ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 5, No. 5, pp. 1192-1197, September 2014

doi:10.4304/jltr.5.5.1192-1197

A Comparative Study of Word Frequency and Text Coverage between English and Chinese for College English Vocabulary Acquisition

Jianping Luo
College English Department, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China

***Abstract*—This comparative study is based on the English and Chinese corpora of more than one million tokens collected by the author in person and finds that English enjoys a large vocabulary but relatively a far low word frequency and text coverage, compared with Chinese. In the English corpus of more than one million tokens, nearly eighty percent types appear fewer than ten times, and the number of the types with the frequency above ten times is too small to reach the 95% text coverage, which is generally seen as the least required for reading comprehension. Then, this paper infers that, for College English learners in China, if they follow the approach of incidental vocabulary acquisition to pick up their new words from reading, they will have to add up their reading outside classroom to a quantity of more than 660,000 words, eleven times as much as the reading in class. That is to say, they will have to read 800 more texts with 800 words each after class, or have to read another ten texts in their free time after they finish learning one text in class. Undoubtedly, this is a reading load too heavy for them to bear, and reveals that the approach of incidental vocabulary acquisition is not feasible for College English teaching and learning.**

***Index Terms*—comparative study between English and Chinese, word frequency, text coverage, vocabulary acquisition, College English learning**

I. Introduction

A. About Word Frequency and Text Coverage Studies between English and Chinese Vocabulary

It is generally believed that English has a large vocabulary, but it is hardly realized that the language accordingly has a small word frequency and text coverage. Few scholars could clearly tell how small the frequency and coverage are, and what effects they have on College English vocabulary acquisition Also, few Chinese scholars of College English learning research could actually tell what is about the Chinese vocabulary, and what frequency and coverage the Chinese vocabulary has. And so, few persuasive papers on vocabulary acquisition study as a good guideline to College English teaching learning in China could be seen really.

In fact, word frequency and text coverage play a much important role on vocabulary acquisition, especially to those learning English as a foreign language (EFL). But unfortunately so far, such studies could still hardly be found in the fields of foreign language vocabulary acquisition in China today.

B. About Vocabulary Acquisition Studies

How to acquire a large vocabulary for College English learners? There are two quite different views on that question. One is incidental vocabulary acquisition (IVA) (Nagy, Heman and Anderson, 1985), and the other is intentional language learning (ILL) (Laufer & Hulstijn, 2001). The so-called IVA is widely claimed that the English learners as a foreign language could incidentally pick up English new words which they have encountered with many times during the natural reading, a reading just for information or message rather than language skills (Laufer, 1998). On the contrary, the approach of ILL requires learners to try hard intentionally to learn and memorize the new words while doing their reading. Those scholars who are in favor of IVA call the new word learned with IVA a by-product, and claim it is much better than the intentional language learning.

It is held that Nagy, Heman and Anderson (1985) mentioned IVA first in 1985. About 2000 or so, some Chinese scholars in the field of English teaching and research began to pay attention to it, and then some IVA research papers have gradually appeared and got more and more attention since. Gai Shuhua (2003), Duan Shiping & Yan Chensong (2004), Hong & Tian Qiuxiang (2005) and so on, are those scholars whose papers have attracted much more attention since then. Gai Shuhua, one of the earlier scholars introducing the IVA researches inside and outside China, conducted an empirical study on English major students (Luo Jian-ping, 2013). Duan & Yan (2004) concluded from their research that IVA would be better approach because it showed a better result. However, some other scholars also gave their different opinions. Li & Tian (2005) suggested that IVA only means to come across new words unintentionally from reading, and there is no reason to put it against the approach of ILL, and for English as foreign language learners, it is better to focus on learning and memorizing new words intentionally when they are doing reading. But compared with IVA, voice for ILL is much smaller.

The idea of IVA has strongly been influencing the way of vocabulary acquisition for College English teaching and learning in China schools at all levels for more than 20 years. About 1980s, Gui Shichun, an influential professor of Guangdong University of Foreign Language Studies, claimed in his research that, the way to improve your test score is to enlarge your English vocabulary and the way to enlarge your English vocabulary is to do large reading (Huang Ling-yan, 2013). Then the question is, how much could the learners have to do their reading? Unfortunately, still no any exact answer. In fact, few learners could do a large English reading during their school years, but on the contrary, much more learners could not pick up their vocabulary large enough by means of IVA to improve their CET test scores?

C. What Poor CET 4 Test Score Feedback Data Reflect

Since the CET Band 4 and 6 began in the mid of 1980s, the average test scores of the test takers, or the College English learners, have long remained in a low state, according to the feedback data of CET 4 in recent years. See the following tables from Luo's study (2013):

TABLE 1.
THE MEAN SCORES OF CET4 IN RECENT YEARS

Test Date	All-M (SD)	Un-M (SD)	211-M (SD)	Non-M (SD)
2012.6	391 (63)	400 (65)	439 (82)	396 (61)
2011.6	390 (62)	399 (65)	433 (79)	396 (62)
2010.12	386 (66)	396 (69)	436 (85)	391 (65)
2010.6	387 (69)	398 (73)	436 (88)	394 (69)

In Table 1, according to Luo (2013), All-M is the *mean of all test-takers*, SD refers to *standard deviation*, Un-M is the *mean of the undergraduates from all test-takers*, 211-M is the *mean of the undergraduates only from 211-universities* (so-called key universities), and so Non-M indicates the *mean of the undergraduates from the non-211-universities* (non-key universities). From the CET4 score system, the CET pass line is 425, the highest score is 710, and the lowest score is 220.

