		**子公人君** 夫爵賓婦		禮門東 北西面 馬南階 賓屍位 觶席辭 阼酒堂 首幣弓 醢	曰為有出 告入立見 **命**取執**主** 獻拜升降 受奠祭射 興洗揖祝 送授答荅 進退酌稽 佐篚	大外俎 長	下**上** 右左 反從
ZL	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	國府邦	禮事馬 服寸掌 徒方物 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	大小正 內長共 胥	下上 中
ZY	二三 五四 六九		子人君		天象位 道咎彖	曰有可行 用貞志孚	大小正 吉利凶 明亨終	未下 上中
ZHZ	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 身治仁	曰為有可 謂死行出 成言知聞 欲見用生 問	大今同 明然	是吾 未下 上中

7.9. Summaries of thematic characters by text

Pre-h lists of autosemantics are good material for thematic analysis of texts in the corpus. This work provides only a very short characteristic of each text according to the associated pre-h list[41].

Chun Qiu

The Chun Qiu has the most characters in numerical, calendrical, and political categories, and in the social section, so almost nothing is left for general groups such as nouns and other. The Chun Qiu contains a limited set of verbs, and just two characters in adjectives-adverbs (only *da* and *zheng*). Numerical and calendar terms are used to denote dates. Social characters include political offices such as *gong*, *hou*, *wang*, *bo*, and ranks such as *shi*, *fu*, and terms such as *ren*, *sun*. There are also names of kingdoms: Qi, Wei, Chu, Chen, Jin, etc. The most used characters in verbs are *sha*, *fa*, *sang*, *meng*, *gui*, etc.

[41] See Appendix 2 for pre-h lists.

Zuo Zhuan

The Zuo Zhuan features the largest pre-h list[42]. The Zuo Zhuan is similar to the Chun Qiu in that it features numeric, calendar, social, and political character groups. It differs from the Chun Qiu, however, in that is has a strong (other) noun group (i.e., its thematic scope is much wider), and a much stronger verb group. It also features many characters in the miscellaneous group. In the nouns, categories close to political area dominate (*li*, *shi*, *wen*, *tian*, *de*, *chao*). The political area is also reflected in verbs, with verbs such as si, sha, ling, zhu, etc. On the other side, there are many verbs, related to the narrative character of the Chun Qiu: *yue*, *gao*, *yan*, *jian*, etc.

Chun Qiu Zuo Zhuan

The Chun Qiu Zuo Zhuan is very similar to the Zuo Zhuan; only few characters appear that are not on the pre-h list of the Zuo Zhuan and belong to the Chun Qiu Zuo Zhuan: three numericals (*qi*, *ba*, *jiu*), a social *cao*, and three verbs *zang*, *qin*, *ju*. The positions of some characters common to both texts increased thanks to the Chun Qiu (e.g., *jia* and *zheng*). The Chun Qiu Zuo Zhuan pre-h list vocabulary is the biggest in our corpus and could be used as the comparison criterion for other vocabularies.

Gongyang Zhuan

The Gongyang Zhuan features thematic content more similar to the Chun Qiu than to the Zuo Zhuan, with just a few noun and other characters. Its vocabulary, however, is a subset of the Chun Qiu Zuo Zhuan (except for a couple of characters).

Guliang Zhuan

The Guliang zhuan is similar to the Gongyang Zhuan.

Li Ji

The Li Ji is a text whose thematic features are similar to the Chun Qiu and to the historical group in that it features many numericals and calendricals on its pre-h list. In addition, it has many socials, most of which, except for three social characters (*mu*, *fu*, *kong*), are part of the Chun Qiu Zuo Zhuan vocabulary. It contains fewer political terms, but a decent list of nouns and verbs, of which almost a quarter appear only in the Li Ji. This fact differentiates the Li Ji from the historical texts.

[42] Except when combined with the Chun Qiu, the Chun Qiu Zuo Zhuan.

Lun Yu

The Lun Yu is a good example of a philosophical text. It has almost no numerical and calendrical characters, has a very restricted set of Social characters (*zi*, *gong*, *ren*, *jun*, *fu*, *kong*), and has practically no political characters. Its verbs are mostly narrative verbs. Its nouns and its social terms provide the theoretical message of the text.

Meng Zi

The Meng Zi is very similar to the Lun Yu; it just has a more extended social lexicon.

