ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 10, No. 3, pp. 509-515, May 2019
DOI: http://dx.doi.org/10.17507/jltr.1003.14

A New Paradigm for the Etymology and Trend Study from the Perspective of Culturomics*

Fei Song
Beijing International Studies University, China

Minghui Xu
Qingdao Experimental School, China

***Abstract*—As an emerging discipline in 2011, culturomics belongs to cultural studies mainly by way of diachronic research and large-scale corpora, so it could be significative for the etymology and trend studies. In this paper, a large-scale diachronic corpus was established based on culturomics, and the condition and quality basic-level category vocabulary (BLCV) is taken as examples for analyses, so as to obtain relevant data and conclusions. It is shown that about 90% of the condition and quality BLCV was originated earlier than Sui, Tang and the Five Dynasties, and the first-level BLCV is earlier than the second-level and the third-level in aspect of origin time. The higher BLCV level, the greater the use rate increase, with more obvious stable development in the diachronic respect. Thus, culturomics, as a new paradigm in linguistic researches, is confirmed to be feasible, distinctive and irreplaceable in an era of big data.**

***Index Terms*—culturomics, etymology, trend study**

I. Introduction

Chinese lexicon has undergone great changes from ancient times to today. Ancient monosyllables are usually used as morphemes to form numerous disyllables, and some of them never cease to broaden, narrow and transfer their meaning, which gradually generate the modern Chinese lexicon system. In this process, some of vocabularies are still popular and significant in Chinese; some are active and invigorated by forming other new vocabulary; and some are ceased and left in history together with ancient Chinese language.

Previously, researches on Chinese lexicon’s diachronic development are inseparable from the classification of diachronic development law in this system. However, emerging "culturomics" provides a revolutionary means of obtaining such data.

Culturomics is an emerging inter-discipline that studies human’s specific ethnic behavior and cultural trends through the quantitative analysis of digitized texts based on big data corpus. This term, first described in a Science article named Quantitative Analysis of Culture Using Millions of Digitized Books by Harvard researchers Jean-Baptiste Michel and his team (2011), is to make up for the limitations and subjectivity of traditional humanities research due to the limited reading and knowledge of researchers themselves.

They cooperate with the Google Books, and based on the digital library of 5,195,769 books with a total of 500 billion vocabularies between 1500 and 2008, conduct quantitative analyses for a range of cultural patterns in language use and trends over time using Google Ngram. Then, Google further develops and opens Google N-gram Viewer, which can directly visualize time-frequency maps of keywords frequencies according to different languages and time scales.

Its arising also attracts other researchers to studying different culture from this perspective, and some significant achievements are also made. In SCOPUS, EBSCO and Whiley Inter Science, 52 academic theses, covering such topics as historical phenomena (Juola, 2013), and writing style (Hughes, 2012) are searched with “Culturomics” as the keywords.

But in China, a few relevant researches are conducted at present, and only 3 theses are searched in CNKI, Wanfang data and VIP database. Shao Peiren and Lin Qun (2012) used culturomics to extract Chinese culture and model its characteristics, and proposed that construction of Chinese culture gene base by culturomics could be a way to protect, inherit and disseminate Chinese culture. Guo Chonghui, Wei, and Ren Xiaoling (2014) reviewed the culturomics-related research at home and abroad, and put forward a model of cloud-based processing platform for digitalized documents. Zhao Haiying, Jia Gengyun, and Pan Zhigeng (2016) suggested that culturomics theory can be applied in cultural computing to study the cultural evolution and development. Nevertheless, there are no researches on applying culturomics to specific Chinese cultural issues.

Chinese BLCV has close and inseparable links with its lexicon in the diachronic development since ancient times. One reason is that many Chinese BLCV are themselves the "living fossils" of Chinese passing down; and the other is

* This project is supported by Beijing Social Science Foundation (Project No.: 16YYC028)

that “productivity” of many today’s BLCV is formed when ancient Chinese monosyllables are evolved to the modern disyllables. Thus, it has vital significance to study Chinese BLCV etymology for understanding Chinese etymology and the trend. On the other hand, culturomics possesses unique advantage in such researches. Based on the diachronic corpus, BLCV generation and development can be studied. According to the generation and length of time, as well as the specific usage conditions in each era, the diachronic literacy and development laws of BLCV can be described and summarized. This complements the traditional paradigm of the etymology research, which usually involves a large amount of literature research and glyph analysis.

