Data Analysis and Improvement Suggestions of Common Words List in Business Chinese

Xu Xinwei[1, a] Zhang Shujuan[2, b] and Ma Zhongwen[3, c]

([1, 2, 3] Huawen college Jinan University,Guangzhou,Guangdong province,China)

[a] Xuxinwei@hwy.jnu.edu.cn,[b] zhangshujuan@hwy.jnu.edu.cn, [c] mazhongwen@hwy.jnu.edu.cn

Keywords：Business Chinese. Domain words. *Common Words List in Business Chinese.*

Abstract.*Common Words List in Business Chinese* is an appendix of *Business Chinese Test Syllabus*, which contains a total of 2457 words. Verified by the author, the number of vocabulary is in fact 2455. Among them, 1038 character species are used by 2455 words .1016 characters and 1022 words are shared with *Chinese Proficiency Test and Chinese Character Syllabus*. To improve the quality of *Common Words List in Business Chinese*, appropriate corpus, a better algorithms and excellent expertise with intervention ability on choosing words are all needed.

1. Introduction

The standardization of domain word selections is of great importance to the repetition rate of words emphasized in textbooks, the proportion control of words beyond syllabus and the efficiency improvement of vocabulary teaching. The vocabulary list of business and trade is an important basis for the overall design, textbook compilations, classroom teaching and tests. Business Chinese test, once called HSK (Business) during the early R&D stage, is the national key scientific research project. The project was confirmed by experts on May 28, 2005. In July 2005, after discussion and modification by experts from home and abroad on the first World Chinese Conference, the name of the exam was changed to Business Chinese Test (short for BCT)[1].In August 2006, *Business Chinese Test Syllabus* (short for *BCTS*) was published by Peking University Press. In October 2006, the business Chinese test was officially put into use. *Common Words List in Business Chinese* (short for *CWLBC*）is an appendix of *BCTS.* In February 2007, *BCTS* points out that there are a few text changes in individual places in the second edition. Our statistics are based on the first edition.*BCTS* for its wide use in all of the world and good results has an important historical position. It has made an important contribution to business Chinese teaching and tests. In addition, words and expressions are divided into everyday life and business categories, which is an important guiding significance for the use of words and also in line with the fact of language.After the publication of the *BCTS*, papers focusing on textbook and choosing words are Xin Ping [2], Zhou Xiaobing&Gan Hongmei [3], An Na&Shi Zhongqi [4]. Authors in these 3 papers tried to use the word frequency information provided by textbooks to obtain a business core vocabulary list, but unfortunately no final syllabus was arranged.

2. Data reports of *CWLBC*

The number of words and expressions in *CWLBC* related to everyday life and work is 2457. According to usage of words, the list is divided into two tables. Table 1 contains 1035 words related to business, social life and working, and 1422 common words in business activities are embodied in table 2. Each word followed by its own pinyin does not mark the part of speech. Those words in two tables are arranged in sequence alignment, without word levels. The importance of *CWLBC* is self-evident, whether from the perspective of language tests or teaching. According to our statistics, there are some mistakes in vocabulary list. For example: No. 289 is missing in 80th page in table 1, and No. 246 in 98th page in table 2 is also missing. Thus there are actually 2455 words in *CWLBC.*

The total 784 character species out of the 1034 words in table 1 share 778 ones with *Syllabus of Graded Words and Characters for Chinese Proficiency*(short for *The Syllabus of old HSK*). Six character species beyond *The Syllabus of old HSK* are佰(bǎi)、莅(lì)、 卯(mǎo)、仟(qiān)、寅(yín)、逾(yú). The number of 778 distributed in *The Syllabus of old HSK* is the following: 408 in Jia level, 240 in Yi level, 71 in Bing level and 59 in Ding level[①].

The total 772 words out of the 1034 words in table 1 are shared with *The Syllabus of old HSK* , which accounts for 74.66% of the proportion in table 1 of *CWLBC* and 8.75％of the proportion in *The Syllabus of old HSK*. The number of shared 772 words distributed in *The Syllabus of old HSK* is the following: 41 in Jia level, 196 in Yi level, 173 in Bing level and 362 in Ding level. There are 262 words beyond *The Syllabus of old HSK*.

