*Sergey V. Zinin**

Vocabulary richness of early Chinese texts: macroanalysis of the Thirteen classics and the Zhuangzi

ABSTRACT: This study analyzes statistical data regarding the vocabulary richness of the Warring States Project CTexts collection of Chinese classics[97]. Vocabulary richness has been primarily used in quantitative linguistics for authorship identification and style analysis, and it has been increasingly applied for various aspects such as language acquisition in other linguistic fields. This study lays the foundation for a quantitative linguistic analysis of the vocabulary of early Chinese texts. It also conducts a macroanalysis of the data, including calculating several vocabulary richness indices and building charts of vocabulary growth. This study finds significant differences in the vocabulary growth of corpus texts. In addition, it reveals that the Shi Jing and Yi Li are two extreme ends of the vocabulary growth spectrum and identifies some historical texts in the middle of the spectrum as a distinct group. Furthermore, the study takes a closer look at specific forms of vocabulary growth such as hapax legomena, dis legomena, and the most frequent characters.

KEYWORDS: Chinese canons, The Thirteen Classics, computational linguistics, quantitative linguistics, vocabulary richness, lexical diversity, type-token ratio, digital corpora, stylometry.

CONTENT

---

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

Zinin, Sergey. Vocabulary richness of early Chinese texts: macroanalysis of the Thirteen classics and the Zhuangzi. In: Учёные записки отдела Китая, выпуск 20, 46 научная конференция Общество и государство в Китае, том XLVI, ч. 1, М., Институт Востоковедения РАН (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

1. Introduction

Several terms in quantitative linguistics "refer to the range of different words used in a text, with a greater range indicating a higher diversity" (McCarthy and Mild, "vocd-D and HD-D," 381). These terms are vocabulary richness, lexical richness[1], lexical diversity, and vocabulary diversity.

[1] The term "lexical diversity," defined as "a complex property that summarizes the range of vocabulary and the avoidance of repetition in the sample" (Malvern and Richards, *Measures of Lexical Richness*, 1), is used intermittently with lexical richness in a text. There are also other terms that are close in meaning, e.g., Pilar

However, since there may not be much difference between the meanings of these terms[2], the term "vocabulary richness" could be used as a suitable representative.

There are various methods to calculate the measures of vocabulary richness (or "indices")[3]. Their values may allow a comparison of texts in various areas including "first and second language acquisition, linguistic input, interaction, demographic influences on language performance, language impairment, delay, aphasia, schizophrenia, stylistics, and forensic linguistics" (Oakes, "Corpus Linguistics and Stylometry," 1073–74). For classical Chinese, vocabulary richness can be useful for comparing texts from the perspective of their vocabulary diversity, lexical sophistication, and so on.

1.1. Importance of Warring States Project (WSP) CTexts vocabulary

The texts in Ctexts may be considered the most important source for studying character vocabulary in classical Chinese. The canon system that had formed from the Han to the Song dynasties had strongly affected all aspects of Chinese literary discourse, including the general character vocabulary of classical Chinese. The Thirteen Classics and its set of characters (i.e., its character vocabulary) have been memorized by generations of Chinese scholars and officials and it definitely has had an effect on most texts produced by these scholars[4]. While the corpus of the WSP Ctexts is not large enough comparing to the entire pre-Qin literature[5], its character vocabulary could be very close to the general character vocabulary of the

---

Duran cites such terms as "flexibility," "verbal creativity," and "lexical range and balance" (Duran et al., "Developmental trends," 221–222). Other authors add terms such as "lexical originality," "lexical sophistication," "lexical density," and "lexical variation" (Laufer and Nation, "A Vocabulary-size Test," 309–320).

[2] They have been used intermittently to describe the same value (e.g., in Malvern and Richards, "Measures of Lexical Richness," and other works).

[3] Tweedie and Baayen in "How Variable May a Constant Be?" (323) use the term "constants," referring to the claim that these variables were thought to be constant by the researchers that created them.

[4] It is reflected in the official status of the late Qin reference work on characters in the Thirteen Classics by Li Hongzao (Li, *Hanyuan Shisanjing ji zi*). According to some calculations, Li Hongzao's work implies that the Thirteen Classics vocabulary contained 6544 unique characters (Qiu, *Written Chinese*, 49–50). Ctexts contains 6055 characters; the Thirteen Classics in Ctext corpus (excluding the Er Ya) contain 5628 characters. The author is grateful to Rodo Pfister who pointed out that fact.

[5] At approximately a half-million characters, it represents about 17% of all pre-Han literature (about three million) and about 6% of the Han and pre-Han literature (eight million), as stated by McLeod (McLeod, "Sinological Indexes," 50).

era. According to Qiu Xigui, "the number of characters in general use during that period would probably fall short of the total number used in the Thirteen Classics" (Qiu, *Written Chinese*, 50).

The present study of The Thirteen Classics and the Zhuangzi[6] will try to lay the foundation for such analysis[7]. Moreover, the WSP Ctexts corpus contains texts of various length (from 2,000 characters to over 200,000 characters), which allows this study to test the validity of various methods.

1.2. Character as type and token (vocabulary unit)

In most languages, the basic unit of texts and vocabularies is the word (word stem). This study utilizes single characters as vocabulary and text units[8]. Linguists generally experience difficulties defining the word for word segmentation purposes. However, in the classical Chinese language, which contains single- and multi-character words, there are certain additional problems with word segmentation and using words as tokens and types, respectively. In the present study, in absence of well-segmented texts, tokens are defined as single characters in a text and types are unique characters in a text[9]. Therefore, the term "vocabulary of text" in this study means the list of unique characters (variants are treated as separate characters) or "character vocabulary"[10], as opposed to "word vocabulary" (similar to the modern Chinese terms *zidian* and *cidian*)[11].

---

[6] The Zhuangzi has been added to offset partly the predominantly "Confucian" character of the WSP corpus. The Er Ya has been omitted since it is not a sample of narrative prose or poetry.

[7] Jun Da describes the general situation with a list of frequencies of characters and concludes that there is not much structured information available (Da, "A Corpus-based Study," 1).

[8] This is examined in more detail in the previous article of this series, i.e., Zinin, "Pre-Qin Digital Classics."

[9] As in the previous study, the texts were cleaned of punctuation and other non-character symbols, and the titles of chapters were removed (see explanation in Zinin, ibid). Character variants, if they include different Unicode representations, are treated as different characters. To be more precise, the Unicode codes of characters serve as types and tokens in this study. It would be a much better situation if the digital versions of The Thirteen Classics with standardized resolved variants were available.

[10] Naturally, it does not mean that the author accepts or prefers the idea of classical Chinese being a monosyllabic language. The character-as-token approach is one possible method and the most feasible approach to study classical Chinese texts.

[11] This approach, using character vocabulary instead of word vocabulary, as showed by Peng et al. could be applied even to modern texts (Peng et al., "Language Independent Authorship", 272).

1.3. Functional and content types of characters

This study will not be distinguishing functional ("empty") and content characters, as it is often done in stylistic analysis in quantitative linguistics. However, in the fifth section, there will be an attempt to conduct separate analyses on hapax legomena, dis legomena, and the most frequent characters.

1.4. WSP corpus sample size and "character stream" abstraction

The vocabulary data for this study has been retrieved from the WSP corpus. The WSP corpus is an online open corpus, built on open source classic Chinese texts, which are considered by the present author to be a sufficient source for a quantitative study of vocabulary richness[12]. Conducting research on an open source corpus ensures its replicability and reproducibility, since any researcher can replicate vocabulary data (first, the numbers of types and tokens) and attempt to reproduce results by applying the same methods[13]. Along with the corpus itself, this study offers all related data (too large to be placed in the Appendix) as an accompanying MS Excel spreadsheet reference, available on Github[14].

The WSP corpus is considered small in relation to some modern Chinese corpora. However, it can be viewed as being large enough for vocabulary richness analysis. Vocabulary richness analysis (especially in practical areas, e.g., in language acquisition and medical studies) is often conducted on short samples of texts (tens or hundreds of words). Popescu suggests the maximum length of vocabulary study sample as 10,000 words (Popescu, "Word Frequency Studies," 3). Many texts in the Ctexts corpus are much larger than this figure.

The reason Popescu suggests this maximum length is text "homogeneity." Not only are the WSP Ctexts long, but also they are not homogeneous narratives created by the same author or even in the same period. In fact, most of the texts in The Thirteen Classics took their current form considerably later and then their subtexts were written. In other words, they are heterogeneous. Moreover, each text in the corpus can be considered a mini corpus for a vocabulary study in itself, especially since it is often a compilation of subtexts of which each one is an independent text in its own right. Thus,

---

[12] As any digital corpus of classical Chinese, the WSP corpus includes some philological problems, the nature of which was discussed by the present author in Zinin, ibid. The corpus can be found at the DOI: http://www.umass.edu/ctexts/index.php (login and password are provided in the pop-up window).

[13] There is a problem with the free availability of reliable classical Chinese corpora for research. See the previous article by the present author in Zinin, ibid.

[14] See the file "Voc_ref.xlsx" at DOI: https://github.com/wsw-ctexts/vocabulary_richness.

the vocabularies of these texts should be investigated separately[15] and such "text patchwork" should be considered as normal for the Chinese tradition.

Actually, studies regarding the assemblages of early manuscripts (e.g., Meyer, "Philosophy on Bamboo") demonstrate that texts, considered be a single unit today, were often broken down into smaller meaningful units and mixed with other texts.

The authorial unity of style, to some degree, can be present in only a few of them[16]. Treating this mix of smaller texts as one large text allows interpreting this large text as a stream of characters[17], which can be sampled at any length. Further analysis will concentrate on the specifics of individual texts and how they relate to the larger body of the text.

---

[15] E.g. the Zhou Li is a compilation of pre-Han texts, probably, assembled by one person (William Boltz in Loewe (Ed.) "*Early Chinese Texts*", 27–29); in the Zhuangzi H.D. Roth indicates presence of five large groupings of heterogeneous collections of chapters, and supposes that it is a collective compilation in early Han (H.D. Roth in Loewe (Ed.) "*Early Chinese Texts*", 56–57); the Chun Qiu, the Gongyang Zhuan, the Guliang Zhuan, the Zuo Zhuan traditionally all were ascribed to one person (Anne Cheng in Loewe (Ed.) "*Early Chinese Texts*", 67–71), but the Gongyang and the Guliang are probably coming from school tradition; while the Zuo Zhuan could have one author-compiler; the Zhou Yi (Edward Shaughnessy in Loewe (Ed.) "*Early Chinese Texts*", 219) always ascribed to one person, the Yi Li (William Boltz in Loewe (Ed.) "*Early Chinese Texts*", 234–237) "detailed and specific descriptions of the ritual ceremonies of a shi" (Loewe (Ed.) "*Early Chinese Texts*", 234), is probably a part of a larger corpus of ceremonial writings (Loewe (Ed.) "*Early Chinese Texts*", 237); the Li Ji (Jeffrey K. Riegel, Loewe (Ed.) "*Early Chinese Texts*", 293–295) — "a ritualist's anthology of ancient usages, prescription, definitions and anecdotes" (Loewe (Ed.) "*Early Chinese Texts*", 293), with "no apparent overall structure" unlike the Zhou Li and the Yi Li, not of same time or origin, its 49 pian (11 groupings) are "extremely diverse and miscellaneous in their style and contents as well as in the origins of the materials of which they are constituted" (Loewe (Ed.) "*Early Chinese Texts*", 295); the Lun Yu (Anne Cheng in Loewe (Ed.) "*Early Chinese Texts*", 314), now considered to be "a composite work of various layers, contributed by different hands"; the Shu Jing is a compilation of texts of "heterogenous nature" (Edward Shaughnessy in Loewe (Ed.) "*Early Chinese Texts*", 376); the Shi Jing's heterogenous nature was not contested by the tradition itself, etc.

[16] It is obviously the Xiao Jing, but also the Zuo Zhuan, the Guliang Zhuan, and the Gongyang Zhuan. In addition, the Chun Qiu and the Zhou Yi contain considerable amounts of formulaic expressions, which create some unity of style. The Lun Yu and the Mengzi, while coming from various sources, have probably been heavily edited in order to appear to have authorial unity. Some researchers, e.g., Dirk Meyer (Meyer "Philosophy on Bamboo") essentially deny the idea of single "authorship" for it.

[17] It should be stressed again that this is a stream of characters, not words.

Meanwhile, The Thirteen Classics were viewed as distinctive stylistic bodies by the Chinese tradition, which often ascribed them to one person as either an author or editor. Definitely, for a reader, the perceived style of the Shi Jing is different from that of the Lun Yu, which is different from that of the Zhou Yi or the Guliang Zhuan. These stylistic differences are often dependent on subject-specific characters or formulaic expressions. Formulaic expressions and repetitive characters strongly affect vocabulary content and growth behavior. The analysis of the vocabulary of these entities is necessary to delineate the area of discussion.

On the larger scale, these text bodies can be considered members of wider genre groups. For example, the Shi Jing represents early Chinese poetry, while the Chun Qiu and the Zuo Zhuan (and two accompanying *zhuan*) belong to historical prose. In addition, the Lun Yu and the Mengzi belong to philosophical prose, while the Li Ji, the Yi Li, and the Zhoy Yi belong to ritualistic prose. One of objectives of this study is to understand if some vocabulary richness measures can be useful for the genre attribution of texts.

In a way, these texts can be compared to the Bible, which consists of texts of different genres. The Bible corpus can also be considered as being more heterogeneous than any of the WSP Ctexts samples. However, the analysis of the Bible vocabulary as a whole still makes sense. The present study is a macroanalysis; i.e., a large-scale investigation of the vocabulary richness of large and heterogeneous texts intended to establish a quantitative framework for further text analysis (it may be more useful to use the term "text richness" instead of "vocabulary richness"[18]).

Therefore, it often treats texts as a stream of characters, which can be sampled at any moment, ignoring subtext borders. The macroanalysis should be followed by a microanalysis of the vocabularies of individual texts as well as sections or chapters of these texts (e.g., the Shi Jing's songs, texts of the Shu Jing) However, this work should be conducted in the future based on the results of this study.