In this paper, Chinese condition and quality BLCV[1] (seen in the Appendix) extracted previously is taken as examples to analyze the feasibility and specific technical details of culturomics into the Chinese etymology and its trend study, so as provide a new etymology paradigm.

II. Establishment of Diachronic Pragmatic Corpus for Condition and Quality BLCV

According to culturomics, a larger-scale Chinese diachronic pragmatic corpus should be established first. Hence, “Online Corpus Website[2]”, is selected as the raw corpus of ancient Chinese in this research. All the data are crawled in a category of condition and quality vocabulary and then undergone a secondary processing, Finally, an expected diachronic corpus with the total volume of 100 million condition and quality Chinese vocabularies is established.

This corpus contains about 100 million condition and quality Chinese vocabularies from Zhou Dynasty to Republic of China, with most Chinese ancient classics compiled in *The Four Branches of Literature* included, for example, *The Book of Songs*, *The Book of History*, *The Book of Changes*, *The Book of Tao and Teh*, *The Analects*, *The Works of Mencius*, *Zuo's Commentary on the Spring and Autumn Annals*, *The Verse of Chu*, *The Book of Rites*, *The Great Learning*, *The Doctrine of the Mean*, *The Spring and Autumn Annals*, *Erya*, *Hung Lieh Chuan*, *Records of the Grand Historian*, *Intrigues of the Warring States*, *Records of the Three Kingdoms*, *A New Account of the Tales of the World*, *The Literary Mind and the Carving of Dragons*, *Complete Tang Poems*, *Zhuziyulei*, *Creation of the Gods*, *The Romance of the Three Kingdoms*, *Heroes of the Marshes*, *Journey to the West*, *A Dream in Red Mansions* and *The Scholars*.

The details of the corpus establishment are shown as below:

A. Data Crawling in a Category of Condition and Quality Vocabulary

Keywords (Chinese condition and quality vocabulary) are input in the website www.cncorpus.org, and the searched data are saved as text files. In the end, 8869 text files are obtained totally, and each file contains the pragmatic examples of one condition and quality vocabulary from all the dynasties. Taking “纤纤” as an example, raw data crawled online are shown as below (in the original format with the first 20 example sentences, and the rest represented by “…”):

“Corpus searching _ Corpus online” (www.cncorpus.org)
Type: ancient Chinese corpus
Searching vocabulary: 纤纤
2014-6-8 15:20:45
No. Sentence Book name Dynasty
1 夫绵绵不绝，必有乱结； [纤纤] 不伐，必成妖孽 the Art of War _ Spring and Autumn and Warring States Period
2 娥娥红粉妆， [纤纤] 出素手 *N*ineteen Ancient Poems _ Wei, Jin and Six dynasties
3 云中的明月；为什么，她红妆艳服，打扮得如此用心；为什么，她牙雕般的 [纤纤] 双 Nineteen Ancient Poems _ Wei, Jin and Six dynasties
4 娥娥与 [纤纤] 同是 Nineteen Ancient Poems _ Wei, Jin and Six dynasties
5 写其容色，而娥娥是大体的赞美， [纤纤] 是细部的刻划，如互易，又必 格不顺 Nineteen Ancient Poems _ Wei, Jin and Six dynasties
……”

It can be shown that the data are basically structuralized, and the searched vocabularies are obviously marked and separated, which is convenient for format adjustment and storage.

B. Structure Adjustment and Storage of the Data

Structure adjustment include two aspects: The structure contains such several fields as "ID, Word, Pragmatic Text, Source and Dynasty"; the searched condition and quality vocabulary can be extracted from the crawled corpus and classified into the "Word" field.

Those condition and quality vocabulary can be thus extracted by nesting the mid function of sql statement into instr function. First, in the "Pragmatic Text" field, the searched vocabularies after "[" are extracted and saved as a temporary field A; second, the text before "]" in the temporary field A are extracted; and last, the extracted vocabulary is saved in the field "Word".