In general, vocabulary characteristic in table 1 is not very prominent in business domain because these words are related to life, social life and working. Therefore, there are relatively more characters and words shared with *The Syllabus of old HSK*.

The total 719 character species out of the 1421 words in table 2 share 704 ones from *The Syllabus of old HSK*. 15 character species beyond *The Syllabus of old HSK* are镑(bàng)、簿(bù)、囤(dùn)、赁(lìn)、募(mù)、讫(qì)、契(qì)、琼(qióng)、赊(shē)、赎(shú)、 萧(xiāo)、蚤(zǎo)、账(zhàng)、圳(zhèn)、仲(zhòng). The number of shared 704 character species distributed in *The Syllabus of old HSK* is the following: 353 in Jia level, 218 in Yi level, 86 in Bing level, 1 in Bing appendix level , 44 in Ding level and 2 in Ding appendix level .

The total 250 words out of the 1421 words in table 2 are shared with *The Syllabus of old HSK*. Words from table 2 for its special relation to business activities share a small percentage and are mainly distributed in the higher level Bing or Ding. The number of shared 250 words distributed in *The Syllabus of old HSK* is the following: 1(经济jīngjì) in Jia level, 31 in Yi level, 43 in Bing level and 175 in Ding level. There are 1171 words beyond *The Syllabus of old HSK*.

In general, there are 1038 character species used in *CWLBC*, including 1016 character species shared with *The Syllabus of old HSK*, which accounts for 34.97% of the characters number of *The Syllabus of old HSK*.

Chart 1 distributions at all levels about shared character species of *CWLBC* & *the Syllabus of old HSK*

Jia level	Yi level	Bing level	Ding level	The character number in *The Syllabus of old HSK*	The character number beyond *The Syllabus of old HSK*
482	316	126+1[②]	89+2	1016	22

There are a total of 1022 shared words between CWLBC and The Syllabus of old HSK at all levels.

Chart 2 distributions at all levels about shared words of *CWLBC* & *The Syllabus of old HSK*

Jia level	Yi level	Bing level	Ding level	The word number in *The Syllabus of old HSK*	The word number beyond *The Syllabus of old HSK*
42	227	216	537	1022	1433

Syllabus of Chinese characters is not included in *BCTS*. In terms of character species in vocabulary, high frequency character species has a strong ability of word formation, such as "价(jià) " , it has a total of 66 words as a word building morpheme. Those words are 半价(bànjià)、减价(jiǎnjià)、讲价(jiǎngjià)、降价(jiàngjià)、漫天要价(màntiānyàojià)、保价(bǎojià)、报价(bàojià)、比价(bǐjià)、起价(qǐjià)、杀价(shājià)、一口价(yìkǒujià)、全价(quánjià)、让价(ràngjià)、实价(shíjià)、市价(shìjià)、收盘价(shōupánjià), and so on. For 产（chǎn）and 资（zī） as word building

① Jia, Yi, Bing and Ding stands for 4 levels from easy to difficult. The following is the same.
② Because The Syllabus of old HSK contains Appendix Bing and Ding. + 1 or + 2 stands for character specie amount in the appendix.

morphemes, there are a total of 28 and 22 words respectively in *CWLBC*. According to their ability from strong to weak of word productivities, the top 10 character species respectively are 价(jià)、产(chǎn)、资(zī)、税(shuì)、商(shāng)、业(yè)、市(shì)、行(héng)、销(xiāo)、金(jīn). Character species productivities are based on formation capacity statistics in *CWLBC*. It must have reference significance for us to determine the order of prior mastery and grade classification.