Shi Jing

The Shi Jing pre-h list is rather short; its verbs and social lists are very similar to that of the Lun Yu and Meng Zi, although shorter. However, its nouns differ, and its other category is rather big for such a text.

Shu Jing

The Shu Jing is similar to the Shi Jing, except that it has a larger numerical list, similar to that of the Chun Qiu. Its social and noun lists are very short for a text of this size. In a way, the Shu Jing has the most skeletal set of characters.

Xiao Jing

The Xiao Jing's pre-h list is naturally limited due to its size; however, it contains social and noun characters, reflecting its genre.

Yi Li

The Yi Li is a highly narrative text. It does not have a large social list, as may be supposed; however, it has very large noun and verb lists.

Zhou Li

The Zhou Li is similar to the Li Ji and the Yi Li, but it has a larger set of numerical characters, some calendrical characters, and balanced nouns and verbs.

Zhou Yi

The Zhou Yi may be grouped together with philosophical texts; but it has disproportional numerical and adjective lists (according to its genre).

Zhuang Zi

Finally, the Zhuang Zi may be grouped with ritualistic texts.

8. Conclusions

This article concentrates on frequency lists of individual texts in the WSP Ctexts corpus in an effort to understand how to extract relevant

topic information from these texts. Frequency lists depend on text length, and many characters high on the lists are "function" (or "empty") characters, which, although they could be used for stylistic analysis, do not give useful information on text topics and genre. This work applies methods developed by Popescu and Altmann for identifying theme-related parts of frequency lists. This task required constructing a custom list of synsemantic characters for Classical Chinese.

The paper begins by analyzing rank-frequency charts and cumulative spectra. The curves are fit to provide slope gradients for text groupings. However, sorting by curve gradients does not bring in interesting groupings (however, these results could be useful in other studies).

The paper next introduces the h-point and a-point values, following the PA methodology. The a-value apparently depends little on text length; however, it does not classify texts along known genre or stylistic groups. One interesting result is that the a-values for Chinese classics indicate, according to Popescu's methodology, that Classical Chinese belongs to the group of analytical languages. The next group of indicators F(k) and G(k), developed within the PA methodology, better classify the texts by genres. Overall, although these indicators may be useful for some other linguistics functions, they do not provide considerable information, which would allow texts to be grouped thematically according to pre-set genre classification.

The main function of the h-point is to provide a "fuzzy border" between synsemantic and autosemantic words or characters on the frequency list. The border is fuzzy because some autosemantic words are higher than the h-point, and some synsemantic words are below the h-point. The autosemantic characters on the pre-h-point list ("thematic characters") have important meaning for thematic features of the text.

Thematic character lists has been developed for texts in the WSW Ctext corpus. These lists could be used for various thematic analyses and genre classification of texts. However, this work does not offer a full ontology for analysis. As an example of prospective analysis, a categorical analysis of texts is presented to identify the historical group (or genre). The categories that could be features for a historical narrative were chosen mostly by observation, because not much research is available on historical-genre vocabularies. The following categories were chosen: numerical, calendar, social, and politico-geographical. We presume that, in the case of the WSP Ctexts corpus, a strong lexical presence in these categories will classify texts as historical narratives as opposed to philosophical, fiction, and ritualistic texts. All other pre-h list autosemantic characters are placed in part-of-speech categories: nouns, verbs, adjectives, and others. In future work, the author may implement a complete ontology (e.g., by applying the sememe system of HowNet or categories of ChineseWordNet).

This approach finally proved productive for thematic analysis and genre attribution in the corpus. It demonstrates that thematic characters could be used for genre classification of texts. The number of autosemantic characters in numerical and calendrical categories, and the strong presence of social and politico-geographical terms in pre-h lists, classify the Chun Qiu, the Zuo Zhuan (and combined the Chun Qiu Zuo Zhuan), the Gongyang Zhuan, and the Guliang Zhuan as historical narratives. The characters in those categories almost exhaust pre-h lists in most of these texts[43]. The Gongyang Zhuan and the Guliang Zhuan contain almost no other nouns, similar to the lapidary chronicle of the Chun Qiu. The ritualistic texts, such as the Yi Li and the Zhou Li, and philosophical texts, such as the Lun Yu and the Zuangzi, contain comparatively more nouns that do not belong to numerical, calendrical, or social terms. The Yi Li also contains the most verbs. Verbs generally prevail over nouns in this classification (naturally, generic verbs, with socio-political terms excluded). The Zhou Yi pre-h list contains the greatest number of adjectives (not belonging to "historical" categories).

