ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 9, No. 4, pp. 405-410, April 2019
DOI: http://dx.doi.org/10.17507/tpls.0904.07

Corpus-based Research to Verify the Hypothesis of Preference for Basic-level Category Vocabulary (BLCV) Acquisition*

Fei Song
Beijing International Studies University, Beijing, China

Qingqing Lan
Marlboro High School, NJ, US

***Abstract*—Many features of basic-level category vocabulary (BLCV) play an important role in international Chinese language teaching, and one of them is preference for BLCV acquisition which has yet remained to be confirmed. In this paper, a written language corpus of Chinese pupils is established, and the usage of the condition and quality BLCV is analyzed. It is found that 212 of 312 those BLCV are used by Chinese pupils in Grade One, the rest 100 are used during the whole primary school years[1]. Besides, learners keep enlarging other vocabularies with their increasing Chinese proficiencies, and thus cause a dilution effect on BLCV. These results show that language learners have grasped most BLCV when they start to use written language, indicating an obvious preference for BLCV acquisition.**

***Index Terms*—BLCV, corpus, acquisition preference**

I. Introduction

The semantic category system is like a hierarchical pyramid. Those in the apex are highly generalized and relatively abstract, with small vocabulary. The lower the level, the weaker the abstraction of this category, with more detailed and larger vocabulary. One of them has a salient status in people's cognition (Rosch, 1973), which is gestalt and has appropriate concreteness. This level is neither too abstract nor too detailed, so it can be regarded as a natural recognitive unit. Compared with other levels, this level has the most obvious differences among different categories, and it is also the basic level for objects classification. Most of human knowledge, including language, is organized at this level. The vocabulary in a language system is mapped to a specific semantic category in a specific hierarchy. Therefore, according to the basic hierarchical category theory, the vocabulary system should also be a hierarchical pyramid structure. Words mapped to the category of the basic level belong to basic words, and the cluster of these words forms a basic-level category vocabulary.

Our previous research results (Song, 2011a) hold that BLCV is in a preference order for language acquisition. The relatively short BLCV word length shortens the memory units of the whole Chinese vocabulary system and reduces the memory burden of the students; Most BLCV are single-morpheme words, so it takes less time to memorize a large number of compound words; BLCV exhibits strong productivity, which enables students to understand and master the derived vocabulary based on the existing one; BLCV has great potential in metaphor and metonymy, giving more extensible pragmatic space; and BLCV has a gestalt that shows more compliance with pupils' acquisition mechanism. According to the above said features, BLCV could play a big role in international Chinese language teaching. BLCV can be viewed as the key nodes to weave the Chinese vocabulary network. Based on the productivity, those relevant new words could be understood without too much effort. The reason is that the concepts reflected by BLCV often coexist in different languages, which makes it especially suitable for international Chinese language teaching. It means that BLCV learning can be an efficient way of vocabulary teaching and acquisition.

However, the foresaid hypotheses and inferences of BLCV acquisition preference still need corresponding empirical research. BLCV "seems" to be the corpus that children begin to learn first in the world, and BLCV also "seems" to be the most of words children begin to use in the first several years. However, no research has been conducted to analyze all the corpora in the children's cognitive process. At present, many researchers analyze foreign learners' writing problems by establishing a written language corpus, and some also study the grammar learning of primary school students. But no such research on the forms of large-scale corpus in children's language learning (consolidation) period has been found, which can verify the use of those vocabularies by learners in the same period.

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

[1] In Chinese primary school system, students are required to finish six years of studies in a row, and each year is regarded as one grade. Namely, you will be in Grade one if in the first year, and the like.

Thus, a written language corpus of Chinese pupils who are still in the language learning period (at least the language consolidation period) will be established to collect their written Chinese lexicon. The condition and quality BLCV (Song, 2015) (see the appendix) extracted in the previous studies are taken as examples to analyze the ways that BLCV could be learned and mastered by learners, as well as changes in their use of those BLCV with increasing Chinese proficiency. By this way, those BLCV acquisition order could be obtained, which could further verify the hypothesis of the BLCV "acquisition preference".