There are many phrase chunks in the *CWLBC* including经济(jīngjì), such as经济舱(jīngjìcāng)、经济杠杆(jīngjìgànggǎn)、经济共同体(jīngjìgòngtóngtǐ)、经济开发区(jīngjìkāifāqū)、经济实体(jīngjìshítǐ)、经济特区(jīngjìtèqū)、经济危机(jīngjìwēijī)、经济效益(jīngjìxiàoyì)、经济学(jīngjìxué)、经济学家(jīngjìxuéjiā)、经济一体化(jīngjìyìtǐhuà) and 经济制裁(jīngjìzhìcái) etc.

In recent years, with the development of corpus linguistics, linguists have discovered that language communication is mostly achieved by fixed or semi-fixed patterning and multi-word-combination structure with computer data analyzing. This fixed or semi-fixed modular structure of words is called chunks or lexical chunks [5].

Through statistical analysis, we find that these words are productive by fixed or semi-fixed modal chunks or lexical chunks. We remember as a whole货运代理（huòyùndàilǐ）,代理商（dàilǐshāng）,代理机构（dàilǐjīgòu）rather than word by word. These words are like 经济(jīngjì)、市场(shìchǎng)、贸易(màoyì)、价格(jiàgé)、证券(zhèngquàn)、资产(zīchǎn)、税(shuì)、卖(mài)、企业(qǐyè)、银行(yínháng)、贷款(dàikuǎn)、货(huò)、盘(pán)、指数(zhǐshù)、帐(zhàng)、市(shì)、人(rén)、交易所(jiāoyìsuǒ)、金(jīn)、账(zhàng)、股(gǔ)、投资(tóuzī)、公司(gōngsī)、价(jià)、财政(cáizhèng)、股票(gǔpiào)、管理(guánlǐ)、技术(jìshù)、委员会(wěiyuánhuì)、国际(guójì)、国有(guóyǒu)、卡(kǎ)、单(dān)、储备(chǔbèi)、信用(xìnyòng)、买(mǎi)、额(é)、零(líng)、金融(jīnróng)、输出(shūchū)、资金(zījīn)、缴(jiǎo)、清(qīng)、款(kuǎn)、期(qī)etc.

BCTS contains 20 idioms, accounting for 0.81% of the total number of collected words and phrases. These 20 idioms are the following.

量入为出（liàngrùwéichū）	货真价实（huòzhēnjiàshí）	讨价还价（tǎojiàhuánjià）
自给自足（zìjǐzìzú）	自私自利（zìsīzìlì）	空头支票（kōngtóuzhīpiào）
漫天要价（màntiānyàojià）	开源节流（kāiyuánjiéliú）	寅吃卯粮（yínchīmǎoliáng）
奇货可居（qíhuòkějū）	囤积居奇（túnjījūqí）	名副其实（míngfùqíshí）
入不敷出（rùbùfūchū）	偷工减料（tōugōngjiǎnliào）	供不应求（gōngbùyìngqiú）
价廉物美（jiàliánwùměi）	买空卖空（mǎikōngmàikōng）	一毛不拔（yìmáobùbá）
一本万利（yìběnwànlì）	一掷千金（yízhìqiānjīn）	

In short, data analysis and contrast serves for the determination of grade parameters about business Chinese characters and words.

3. The recognition of words in business domain and the uncertainty of word quantity

So far, the number of difficulty level about business domain words and the proportion between generic and field words is a matter of preference.Collection word criterion like *manager Chinese*, mainly concentrating in Jia and Yi levels, has limited those common words from *Old HSK Syllabus*. Zhou Xiaobing (2008) believes that business Chinese mainly involves economic knowledge, business activities and business etiquette and so on. He obtains 543 words based on statistics which may be associated with business in *The Syllabus of old HSK* scope. (actually 542 because 转配（zhuǎn pèi）is not in the list of *The Syllabus of old HSK*) .542 words only account for 6.16% of the total vocabulary in *The Syllabus of old HSK*.Zhou Xiaobing（2008）obtains 543 business domain words based on preparation by screening from 8822 words in *Old HSK Syllabus*. Meanwhile, the selection of commonly used expressions in *CWLBC* is based on the dynamic word frequency statistics in modern Chinese. It should be said that the scope of the corpus of the two ways is different. However, only 399 out of 1022 shared words between *Old HSK Syllabus* and *CWLBC* are

considered business domain words, meanwhile 623 words in *CWLBC* are not to be determined by Zhou Xiaobing. What's more, 143 words determined to be business words from *Old HSK Syllabus* by Zhou Xiaobing are not collected in *CWLBC.* 623 words, not being regarded as business domain words by Zhou Xiaobing's second statistics based on vocabulary list from *Old HSK Syllabus*, account for a big proportion in *CWLBC.* The deviation of domain words cognition is caused by three reasons:

3.1, Due to the different theoretical frameworks, the difference between generic and domain words is determined. Zhang Li [6] proposes a theoretical framework about the internal structure of business Chinese communication skills. He believes that the structure of business Chinese communicative competence is the Pyramid, from low to high order "basic etiquette and communication -- basic life -- general business information exchange -- business negotiation". BCT R&D center absorbed Zhang Li's the concept of communicative competence in business Chinese. BCT's involvement related to daily and social activities is due to demand analysis. At present, the use of Chinese in the business activities mainly includes two major categories: business activities, daily life and social communication.

At present, the understanding of "business Chinese" as a specific language is more consistent, but the specific content of "business" is difficult to achieve unity. Zhou Xiaobing's business concept is stricter than Zhang Li's. Due to the different theoretical frameworks, the difference between generic and domain words is determined.

3.2, the selection difference of corpus. The corpus determines the content and quantity of words. The selection of corpus includes two aspects: the scale and content of language materials.

In comparison, words in the natural language every day can be considered infinite. How to select the most valuable words and characters became the focus of contradictions. According to Xin Ping (2007), the business corpus for *CWLBC* consists of 140 million characters in economic field and, in contrast, 590 million characters in other 14 categories. We found the following defects exist in the corpus. First the corpus is relatively single, only the written style, without speaking; second, the corpus excludes content related to science and technology, real estate, automobile; third there are no statistical data from present business textbooks and student writing content.

Zhou Xiaobing analyzed business domain words based on vocabulary list in *Old HSK Syllabus*, however, *Old HSK Syllabus* and *CWLBC*, was developed in 1992. Some of the high frequency words are out of date in today's society, which cannot reflect the true frequency of vocabulary use; what's more, it will cause some valuable information loss based on the second processing of vocabulary list.

Word frequency statistics require the balance and dynamic property of corpus, not only focusing on the text but also the style and the time limit.

3.3 Different experts, different views. Because of the choice of language materials and the diversity of statistical methods, it is easy to form the uncertainty of low-frequency phrase selection at low frequency. The manual intervention has become essential in the process of vocabulary list. In the process of intervention, the teaching experience, theoretical accomplishment, the sensitivity of words identification and the attitude of scholarly research determined the individual quantity and overall quality of vocabulary.

Based on the word frequency data, the word list completes the selection of 2500 words, which lays a foundation for the domain vocabulary. With the further development of business Chinese, it is beneficial to revise the original vocabulary syllabus.

4. Suggestions on improving *CWLBC*

The process of R&D the vocabulary list involves the following aspects:（1）the construction of corpus; word segmentation, word frequency statistics; weight calculation and domain clustering;（2）vocabulary comparative analysis;（3）final expert intervention. Efforts must also be carried out from those three aspects to improve the quality of *CWLBC.*

4.1 source of corpus. We have pointed out defects in the source of the corpus of *CWLBC*. The core of the vocabulary syllabus is the corpus. Daily financial and economic materials are an important source of domain words selection. Words used by students in writing can reflect their needs in communication and expression to a certain extent. The purpose of business textbooks is to guide students to use Chinese for business activities. Textbooks reflect the real business activities to some extent. So daily financial and economic materials, written materials and textbooks are an important source of our absorption of domain words. spoken materials, documents, and forms in business cases, negotiations and other real-life activities are a useful supplement to the corpus. Corpus of multivariate and close to the use of the environment can guarantee the richness and coverage of words and make the syllabus more effective in guiding the role.