The analysis of the pre-h list autosemantic vocabulary also allows us to identify the Zuo Zhuan as the most fundamental text in the corpus, from the vocabulary point of view: its vocabulary could be used as the basis for analyzing the vocabulary of other texts. It is not surprising that the vocabulary of historical narratives is almost entirely contained within the list of the Zuo Zhuan, although philosophical texts (i.e., the Lun Yu, the Mengzi, the Yi Jing, and the Zhuangzi) and the Shi Jing and the Shu Jing also have the basic pre-h list vocabulary contained in the Zuo Zhuan. Still, the thematic characters indicate content differences between texts. The pre-h lists of characters could be used for thematic and genre analysis of classical Chinese texts.

Another direction of future work could be analysis of the phenomenon what Meyer[44] called "macrolevel consistency" (Meyer, "*Philosophy on Bamboo*", 185) or "macrocoherence" (Meyer "*Philosophy on Bamboo*", 206) of classic Chinese texts. Many classic texts are not consistently "narrative', being composed of multiple "units of thought". Being a cumulative (non-monolithic) work (Brooks and Brooks, *The Original Analects*, 4–5), what does make an entire text an independent unity? Brooks and Brooks develop the accretion theory of evolutionary development of such texts (Brooks and Brooks, *The Original Analects*, 201ff),

[43] The numerical and calendrical terms in the corpus have a strong presence of social and politico-geographical terms.

[44] Meyer cites Kern as the origin of this concept (Kern, "Quotation and the Confucian Canon", 35–36), though Meyer elaborates this concept considerably.

using formal stylistic analysis as well as study of distribution patterns to discover timeline of accretion in the Lun Yu.

Studying various versions of the Zi yi (currently a chapter in the Li Ji), Meyer tries to find balance between single units of the manuscript versions, and text as a whole. He notes that even though "the different units are not blended into a unified whole structurally (formally, they remain fully isolated, and hence specific answers to a given concern), the uniform pattern of each of these units nevertheless creates a sense of consistency and allows the recipient to identify with the work as a whole" (Meyer, "*Philosophy on Bamboo*", 185). The macrolevel consistency allows the Zi Yi, consisting of "units of thought" to be perceived as one work: "the macrolevel consistency of the "Zi yi" suggests that it is unlikely that this anthology is an accidental collection of otherwise-unrelated materials" (Meyer, 206).

This author presumes that the concepts of accretion and macrolevel consistency could be extended to larger collections of individual texts, like the Li Ji, and comparative analysis of pre-h list vocabularies of these collections and their modules could help to understand the nature of this perceived coherence.

REFERENCES

ABBREVIATIONS

Chun Qiu	CQ
Chun Qiu Zuo Zhuan	CQZZ
Gongyang Zhuan	GY
Guliang Zhuan	GL
Li Ji	LJ
Lun Yu	LY
Mengzi	MZ
Shi Jing	SHI
Shu Jing	SHU
Xiao Jing	XJ
Yi Li	YI
Zhou Yi	ZY
Zhuangzi	ZHZ
Zhou Li	ZL
Zuo Zhuan	ZZ

GITHUB RESOURCES

https://github.com/wsw-ctexts/vocabulary_richness/character_frequency_curve_fitting.xls

https://github.com/wsw-ctexts/vocabulary_richness/character_frequency_reference_data.xls

Appendix 1 Synsemantics ("empty characters" and stop-words)

Class 1 Synsemantics (empty characters)