1.5. Previous work

As to the present author's knowledge, there have been few vocabulary richness studies of classic Chinese literature. More specifically, the majority of the studies regarding the vocabulary of classic texts have consisted of character frequencies studies[19] with no systematic analysis of the vocabu-

[18] As Gejza Wimmer writes about the main index used as vocabulary richness measure in this study, TTR, "the TTR as a measure of vocabulary richness is a misnomer. As a measure of the richness of the text it can perhaps function if some problems could be solved" (Wimmer, "Type-token Relation," 362).

[19] See the review of this literature in Zinin, ibid.

lary richness for Chinese classics[20]. Therefore, the main objective of the present study is to lay the statistical foundation for further analysis of the vocabulary richness of classical Chinese texts.

The remainder of this study is as follows. In the second and third sections, vocabulary indices (or "constants" according to Tweedie and Baayen in "How Variable May a Constant Be?") will be introduced. The final or partial values of these constants for the corpus will also be presented, along with diagrams of the hierarchical clustering of texts. In the fourth section, developmental profiles (mostly for the type-token ratio (TTR) index) for entire samples and normalized lengths will be displayed. In the fifth section, some introduction into the developmental analysis of hapax legomena (V1) and dis legomena (V2) as well as the most frequent words (V50+) will be conducted. Finally, a discussion of the results and conclusions will be presented.

1.6. Acknowledgments

E. Bruce Brooks, who has supported the WSP Ctexts project from its beginning, has read the initial draft of the manuscript, made many important suggestions, and the present author has enjoyed an extremely fruitful discussion with him. Brooks as well as Rodo Pfister (to whom the present author is also grateful for reading the early draft) raised a very important issue regarding the validity of "character counting" in the study of early Chinese texts. The ensuing discussion with Brooks and Pfister made this author review several important views on the statistical approach to the texts in the WSP Ctexts corpus. Their comments also helped improve many academic aspects of this study's initial draft[21].

2. Measuring Vocabulary Richness

The most basic approach to measure vocabulary richness is to use final value indices in which a formula is applied to the entire sample. The simplest form of such indices, the TTR, is known to be dependent on text length, which makes an impractical comparison of static value indices for texts of different length. However, there are other indices that claim independency of this parameter (Tweedie and Baayen call them "constants"[22]). Thus, this study will first test this final value approach and discuss the results[23].

---

[20] Nevertheless, it is worth mentioning the Le Guan Ha article that examines the Zipf's rank distribution on large modern Chinese corpora (and compares the curves with English) (Ha et al., "Extension of Zipf's Law").

[21] All of the remaining factual, typographical and grammatical errors are the sole responsibility of the present author.

[22] Tweedie and Baayen, "How Variable May a Constant Be?" 343.

[23] There are many methods.

2.1. TTR as an Index of Vocabulary Richness

The most basic quantitative index of vocabulary richness is the TTR, which is the ratio of the number of types to the number of tokens in a given text sample. It is a well-established fact that the final values of the TTR (values calculated for the entire text sample) are not permanent vocabulary richness characteristics. Moreover, as Vulanovic and Koehler note, "statistical distribution of this index is unknown and, therefore, tests of significance of differences in the TTR between authors or texts cannot be conducted" (Vulanovic and Koehler, "Syntactic Units and Structures," 284). However, TTR values can be helpful to compare similarities in origin and sample size texts. Hence, this index is still used in authorship forensic and style studies.

The main problem with the TTR is its dependency on sampling size. Table 2.1 features the WSP texts, ordered by their TTR final values. It is clear that the texts could have also been ordered by their lengths; i.e., the shorter the text, the higher it is on the list[24]. This means that the TTR values of complete texts will not be helpful in a comparative style analysis.

Table 2.1. TTR final values for the WSP corpus[25]

TEXT	N	V	TTR
XJ	1800	374	0.207778
SHI	29622	2833	0.095638
LY	15923	1361	0.085474
SHU	24537	1910	0.077842
ZY	13348	1030	0.077165
CQ	16791	941	0.056042
MZ	35354	1892	0.053516
ZHZ	65251	2968	0.045486
ZL	49410	2212	0.044768
GL	40835	1594	0.039035
GY	44224	1640	0.037084

[24] With notable exceptions such as the Shi Jing, the Shu Jing, the Zhuangzi, and the Yi Li.

[25] The first column features the abbreviated text name (here and thereafter, see Abbreviations section for full text names); the second column is N the number of tokens (characters) in the text, or, the sample size; the third column, V, features the number of types in the text, and the fourth column is the TTR, calculated as the ratio V/N, where V is the number of types in the complete text, and N is the number of tokens [25]. The structure of all further index tables is same.

LJ	97994	3041	0.031033
YL	53882	1536	0.028507
ZZ	178563	3235	0.018117
CQZZ	195354	3251	0.016642

2.2. Other indices

The TTR issues have been known for long time and many researchers have attempted to create a length-independent measure of lexical richness ("length-invariant statistics")[26]. Tweedie and Baayen conveniently summarized these attempts in their article, "How Variable May a Constant Be?," and conducted a study to demonstrate that these indices still depend on text length, although some of them are "less dependent" than others. The present study follows Tweedie and Baayen's approach by applying it to the texts in the WSP Ctexts corpus[27].

Table 2.2 presents a list of several nonparametrical and parametrical indices[28].

Table 2.2. List of static indices of lexical diversity[29]

Index Full Name	**Short Name**	**Calculation method**
Guiraud	R[30]	$R = \frac{V(N)}{\sqrt{N}}$

[26] Fiona Tweedie and Harald Baayen use the term "lexical constants" (Tweedie and Baayen, "How Variable May a Constant Be?").

[27] Since the publication of their article (Tweedie and Baayen, "How Variable May a Constant Be?"), several more indices were invented with varying degree of success. The present study will only use those in the original Tweedie and Baayen article. Some newer articles will be mentioned, but they do not add much progress to the already known methods. David Mitchell (Mitchell, "Type-token Models: S Comparative Study") provides an even larger list.

[28] Nonparametrical models usually depend on the sample size (number of tokens) N and the number of types V, while parametrical models introduce extra textual parameters (e.g., Brunet's formula for W includes the parameter "a," which is usually set to 0.172; see Table 2.2).

[29] The first column contains the name(s) of the researcher(s), the second one includes an abbreviated index notation, and the third one presents its formula, following Tweedie and Baayen's "How Variable May a Constant Be?" 326–331. "N" is the sample size in tokens (characters) and "V" is the vocabulary size in tokens. V(N) is the number of types in the sample of size N. It is usually more convenient to simply use "V" when "N" is obvious. V(1,N) is the number of types that are hapax legomena in the sample of size N, while (V2,N) is the number of dis legomena in the sample.

Herdan	C	$C = \frac{\log V(N)}{\log N}$
Rubet	k	$k = \frac{\log V(N)}{\log(\log N)}$
Maas	A^2	$a^2 = \frac{\log N - \log V(N)}{\log^2 N}$
Luk'janenkov & Nesitoj	LN	$LN = \frac{1 - V(N)^2}{V(N)^2 \log N}$
Brunet	W	$W = N^{V(N)^{-a}}$
Honoré	H	$H = 100\frac{\log N}{1 - \frac{V(1,N)}{V(N)}}$
Sichel	S	$S = \frac{V(2,N)}{V(N)}$
Michéa	M	$M = \frac{V(N)}{V(2,N)}$
Yule	K	$K = 10^4 \frac{[\sum_{i=1}^{N} V(i,N)(i/N)^2] - N}{N^2}$ $= 10^4\left[-\frac{1}{N} + \sum_i V(i,N)\left(\frac{i}{N}\right)^2\right]$.
Herdan	Vm	$V_m = \sqrt{\sum_{i=1}^{V(N)} V(i,N)(i/N)^2 - \frac{1}{V(N)}}$.

Table 2.3 contains the values of these indices for the entire corpus[31]. The reason why this study calculated the values of these indices is that the indices were presumed to reflect the intrinsic inner characteristics of the texts expressed in their vocabulary, some of which are still popular in research. Based on the material of English prose, Tweedie and Baayen demonstrated that constants are not actually "constant." However, it is interesting to test them against classical Chinese texts, especially groups of texts with varying lengths such as those from the WSP Ctexts corpus.

[30] This formula counts in all tokens in sample as N. In case if only nouns, etc. real words (no function words) are counted, there could be V/SQR(2N) formula. See Daller, "Guirad's Index" and Van Hout and Vermeer, "Comparing measures of lexical richness".

[31] Yule's K and Herdan's Vm are not presented in Table 2.3.

Table 2.3. Lexical diversity indices[32]. The first column features text names, second and third column — such numerical indices as number of tokens in text sample (N) and the number of types (V), and other columns are featuring other indices, presented by their abbreviation.

Text	N	V	TTR	R	C	k	A2	LN	W	H	S	M	K	Vm
			TTR=V/N											
Chunqiu	16791	941	0.056	7.262	0.7038	4.7514	0.0701	-0.237	20.012	647.5239	0.1605	6.231788	119.086	0.1028
Chunqiu Zuozhuan	195354	3251	0.017	7.355	0.6638	4.854	0.0635	-0.189	20.73	640.3747	0.1003	9.972393	63.5706	0.0789
Gongyang-zhuan	44224	1640	0.037	7.799	0.692	4.8195	0.0663	-0.215	19.973	611.467	0.1421	7.038627	82.4085	0.0872
Guliang-zhuan	40835	1594	0.039	7.888	0.6945	4.8245	0.0662	-0.217	19.819	622.8802	0.1255	7.97	95.4268	0.0954
Liji	97994	3041	0.031	9.714	0.6978	4.9885	0.0605	-0.2	18.047	656.4982	0.1256	7.960733	71.3906	0.0831
Lunyu	15923	1361	0.085	10.79	0.7458	5.0266	0.0605	-0.238	16.391	646.9407	0.1646	6.075893	135.486	0.1122
Mengzi	35354	1892	0.054	10.06	0.7204	4.9812	0.0615	-0.22	17.471	623.146	0.1543	6.479452	105.034	0.1
Shi-jing	29622	2833	0.096	16.46	0.772	5.3074	0.051	-0.224	13.785	622.8163	0.1719	5.817248	49.1837	0.063
Shu-jing	24537	1910	0.078	12.19	0.7474	5.1071	0.0575	-0.228	15.741	621.5388	0.1466	6.821429	54.0901	0.0625
Xiao-jing	1800	374	0.208	8.815	0.7904	5.0194	0.0644	-0.307	14.964	571.5831	0.1872	5.342857	114.58	0.0355
Yili	53882	1536	0.029	6.617	0.6735	4.7206	0.069	-0.211	21.851	613.2909	0.1335	7.492683	66.3595	0.077
Zhouli	49410	2212	0.045	9.951	0.7126	4.9809	0.0612	-0.213	17.702	631.5522	0.1356	7.373333	84.989	0.0901
Zhouyi	13348	1030	0.077	8.915	0.7303	4.8952	0.0654	-0.242	17.824	533.8164	0.1932	5.175879	102.313	0.0933
Zhuangzi	65251	2968	0.045	11.62	0.7212	5.0874	0.0579	-0.208	16.482	679.8142	0.1435	6.967136	90.839	0.0937
Zuozhuan	178563	3235	0.018	7.656	0.6683	4.8727	0.0632	-0.19	20.324	642.0841	0.0998	10.01548	67.0797	0.0811

[32] See "voc_ref.xlsx/MAIN_LOOKUP/Lexical diversity indices".

3. Final Value Index Approach

Quantitative linguistics is not very popular in philological studies, partly because the results of this discipline cannot be immediately applicable or interpreted in philological analyses, e.g., stylistically. Particularly, it relates to "word counts," which were even determined to be useless[33]. It is true that indices' final values do not elucidate much about these texts, and, by themselves, are not very useful[34]. As Van Hout and Vermee formulate, "does a higher outcome really reflect a richer underlying lexicon? Can we be certain and happy about the values produced by lexical measures?"[35]

However, these indices can be used for a comparison of texts of similar length; i.e., why vocabulary richness indices are still used in authorship identification and style characterization. Further in this study, each index (or "constant") will be discussed separately, and its values will be presented both numerically (as a sorted list of values) and visually (as a dendrogram).

3.1. Clustering TTR values

The simplest (and most controversial) index, the TTR, is presented in Column 4 in Table 2.3. It is easy to see that the highest TTR (0.208) is produced by the Xiao Jing, while the lowest TTR (0.017) is produced by the combined Chun Qiu Zuo Zhuan. This does not mean that the vocabulary of the Xiao Jing is "richer" than that of the Chun Qiu. The Xiao Jing only includes 374 unique characters (types), but it is very short (1,800 characters in the WSP Ctexts version). The Chun Qiu Zuo Zhuan includes 3,251 unique characters, but it is the longest text in the WSP corpus at 195,354 characters. The number of types increases with the sample size, but as the sample size changes, so do their ratios. Therefore, the TTR value for the entire text (the final value) depends on the sample size.

While the TTR generally diminishes with sample size, some larger texts include higher TTR values than other smaller ones. It could be useful to group these texts by such values based on certain "similarity" metrics. One of the ways to group the items is through hierarchical clustering.

---

[33] "A word frequency analysis of a text can reveal nothing about its characteristics (e.g., author, language, style, type of literature). The only exception appears to be Shakespeare" (Naranan and Balasubrahmanyan, "Models for Power Law Relations," 38). See the critique of this position by Sampson (Sampson, "Review of Harald Baayen.")

[34] Duran et al., with their D (vocd), attempt to introduce a new index of lexical diversity. However, McCarthy et al. ("vocd-D and HD-D") argue that this index also depends on sample length (McCarthy et al., ibid, 382). The present study also includes a review of post-Baayen indices (McCarthy et al., ibid, 382).

[35] Hout van and Vermee, "Comparing Measures," 94.

In addition, the results of hierarchical clustering can be graphically presented as a cluster dendrogram[36].

The cluster dendrogram presents the same data as a regular table, but a clustering algorithm attempts to combine the texts (as "geometrical points") into largers groups (clusters) based on their closeness as "points" (starting from two). Moreover, it further combines smaller groups of points into largers clusters based on the Euclidian distance between the centers of the clusters[37].