[1] Song Fei. Construction of basic-level condition and quality category vocabulary lexicon in international Chinese Teaching [D]. Minzu University of China. 2015

[2] website: http://www.cncorpus.org/index.aspx

After data storage, the field “ID” is for the serial number of the pragmatic text; "Word" for the condition and quality vocabulary; "Pragmatic Text" for the specific context in which the searched vocabulary appears; "Source" for the text source; and "Dynasty" for the dynasty in a diachronic order.

The total volume of data storage reaches 9,368,062, and one pragmatic text represents one piece of data, containing one condition and quality vocabulary.

III. Extraction and Analysis of Diachronic Pragmatic Data for BLCV

According to the age labels, the chronological distribution of the corpus is divided into nine periods: Zhou Dynasty; Spring and Autumn and the Warring States; Han Dynasty, Wei, Jin and Six dynasties; Sui, Tang and Five dynasties; Song Dynasty; Yuan and Ming dynasties; Qing Dynasty; and the early Republic of China. Based on the statistics, overall distribution of the condition and quality vocabulary in the corpus is obtained, of which the diachronic order and usage trend are analyzed in this study.

A. Analysis of the Diachronic Order

The dynasties that 312 Chinese BLCV[3] first appear in the corpus are shown as below (excerpt):

Table. I
Dynasties of BLCV for the first appearance in the corpus (excerpt)

ID	Vocabulary	Dynasty	Level
1	大	Zhou Dynasty	First level
2	多	Zhou Dynasty	First level
3	好	Zhou Dynasty	First level
4	新	Zhou Dynasty	First level
5	长	Zhou Dynasty	First level
6	快	Zhou Dynasty	First level
7	近	Zhou Dynasty	First level
8	深	Zhou Dynasty	First level
9	高	Zhou Dynasty	First level
10	热	Zhou Dynasty	First level
…	…	…	…
31	乱	Zhou Dynasty	Second level
32	薄	Zhou Dynasty	Second level
33	严	Zhou Dynasty	Second level
34	旧	Zhou Dynasty	Second level
35	满	Zhou Dynasty	Second level
36	白	Zhou Dynasty	Second level
37	粗	Spring and Autumn and the Warring States	Second level
38	假	Zhou Dynasty	Second level
39	低	Zhou Dynasty	Second level
40	远	Zhou Dynasty	Second level
…	…	…	…
197	灵活	Yuan and Ming dynasties	Third level
198	真诚	汉 Han Dynasty	Third level
199	清晰	Wei, Jin and Six dynasties	Third level
200	艳	Spring and Autumn and the Warring States	Third level
201	丑	Zhou Dynasty	Third level
202	臭	Zhou Dynasty	Third level
203	国产	Han Dynasty	Third level
204	便宜	Spring and Autumn and the Warring States	Third level
205	懒	Wei, Jin and Six dynasties	Third level
206	崇高	Zhou Dynasty	Third level
…	…	…	…

As is shown in the table, BLCV is classified into three levels, with the first ten pieces of BLCV data excerpted from each level. The statistics suggest that ten BLCV of the first level start to appear in Zhou Dynasty. As for the ten BLCV of the second level, one first appears in the period of Spring and Autumn and the Warring States and the rest nine in Zhou Dynasty; And as for the ten of the third level, many vocabularies start to appear in later eras such as Spring and Autumn and the Warring States; Han Dynasty; Wei, Jin and Six dynasties; as well as Yuan and Ming dynasties.

The overall statistics of dynasties that BLCV in each level first appear are shown as below:

[3] Song Fei. Construction of basic-level condition and quality category vocabulary lexicon in international Chinese Teaching[D]. Minzu University of China. 2015

TABLE. II
STATISTICS OF DYNASTIES FOR THE FIRST APPEARANCE OF BLCV IN EACH LEVEL

	First level	Percentage	Second level	Percentage	Third level	Percentage
Zhou Dynasty	27	90.00%	97	58.43%	22	18.97%
Spring and Autumn and the Warring States	1	3.33%	31	18.67%	41	35.34%
Han Dynasty,	0	0.00%	12	7.23%	11	9.48%
Wei, Jin and Six dynasties	1	3.33%	9	5.42%	18	15.52%
Sui, Tang and Five dynasties	1	3.33%	7	4.22%	9	7.76%
Song Dynasty	0	0.00%	4	2.41%	4	3.45%
Yuan and Ming dynasties	0	0.00%	2	1.20%	5	4.31%
Qing Dynasty	0	0.00%	3	1.81%	1	0.86%
the early Republic of China	0	0.00%	0	0.00%	2	1.72%
(not appearing)	0	0.00%	1	0.60%	3	2.59%