不与且丕並个乃久么之乍乎也了于云互些亟亦介仍从他代以们任伊伏休似但何佯使例依便促俄俱個倏們倘借假偏偕做偶偷偽備僅僉儂儻儿兀先克兒全兩兮其具兼再几凡凭切则初别別到則前剛劇加動務勝勿匆匪半卒博即却卻又及另叨只叵吁向否吧呀呢呼咄咨咱咸哇哉哟員哦哩哪哼唉唯唷啊啥啦喂善喏喔哟喳嗎嗚嗟嗡嗨嘎嘗嘛嘟嘻嘿噓噗噫噯噹嚕在地坐垂堪塊壹夙多奄奇奉奚如妄妳始姑孰它容密實寧審將專尋對尚就屢屬差已希平幾底庶庸廈弗弟強彌彼往很後徐徑得從復循微必忌忝忽怎怡急恆恐恭悄悉惠惡愈意慎慮憑應戎我或所才承把抑投披拿据擅據故敬數斯於旃旅旋无既旦早時普暗暨暫暴曩更曷曼曾替最會朕杳枉某極横權欠歟止此歷殆殊殫毋每比汔沒沓沿況洎洵浸深渠溘滋漫漸潛濫焉無煞爭爾特猗猝猥猶獨率甚甫由畢略疇疾痛的皆盍盛盡直相着矣矧稀稍窮竊竟端競第等籲粗粵約純素索累給綦緣縱總繆繇繼给罔罕罷羌習翻者而耶聊胡能脫臨自至致與舉良苟若茲莫萬著蒙蓋薦虽蚤蛾被裁裹要親覲訖設許訾詎該詳誰請諸諾謇謬謹让该诸谁豈豫賴起趁越趣趨距跟較輒輩辱迄还这迪迭逆這速遂遄遍遞適邊那邪都鄉酷錯闔阿随隨雖雜非靠靡頃順預頗頻顓類顧驟鮮黨鼎更不若了裡

Class 2 synsemantics (stop-words)

一上下中为亡交人今令作你來信元入公共原去反取可各同后和因固大太夫她好孔安定小少尔居巨常并当徒忘思您情手打政斷方日是有未本来果業正殺永泰洪浪為烏然甘生用當白益看真私空立終絕經蔑行見誠財足身通連過還重長間陰陽雅首行

Appendix 2 Thematic characters in the WSP Ctexts corpus

Chun Qiu

Category	**Character**
Numerical	十二三七四八六五九
Calendar	月年夏春冬秋
Social titles/names	公人侯子師孫王夫伯叔曹
Political geography	齊晉宋鄭衛楚陳邾
Verbs	有伐來葬盟奔莒歸蔡殺杞帥侵
Adjectives	大正
Miscellaneous	

Zuo Zhuan

Category	Character
Numerical	二三十一五四六
Calendar	月年日秋春冬夏
Social titles/names	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
Political geography	晉楚齊國鄭衛宋陳吳周城秦趙魯邑
Nouns	禮事文書天德朝罪門馬亂女謀服
Verbs	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
Adjectives/adverbs	大寡今武小難
Miscellaneous	是吾未下上中右

Chun Qiu Zuo Zhuan

Category	Character
Numerical	二三十一五四六七八九
Calendar	月年日秋春冬夏
Social titles/names	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曹
Political geography	晉楚齊國鄭衛宋陳吳周城秦趙魯邑邾
Nouns	禮事文書天德朝罪門馬亂女謀服甲
Verbs	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葬圍侵莒申食戰
Adjectives/adverbs	大寡今武小難**正**
Miscellaneous	是吾未下上中右

Gongyang Zhuan

Category	Character
Numerical	二三十一五四七
Calendar	**月**年日秋春冬夏
Social titles/names	**子公人**君**侯**王師夫伯氏叔孫歸季父帥曹婁
Political geography	晉楚**齊**國鄭衛宋陳邾
Nouns	書天
Verbs	曰**為**有可伐盟殺出**言**入來立取執葬莒譏弒稱
Adjectives/adverbs	大**正**然
Miscellaneous	未是吾

Guliang Zhuan

Category	Character
Numerical	二三十一五四七
Calendar	月年日秋春冬夏
Social titles/names	子公人君侯王師夫伯叔孫歸父帥蔡曹
Political geography	晉楚齊國鄭衛宋陳邾
Nouns	事天
Verbs	曰為有可伐盟殺出言入來辭葬莒弒志
Adjectives/adverbs	大正內
Miscellaneous	是未

Li Ji

Category	Character
Numerical	二三十一五四
Calendar	月年日夏
Social titles/names	子公人君侯王夫臣民士母婦孔
Political geography	楚國
Nouns	禮事文天德朝門服世賓喪東樂方道廟義衣宗位
Verbs	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
Adjectives/adverbs	大小正內貴外然明反長尊
Miscellaneous	是未上中