The standard cluster dendrogram algorithm (Euclidian metrics with the "average method") produced the graph in Figure 3.1. If the dendrogram is cut at the 0.04 level on y-axis[38], then the algorithm groups texts into three wide groups: Group 1, consisting of the outlier the Xiao Jing; Group 2, consisting of the Shi Jing, the Shu Jing, the Lun Yu, and the Zhou Yi; and Group 3, consisting of all other texts. Group 2 features texts with a higher TTR level, so it is not surprising that small- to medium-sized texts belong there. Group 3 features texts with a lower TTR level. What is surprising, e.g., is that the Chun Qiu and the Mengzi were also placed in Group 3, while the Shi Jing falls into Group 2 (i.e., it is treated as a shorter text).

$$TTR = V(N)/N$$

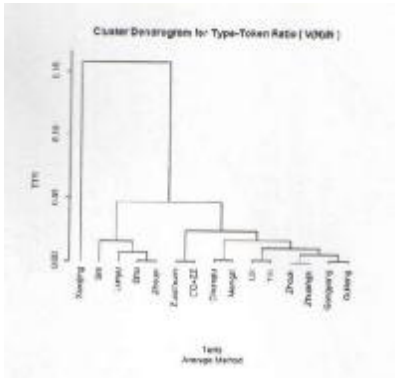

Figure 3.1. TTR dendrogram

[36] This study utilized clustering software provided by the standard R language package. It also used agglomerative clustering with Euclidian average distance for metrics.

[37] In other words, "the average method." Other methods were attempted in this study, but they did not produce a significant difference.

[38] Hereafter, the value for the horizontal cut is chosen in the way to identify the most meaningful largest groups.

Table 3.1. TTR values (sorted by TTR in decreasing order)[39]

Text	N	V	TTR
XJ	1800	374	0.2078
SHI	29622	2833	0.0956
LY	15923	1361	0.0855
SHU	24537	1910	0.0778
ZY	13348	1030	0.0772
CQ	16791	941	0.056
MZ	35354	1892	0.0535
ZHZ	65251	2968	0.0455
ZL	49410	2212	0.0448
GL	40835	1594	0.039
GY	44224	1640	0.0371
LJ	97994	3041	0.031
YL	53882	1536	0.0285
ZZ	178563	3235	0.0181
CQZZ	195354	3251	0.0166

It is difficult to see much "stylistical meaning" in grouping together, e.g., the Zhou Li, the Zhuangzi, the Gongyang Zhuan, and the Guliang Zhuan, except for their ordering according to the TTR final values. In addition, combining the Chun Qiu and the Mengzi definitely contradicts stylistic expectation. Otherwise, the results of the TTR approach are basically what could be expected and they mostly reflect text sample size. However, the indices to be discussed below claimed independency of text length. Therefore, they will be reviewed in the order of their position in Table 2.2, which is the order of their presentation in Tweedie and Baayen's article.

3.2. Guiraud's R

$$R = \frac{V(N)}{\sqrt{N}}$$

[39] See "voc_ref.xlsx/MAIN_LOOKUP/TTR values for WSP corpus".

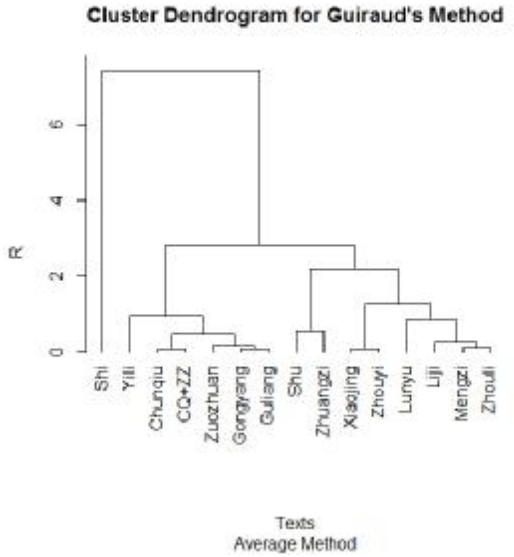

Figure 3.2. Guiraud's R dendrogram

Table 3.2. Guiraud's R values (sorted by R in decreasing order)[40]

Text	N	V	R
SHI	29622	2833	16.46036
SHU	24537	1910	12.19334
ZHZ	65251	2968	11.61904
LY	15923	1361	10.78563
MZ	35354	1892	10.06241
ZL	49410	2212	9.951251
LJ	97994	3041	9.714416
ZY	13348	1030	8.915160
XJ	1800	374	8.815265
GL	40835	1594	7.888093
GY	44224	1640	7.798568
ZZ	178563	3235	7.655588
CQZZ	195354	3251	7.355392
CQ	16791	941	7.261918
YL	53882	1536	6.617125

[40] See "voc_ref.xlsx/MAIN_LOOKUP Guiraud's R values (sorted by R in decreasing order)".

Guiraud's R demonstrates less dependency on text length since text order according to R is not the text size order. The Shi Jing goes to the top of the ordered list, which is closed by the Yi Li. The longest texts, such as the Zuo Zhuan and the Zuo Zhuan with the Chun Qiu, are still placed closer to the end, while other long texts, such as the Zhuangzi and the Li Ji, are placed in the first half of the list. The dendrogram cut at the 2.0 level provides four groups: 1) the singular Shi Jing; 2) the Shu Jing and the Zhuangzi; 3) the Yi Li; and 4) the Chun Qiu, the Zuo Zhuan with the Chun Qiu, the Zuo Zhuan, the Gongyang Zhuan, and the Guliang Zhuan (i.e., mostly "historical"[41] prosaic texts). This arrangement indicates some relationship to stylistic characteristics[42].

However, this clustering does not offer meaningful stylistic grouping, especially since the Xiao Jing is paired with the Zhou Yi and the Mengzi is paired with the Zhou Li. Yet, Hoet Van and Vermee consider (Hout van and Vermee, "Comparing Measures," 100) Guiraud's R to be the most productive measure of lexical richness (measuring proficiency of second language learning).

3.3. Herdan's C

$$C = \frac{\log V(N)}{\log N}$$

[41] Here, the adjective "historical" is not a genre definition. Some of these texts are not really "historic," e.g., the Chun Qiu "chronicle itself may be seen as a developed form of omen record" (Brooks and Brooks, "Emergence of China," 22), not a consciously written historical text that could be extended to the Gongyang Zhuan and the Guliang Zhuan. These texts are not even "narratives," according to the following popular definition: "[N]either does narrative exist without integration into the unity of a plot, but only chronology, an enunciation of a succession of uncoordinated facts" (Bremond, "Logic of Narrative Possibilities," 390). The Zuo Zhuan contains some narratives and historical prose. However, these texts are not only close stylistically but they also record and interpret events in history.

[42] However, this author does not want to state that vocabulary richness values can be directly linked to genre stylistics. This issue will be discussed more in Section 4.3. However, it is worth noting any discovered correlation between quantitative indices and genre stylistics.

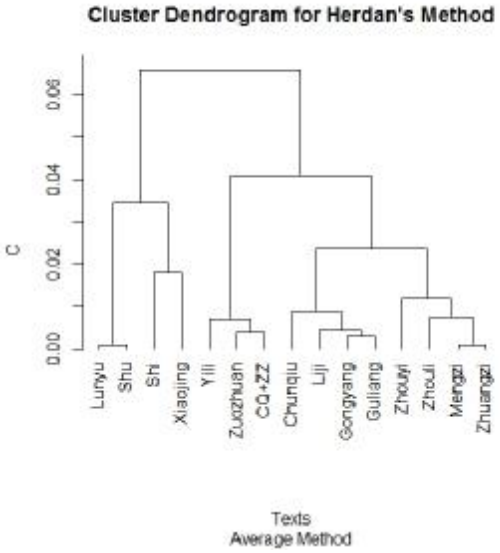

Figure 3.3. Herdan's C dendrogram

Table 3.3. Herdan's C values (sorted by C in decreasing order)[43]

Text	N	V	C
XJ	1800	374	0.790371
SHI	29622	2833	0.772036
SHU	24537	1910	0.747418
LY	15923	1361	0.745797
ZY	13348	1030	0.730311
ZHZ	65251	2968	0.721238
MZ	35354	1892	0.72045
ZL	49410	2212	0.712594
CQ	16791	941	0.703795
LJ	97994	3041	0.697832
GL	40835	1594	0.694527
GY	44224	1640	0.69201
YL	53882	1536	0.67345
ZZ	178563	3235	0.668319
CQZZ	195354	3251	0.663794

[43] See "voc_ref.xlsx/MAIN_LOOKUP/ Herdan's C values (sorted by C in decreasing order)".

Unlike Giraud's R, Herdan's C follows texts' size closer, although not as close as the TTR. If a cut on its dendrogram is made at the 0.03 level, then the clustering produces several groups, vaguely depending on size. That is, it groups the Chun Qiu, the Gongyang Zhuan, and the Guliang Zhuan with the Li Ji, but it groups the Zuo Zhuan and the Chun Qiu Zuo Zhuan together with the Yi Li. It also groups the Lun Yu and the Shu Jing, while placing the Mengzi into one group with the Zhuangzi.

3.4. Rubet's k

$$k = \frac{\log V(N)}{\log(\log N)}$$

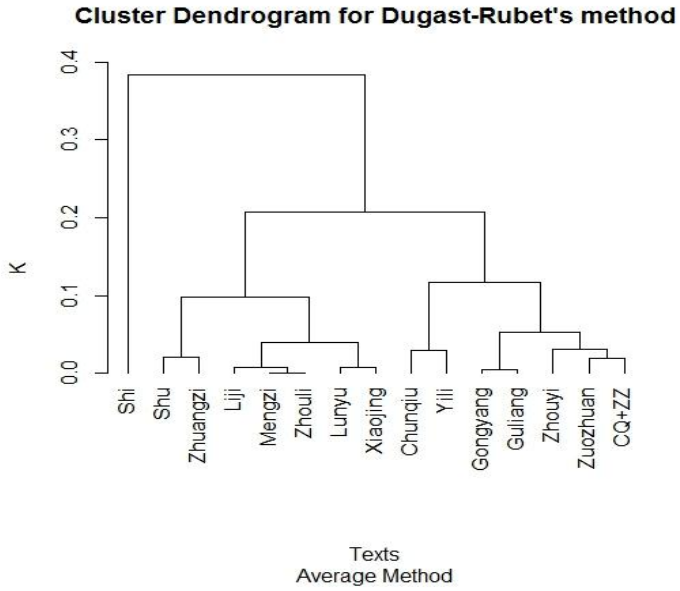

Figure 3.4. Rubet's k dendrogram

Table 3.4. Rubet's k values (sorted by k in decreasing order)[44]

Text	N	V	K
SHI	**29622**	**2833**	5.307357
SHU	**24537**	**1910**	5.107089
ZHZ	**65251**	**2968**	5.087419
LY	**15923**	**1361**	5.02657
XJ	**1800**	**374**	5.019382
LJ	**97994**	**3041**	4.98853
MZ	**35354**	**1892**	4.981165
ZL	**49410**	**2212**	4.980872
ZY	**13348**	**1030**	4.895199
ZZ	**178563**	**3235**	4.872745
CQZZ	**195354**	**3251**	4.854049
GL	**40835**	**1594**	4.824491
GY	**44224**	**1640**	4.819515
CQ	**16791**	**941**	4.751399
YL	**53882**	**1536**	4.720624

Rubet's k is similar to Guiraud's R in four characteristics that were indicated above: 1) the order of the text, structured by decreasing k, does not follow the text lengths' ordering; 2) it groups "historical texts" together; 3) it places the Shi Jing at the top; and 4) it places the Yi Li at the bottom of the k-ordered list.

3.5. Maas' A^2

$$a^2 = \frac{\log N - \log V(N)}{\log^2 N}$$

[44] See "voc_ref.xlsx/MAIN_LOOKUP/ Rubet's k values (sorted by k in decreasing order)".

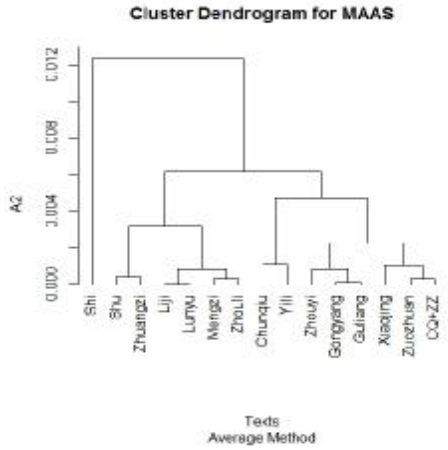

Figure 3.5. Maas's A^2 dendrogram

Table 3.5. Maas's A^2 values (sorted by A^2 in decreasing order)[45]

Text	N	V	A^2
CQ	**16791**	**941**	0.070106
YL	**53882**	**1536**	0.069017
GY	**44224**	**1640**	0.066296
GL	**40835**	**1594**	0.066248
ZY	**13348**	**1030**	0.065373
XJ	**1800**	**374**	0.064397
CQZZ	**195354**	**3251**	0.063545
ZZ	**178563**	**3235**	0.063156
MZ	**35354**	**1892**	0.061461
ZL	**49410**	**2212**	0.061231
LJ	**97994**	**3041**	0.06054
LY	**15923**	**1361**	0.060495
ZHZ	**65251**	**2968**	0.057899
SHU	**24537**	**1910**	0.057538
SHI	**29622**	**2833**	0.05098

Maas' A^2, like Rubet's k and Guiraud's R, does not display a correlation of text lengths and index values. In addition, it does not offer any specific genre grouping (e.g., historic texts). Meanwhile, it selects the Shi Jing as a text at the list's extreme and places the Yi Li close to this extreme.

[45] See "voc_ref.xlsx/MAIN_LOOKUP/ Maas's A^2 values (sorted by A^2 in decreasing order)".