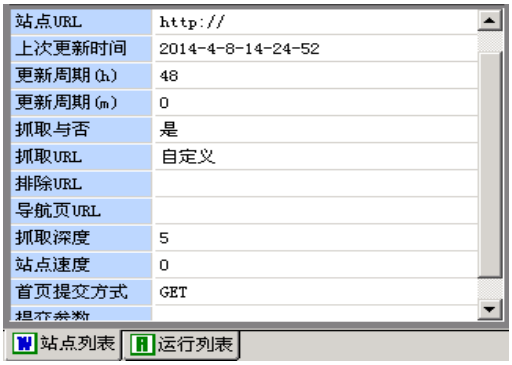
II. The Establishment of a Written Language Corpus of Chinese Pupils

A. Selecting the Target Website for Crawling

The "Website of Chinese Pupils' Excellent Essays" (http://www.eduxiao.com/) is finally selected as the target website to establish a written language corpus of Chinese pupils, due to the following advantages: clear website structure, with an orderly organization manner according to the grade and genre, and high URLs similarity among different categories, which is convenient for analysis; clean storage page without obvious text segmentation; and appropriate corpus size and volume, with totally about 20,000 articles and 10 million magnitude of corpus, which are enough to objectively reflect the situation.

B. Configuring and Grabbing the Corpus with the Web Crawler Tool

After selecting the target website, such parameters as site, crawling depth and cycle, and site speed with the web crawler according to the website structure are configurated. As is shown below:



站点 URL	Site URL
上次更新时间	Last updating time
更新周期(h)	Updating cycle (h)
更新周期(m)	Updating cycle (m)
抓取与否	Crawling or not
抓取 URL	Crawling URL
排除 URL	Removing URL
导航页 URL	Start page URL
抓取深度	crawling depth
站点速度	Site speed
首页提交方式	Submission way of the home page
提交参数	Submission parameters

Figure 1. Web crawler configuration dialog (partial).

Several levels of pages under URL can be crawled by setting the crawling depth. According to the webpage storage structure and essays classification, the crawling depth for the target website is set to 1.

Then, with the web crawler tool, 96,2979,0 Chinese characters, including 17,779 pieces of essays, are crawled, and the authors cover from Grade One to Six.

C. Removing Webpage Tags from the Crawling Results

Removing webpage tags is the most critical step to establish a web-crawling-based corpus, because of various webpage tags and complex structures in the crawled documents. After those tags are removed in batches, only a small part of the texts can be used as a corpus, and then, the subsequent structural adjustment can be performed.

D. Segmenting and Storing Lexicon

Next is the word segmentation for the corpus, and then, the body data can be stored in two structures: one is based on the order of the original articles, with the text of an essay as a piece of data; the other is to break all the texts and store the words of different categories according to the previous classification of essays, such as the words and relevant frequencies in "Grade One". Thus, every word can be stored without repetition in the latter structure.

The former structure is suitable for examining the "general usage" (distribution) of specific vocabulary in the corpus, and the latter is the "frequency" of a specific word.

After the above four steps, the corpus in a moderate size is established, with the grade distribution shown as bellow:

TABLE 1
GRADE DISTRIBUTION OF THE CORPUS

Classification	Quantity (pieces)
Grade One	413
Grade Two	1,505
Grade Three	3,101
Grade Four	3,587
Grade Five	3,803
Grade Sixth	2,606

III. EXTRACTING BLCV ACQUISITION ORDER DATA

BLCV acquisition order in Grade One to Six can be sequenced from the occurrence frequency in the corpus of each grade. If the frequency exhibits an positively increasing trend from Grade one to Six, this BLCV might be acquired at a later period and ranked backward. However, in this case, the corpus size of each grade should be exactly the same. Or otherwise, the occurrence frequency can be restricted by the corpus volumes. For example, Grade One has only 10,000 words and Grade Six has 100,000 words. Under this condition, it requires dividing the frequency by the total number of words in that grade, to obtain the average BLCV number and the frequency in each grade. Hence, the data of the condition and quality BLCV acquisition order mainly include the number of occurrences of each BLCV word in each grade, as well as the frequency division according to the grades. The total number thus extracted reaches 1,647. In the first level of the hierarchical BLCV corpus, Chinese characters "大, 多, 好" rank the first three; and in the third level, " 博学, 没用, 本分" rank the last three, which is shown as below:

TABLE 2
DATA EXAMPLES OF BLCV ACQUISITION ORDER

ID	Words	Frequency	Grade	The total frequency of the grade	Frequency	BLCV level
1	大	0.004341456	1	74,399	323	First level
1	大	0.002855515	2	284,362	812	First level
1	大	0.002754268	3	732,318	2017	First level
1	大	0.002545564	4	1,086,989	2767	First level
1	大	0.00217363	5	1,329,573	2890	First level
1	大	0.001942037	6	930,981	1808	First level
2	多	0.001612925	1	74,399	120	First level
2	多	0.001427758	2	284,362	406	First level
2	多	0.001577184	3	732,318	1155	First level
2	多	0.001613632	4	1,086,989	1754	First level
2	多	0.001644889	5	1,329,573	2187	First level
2	多	0.001572535	6	930,981	1464	First level
3	好	0.002755413	1	74,399	205	First level
3	好	0.002908265	2	284,362	827	First level
3	好	0.002511204	3	732,318	1839	First level
3	好	0.002558444	4	1,086,989	2781	First level
3	好	0.002553451	5	1,329,573	3395	First level
3	好	0.002284687	6	930,981	2127	First level
…	…	…	…	…	…	…
311	博学	0.0000013655	3	732,318	1	Third level
311	博学	0.0000045999	4	1,086,989	5	Third level
311	博学	0.0000015042	5	1,329,573	2	Third level
311	博学	0.0000021483	6	930,981	2	Third level
310	没用	0.0000268821	1	74,399	2	Third level
310	没用	0.0000070333	2	284,362	2	Third level
310	没用	0.0000218484	3	732,318	16	Third level
310	没用	0.0000266792	4	1,086,989	29	Third level
310	没用	0.0000255721	5	1,329,573	34	Third level
310	没用	0.0000279275	6	930,981	26	Third level
312	本分	0.0000027311	3	732,318	2	Third level
312	本分	0.0000007521	5	1,329,573	1	Third level
312	本分	0.0000032224	6	930,981	3	Third level

It can be shown that "大, 多, 好" appear in the corpus from Grade One to Six. The "Frequency" field represents the total occurrence frequency of BLCV in each grade, and when this frequency is divided by the total frequency of that

grade corpus, this result represents the average frequency of that BLCV in each grade. Taking "大" as an example, the frequency decreases from Grade One to Grade Six, indicating that it belongs to the first batch of condition and quality BLCV. After being mastered skillfully in Grade One, this Chinese character is less used as the vocabularies and expression capabilities increase. The same is to the Chinese character "好". However, as for another Chinese character "多", though it ranks in the front and has been widely used in Grade One, this character shows a stable trend of usage from Grade One to Six. The reason is temporarily unexplainable. It can only be guessed that it is not easy to be replaced by other more vivid synonyms as the language proficiency grows.

As for "博学, 没用, 本分", the first one "博学" does not appear in Grade One and Two, and the second one "本分" does not appear in Grade One, Two and Four, showing that these two characters rank relatively backward in the acquisition order, and they are not used until Grade Three. The third one "没用" appears from Grade One to Six, but it is not used that frequently.

IV. The Data Analysis of the BLCV Acquisition Order

A. Mastering Sequence Analysis

According to statistics, among 312 BLCV in the corpus, 212 are used in Grade One, and the rest 100 are first used at least in Grade Two.

These 100 BLCV include one first-level Chinese character, 37 second-level Chinese characters and 62 third-level Chinese characters. Specifically, the first-level character "富" was used from Grade Two, and 37 second-level and 62 third-level characters were used at least in Grade Two.

Among the 37 second-level characters, 30 (严, 广, 男, 古, 实, 具体, 明显, 民主, 偏, 弱, 友好, 野, 全面, 公开, 纯, 齐, 公共, 高级, 彩, 狂, 灵, 盲目, 脆, 适当, 详细, 随便, 俗, 难得, 稀, 严肃) are used from Grade Two; 3 (均, 虚, 牢) from Grade Three; 3 (荒, 旱, 外来) from Grade Four; and 1 (切实) from Grade Five.

Among the 62 third-level characters, 22 (匆匆, 持久, 窄, 客气, 腐败, 模糊, 业余, 笨, 传统, 平和, 时髦, 简陋, 无知, 深沉, 单调, 威风, 凹, 稠, 无能, 无理, 世故, 小气) are used from Grade Two; 20 (真诚, 国产, 俊, 钝, 廉洁, 抽象, 顽固, 空虚, 悲观, 不妥, 恭敬, 倔强, 生硬, 委婉, 没出息, 迟钝, 轻浮, 不和, 博学, 本分) from Grade Three; 16 (初级, 反动, 腥, 片面, 含糊, 糙, 过时, 酥, 奢侈, 可耻, 断断续续, 专制, 冒失, 无礼, 轻薄, 褐) from Grade Four; and 4 (消极, 中型, 矜持, 痴情) from Grade Five.

On the whole, BLCV usage exhibits a chronological order: the earliest is use of the first-level characters; then is the second-level, and the last is the third-level. This also confirms the theoretical motivation of BLCV hierarchy in the previous study from another perspective.