Lun Yu

Category	Character
Numerical	三
Calendar	
Social titles/names	子公人君夫孔
Political geography	
Nouns	禮事仁道路學
Verbs	曰為有可謂行言知聞見問
Adjectives/adverbs	好
Miscellaneous	是吾未

Meng Zi

Category	Character
Numerical	一百

Calendar	
Social titles/names	子公人君王夫民父士孟孔舜
Political geography	齊國
Nouns	事天仁道樂心義
Verbs	曰為有可謂行言知聞欲見問食
Adjectives/adverbs	大今然
Miscellaneous	是吾未下

Shi Jing

Category	**Character**
Numerical	四百
Calendar	月日
Social titles/names	子公人君王歸民孔
Political geography	國
Nouns	天德女心山方思樂南
Verbs	曰為有可行言來命維
Adjectives/adverbs	大憂
Miscellaneous	是**中**予載

Shu Jing

Category	**Character**
Numerical	三**一五四**百
Calendar	
Social titles/names	公人王民帝殷
Political geography	**周**邦
Nouns	事**文天**德方
Verbs	曰有言**敢**命**用**作刑
Adjectives/adverbs	大**今小**明
Miscellaneous	**下上**惟予汝厥

Xiao Jing

Category	**Character**
Numerical	
Calendar	
Social titles/names	子人民**父**
Political geography	
Nouns	事**天**孝

Verbs	
Adjectives/adverbs	
Miscellaneous	

Yi Li

Category	**Character**
Numerical	二三一
Calendar	
Social titles/names	子公人君夫爵賓婦
Political	
Nouns	禮門東北西面馬南階賓屍位觶席辭阼酒堂首幣弓醢
Verbs	曰為有出告入立見命取執主獻食拜升降受奠祭射興洗揖祝送授答荅進退酌稽佐篚
Adjectives/adverbs	大外俎長
Miscellaneous	下上右左反従

Zhou Li

Category	**Character**
Numerical	二三一十五八九百
Calendar	日歲
Social titles/names	子人侯王夫氏民司士師史賓帥客官
Political geography	國府邦
Nouns	禮事馬服寸掌徒方物車喪法器田刑
Verbs	曰為有謂行入命用政令食祭祀治禁辨鼓受教
Adjectives/adverbs	大小正內長共胥
Miscellaneous	下上中

Zhou Yi

Category	**Character**
Numerical	二三五四六九
Calendar	
Social titles/names	子人君
Political geography	
Nouns	天象 zy 位道咎彖
Verbs	曰有可行用貞志孚
Adjectives/adverbs	大小正吉利凶明亨終
Miscellaneous	未下上中

Zhuangzi

Category	Character
Numerical	三一
Calendar	日
Social titles/names	子人君王夫民
Political geography	國
Nouns	事天道物神方德心形聖 zhz 足名 zhz 世樂義身治仁
Verbs	曰為有可謂死行出成言知聞欲見用生問
Adjectives/adverbs	大今同明然
Miscellaneous	是吾未下上中

Literature

Baayen R. Harald. Word frequency distributions. Dordrecht: Text, speech, and language technology No. 18, Kluwer Academic, 2001.

Bai Yulin, Chi Duo 白玉林, 迟铎 (eds.). Gu Hanyu xuci cidian 古汉语虚词词典 [The Dictionary of Classical Chinese Function Words]. Beijing: Zhonghua shuju, 2004.

Bei Guiqin, Zhang Xuetao 贝贵琴, 张学涛. Hanzi pindu tongji 汉字频度统计 [Chinese Character Frequency Statistics]. Beijing: Electronic Industry Press, 1988.

Brooks E. Bruce, Brooks A. Taeko. The Original Analects: Sayings of Confucius and his Successors. New York: Columbia University Press, 1998.

Chen Xiaohe, Feng Minxuan, Xu Runhua 陈小荷, 冯敏萱, 徐润华. Xian Qin wenxian xinxi chuli 先秦文献信息处理 [Information management of pre-Qin literature]. Beijing: Shijie tushu chuban gongsi, 2013.

Chi Duo 迟铎. Gǔdài hànyǔ xūcí cídiǎn 古代汉语虚词词典 [Dictionary of function words of Classical Chinese] n.p.: Shāngwù yìn shūguǎn guójì yǒuxiàn gōngsī, 2010.