3.6. Lukyanenko–Nesytoj's LN

$$LN = \frac{1 - V(N)^2}{V(N)^2 \log N}$$

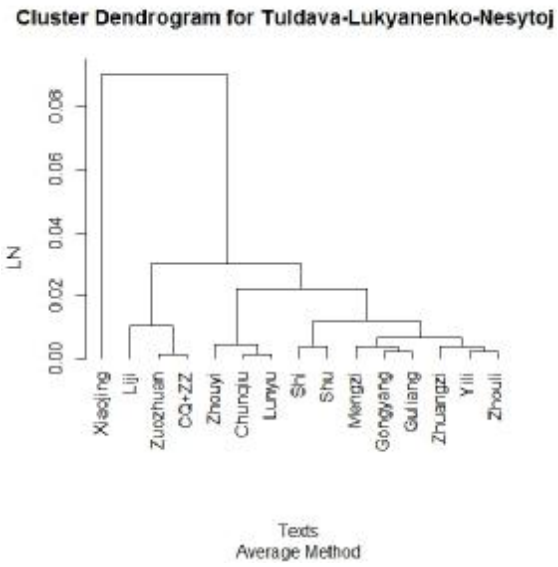

Figure 3.6. Lukyanenko and Nesitoj's LN dendrogram

Table 3.6. LN values (sorted by LN in inreasing order)[46]

Text	N	V	LN
XJ	**1800**	**374**	-0.30719
ZY	**13348**	**1030**	-0.2424
LY	**15923**	**1361**	-0.23798
CQ	**16791**	**941**	-0.23668
SHU	**24537**	**1910**	-0.2278
SHI	**29622**	**2833**	-0.22363
MZ	**35354**	**1892**	-0.21986
GL	**40835**	**1594**	-0.21687
GY	**44224**	**1640**	-0.21525
ZL	**49410**	**2212**	-0.21305
YL	**53882**	**1536**	-0.21135

[46] See "voc_ref.xlsx/MAIN_LOOKUP/ LN values (sorted by LN in inreasing order)".

ZHZ	**65251**	**2968**	-0.2077
LJ	**97994**	**3041**	-0.20035
ZZ	**178563**	**3235**	-0.19041
CQZZ	**195354**	**3251**	-0.18901

The Lukyanenko–Nesytoj's LN, similar to the TTR, basically displays a correlation of text lengths and index values. Moreover, it separates "historical" texts. Here the Shi Jing is placed in the middle of the ordered list, while groupings in the dendrogram (cut at the 0.02 level) do not offer much stylistic meaning.

3.7. Brunet' W

$$W = N^{V(N)^{-a}}$$

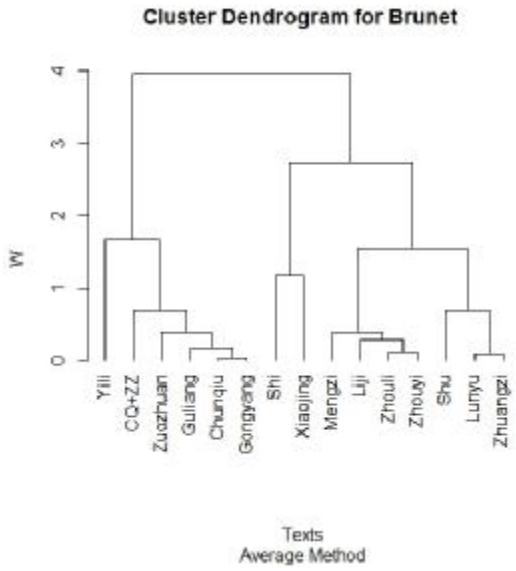

Figure 3.7. Brunet's W dendrogram

Table 3.7. Brunet's W values (sorted by W in increasing order)[47]

Text	N	V	W
SHI	**29622**	**2833**	13.78493
XJ	**1800**	**374**	14.96381
SHU	**24537**	**1910**	15.74132
LY	**15923**	**1361**	16.39098
ZHZ	**65251**	**2968**	16.48212
MZ	**35354**	**1892**	17.47091
ZL	**49410**	**2212**	17.70207
ZY	**13348**	**1030**	17.82408
LJ	**97994**	**3041**	18.04657
GL	**40835**	**1594**	19.81939
GY	**44224**	**1640**	19.97337
CQ	**16791**	**941**	20.0124
ZZ	**178563**	**3235**	20.32376
CQ Zuozhuan	**195354**	**3251**	20.73038
YL	**53882**	**1536**	21.85115

If the cut is made at the 1.0 level, then Brunet's W provides five sub-groups of which one of them groups historical texts together. In addition, it places the Yi Li and the Shi Jing at the extreme ends of the ordered list.

3.8. Honore's H

$$H = 100\frac{\log N}{1 - \frac{V(1,N)}{V(N)}}$$

[47] See "voc_ref.xlsx/MAIN_LOOKUP/ Brunet's W values (sorted by W in increasing order)".

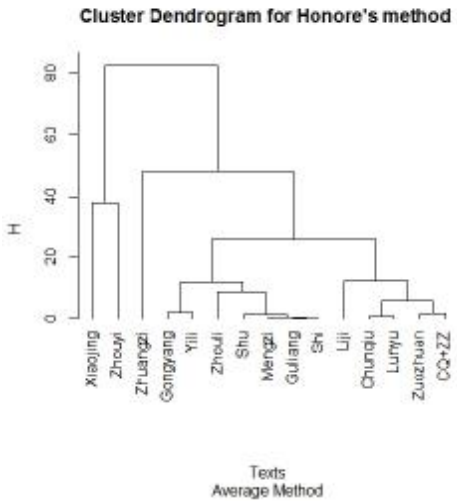

Figure 3.8. Honore's H dendrogram

Table 3.8 Honore's H values (sorted by H in increasing order)[48]

Text	N	V(N)	H
ZY	**13348**	**1030**	533.8164
XJ	**1800**	**374**	571.5831
GY	**44224**	**1640**	611.467
YL	**53882**	**1536**	613.2909
SHU	**24537**	**1910**	621.5388
SHI	**29622**	**2833**	622.8163
GL	**40835**	**1594**	622.8802
MZ	**35354**	**1892**	623.146
ZL	**49410**	**2212**	631.5522
CQ Zuozhuan	**195354**	**3251**	640.3747
ZZ	**178563**	**3235**	642.0841
LY	**15923**	**1361**	646.9407
CQ	**16791**	**941**	647.5239
LJ	**97994**	**3041**	656.4982
ZHZ	**65251**	**2968**	679.8142

Honore's H does not correlate text lengths and index values, but it is difficult to find stylistic meaning in its groupings.

[48] See "voc_ref.xlsx/MAIN_LOOKUP/ Honore's H values (sorted by H in increasing order)".

3.9. Sichel's S

$$S = \frac{V(2, N)}{V(N)}$$

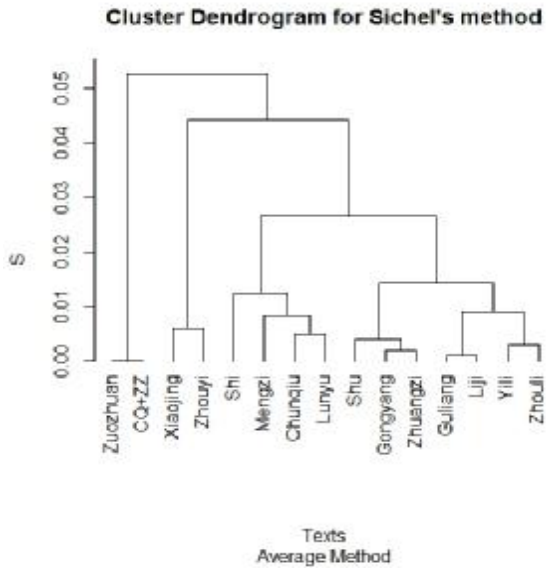

Figure 3.9. Sichel's S dendrogram

Table 3.9. Sichel's S values (sorted by S in decreasing order)[49]

Text	N	V	S
ZY	**13348**	**1030**	0.193204
XJ	**1800**	**374**	0.187166
Shi	**29622**	**2833**	0.171903
LY	**15923**	**1361**	0.164585
CQ	**16791**	**941**	0.160468
MZ	**35354**	**1892**	0.154334
SHU	**24537**	**1910**	0.146597
ZHZ	**65251**	**2968**	0.143531
GY	**44224**	**1640**	0.142073

[49] See "voc_ref.xlsx/MAIN_LOOKUP/ Sichel's S values (sorted by S in decreasing order)".

ZL	**49410**	**2212**	0.135624
YL	**53882**	**1536**	0.133464
LJ	**97994**	**3041**	0.125617
GL	**40835**	**1594**	0.125471
CQZZ	**195354**	**3251**	0.100277
ZZ	**178563**	**3235**	0.099845

If the dendrogram is cut at the 0.02 level, then Sichel's S clustering produces four groups, grouping together (among others) the two longest texts and then the Xiao Jing and the Zhou Yi. While Sichel's S order is not exactly the TTR order, it vaguely correlates to text size.

3.10. Michea's M

$$M = \frac{V(N)}{V(2, N)}$$

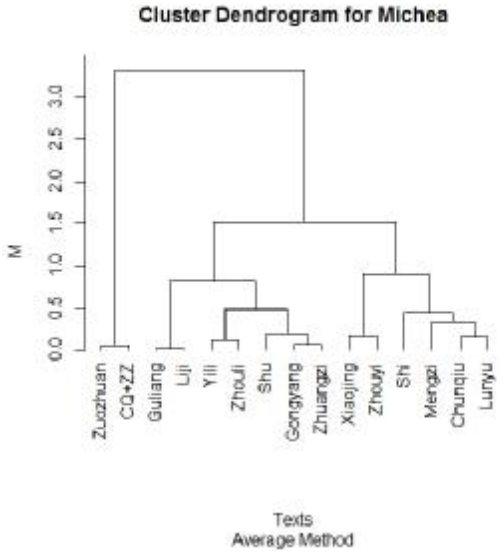

Figure 3.10. Michea's M dendrogram

Table 3.10. Miches's M values (sorted by M in increasing order)[50]

Text	N	V	M
ZY	**13348**	**1030**	5.175879
XJ	**1800**	**374**	5.342857
SHI	**29622**	**2833**	5.817248
LY	**15923**	**1361**	6.075893
CQ	**16791**	**941**	6.231788
MZ	**35354**	**1892**	6.479452
SHU	**24537**	**1910**	6.821429
ZHZ	**65251**	**2968**	6.967136
GY	**44224**	**1640**	7.038627
ZL	**49410**	**2212**	7.373333
YL	**53882**	**1536**	7.492683
LJ	**97994**	**3041**	7.960733
GL	**40835**	**1594**	7.97
CQZZ	**195354**	**3251**	9.972393
ZZ	**178563**	**3235**	10.01548

Michea's M index, in some ways, is similar to Sichel's S and other indices that correlate index values and text sizes.

3.11. Yule's K

$$K = 10^4 \frac{[\sum_{i=1}^{N} V(i,N)(i/N)^2] - N}{N^2}$$
$$= 10^4 \left[-\frac{1}{N} + \sum_i V(i,N) \left(\frac{i}{N}\right)^2 \right],$$

[50] See "voc_ref.xlsx/MAIN_LOOKUP/ Miches's M values (sorted by M in increasing order)".

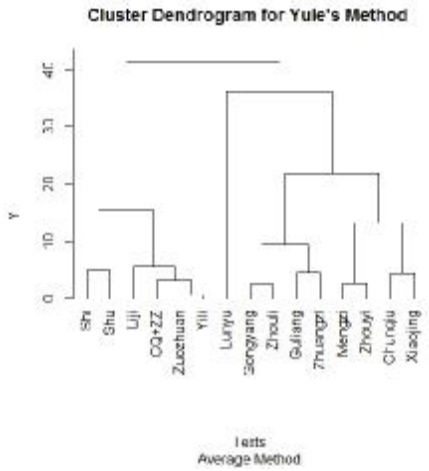

Figure 3.11. Yule's K dendrogram

Table 3.11. Yule's K values (sorted by K in decreasing order)[51]

Text	N	V	K
LY	15923	1361	135.486
CQ	16791	941	119.086
XJ	1800	374	114.580
MZ	35354	1892	105.034
ZY	13348	1030	102.313
GL	40835	1594	95.4268
ZHZ	65251	2968	90.8390
ZL	49410	2212	84.9890
GY	44224	1640	82.4085
LJ	97994	3041	71.3906
ZZ	178563	3235	67.0797
YL	53882	1536	66.3595
CQZZ	195354	3251	63.5706
SHU	24537	1910	54.0901
SHI	29622	2833	49.1837

Yule's K index does not display direct dependency on the text length. Its groupings, provided by clustering (cut at 14), are very different than those of other indices.

[51] See "voc_ref.xlsx/MAIN_LOOKUP/ Yule's K values (sorted by K in decreasing order)".

3.12. Herdan's Vm

$$V_m = \sqrt{\sum_{i=1}^{V(N)} V(i,N)(i/N)^2 - \frac{1}{V(N)}}.$$

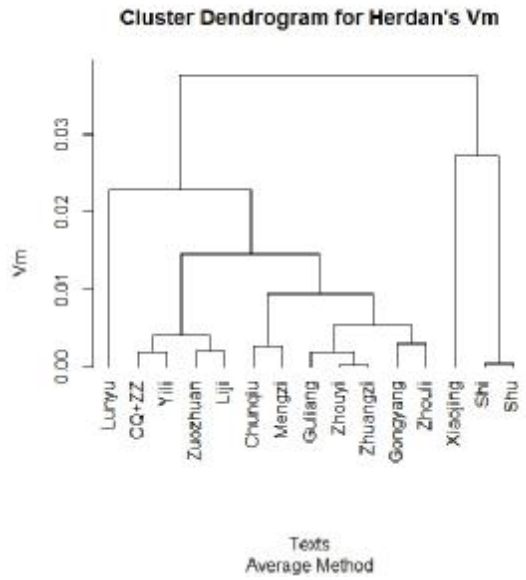

Figure 3.12. Herdan's Vm dendrogram

Table 3.12. Herdan's Vm values (sorted by Vm in increasing order)[52]

Text	N	V(N)	Vm
XJ	1800	374	0.035509
SHU	24537	1910	0.062502
SHI	29622	2833	0.063014
YL	53882	1536	0.077046
CQZZ	195354	3251	0.078906
ZZ	178563	3235	0.081092
LJ	97994	3041	0.083106

[52] See "voc_ref.xlsx/MAIN_LOOKUP/ Herdan's Vm values (sorted by Vm in increasing order)".