4.2 better technical means. The principle of field clustering is mentioned by Liu Hua [7]. We can use the formula to calculate the weight of each word. $w(w_i, c_j) = \sqrt{\sqrt{\sum_j \left(p_{ij} - \overline{p_i}\right)^2} \Big/ \sum_j p_{ij}} \times \left(\log(N(w_i) \Big/ N)\right)^2 \times \sqrt[n]{p_{ij}}$ .If a word is relatively rare but it has appeared many times in this article, it probably reflects the characteristics of the article. It is the domain key words that we need.

In statistical language, the importance of weight is assigned to each word on the basis of word frequency. The most common words ["的（de）", "是（shì）" and "在（zài）"] give the least weight, those more common words give less weight and the rarer words give greater weight. This weight is called inverse document frequency (IDF), and its size is inversely proportional to the common degree of a word.

When knowing the word frequency (TF) and the inverse document frequency (IDF), we can multiply the two values and get the TF-IDF value of a word. The higher the importance of a word to an article, the greater its TF-IDF value. So the top few words are the domain keywords of the article.

4.3 manual intervention. As experts on vocabulary, they should improve their teaching experience and theoretical accomplishment and strengthen their sensitivity to word identification with rigorous research attitude so as to improve the overall quality of domain words.

First, in the course of vocabulary selection, it is necessary to interfere with the choice of words through subjective association, but it is unavoidable to encounter the problem of quantity of vocabulary. Xin Ping (2007) defined the number of words in business domain as 2500 or so. The vocabulary list we have developed should satisfy the coverage requirement and make the number of words as reasonable as possible. In our opinion, a more comprehensive vocabulary list can be controlled at around 8000 words. When mastering 8000 or so words, you can acquaint more than 99% of those words in an article [8].

Second, scholars in the manual screening must be experienced in teaching business Chinese, which includes not only language teachers but also teachers with professional knowledge in business. They need to know language skills and vocabulary of every stage well in business Chinese teaching.

The perfection of the vocabulary list in the business domain is a subject that needs to be studied. As Liu Runqing [9] said, "How to determine the grade and select the standard according to word frequency so as to facilitate the development of teaching materials is a significant subject".

How to establish a scientific link between characters and words is a complex application of systematic project. *CWLBC* is a periodic result, but not perfect. We expect a better *CWLBC*.

Acknowledgement

In this paper, the research was sponsored by "innovation platform" of Overseas Chinese Education Research Institute, Jinan University. (Project No. CXPTYB201305).

Reference

[1] China National Chinese International Promotion Leading Group Office, Peking University business Chinese Test R&D office.*Business Chinese Test Syllabus*, Peking University press. 2006.

[2] Xin Ping. "Research on the Rank Parameters of Business Terms in Business Chinese Textbooks". *Applied Linguistics*, 2007 (03): 70-77.

[3] Zhou Xiaobing, Gan Hongmei. "A Survey on Business Chinese Word Selection of Teaching Materials and Vocabulary Syllabus Compilation". *Chinese Teaching in the World*, 2008 (1): 77-84.

[4] Anna, Shi Zhongqi, "Study on Business Chinese Word Rates of Teaching Materials and Core Vocabulary Words". *Applied Linguistics*. 2012 (02): 122-130.

[5] Lewis M. *The Lexical Approach*. Language Teaching Publications, 1997.

[6] Zhang Li. "Business Chinese Teaching Needs Analysis". *Language Teaching and Research*. 2006 (03): 55-60.

[7] Liu Hua. "A Field Words Clustering System in Corpus with C#". *Computer Engineering and Applications*. 2005 (36):167-169.

[8] Institute of language teaching at Beijing Language Institute. *Modern Chinese Frequency Dictionary*, Beijing Language Institute Press, 1986.

[9] Liu Runqing, "on the Reform of English Syllabus from Separate Syllabus to Uniform Curriculum Standard". *Foreign Language Teaching and Research*, 2002 (06): 403-404.