Chou Wen-hui. "On the Lexical Differences between South and North as Revealed by Diachronic Substitutions of Commonly Used Body-Part Terms Chinese Lexical Semantics". In: Chinese Lexical Semantics – 14th Workshop (CLSW'14), Zhengzhou, China, Lecture Notes in Computer Science, 2013. Vol. 8229: 196–207.

Da Jun. "A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction". In: *Zhang Pu, Xie Tianwei, Xu Juan* (eds.). 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. Beijing: Tsinghua University Press, 2004: 501–511.

Dobson William Arthur Charles Harvey. A Dictionary of the Chinese Particles. Toronto: University of Toronto Press, 1974.

Evert Stefan. The Statistics of Word Co-occurrences: Word Pairs and Collocations. PhD Thesis, Institut für maschinelle Sprachverarbeitung, Stuttgart: Universität Stuttgart, 2005.

Fang Chengyu Alex, Cao Jing. Text Genres and Registers: The Computation of Linguistic Features. Berlin, Heidelberg: Springer, 2015.

Fries Charles C. The structure of English. New York: Harcourt Brace, 1952.

Golcher Felix. "A New Text Statistical Measure and its Application to Stylometry". In: *Davies Matthew, Rayson Paul, Hunston Susan, Danielsson Pernilla* (eds.). Proc. of the Corpus Linguistics conference (CL'07). Birmingham, U.K., 2007. No. 71: 1–15.

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.

Hai Liuwen 海柳文. Shisanjing zipin yanjiu 十三经字频研究 [Character frequency study in the Thirteen Classics]. Beijing: Gaodeng jiaoyu chubanshe, 2011.

Hao Lili, Hao Lizhu. "Automatic Identification of StopWords in Chinese Text Classification". In: 2008 International Conference on Computer Science and Software Engineering, Wuhan, 2008. Vol. 1: 718–722.

Hirsch Jorge E. "An Index to Quantify an Individual's Scientific Research Output". Proceedings of the National Academy of Sciences of the United States of America 102.46 (2005): 16569–16572.

Kern Martin. "Quotation and the Confucian Canon in Early Chinese Manuscripts: The Case of 'Zi Yi' (Black Robes)". Asiatische Studien / Études Asiatiques. 59.1 (2005): 293–332.

Kim Yunhyong, Ross Seamus. "Variations of Word Frequencies in Genre Classification Tasks". In: Proc. of the DELOS Conf. on Digital Libraries, Tirrenia, Italy, 2007.

Klammer Thomas P., Schulz Muriel R., Volpe Angela D. Instructor's Manual to accompany Analyzing English Grammar, Sixth Edition, n.p.: Pearson Education, 2012.

Köhler Reinhard, Altmann Gabriel, Piotrowski Rajmund G. (eds.). Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook. Walter de Gruyter, 2005.

Kytö Merja, Lüdeling Anke (eds.). Corpus linguistics: an international handbook. Berlin, New York: Walter de Gruyter: Handbooks of linguistics and communication science, 29.1–29.2 Handbücher zur Sprach- und Kommunikationswissenschaft Bd. 29.1–29.2, 2008–2009.

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

Li Bo 李波. Shiji zipin yanjiu 史记字频研究 [Study of Character Frequencies in the Shi Ji]. Beijing: Shangwu yinshuguan, 2006.

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 High Frequency Character Distribution and Word Formation in *Shisanjing*]". MA Diss., University of Heilongjiang, 2009.

Mačutek Ján, Popescu Ioan-Iovitz, Altmann Gabriel. "Confidence intervals and tests for the h-point and related text characteristics". Glottometrics 15 (2007): 45–52.

Mahalakshmi S., Sivasankar E. "Cross Domain Sentiment Analysis Using Different Machine Learning Techniques". In: Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO – 2015): Volume 415 of the series Advances in Intelligent Systems and Computing, Springer International Publishing, Switzerland, 77–87, 2015.

Meyer Dirk. Philosophy on Bamboo: Text and the Production of Meaning in Early China. Leiden: Brill, 2012.