According to the overall statistics, except these three Chinese vocabularies “稳” (Spring and Autumn and the Warring States), “重要”(Wei, Jin and Six dynasties) and “硬” (Sui, Tang and Five dynasties) among 30 BLCV of the first level, the rest 27 first appear in Zhou Dynasty, accounting for 90% of the first level. Among 166 BLCV of the second level, 97 vocabularies first appear in Zhou Dynasty, accounting for 56.43% with a little bit decrease compared with the first level, and besides, one does not appear in the corpus. As for the third level, this proportion only accounts for 18.97%, showing a further decrease. In addition, 41 vocabularies, being the most in this level, start to appear in the period of Spring and Autumn and the Warring States, accounting for 35.34%. BLCV of this level also is the only one that new vocabulary appears in each historic period.

Thus, the diachronic order of those BLCV appearance exhibits a trend that the first level is earlier than the second and the third level, and is highly consistent with the hierarchic BLCV. All of those indicate that these extracted and hierarchic BLCV not only possess the aforesaid features, but also have earlier etymologies.

B. Analysis of Usage Trend

The so-called usage trend refers to the fluctuation of every condition and quality BLCV in different dynasties. Because of the unbalanced lexicon vocabulary in every era as well as long time-span in the diachronic corpus, with great changes of the vocabulary and grammar, it is difficult to find a general method and tool for word segmentation, and thus, the frequency used in the modern Chinese corpus cannot reflect the frequency of the text per unit. In this research, the ratio (also called "frequency" for the convenience) of the number of texts with condition and quality BLCV to the total number of texts with condition and quality vocabulary is employed to make it comparable among BLCV usage in every era. According to statistics, the total number of texts with condition and quality vocabulary in every era is shown below:

TABLE. III
THE NUMBER OF TEXTS WITH CONDITION AND QUALITY VOCABULARY IN EVERY ERA

Era	*Text number*
Zhou Dynasty	53834
Spring and Autumn and the Warring States	341009
Han Dynasty	772296
Wei, Jin and Six dynasties	1350315
Sui, Tang and Five dynasties	3398536
Song Dynasty	3614784
Yuan and Ming dynasties	2189302
Qing Dynasty	1608064
the early Republic of China	1077925

The means of “adding the difference between neighboring frequencies”, which was used for the usage trend analysis of the BLCV acquisition order previously, is still applicable here. Namely, the frequency of one BLCV in an era subtracts from the last era, and then all the resulting differences are added together. A positive sum displays an overall increase trend of usage for that BLCV over times, or otherwise. The sums calculated are shown below (excerpt):

TABLE. IV
SUM OF THE DIFFERENCE BETWEEN THE NEIGHBORING FREQUENCIES

ID	*Vocabulary*	*Sum of the neighboring difference*	*Level*
1	大	-0.0129423664	First level
2	多	0.0016954401	First level
3	好	-0.0007132258	First level
4	新	0.0032481051	First level
5	长	-0.0005766758	First level
6	快	0.0008652944	First level
7	近	0.0018971524	First level
8	深	0.0023264076	First level
9	高	0.0034097778	First level
10	热	0.0006085338	First level
…	…	…	…
32	薄	-0.0005755121	Second level
33	严	0.0033981364	Second level
34	旧	0.0042978418	Second level
35	满	0.0075252267	Second level
36	白	0.0014476421	Second level
37	粗	-0.0000914321	Second level
38	假	-0.0003311653	Second level
39	低	0.0001335042	Second level
40	远	0.0005383437	Second level
41	正	-0.0013260831	Second level
…	…	…	…
197	灵活	0.0000000000	Third level
198	真诚	-0.0000003671	Third level
199	清晰	0.0000038980	Third level
200	艳	0.0001302411	Third level
201	丑	0.0026250550	Third level
202	臭	-0.0004226437	Third level
203	国产	-0.0000018357	Third level
204	便宜	0.0000538387	Third level
205	懒	0.0000011228	Third level
206	崇高	-0.0000343681	Third level

The above table shows the results of adding the neighboring difference among condition and quality BLCV of these three levels. It can be found that, different from the acquisition order, an increasingly obvious uptrend of diachronic usage for these BLCV from the third level to the first level in the process of emergence and development. For example, the first ten BLCV of the third level has more negative sums than the other two levels, so the total value is correspondingly smaller; then is the second level; and the biggest is the first level.