GY	44224	1640	0.087152
ZL	49410	2212	0.090117
ZY	13348	1030	0.093307
ZHZ	65251	2968	0.093677
GL	40835	1594	0.095377
MZ	35354	1892	0.10003
CQ	16791	941	0.102799
LY	15923	1361	0.112172

Herdan's Vm is interesting since it singles out the Lun Yu (similar to Yule's K). Otherwise, it does not offer any interesting stylistic grouping.

3.13. Partial TTR measurements

The results of the analysis based on the final values of indices seem to be extremely diverse. Only a few constants allowed the grouping of texts (by clustering) in a way that could be remotely interpreted as stylistically meaningful.

It is possible to measure WSP texts at equal sample sizes. The TTR values could be calculated at some fixed sample intervals, e.g., at 15,000 and 30,000 characters. The results for the 30,000 token samples are presented in Dendrogram 3.13 and Table 3.1, while the results for the 15,000 token samples are presented in Dendrogram 3.14 and Table 3.14. These points have been selected due to the 30,000 tokens being the sample size, which include all large- and medium-sized texts, and the 15,000 tokens since this sample size includes all texts (except the Xiao Jing).

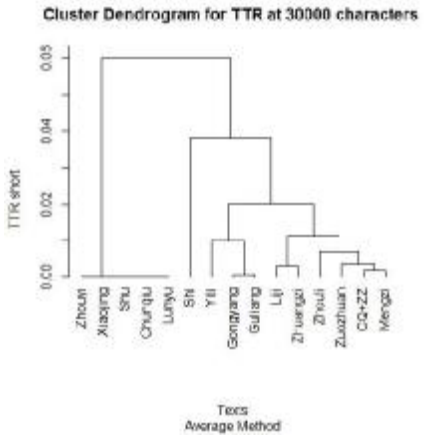

Figure 3.13. Partial TTR's dendrogram. Taken for sample lenghts of 30 000 characters (shorter texts are assigned 0 values)

Table 3.13. Partial TTR values (sorted in decreasing order by TTR), for sample lenghts of 30 000 characters (shorter texts are assigned n/a values)[53]

Text	N	V	V(30000)	TTR(30000)
SHI	29622	2833	2833	0.094433
ZHZ	65251	2968	2161	0.072033
LJ	97994	3041	2069	0.068967
ZZ	178563	3235	1902	0.0634
CQZZ	195354	3251	1825	0.060833
MZ	35354	1892	1768	0.058933
ZL	49410	2212	1627	0.054233
GL	40835	1594	1392	0.0464
GY	44224	1640	1381	0.046033
YL	53882	1536	1082	0.036067
CQ	16791	941	n/a	n/a
LY	15923	1361	n/a	n/a
SHU	24537	1910	n/a	n/a
XJ	1800	374	n/a	n/a
ZY	13348	1030	n/a	n/a

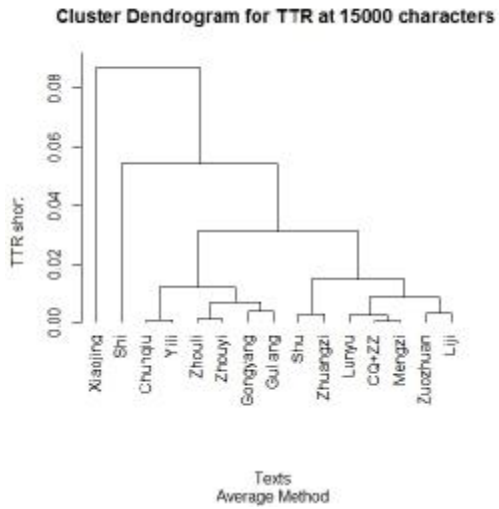

Figure 3.14. Partial TTR's dendrogram. Taken for sample lenghts of 15 000 characters (shorter texts are assigned 0 values).

[53] See "voc_ref.xlsx/MAIN_LOOKUP/ Partial TTR values (sorted in decreasing order by TTR)".

Table 3.14. Partial TTR values for 15000 samplees (sorted in decresing order by TTR), for sample lenghts of 15 000 characters (shorter texts are assigned n/a values)[54]

Text	N	V	V(15000)	TTR(15000)
SHI	29622	2833	2067	0.1378
SHU	24537	1910	1658	0.110533
ZHZ	65251	2968	1611	0.1074
LJ	97994	3041	1510	0.100667
ZZ	178563	3235	1461	0.0974
MZ	35354	1892	1375	0.091667
CQZZ	195354	3251	1364	0.090933
LY	15923	1361	1328	0.088533
ZL	49410	2212	1125	0.075
GL	40835	1594	1037	0.069133
ZY	13348	1030	1030	0.068667
GY	44224	1640	978	0.0652
CQ	16791	941	888	0.0592
YL	53882	1536	872	0.058133
XJ	1800	374	n/a	n/a

While some texts are shorter than 30,000 tokens, it is possible to compare the results with clustering at 15,000 tokens. Cutting both dendrograms at the 0.02 level provides similar results. Unlike the final values dendrogram, the Shi Jing (the Xiao Jing is missing) is singled out in clustering and placed at the top of the list, while the Yi Li is placed at the bottom. In both cases, historical texts are split. The values of the TTR in both cases of partial samples do not correlate with text lengths, unlike the situation with final values[55]. This supports the idea that the TTR index (under certain conditions) can be beneficial for evaluating vocabulary richness.

3.14. Discussion of the results for the final value and partial index approaches

As Hoover notes, "various measures of vocabulary richness produce further interesting differences in how they rank texts on the basis of vocabulary richness – differences that reflect their radically different bases and methods of calculation" (Hoover, "Another Perspective," 169). Hoover, similar to Tweedie and Baayen (336), had the benefit of controlling

[54] See "voc_ref.xlsx/MAIN_LOOKUP/ Partial TTR values for 15000 samplees (sorted in decresing order by TTR)".

[55] In Section 4, these results will be discussed in more detail.

these differences by authorship of texts in their sample and grouping the constants based on correct ranking[56]. In the case of the WSP corpus, other criteria could be used for grouping indices.

Table 3.14. Indices matching four criteria

Index	Short name	Length-ordered	Grouping historical texts	Shi Jing/ top	Yi Li / bottom
TTR	TTR	yes	no	(second)	no
TTR at 30000 sample	TTR	no	no	yes	yes
TTR at 15000 sample	TTR	no	no	yes	yes
Guiraud	R	no	yes	yes	yes
Herdan	C	almost	no	(second)	no
Rubet	k	no	yes	yes	Yes
Maas	A2	no	no	yes	almost
Luk'janenkov-Nesitoj	LN	yes	no	no	no
Brunet	W	no	yes	yes	yes
Honoré	H	no	no	no	No
Sichel	S	almost	no	no	No
Michéa	M	almost	no	no	No
Yule	K	no	no	yes	No
Herdan	VM	no	no	no	no

Four recurrent binary features have been noted earlier. First, the order of the final values may or may not reflect text lengths. Second, some indices group "historical" texts together (which might be seen as stylistic selection), while others do not. Third, some indices could definitely favor the Shi Jing, placing it at the top of the list. Fourth, similarly, some indices place the Yi Li at the opposite end (of the Shi Jing) of the list. These four binary features could form the criteria for grouping indices. Table 3.15 presents the breakdown. Finally, Guiraud's R, Rubet's k, and Brunet's W satisfy all four criteria. The partial TTR values match three of the four criteria, excluding genre grouping. This result will be discussed in more detail in the fourth section.

[56] Hoover further notes that "these variations in richness order merely emphasize the fact that different measures of vocabulary richness measure different aspects of vocabulary structure" (Hoover, "Another Perspective," 169).

4. Vocabulary Development Profiles

The TTR index could still be used for the description of vocabulary richness and growth by charting vocabulary growth dynamically as developmental profiles, which was demonstrated by, e.g., Tweedie and Baayen's "How Variable May a Constant Be?"[57].

Unlike most other indices, the TTR has an immediate and clear meaning of measuring the relative rate of vocabulary growth. If a text demonstrates a comparatively faster growth of vocabulary, then its developmental curve remains higher on the chart compared to the curves of the texts with a lower rate of vocabulary growth. The TTR developmental profile allows analyzing the dynamics of the rate of adding new characters to the existing vocabulary. Analysis of developmental profiles helps visually (as well as numerically[58]) estimate how fast some texts add new characters to their vocabulary compared to other texts[59]. Figure 4.1 displays a view of the TTR's complete developmental profile for the WSP Ctexts texts[60].

Developmental profiles are based on the abstraction of the character stream. This means that texts are perceived as one long string of characters, presumably stochastically generated by one source for each text (i.e., which vocabulary is being evaluated). The TTR values are calculated at even intervals in the sample and the subtext borders are not taken into consideration. Considering the average size of text samples, a 1,000-character interval was chosen as the interval for this study.

4.1. Complete TTR developmental profile

The curves in Figure 4.1 demonstrate that the longer WSP texts gradually converge into an asymptote, flattening out at approximately

[57] These authors offer a more advanced study (Tweedie and Baayen, "How Variable May a Constant Be?"). The present study does not investigate randomization, intervals, coherent prose, and so on.

[58] This difference could further be the foundation for evaluating text stylistic differences, while it is too early to make definite quantitative suggestions. This is why there are no quantitative indices for curve slopes in the present study, especially since it is still unclear how they could be used. Therefore, this study will only perform some visual observations.

[59] "Trimming the texts to equal size allows the number of types to be used as a direct measure of vocabulary richness and lays the groundwork for an examination of intratextual variability" (Hoover, "Another Perspective," 159).

[60] The TTR values (y-axis) are presented for every 1,000 characters of texts (x-axis).

70,000 tokens[61], meaning that there is no significant change in the rate of vocabulary growth beyond this sample size; i.e., after this point, character vocabularies are saturated. However, the curve slopes (i.e., rates of vocabulary growth) vary in the zone preceding this area.

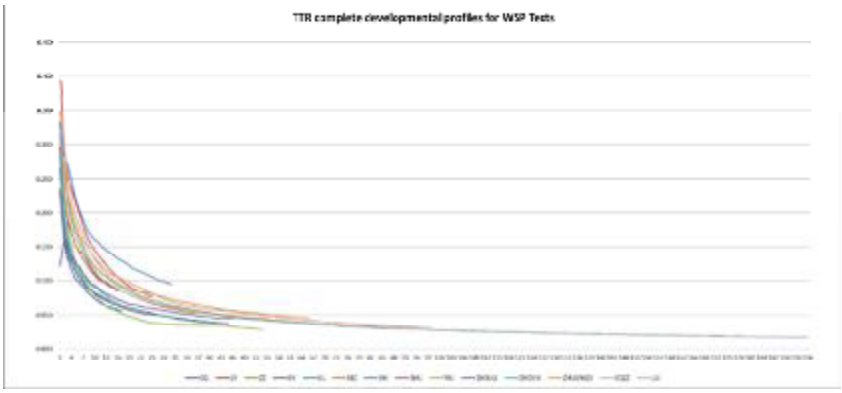

Figure 4.1. TTR complete developmental profiles for WSP Texts[62]

As David Malvern and Bryan Richards state, "[t]he more slowly a curve falls and the higher up in the space it is, the nearer that sample is to the most diverse text possible. The more quickly a curve falls, the lower in the space it lies and the nearer it is to the least diverse text there can be" (Malvern and Richards, "Measures of Lexical Richness," 3). The Shi Jing clearly remains at the top. All of this explains why the ability to select the Shi Jing and the Yi Li as two extreme points of the spectrum was earlier referred to as a "criterion for classifying indices." The indices that placed them on the top and bottom positions need to reflect the tendency.

To better understand the slopes of the curves in Figure 4.1, they can be fitted to some function. It is not exactly known what is the law behind the TTR developmental profile curves, but it is generally presumed that power law (y = ax^b) is a good approximation. The fitting of the TTR

[61] There have been several hypotheses regarding the nature of TTR curves by several researchers, from Poisson to inverse Gaussian distribution (see a review in Mitchell, "Type-token Models," 2).

[62] This chart is built based on data presented in "voc_ref.xlsx/TTR_ALL". The main table presents V(N) values for all WSP texts, take at sample sizes each thousand, and respective TTR value for this V(N). (The final values for some texts could be slightly different from exact values, as N is a multiple of 1000.) The scalable chart TTR complete developmental profiles for WSP Texts are situated right below the main table.

curves was implemented using a free, online curve-fitting package, MyCurveFit[63], to increase the reproducibility of this study. The main results (the curve images and the numerical parameters) are presented in the accompanying Excel spreadsheet[64].

The power law curves can be converted into straight lines, when the X- and Y-axes are converted to logarithmic scale. In this case, the parameter "a" becomes the Y-axis intercept and the parameter "b" becomes the line's slope. These numbers are presented on the Main Lookup sheet ("voc_ref.xlsx/MAIN_LOOKUP").

Table 4.1 presents these numbers for the corpus texts, sorted by the increasing "b" parameter. It shows that the Yi Li and the Xiao Jing demonstrate the steepest drop in the TTR curves, while the Shi Jing, the Lun Yu, and the Zhou Li demonstrate more vocabulary diversity.

Table 4.1. Power Method Curve Fitting Parameters y = ax^b (complete samples)[65]

Text	a	b
SHI	0.364123	-0.36189
LY	0.303624	-0.43227
ZL	0.268362	-0.46971
SHU	0.415781	-0.47475
MZ	0.321539	-0.47974
GY	0.239313	-0.48198
ZHZ	0.383938	-0.48299
ZY	0.282812	-0.50061
CQ	0.234874	-0.50537
GL	0.268645	-0.50955
CQZZ	0.353877	-0.51907
ZZ	0.370022	-0.51933

[63] DOI: www.mycurvefit.com. This free site will not accept sequences larger than 100 points. Thus, several larger texts were truncated to 90–100 points. In a few cases, e.g., the Zhou Li, a few points at the beginning were definitely out of range, which skewed the fitting. Therefore, a couple of points were taken out of the sequence (in the case of the Zhou Li) to better fit the other points. Otherwise, the fitting process was very basic. Besides the power law, other fitting methods were tested such as the polynomial and exponential functions. They also demonstrated good results (especially for some curves, see the spreadsheet), but the power law was the best for most of the curves.