B. Use Trend Analysis

The trend analysis of the BLCV acquisition order refers to the analysis of the frequency increase and decrease trend appearing in the corpus from Grade One to Six for each character, by way of "adding the difference between neighboring frequencies". Namely, the occurrence frequencies of one BLCV character (the entire BLCV of a level) in the corpus are first sequenced from Grade One to Six; then, the former frequency is subtracted from the latter one; and finally, those differences are added (for example, frequency of Grade Two subtracts from the Grade One; Grade Three subtracts from Grade Two, and the like). Thus to observe the general frequency increase and decrease trend from Grade One to Six, and further to judge the approximate acquisition order.

As is expected that BLCV is first learned, and the appropriate frequency is very high at the early stage of language acquisition (such as in Grade One and Two). The word categories exhibit a significant increase trend from Grade One to Six, so the frequency of each word is relatively small and shows a decrease trend. Results of adding the difference between neighboring frequencies should generally be negative. Because of small frequency value, the value for calculation here is multiplied by 100,000.

From the perspective of the BLCV level, the frequency difference sum of the first, second and third level is -8.431147696, -0.37303244, and 0.098702854, respectively. As predicted before, frequency of condition and quality BLCV decreases to a certain degree due to the dilution effect; the sums of the first and second level are both negative; and only the third level is barely positive.

The above data show that BLCV generally are the earliest to be learned, showing a preference for acquisition. Among them, the first-level BLCV is almost all used since Grade One, and has a clear decline in the frequency of use from Grade One to Six. This indicates that the first-level BLCV is the first to be learned and used, but the number does not rise as the language level rises, so the total frequency value drops relatively fast. Apart from the dilution effect, the use frequency of the second-level BLCV basically remains the same from Grade One to Six. The reason is that some new words will be acquired by Chinese language learners while the use frequency of the others learned earlier is reduced. Thus, both parts offset by each other, maintaining a basic equilibrium state. The overall use frequency of the third-level BLCV increases as the new vocabulary is enlarged with the improvement of language levels.

V. Conclusion

Most of the 312 condition and quality BLCV in the corpus begin to appear in Grade One, and only 100 are first used at least in Grade Two. This shows that condition and quality BLCV is acquired earlier by language learners, and most of them have been mastered when learners start to use written language, indicating an obvious preference for BLCV acquisition.

In respect of different BLCV levels, the acquisition exhibits a chronological order: the earliest is use of the first-level BLCV; then is the second-level, and the last is the third-level.

Besides, learners continue to improve their Chinese proficiency, and enlarge their vocabulary, which has a "dilution effect" on the use frequency of the entire BLCV; at the same time, learners show a stronger tendency to use non-BLCV that is more accurate than BLCV. As a result, the BLCV use tends to decrease or slow down with the improvement of Chinese proficiency (limited to Grade One to Six). Among them, the first-level BLCV declines significantly, the second-level is relatively stable, and the third-level slow down.

Appendix

Table A
Hierarchy Corpus of the Basic-Level Vocabulary

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	

REFERENCES

[1] Lu Caihong. (2004). The acquisition priority of basic level categories. *Journal of Qiqihar University (Phi & Soc Sci)*, (5), 96-98.
[2] Qi Shuling.(2014). The cognitive sequence of basic-level category vocabulary in international Chinese language teaching—An example of words related to human body. *Applied Linguistics*, (4), 85-90.
[3] Rosch, Ɛ. (1973). Natural categories. *Cognitive Psychology*, (4), 328-350.
[4] Rosch, Ɛ. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382-439.
[5] Song Fei. (2011a). Research on basic-level category vocabulary of modern Chinese.Ph.D. dissertation, Minzu University of China.
[6] Song Fe. (2011b). 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 Fe. (2014). Research on large-scale corpus-based relative frequency location method for modern Chinese basic-level vocabulary. *Applied Linguistics*, (4), 78-84.
[8] Song Fe. (2015). Construction of basic-level condition and quality category vocabulary lexicon in international Chinese language teaching. Ph.D. dissertation,Minzu University of China.
[9] Yang Jichun. (2011). Focusing on teaching vocabulary of basic-level category in teaching chinese vocabulary to foreign students. *Journal of Research on Education for Ethnic Minorities*, (3), 39-44.
[10] Yang Jichun. (2014) Theories and methods on the construction of basic-level category lexicon for international Chinese language teaching. *Applied Linguistics*, (4), 68-76.

Fei Song, Ph.D. of linguistics and applied language, Associate Professor of School of Chinese, Beijing International Studies University, Beijing, China. He focuses on Chinese language processing and international Chinese language teaching.

Qingqing Lan, Master of Teaching Chinese to Speakers of Other Languages (MTCSOL), a mandarin teacher of Marlboro High School, NJ, US, Performing arts in Chinese Class in US. She focuses on performing arts in Chinese Class in US.