From the perspective of acquisition order, the higher the BLCV level, the lower its use rate with the improvement of learners' language proficiency. However, in respect of diachronic etymology and the development, the higher the level, the higher the use rate increase over time. Hence, BLCV level could be linked to its diachronic stableness, with the higher the level, the greater the diachronic stableness.

IV. CONCLUSION

Data obtained by means of culturomics suggest that about 90% of BLCV appears earlier than Sui, Tang and Five dynasties, and is featured with an earlier origin in respective of condition and quality BLCV etymology. In addition, it has a ladder pattern for the emerging times with the earliest being the first level BLCV and the latest being the third level.

During the emergence and development of condition and quality BLCV, the biggest use rate increase is the first level of the BLCV, then the second level, and the smallest is the third level. Hence, in respect of diachronic development of lexicon, the higher the BLCV level, the higher the use rate increase over time, and the higher BLCV level, the greater the diachronic stableness.

It is also the first try to apply culturomics into linguistic researches and thus make some valuable linguistic conclusions, with those distinctive vocabulary-related data that is hard to acquire in other paradigms. By this way, culturomics is confirmed to be feasible, distinctive and irreplaceable in linguistic research and provides a new paradigm for relevant studies.

REFERENCES

[1] Guo Chonghui, Wei Wei and Ren Xiaoling. (2014). A Review on Culturomics. *Journal of the China Society for Scientific and*

Technical Information, 7, 765-774.

[2] Hughes J M,Foti N J,Krakauer D C,et a1. (2012). Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *Proceedings of the National Academy of Sciences*, 109(20), 7682-7686.

[3] Juola P. (2013) Using the Google N-gram Corpus to Measure Cultural Complexity. *Literary and Linguistic Computing*, 4, 1-8.

[4] Kumar N,Sahu M.(2011) The Evolution of Marketing History：a Peek Through Google N-gram Viewer. *Asian Journal of Management Research,* 2, 415-426.

[5] Michel J B,Aiden E L,Shen Y K,et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), 176-182.

[6] Song Fei. (2011). The method of extracting and classifying vocabulary in basic category in international teaching of Chinese language and its future application. *Chinese Language Globalization Studies, 2,* 171-184.

[7] Song Fei. (2011). Research on basic-level category vocabulary of modern Chinese. Minzu University of China.

[8] Song Fei. (2014). Research on large-scale corpus-based relative frequency location method for modern Chinese basic-level vocabulary. *Applied linguistics*, 4, 78-84.

[9] Song Fei. (2015). Construction of basic-level condition and quality category vocabulary lexicon in international Chinese Teaching. Minzu University of China.

[10] Shao Peiren, Lin Qun. (2012). Exploration of Extracting Chinese Cultural Genes and Modeling Its Characteristics. *Journal of Xuzhou Normal University (Philosophy and Social Sciences Edition)*, 2, 107-111.

[11] Zhao Haiying, Jia Gengyun and Pan zhigeng. (2016). Review on the Methods and Applications in Cultural Computing. *Computer Systems & Applications*, 6, 1-8.