[64] See "voc_ref.xlsx/ FIT_ALL".

[65] See "voc_ref.xlsx/MAIN_LOOKUP/ Power Method Curve Fitting Parameters".

LJ	0.413892	-0.52272
YL	0.276727	-0.55608
XJ	0.276	-0.56163

4.2. Partial TTR developmental profiles

The sample size of 70,000 tokens is the approximate area beyond which the curves of texts in the corpus visibly flatten. Meanwhile, 30,000 tokens is another interesting sample size, allowing a better scale for comparison of most texts, except for the shortest (Figure 4.2). Numerical values of 30,000 sample end point values are presented in Table 4.2.

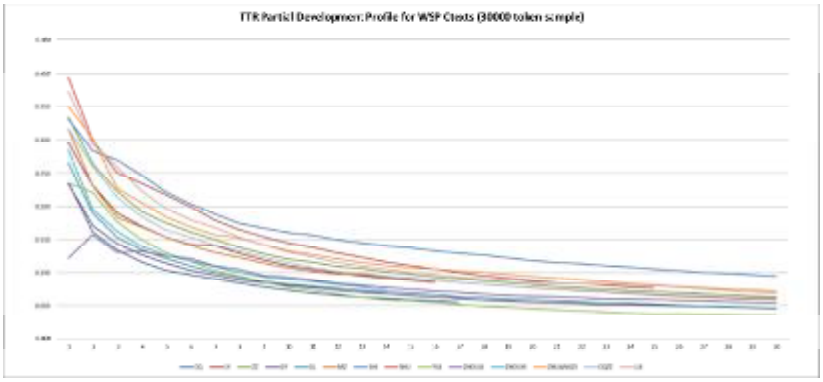

Figure 4.2. TTR profile for WSP text samples at 30000 characters[66]

Table 4.2. TTR values for WSP text samples at 30000 characters[67]

Text	N	V	V(30000)	TTR(30000)
SHI	29622	2833	2833	0.094433
ZHZ	65251	2968	2161	0.072033
LJ	97994	3041	2069	0.068967
ZZ	178563	3235	1902	0.063400
CQZZ	195354	3251	1825	0.060833
MZ	35354	1892	1768	0.058933
ZL	49410	2212	1627	0.054233
GL	40835	1594	1392	0.046400

[66] This chart is also built based on data presented in "voc_ref.xlsx/ TTR_ALL". The scalable TTR profile for WSP Text samples at 30000 characters is situated below the main table.

[67] See "voc_ref.xlsx/MAIN_LOOKUP/ TTR" values for WSP Text samples at 30000 characters.

GY	44224	1640	1381	0.046033
YL	53882	1536	1082	0.036067
CQ	16791	941	n/a	n/a
LY	15923	1361	n/a	n/a
SHU	24537	1910	n/a	n/a
XJ	1800	374	n/a	n/a
ZY	13348	1030	n/a	n/a

It is possible to group texts by curve gradients in three subgroups[68]. On the top, there is the Shi Jing (with a TTR of 0.094, which is almost three times larger than the lowest one; i.e., the Yi Li with a TTR of 0.036). Then the Shu Jing should have come next, but it is slightly shorter than 30,000 tokens in the WSP version)[69]. After that, come the Zuo Zhuan, the Chun Qiu Zuo Zhuan, the Lun Yu, and the Mengzi. Finally, there are the others, which include the Zhou Li, the Guliang Zhuan, the Gongyang Zhuan, the Xiao Jing, the Yi Li, and the Chun Qiu. This is close to what the cluster algorithm displays, except that the Shi Jing represents the top group, not the Xiao Jing.

It is also possible to group the texts as follows:

1) The Shi Jing;
2) The Shu Jing, the Zhuangzi, the Li Ji, the Zuo Zhuan, the Chun the Qiu Zuo Zhuan, the Mengzi, the Zhou Li, the Guliang Zhuan, and the Gongyang Zhuan;
3) The Yi Li.

The sample size of 15,000 tokens allows a clearer picture, including practically all of the WSP corpus texts (Figure 4.3). This profile allows the identification of a more articulate grouping:

1) The Shi Jing;
2) The Shu Jing, the Zhuangzi, the Li Ji, the Zuo Zhuan, the Chun Qiu Zuo Zhuan, the Mengzi, and the Zhou Li;
3) The Zhou Yi, the Chun Qiu, the Guliang Zhuan, and the Gongyang Zhuan;
4) The Yi Li.

[68] Even if the text samples are smaller than 30,000 token size. These sub-groups are not "clusters" since the grouping is based on simple visual sorting.

[69] However, for a continued curve, the values will be less than the Zhuangzi and the Li Ji.

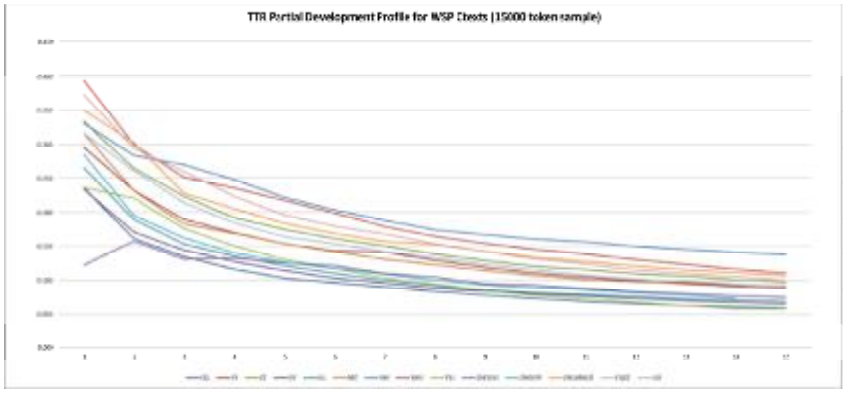

Figure 4.3. TTR profile for WSP Text samples at 15000 characters[70]

Table 4.3. TTR values for WSP Text samples at 15000 characters[71]

Text	N	V	V(15000)	TTR(15000)
SHI	29622	2833	2067	0.1378
SHU	24537	1910	1658	0.110533
ZHZ	65251	2968	1611	0.1074
LJ	97994	3041	1510	0.100667
ZZ	178563	3235	1461	0.0974
MZ	35354	1892	1375	0.091667
CQZZ	195354	3251	1364	0.090933
LY	15923	1361	1328	0.088533
ZL	49410	2212	1125	0.075
GL	40835	1594	1037	0.069133
ZY	13348	1030	1030	0.068667
GY	44224	1640	978	0.0652
CQ	16791	941	888	0.0592
YL	53882	1536	872	0.058133
XJ	1800	374	n/a	n/a

[70] This chart is also built based on data presented in "voc_ref.xlsx/ TTR_ALL". The scalable chart TTR profile for WSP Text samples at 15000 characters is situated right below the main table.

[71] See "voc_ref.xlsx/MAIN_LOOKUP/ TTR values for WSP Text samples at 15000 characters".

Table 4.4. Power Method Curve Fitting Parameters y = ax^b (samples of 15000 characters)[72]

Text	a	b
shu	0.351897	-0.32366
shi	0.351897	-0.32366
ly	0.303005	-0.42924
Zhz	0.368025	-0.43461
ZZ	0.346208	-0.443
cqzz	0.333212	-0.44924
Zl	0.256942	-0.44937
lj	0.392175	-0.45649
mz	0.316927	-0.46192
Gy	0.236047	-0.46464
Yl	0.265421	-0.48969
Zy	0.282812	-0.50061
gl	0.267284	-0.50331
cq	0.234784	-0.50475
xj	n/a	n/a

The curve fitting was also conducted for the partial sample of 15,000 characters. The results are presented in Table 4.4. This table better demonstrates that the Shu Jing, the Shi Jing, and the Lun Yu could have the most diverse vocabularies, while the Chunqiu and the Xiaojing have the least (the Yi Li is not the last text on this list).

4.3. Discussion of the results of developmental profiles

The results of the developmental profiles demonstrate that the WSP texts display varying TTR tendencies. The Shi Jing's vocabulary grows consistently stronger than any other text, i.e., this text is the most lexically diverse. The Yi Li's vocabulary demonstrates distinctively weaker growth in the WSP corpus. The texts that could be grouped as "historical" (the Zuo Zhuan, the Chun Qiu, the Guliang Zhuan, and the Gongyang Zhuan) tend to be close to one another in the middle of the spectrum (or slightly lower than the average). This is why the final value indices that placed the Shi Jing and the Yi Li at opposite ends of the spectrum as well as grouped "historical" texts together represent more interest.

Is it possible to associate these observations regarding the TTR values with stylistic or genre characteristics such as texts being historical or po-

[72] See "voc_ref.xlsx/MAIN_LOOKUP/ Power Method Curve Fitting Parameters".

etic? Is it possible to associate specific vocabulary richness values in the WSP corpus with texts being historical or poetic? Is it possible that vocabulary grows differently for historical texts ascribed to one author, while this index for a historical text by another author could be closer to a poetic text? In other words, could vocabulary richness be an index of a genre or stylistic feature unlike individual style? How do we evaluate vocabulary richness data for large heterogeneous collections like the texts in the WSP corpus?

At this point, the present author can only admit the complexity of the problem. However, the author cannot agree with the statement that "word count" is irrelevant, i.e., it cannot be used for any stylistic or authorship analysis. The developmental profiling provides information on texts' comparative vocabulary richness, singling out the Shi Jing as the richest text and the Yi Li as the simplest, while other texts took an intermediate position between them.

5. Developmental Profiles of Rare and Frequent Characters

There are two main approaches to utilize word frequencies for stylistic analysis. In one approach, infrequent words (hapax legomena, and legomena, i.e., the characters that have only two samples) are analyzed to evaluate the vocabulary richness of texts or compare texts. In the opposite approach, the most frequent words (functional, "empty" or "noncontent") are considered the key to stylistic analysis. This article will analyze the developmental profiles of hapax legomena (V1) and dis legomena (V2) as well as characters with a frequency of 50 or higher[73].

5.1. Hapax legomena (V1) and dis legomena (V2)

One of important objects of quantitative linguistics analysis is hapax legomena (singletons, unique words, V(1,N) or V1). According to the large number of rare events (LNRE) model of word frequency distributions developed by Baayen[74], they play an important role in defining vocabulary growth as well as dis legomena (V2). The TTR method can be applied to V1 and V2 to create V1 TTR[75] and V2 TTR indices. V1 TTR

[73] These are different for most texts, but they definitely include all function characters as well as some of the most frequent content words.

[74] Baayen, following Khmaladze, describes word frequency distributions as "Large Number of Rare Events (LNRE) distributions, distributions characterized by the presence of large number of words with very low probabilities of occurrence"(Baayen, "Word Frequency," 54–55), the outcome of which is that "sample relative word frequencies cannot be used to obtain the expected values of the vocabulary size" (ibid, 57).

[75] It is sometimes called "the index of diversity" (Tuldava, "Stylistics, Author Identification," 375).

curves demonstrate hapax legomena tendencies during text growth[76]. The question is "Should they follow the general TTR distribution?"

Table 5.1 contains V1 TTR numbers for 30,000 token samples. There is roughly the same order of texts as that for regular TTR values at this sample size, but there are definitely no pronounced groups. The V1 TTR chart (Figure 5.1) shows that most curves are extremely close to one another. Unlike the regular TTR chart, the V1 curves in Figure 5.1 converge very close to 30,000 tokens and they do not have considerable differences in their slopes. Yet, the V1(30,000) TTR value of the Shi Jing (0.027) is roughly three times higher than that of the Yi Li (0.009). Thus, the ratio between the extremes remains.

Table 5.1. TTR V1 values for WSP Text (samples of 30000 characters)[77]

Text	**N**	**V**	**V1(30000)**	**TTR(30000)**
SHI	29622	2833	799	0.026633
ZHZ	65251	2968	713	0.023767
LJ	97994	3041	581	0.019367
ZZ	178563	3235	556	0.018533
CQZZ	195354	3251	515	0.017167
MZ	35354	1892	487	0.016233
ZL	49410	2212	442	0.014733
GL	40835	1594	370	0.012333
GY	44224	1640	341	0.011367
YL	53882	1536	274	0.009133
CQ	16791	941	n/a	n/a
LY	15923	1361	n/a	n/a
SHU	24537	1910	n/a	n/a
XJ	1800	374	n/a	n/a
ZY	13348	1030	n/a	n/a

[76] Some interesting statistics on V1 distribution in modern Chinese web corpora is presented in Hsieh, "Why Chinese Web-as-Corpus is Wacky?".

[77] See "voc_ref.xlsx/MAIN_LOOKUP/ TTR V1 values for WSP Text (samples of 30000 characters)".

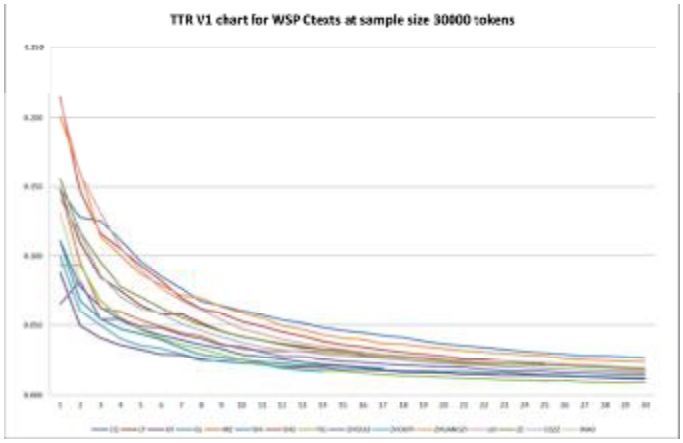

Figure 5.1. TTR chart for V1 WSP Text samples at 30000 characters[78]

Dis legomena (V2), the words that are encountered exactly twice, should be, unlike V1, "real content words." Table 5.2 presents the V2(30,000) TTR values, while Figure 5.2 displays the V2(30,000) TTR curves. However, the dynamic curves view for V2 TTR is similar to the general TTR distribution. This raises the following questions. How consistent is this growth? Is it possible that, since the Shi Jing consists of many different pieces of poetry with flowery language and rare characters, there is constant growth in hapax legomena that does not become dis legomena quickly enough (and with it, diminishing TTR)?