Nakagawa Hiroshi, Kojima Hiroyuki, Maeda Akira. "Chinese Term Extraction from Web Pages Based on Compound word Productivity", 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004), Third SIGHAN Workshop on Chinese Language Processing, 79–85, Barcelona, Spain, July, 2004.

Pan Xiaxing, Hui Qiu, Liu Haitao. "Golden section in Chinese contemporary poetry". In: Glottometrics 32 (2015): 55–62.

Pines Yuri. "Lexical Changes in Zhanguo Texts". In: Journal of the American Oriental Society 122, No. 4 (2002): 691–705.

Popescu Ioan-Iovitz. Word Frequency Studies, Quantitative linguistics 64, Berlin, New York: Walter de Gruyter, 2009.

Popescu Ioan-Iovitz, Altmann Gabriel, Grzybek P., Jayaram B.D., Köhler R., Krupa V., Mačutek J., Pustet R., Uhlířová L., Vidya M.N. Word frequency studies. Quantitative linguistics 64, Berlin, New York: Mouton de Gruyter, 2009.

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

Qin Qin 覃勤. "Xianqin guji zipin 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.

Qiu Bing, Zhu Qingzhi. "Corpus Building for the Outcome-Based Education of the Ancient Chinese Courses Chinese Lexical Semantics". In: Chinese Lexical Semantics, 8922 (2014): 358–368.

Rajaraman Anand, Ullman Jeffrey. Mining of Massive Datasets. Stanford: Cambridge University Press, 2011.

Stamatatos Efstathios, Fakotakis Nikos, Kokkinakis George. "Text Genre Detection Using Common Word Frequencies". In: 18th International Conference on Computational Linguistics, Proceedings of the Conference (COLING 2000), Universität des Saarlandes, Saarbrücken, Germany, 2000: 808–814.

Tweedie Fiona J., Baayen R. Harald. "How Variable May a Constant be? Measures of Lexical Richness". In: Perspective Computers and the Humanities 32:5 (1998): 323–352.

Vaismoradi Mojtaba, Turunen Hannele, Bondas Terese. "Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study". Nursing and Health Sciences, 15:3 (2013): 398–405.

Wang Haifen 王海锋. Gu Hanyu xuci cidian 古代漢語虛詞詞典 [Dictionary of Classical Chinese Function Words]. Beijing: Beijing Daxue chubanshe, 1996.

Wang Zhengbai 王政白 (ed.). *Gu Hanyu xuci cidian* 古代汉语虚词 词典 [Dictionary of Classical Chinese Function Words]. Hefei, 1986.

Yang Bojun 楊伯峻. *Gu Hanyu xuci* 古漢語虛詞 [Function Words of Classical Chinese]. Beijing: Zhonghua, Third edition, 2000.

Yu Fang, Liu Haitao. "Comparison of vocabulary richness in two translated Hongloumeng". Glottometrics, 31 (2015): 54–75.

Zinin Sergey. "Pre-Qin Digital Classics: Study of Text Length Variations". In: Учёные записки отдела Китая, выпуск 15. 44-я научная конференция «Общество и государство в Китае». Т. 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.

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, 2016: 197–253.

Sergey Zinin*

Analysis of character-frequency lists of Chinese classics and its application to content analysis and genre attribution

ABSTRACT: This paper analyzes the frequency distributions and spectra of the characters of texts in the Warring States Project (WSP) Ctexts corpus to extract information on genre and thematic aspects of the texts. It uses the methodology developed by Popescu and Altmann (PA) to identify important parts of word-frequency lists. The PA approach introduces the concept of "h-point" to separate synsemantic and autosemantic parts of frequency lists. This analysis is enabled by developing a list of synsemantic characters for classical Chinese. Although not all PA methods provide useful information, the analysis of autosemantic characters in the pre-h-point list (pre-h list) may provide results on genre classification and text topics. This article analyzes a few particular cases in which autosemantic characters in pre-h lists could be useful for identifying historical narratives in the corpus and thematic analysis. The results are important for linguistic analysis of Chinese texts and for a better understanding of the texts.

KEYWORDS: quantitative linguistics; Chinese corpora; Thirteen Classics; Zhuangzi; frequency lists; frequency spectrum; vocabulary richness; vocabulary exploitation; synsemantic characters; autosemantic characters; thematic characters; h-point; k-point; thematic concentration; macrolevel consistency; accretion theory.

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