APPENDIX

TABLE I
HIERARCHY CORPUS OF BLCV

First level	1 大	7 近	13 坏	19 少	25 难
	2 多	8 深	14 静	20 真	26 重要
	3 好	9 高	15 早	21 老	27 美
	4 新	10 热	16 稳	22 细	28 短
	5 长	11 小	17 轻	23 晚	29 重
	6 快	12 强	18 富	24 硬	30 冷
Second level	1 乱	35 黄	69 及时	103 软	137 轻松
	2 薄	36 中	70 公	104 残	138 松
	3 严	37 久	71 弱	105 紧张	139 生动
	4 旧	38 充分	72 行	106 高级	140 详细
	5 满	39 光	73 突出	107 净	141 荒
	6 白	40 努力	74 暗	108 穷	142 浑
	7 粗	41 严重	75 合理	109 慢	143 危险
	8 假	42 古	76 直	110 复杂	144 紫
	9 低	43 有些	77 友好	111 虚	145 瘦
	10 远	44 一切	78 香	112 女	146 浅
	11 正	45 安全	79 厚	113 暖	147 湿
	12 生	46 实	80 野	114 彩	148 母
	13 紧	47 死	81 全面	115 淡	149 旱
	14 空	48 先进	82 阴	116 狂	150 甜
	15 积极	49 一般	83 丰富	117 清	151 胖
	16 特别	50 精	84 公开	118 灵	152 傻
	17 青	51 自然	85 切实	119 挺	153 灰
	18 红	52 健康	86 宽	120 曲	154 随便
	19 忙	53 熟	87 年轻	121 圆	155 俗
	20 外	54 认真	88 伟大	122 凉	156 格外
	21 欢	55 具体	89 正常	123 成熟	157 平静
	22 余	56 便	90 纯	124 公正	158 外来
	23 原	57 正式	91 怪	125 蓝	159 聪明
	24 密	58 只	92 强烈	126 牢	160 脏
	25 亲	59 副	93 平	127 尖	161 热闹
	26 均	60 错	94 齐	128 烂	162 光荣
	27 易	61 明显	95 方	129 幸福	163 难得
	28 广	62 贵	96 必要	130 热情	164 扎实
	29 对	63 响	97 公共	131 发达	165 稀
	30 全	64 民主	98 文明	132 惨	166 严肃
	31 男	65 亮	99 零	133 美好	
	32 苦	66 偏	100 简单	134 盲目	
	33 黑	67 恶	101 坚决	135 脆	
	34 华	68 绿	102 破	136 适当	
Third level	1 灵活	25 腐败	49 廉洁	73 高贵	97 委婉
	2 真诚	26 勇敢	50 大方	74 单调	98 断断续续

	3 清晰	27 模糊	51 传统	75 悲观	99 专制
	4 艳	28 业余	52 片面	76 酥	100 没出息
	5 丑	29 酸	53 皱	77 粗暴	101 迟钝
	6 臭	30 谨慎	54 抽象	78 威风	102 矜持
	7 国产	31 节约	55 咸	79 凹	103 幼小
	8 便宜	32 温柔	56 平和	80 谦虚	104 冒失
	9 懒	33 消极	57 好听	81 自私	105 世故
	10 崇高	34 秘密	58 顽固	82 稠	106 小气
	11 人为	35 俊	59 朦胧	83 不妥	107 下贱
	12 坚强	36 笨	60 时髦	84 恭敬	108 无礼
	13 匆匆	37 乖	61 简陋	85 奢侈	109 痴情
	14 冷静	38 涩	62 含糊	86 有为	110 轻薄
	15 老	39 反动	63 贪	87 无能	111 褐
	16 宏伟	40 不平	64 朴实	88 倔强	112 轻浮
	17 乐观	41 腥	65 有用	89 狡猾	113 不和
	18 辣	42 鼓	66 空虚	90 不起眼	114 没用
	19 有趣	43 钝	67 无知	91 胆小	115 博学
	20 持久	44 孤独	68 无私	92 可耻	116 本分
	21 窄	45 短暂	69 冷淡	93 生硬	
	22 繁荣	46 无情	70 糙	94 无理	
	23 客气	47 天真	71 过时	95 中型	
	24 初级	48 骄傲	72 深沉	96 难听	

Fei Song was born in Linyi, China in 1986. He received his PH.D. degree in linguistics from Minzu University of China in 2015.

He is currently an associate professor of Beijing International Studies University. He focuses on Chinese language processing and international Chinese teaching.

Minghui Xu was born in Hebei, China in 1986. He received Master degree of Teaching Chinese to Speakers of Other Languages from Beijing International Studies University in 2018.

He is a Chinese language teacher of Qingdao Experimental School, Qingdao, China.