Table 5.2. TTR V2 values for WSP Text (samples of 30000 characters)[79]

Text	N	V	V2(30000)	TTR(30000)
shi	29622	2833	487	0.0162
ZHZ	65251	2968	361	0.0120
LJ	97994	3041	293	0.0098
zz	178563	3235	246	0.0082
cqzz	195354	3251	253	0.0084
mz	35354	1892	287	0.0096
ZL	49410	2212	225	0.0075
gl	40835	1594	196	0.0065
gy	44224	1640	222	0.0074

[78] See "voc_ref.xlsx/V1 spreadsheet".

[79] See "voc_ref.xlsx/MAIN_LOOKUP/ TTR V2 values for WSP Text (samples of 30000 characters)".

YL	53882	1536	146	0.0049
cq	16791	941	n/a	n/a
ly	15923	1361	n/a	n/a
shu	24537	1910	n/a	n/a
xj	1800	374	n/a	n/a
ZY	13348	1030	n/a	n/a

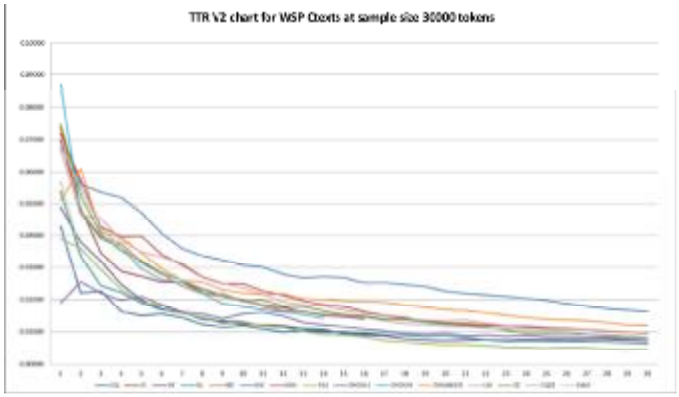

Figure 5.2. TTR chart for V2 WSP Text samples at 30000 characters[80]

The V2 TTR distribution is visually closer to the general TTR distribution. The rate of hapax legomena growth is similar for all of the texts. The dis legomena, which could be the content descriptors of the texts, behave similarly to general TTR curves.

According to Baayen, the distribution of hapax legomena is important for understanding when text vocabulary is nearing saturation or is moving from the central LNRE zone to the late LNRE zone. The central LNRE zone is the range of sample sizes "where the expected number of hapax legomena is increasing" (Baayen, "Word Frequency," 56). In the late LNRE zone, the growth of hapax legomena stops, and its curve first flattens and then decreases. The stalling of the growth of hapax legomena is only observed in three texts of the WSP corpus (Figure 5.3). First, for the Zuo Zhuan, when it reaches sample sizes of more than 60,000 (where its TTR curve becomes closer to horizontal asymptote). Second, for the Yi Li, where it occurs comparatively early, at 40,000 tokens. This fact also places the Yi Li into a category of texts with rather a poor vocabulary growth[81]. The flattening of the V1 curve for the Zuo Zhuan after 60,000,

[80] See "voc_ref.xlsx/V2 spreadsheet".
[81] Naturally, this only refers to the WSP corpus.

together with the flattening of its TTR curve, should allow one to make some projections regarding the general vocabulary size of the writers in the Warring States period.

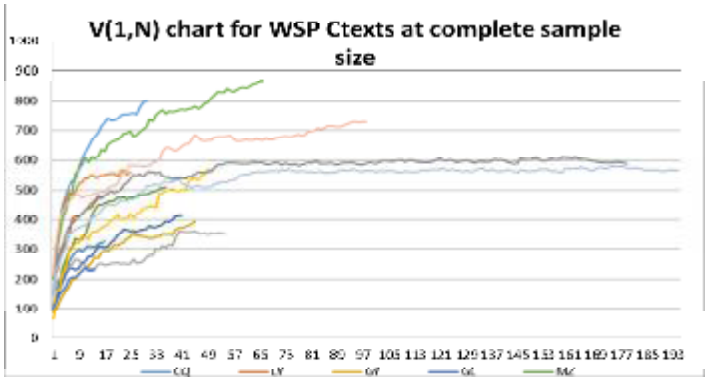

Figure 5.3. Hapax legomena for WSP corpus

5.2. Frequent words (V50+)

The final index reviewed in this study is based on the distribution of the most frequent words. What part of the corpus is covered by the most frequent words?[82] Researchers often use parameters such as "corpus coverage," "lexical coverage," or "cumulative frequency" (Da, "A Corpus-based Study of Character and Bigram Frequencies"). The studies that implement them mostly operate with frequency lists, e.g., the top 1000 characters (ordered by frequency). This approach seems to be not well adjusted in the case of the WSP corpus, which is so heterogeneous that it needs an individual list of frequent characters for each text. Moreover, such lists can be difficult to merge in order to create a single-frequency word list for the entire corpus[83].

[82] Van Hout and Vermeer describe LFP – Lexic frequency profiles, dividing words into nine groups (Hout van and Vermee, "Comparing Measures of Lexical Richness," 107).

[83] Smith and Witten recommend starting from the top 1% of the frequency list, merging the lists for the corpus texts. However, this approach was not chosen for the present study (Smith and Witten, "Language Inference from Function-Word"). As Bin Li et al. notice, of the 100 most frequent characters for their corpus, "25 characters surprisingly do not occur in all the literatures" (Li et.al., "Corpus-based Statistics," 148). Moreover, "The general characters are themselves of high frequency, but they are not necessarily distributed uniformly" and "this non-uniform distribution reflects the diversity of these literatures in domains, ages, and the writing styles" (Ibid.).

Therefore, for this study, another method was selected: the analysis of the distribution of characters that are found in each individual text fifty times or more (V50+)[84]. This study will not reject functional words since it is searching for general character frequencies not just the frequencies of content words[85].

Figure 5.4[86] presents the chart of the complete V50+ TTR distributions, while Figure 5.5[87] displays the charts for V50+ TTR at 30,000 token samples[88]. Table 5.3 presents V50+ data for the complete WSP Ctexts, including the ratios of the number of V50+ tokens to all vocabulary as well as the V50+TTR[89]. Table 5.4 presents the V50+ TTR at the sample size of 30,000 tokens[90]. Finally, Table 5.5 presents the data regarding complete text coverage by V50+ characters[91].

According to Table 5.5, V50+ characters tend to cover, on average, more than 70% of texts longer than 30,000 tokens. In addition, this is comparable to using the most frequent words in other methods[92]. This also justifies the V50+ approach.

---

[84] See "Voc_ref.xlsx/V50+CHARS spreadsheet," which lists these V50+ characters for each text in the WSP corpus, along with the number of entries and the total sum of the entries of V50+ characters for each text. These sums allow calculating text coverage.

[85] Another approach is to reject content words: "A near must of stylometric investigations is to exclude content words from the start. The reason for this is obvious: the use of content words depends on content, and the content of a text (_topic_) is, by definition, not covered by stylometry" (Golcher, "A New Text Statistical Measure," 3), although Felix Golcher himself offers a content-word ignorant method. On keeping content words and the two urns method, see Kornai, "How Many Words," 50). See also Evert on hapax legomena and dis legomena (Evert, "The Statistics of Word Co-occurrences").

[86] See "Voc_ref.xlsx/V50+ spreadsheet."

[87] See "Voc_ref.xlsx/V50+ spreadsheet."

[88] V50+ characters are not all functional words, naturally. See the list of characters for all of the texts at "Voc_ref.xlsx/V50+CHARS spreadsheet."

[89] See "Voc_ref.xlsx/MAIN_LOOKUP/TTR" for V50+ characters in the WSP CTexts (complete samples, sorted by V50+/V(N) decreasing).

[90] See "Voc_ref.xlsx/MAIN_LOOKUP/" regarding the TTR V50+ values for the WSP CTexts samples at 30,000 characters (sorted by decreasing TTR).

[91] See "Voc_ref.xlsx/MAIN_LOOKUP/" regarding the TTR V50+ for the WSP CTexts complete size (sorted by decreasing V50+ coverage).

[92] Jun Da (Da, Jun, "A Corpus-based Study," 6) reports that cumulative coverage of the top 705 characters in their study is approximately 75%.

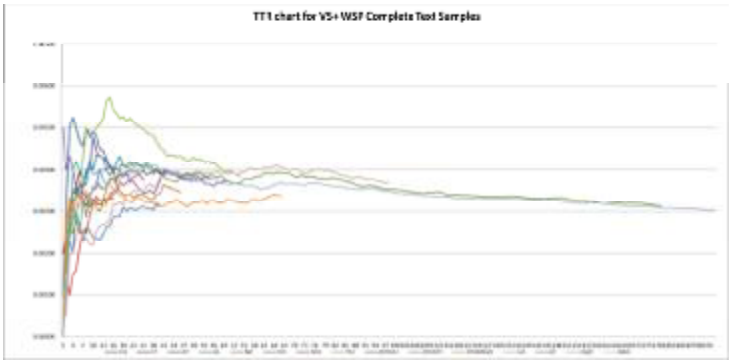

Figure 5.4. TTR chart for V50+ WSP Complete Text Samples

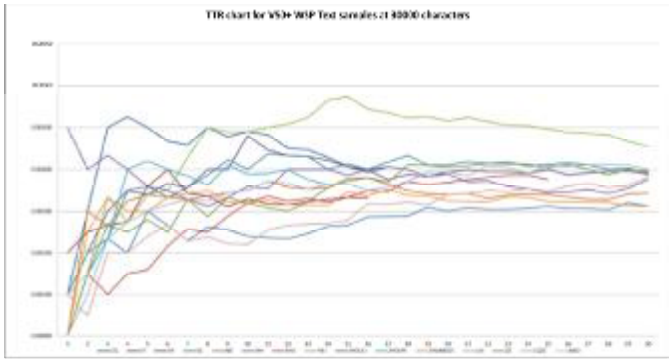

Figure 5.5. TTR chart for V50+ WSP Text samples at 30000 characters

Table 5.3. TTR of V50+ characters for WSP CTexts (complete samples, sorted by V50+/V(N) decreasing)

Text	N	V	V50+	V50+/V(N)	V50+TTR
CQZZ	195354	3251	593	0.18240541	0.0030355
ZZ	178563	3235	560	0.17310665	0.0031361
YL	53882	1536	209	0.13606771	0.0038788
LJ	97994	3041	364	0.11969747	0.0037145
GY	44224	1640	170	0.10365854	0.0038441
GL	40835	1594	155	0.09723965	0.0037958
ZL	49410	2212	187	0.08453888	0.0037847
ZHZ	65251	2968	222	0.07479784	0.0034022
CQ	16791	941	69	0.07332625	0.0041093
MZ	35354	1892	125	0.06606765	0.0035357

ZY	13348	1030	51	0.04951456	0.0038208
SHU	24537	1910	94	0.04921466	0.0038309
LY	15923	1361	53	0.03894195	0.0033285
SHI	29622	2833	94	0.03318037	0.0031733
XJ	1800	374	3	0.00802139	0.0016667

Table 5.4. TTR V50+ values for WSP CTexts complete samples at 30000 characters (sorted by TTR decreasing)

Text	**N**	**V**	**V50+ TTR**
YL	53882	136	0.004533
CQZZ	195354	120	0.004000
ZZ	178563	119	0.003967
GL	40835	118	0.003933
GY	44224	117	0.003900
LJ	97994	115	0.003833
ZL	49410	113	0.003767
MZ	35354	103	0.003433
SHI	29622	94	0.003133
ZHZ	65251	94	0.003133
SHU	24537	n/a	0
LY	15923	n/a	0
ZY	13348	n/a	0
CQ	16791	n/a	0
XJ	1800	n/a	0

Table 5.5. TTR V50+ for WSP CTexts complete size (sorted by V50+ coverage decreasing)

Text	N	V50+	V50+sum	V50+ coverage
cqzz	195354	593	170269	0.8716
zz	178563	560	153336	0.8587
YL	53882	209	43366	0.8048
LJ	97994	364	76523	0.7809
gy	44224	170	33745	0.7630
gl	40835	155	30718	0.7522
ZHZ	65251	222	47079	0.7215
cq	16791	69	11826	0.7043
ZL	49410	187	34388	0.6960
mz	35354	125	23454	0.6634
ly	15923	53	8902	0.5591
ZY	13348	51	7314	0.5479

shu	24537	94	12818	0.5224
shi	29622	94	13329	0.4500
xj	1800	3	206	0.1144

Figure 5.4, and especially Figure 5.5, demonstrate a curve order in reverse to what was observed at the regular TTR at 30,000 token samples (Figure 4.2). The Yi Li's curve comes at the top of the most frequent word curves, while the Shi Jing tends to be at the bottom. It appears that the more frequent words (V50+) there are in a text, the lower its TTR score and curve position[93].

5.3. Discussion of results for rare and the most frequent words

The data analysis for V1, V2, and V50+ demonstrates that content words (words found in texts two times or more but not too often) provide considerable input into the separation of texts into groups based on the TTR developmental profiles. Hapax legomena tend to be smoother models than dis legomena (and probably include a higher degree of words). For the most frequent words, which are responsible for most coverage of the samples, the situation is in contrast to what was observed for the TTR. The highest ratio of such words at the sample size of 30,000 is found in the Yi Li[94], while the Shi Jing has the smallest V50+ TTR index value. This could explain the positioning of their regular TTR curves.

6. Conclusions

This study examined the vocabulary richness of the WSP Ctexts corpus with the main objective of establishing the quantitative foundation for a general analysis of text vocabularies, mostly based on developmental profiles of TTR. The WSP Ctexts corpus is an open corpus of classical Chinese texts, which allows downloading and independent processing of data. All of the numerical data used in this study is available on Github. This study is an attempt to create reproducible research, and all of its components are available for independent processing.

In the first section of this study, traditional final value approaches were utilized to identify whether vocabulary richness indices (constants) could

[93] This could be a similar result to the "law of decreasing new vocabulary growth" described, e.g., by Feng Zhiwei (Feng, "Introduction of Modern Terminology," 1996) and formulated by Li and Zhang as "The repeated occurrences of high frequency words indicate a tendency of decreasing new vocabulary growth" (Li and Zhang, "Inter-textual Vocabulary," 14).

[94] According to Table 5.3, the Yi Li is close to the absolute top regarding the V50+TTR, directly behind the Chun Qiu Zuo Zhuan and the Zuo Zhuan.

supply important information regarding the stylistic groupings of texts. As previously shown by many researchers (particularly by Tweedie and Baayen), all such indices are not really "constants," but they depend on sample size. However, this approach was almost never applied to classical Chinese texts, and that is one of the reasons why this article presents these indices.

The final value indices' analysis did not analyze the dependency of indices of sample size, but it demonstrated that a majority are not very useful for stylistic grouping of texts. However, the comparison of indices was still valuable since it revealed that not all of them were directly related to sample size. Some of the indices allowed the grouping of prosaic historical texts and demonstrated the proclivity of placing the Shi Jing and the Yi Li on opposite ends of the vocabulary richness spectrum.

The first result of this study is the identification that Guiraud's R, Rubet's K, and Brunet's W match all of these criteria. Consequently, they can be considered as suitable candidates for further stylistic analysis.

In the second section, instead of the final value approach, development profiles were analyzed, mostly for the TTR[95]. The TTR developmental profiles allowed observing at what rate new words were added to the existing vocabulary. It also allowed observation of the direct change of vocabulary over comparable text sizes and comparing texts from this relative viewpoint.

The WSP Ctexts is a collection of large-sized heterogeneous texts, usually consisting of many chapters that must be considered independent texts themselves. In addition, there is no single narrative structure. Therefore, the texts were abstracted as streams of characters. The analysis of the complete length curves showed that some type of vocabulary saturation occurs around the sample size of 60,000 characters for the longest texts. Furthermore, their TTR curves approach a horizontal asymptote, and their hapax legomena numbers stop increasing. Therefore, the sample sizes of 15,000 and 30,000 tokens were chosen as cross-cut points.

The comparison of the TTR developmental curves showed that the Shi Jing and the Yi Li create upper and lower borders for other curves, serving as extremes of the spectrum of developmental curves. The texts between them can also be separated (at the 15,000 sample size) into a group of "historical texts" (the Chun Qiu, the Zuo Zhuan, the Guliang Zhuan, the Gongyang Zhuan, and the remainder).

---

[95] Similar analysis is possible for other indices, especially, Guiraud's R, Rubet's k, and Brunet's W. However, the TTR is "transparent" and it was good enough for this study.

Among the WSP corpus texts, the Shi Jing demonstrated the most diverse vocabulary with the highest rate of growth. The Yi Li included the lowest rate of inflow of new characters among the entire corpus as well as the least diverse vocabulary. At the 30,000 sample size, the Shi Jing's TTR value was practically three times larger than that of the Yi Li, i.e., its vocabulary was three times larger than that of the Yi Li. The same ratio was observed for the first two spectral elements (V1 and V2).

However, this ratio was reversed for most frequent characters, (V50+, found 50 times or more in a text). These numbers, presented in the third section, focused on hapax legomena and dis legomena influx as well as words that were most frequent. The Yi Li had the highest rate of accumulating frequent words per 1,000 tokens, while the Shi Jing had the lowest. In other words, the Yi Li utilized functional characters and high-frequency characters at a much higher degree than the Shi Jing, while the Shi Jing included a vocabulary that was more diverse. Hapax legomena in the Yi Li stopped increasing and even began decreasing from about 70% of the sample size. This signaled stagnation of vocabulary growth, while in the Shi Jing, they did not slow the rate of increase.

The second result of this study is the discovery that, in regard to vocabulary richness, the Shi Jing and the Yi Li formed two extremes of The Thirteen Classics[96]. Unfortunately, there is no clear separation (based on the TTR developmental profiles) of the texts into genre categories. The Shi Jing is outstanding as the only poetic texts. In addition, the "historical" texts such as the Chun Qiu, the Zuo Zhuan, the Guliang Zhuan, and the Gongyang Zhuan display very similar types of vocabulary growth, which makes them a special group in the corpus. However, the other texts such as the Zhuangzi, the Shu Jing, and the Mengzi also demonstrate higher vocabulary growth, but do not form a special "philosophical" group. The same is true for "ritualistic texts."

It was mentioned, a few times in this study, that a vocabulary richness analysis of large texts can be used in stylistic analysis. Is it possible to stylistically interpret the results of the analysis presented in this study? Is it possible to interpret the data regarding the growth of the Shi Jing vocabulary? Can this interpretation be performed in comparison with the Yi Li? This could possibly be evidence that the Shi Jing tends to incorporate the vocabulary of many diverse poems, while the Yi Li tends to be formulaic and utilizes many standard expressions and function characters. The present author agrees with Hoover's opinion:

[96] Without the Er Ya.

> Such measures cannot provide a consistent, reliable, or satisfactory means of identifying an author or describing a style. There is so much intratextual and intertextual variation among texts and authors that measures of vocabulary richness should be used with great caution, if at all, and should be treated only as preliminary indications of authorship, as rough suggestions about the style of a text or author, as characterizations of texts at the extremes of the range from richness to concentration. Perhaps their only significant usefulness is as an index of what texts or sections of texts may repay further analysis by more robust methods. (Hoover, "Another Perspective,"173)

Further analysis of the individual vocabularies of these subtexts is necessary. Combined with the general framework presented in this study, this analysis should clarify the lexical landscape of The Thirteen Classics. The continuation of this study will conduct a more detailed examination regarding the statistical setup of the vocabularies of the corpus.

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

Literature

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

Bremond, Claude. "The Logic of Narrative Possibilities." 1966. Trans. Elaine D. Cancalon. *New Literary History: A Journal of Theory and Interpretation* 11:3 (1980): 387–411.

Brooks, Bruce E., Brooks, Taeko A. *The emergence of China: from Confucius to the empire.* Amherst: *Ancient China in Context*, University of Massachusetts at Amherst, 2015.

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, Tianwei Xie and Juan Xu (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.

Daller, Michael H. "Guirauds index of lexical richness". In: British Association of Applied Linguistics, September 2010 (DOI: http://eprints.uwe.ac.uk/11902/).

Durán, Pilar, D. Malvern, B. Richards and N. Chipere. "Developmental trends in lexical diversity." *Applied Linguistics* 25:2 (2004): 220–242.

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

Feng, Zhiwei. *Introduction of Modern Terminology.* Beijing: The Language Publishing House, 1996.

Feng, Zhiwei. "Evolution and present situation of corpus research in China." *International Journal of Corpus Linguistics* 11:2 (2006), 173–207.

Golcher, Felix. "A New Text Statistical Measure and its Application to Stylometry." In *Proc. of the Corpus Linguistics conference (CL'07)*, Article 71, 2007.

Herdan, Gustav. *Type-token mathematics; a textbook of mathematical linguistics.* 'S-Gravenhage Mouton: *Janua linguarum, studia memoriae Nicolai van Wijk dedicate* No. 4, 1960.

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

Hoover, David L. "Another Perspective on Vocabulary Richness." *Computers and the Humanities* 37 (2003) 151–178.

Hout, Roeland van and Anne Vermee. "Comparing measures of lexical richness" In: Eds. Helmut Daller, James Milton and Jeanine Treffers-Daller , *Modelling and Assessing Vocabulary Knowledge*. Cambridge, UK: Cambridge University Press, Ch.5, 93–115.

Hsieh, Shu-Kai. "Why Chinese Web-as-Corpus is Wacky? Or: How Big Data is Killing Chinese Corpus Linguistics." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), eds. Nicoletta Calzolari (Conference Chair) ,Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis, 2014, May*, 26–31, Reykjavik, European Language Resources Association (ELRA).

Kornai , András. "How many words are there?" *Glottometrics 4 (2002), 61–86.*

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

Kytö, Merja and Anke Lüdeling (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.

Laufer, Batia and Paul Nation. "A vocabulary-size test of controlled productive ability." *Language Testing* 16:1 (1999), 33–51.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming and F. J. Smith. "Extension of Zipf's Law to Word and Character N-grams for English and Chinese". *Computational Linguistics and Chinese Language Processing* 8:1 (2003), 77–102.

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

Li, J. and F. Zhang. *Inter-textual vocabulary growth patterns for marine engineering English.* Bejing: Editorial office for contemporary foreign languages, 2011.

Li Hongzao's (李鸿藻). *Hanyuan Shisanjing ji zi* 翰苑十三經集字 (A collection of characters in the Thirteen Classics compiled by the National Academy), 1889.

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

Malvern, David D., Ngoni Chipere, Brian J. Richards and Pilar Durán. *Lexical Diversity and Language Development.* Houndmills, Basingstoke, Hampshire, New York: Palgrave Macmillan, 2004.

Malvern, David D. and Bryan Richards. "Measures of Lexical Richness." *In: The Encyclopedia of Applied Linguistics*. Ed. by Carol A. Chapelle. Blackwell Publishing Ltd., 2013.

McCarthy, Philip M. and Scott Jarvis Mild. "vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment." *Behavior Research Methods*, 42:2 (2010), 381–392.

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

Meyer, Dirk. *Philosophy on Bamboo: Text and the Production of Meaning in Early China*. Leiden: *HCT 2*, Brill, 2012.

Michell, Colin Simon. *Investigating the Use of Forensic Stylistic and Stylometric Techniques in the Analysis of Authorship on a Publicly Accessible Social Networking Site (Facebook)*. MA Thesis 2013.

Mitchell,David. "Type-token models: a comparative study." *Journal of Quantitative Linguistics*, 22:1 (2015), 1–21. —

Naranan, S. and V.K. Balasubrahmanyan. "Models for Power Law Relations in Linguistics and Information Science." *Journal of Quantitative Linguistics* 5:1–2 (1998), 35–61.

Nelson, Robert. "Issues with the capture-recapture measure of vocabulary size." *The Mental Lexicon* 10:1 (2015), 168–179.

Oakes, Michael Philip**.** "Corpus Linguistics and Stylometry". In Eds. A LÃ¼deling and M. KytÃ.. *Corpus Linguistics: An International Handbook*. Mouton de Gruyter, 2009: 1070–1090.

Peng, Fuchun, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. "Language independent authorship attribution using character level language models".

In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, 1 (2003), 267–274.

Piantadosi, Steven T. "Zipf's word frequency law in natural language: A critical review and future directions." *Psychonomic Bulletin & Review,* 21 (2014), 1112–1130.

Popescu, Ioan-Iovitz. *Word Frequency Studies,* Quantitative linguistics 64, Walter de Gruyter, 2009.

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

Qiu Xigui 裘锡圭. *Chinese Writing*. Translated by Gilbert L. Mattos and Jerry Norman. Early China Special Monograph Series No. 4. Berkeley: The Society for the Study of Early China and The Institute of East Asian Studies, University of California, 2000.

Read, John. *Assessing vocabulary*. Cambridge, England: Cambridge University Press, 2000.

Sampson, Geoffrey. "Review of Harald Baayen: Word Frequency Distributions." *Computational Linguistics* 28 (2002): 565–569.

Smith, Tony C. and Ian H.Witten. "Language Inference from Function-Word". Working Paper 93/3, University of Waikato, New Zealand, 1993.

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

Tuldava, Juhan. "Stylistics, author identification." Gabriel Altmann and Rajmund G. Piotrowski (Eds.) *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook.* Walter de Gruyter, Berlin, New York, 2005, 368–387.

Vulanovic , Relja and Köhler, Reinhard. "Syntactic units and structures". In Köhler, Reinhard , Gabriel Altmann and Rajmund G. Piotrowski (Eds.) *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook.* Walter de Gruyter, Berlin, New York, 2005, 274–291.

Wang Dahui, Menghui Li and Zengru Di. "True reason for Zipf's law in language." *Physica A* 358 (2005): 545–550.

Wimmer, Gejza. "The Type-Token relation." In Köhler, Reinhard , Gabriel Altmann and Rajmund G. Piotrowski (Eds.) *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook.* Walter de Gruyter, Berlin, New York, 2005, 361–368.

Wimmer, Gejza and Gabriel Altmann. "On Vocabulary Richness." *Journal of Quantitative Linguistics* 6:1 (1999), 1–9.

Xiao, Hang. "On the Applicability of Zipf's Law in Chinese Word Frequency Distribution." *Journal of Chinese Language and Computing* 18:1 (2008). 33–46.

Yang, Yuting, Yunhua Qu, Chenyao Bao and Xiaowen Zhang. "A Model-based Feature Optimization Approach to Chinese Language." *Processing Journal of Quantitative Linguistics,* 22, No. 1 (2015): 55–81.

Zhang, Dongbo and Shouhui Zhao. "The Totality of Chinese Characters — A Digital Perspective." *Journal of Chinese Language and Computing* 17:2 (2007), 107–125.

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

S.V. Zinin*

Vocabulary richness of early Chinese texts: macroanalysis of the Thirteen classics and the Zhuangzi

ABSTRACT: This study analyzes statistical data regarding the vocabulary richness of the Warring States Project CTexts collection of Chinese classics[97]. Vocabulary richness has been primarily used in quantitative linguistics for authorship identification and style analysis, and it has been increasingly applied for various aspects such as language acquisition in other linguistic fields. This study lays the foundation for a quantitative linguistic analysis of the vocabulary of early Chinese texts. It also conducts a macroanalysis of the data, including calculating several vocabulary richness indices and building charts of vocabulary growth. This study finds significant differences in the vocabulary growth of corpus texts. In addition, it reveals that the Shi Jing and Yi Li are two extreme ends of the vocabulary growth spectrum and identifies some historical texts in the middle of the spectrum as a distinct group. Furthermore, the study takes a closer look at specific forms of vocabulary growth such as hapax legomena, dis legomena, and the most frequent characters.

KEYWORDS: Chinese canons, The Thirteen Classics, computational linguistics, quantitative linguistics, vocabulary richness, lexical diversity, type-token ratio, digital corpora, stylometry.

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

[97] It contains The Thirteen Classics (excluding the Er Ya) and adds the Zhuangzi to balance the "Confucian" texts.