TABLE 2.
THE NUMBERS AND RATES OF THE UNDERGRADUATES WITH A NON-ZERO SCORE

Test Date	All U-takers	≥430 / %	630-710 / %	330-220 / %
2012.6	3614882	323764 / 29.5%	3898 / 0.1%	456826 / 12.6%
2011.6	3420565	286042 / 28.3%	4277 / 0.1%	430153 / 12.6%
2010.12	3572224	277713 / 28.6%	3107 / 0.1%	594345 / 16.6%
2010.6	3313653	256240 / 30.3%	4792 / 0.1%	582239 / 17.6%

Table 2 is about the numbers and rates of the test takers who are the undergraduates with a non-zero score from the test. In that table, "*All U-takers*" refers to the number of test takers who are four-year undergraduates from universities and colleges with scores above zero, and "*≥430 / %*" shows the number and rate of the those whose scores are over 430, which suggests that they have passed the test, and "*630-710 / %*" refers to the number and rate of the test takers who get the top score, and "*330-220 / %*" suggests the number and rate of those whose scores stay on the bottom in the grade system.

What deep and serious matter can be seen clearly from the tables above? In China, English teaching and learning have long been staying in a low level, and what's more, it would be much surprising to see that far more than half of the undergraduates fail the test every time, even though they have learned that foreign language for at least eight years. So, what's wrong? What does it strongly reflect?

Again, according to the general viewpoints, the low test score is due to a small vocabulary, and the small vocabulary results from a small reading. And so it can be reasonably inferred that learners could not enlarge their reading actually, which causes them fail to pick up their vocabulary large enough by means of IVA, and then that causes them to get a low test score. If so, then questions rise again: Why are the learners not able to have enough reading? Is there anything wrong with the idea of IVA for College English teaching and learning?

D. Wrong Idea with IVA for College English Teaching and Learning

There are at least two wrong things for the matter. One is their neglecting the influence of English small word frequency and text coverage on IVA, compared with Chinese. Another is the negative transfer effect of the idea of their mother language learning.

To pick up a new word from IVA, some scholars (Zahar, Cobb & Spada, 2001) reveal that the word frequency needs from 6 to 20 times' encountering, and the average is 10 times (Saragi, Nation & Meister, 1978), and 8 times at least is needed (Horst, Cobb & Meara, 1998; Waring & Takaki, 2003). But few failed to study the question further that how large a reading is needed to meet the average required frequency of 10 times because they hold that it is just the way that they have learned their mother tongue. So, it is not strange that quite a few scholars hold that IVA is feasible to English learning as foreign language. They take it for granted that a considerable part of the EFL learner' s vocabulary is made up of such by-products from reading, and IVA becomes an only way to enlarge their vocabulary (Nagy, Hermann & Anderson, 1985; Nation, 2001; Wu Wei & Xu Hong, 2006). They did not realize that a much large English reading would actually be an insurmountable obstacle to meet the frequency requirement in EFL learning.

II. Research Design

A. Questions of this Comparative Study

(1) What is about the English vocabulary, word frequency and text coverage, compared with the Chinese ones?

(2) How much reading should the College English learners do in order to pick up a required vocabulary by means of IVA?

B. Purposes of this Comparative Study

Usually, it is hard to tell what is large and what is small without comparison. So this is one of the paper's purposes. By a comparative study between English and Chinese, it will not be difficult to find what are the English word frequency and text coverage, and it will also not be difficult to find whether it is suitable to apply the IVA to College English learning. Besides, even though VIA might be feasible to mother tongue, or Chinese, it would not mean that it also feasible to College English learning, which consists of another purpose of this paper.

C. Methodologies of this Comparative Study

(1) Corpus used

Two corpora are made up for this comparative study, one is English corpus, and the other is Chinese one. English and Chinese reading materials are collected according to the principles: texts used in textbooks of College English, Chinese texts used in middle school textbooks, and both English and Chinese stories, news, reports, academic literatures and so on. The total tokens of both English and Chinese reach to more than one million respectively.

(2) Word counts

Chinese words are somehow different from the English ones. Chinese words are made up of Chinese characters. Sometimes, one Chinese word is made up from one Chinese character, but sometimes one Chinese word is from more than one characters. So, to made it simple or clear, some Chinese scholars prefer to apply one theory called character-based theory to show the difference from the western languages. Xu Tongqiang (2005), a professor of Beijing University, argues, "The Chinese character is a basic unit of Chinese structure". Also, "The frequency of character is one important characteristic for use of Chinese" (Li Guoying & Zhou Xiaowen, 2011). And so, "the statistics of Chinese character frequency is of much value to language teaching" (Fu Yonghe, 1985).

Thus, in this paper, the word counts for Chinese corpus are based on character units, and for English are based on word units, not on lemmas nor word families.

(3) Statistics of word frequency and text coverage

This paper uses the corpus software called Antconc to count the word or character frequency, and uses another tool Excel to count the test coverage. Both tools turn out sound and persuasive statistic data for this comparative study.

III. Statistic Results of this Comparison

A. Word Frequency Compared between English and Chinese

The following table shows the statistic outcomes of frequency of both English words and Chinese characters. The percentage in the bracket refers to the rate of the types which are used either in the English corpus or in the Chinese corpus, and both numbers and rates are the types which are accumulated other than those in the line "1-f ws".

TABLE 3.
FREQUENCY OF ENGLISH WORDS AND CHINESE CHARACTERS

	English	Chinese
Tokens	1,015,941	1,025,527
Types	35416	4513
F ≥50	2108 (5.9%)	1704 (37.7%)
F ≥15	5494 (15.5%)	2574 (57.0%)
F ≥10	7370 (20.8%)	2887 (63.9%)
F ≥5	12034 (33.9%)	3392 (75.1%)
1-f ws	13096 (36.9%)	542 (10.1%)

The table above is also from a study of this author (Luo Jian-ping, 2013), in which the line "1-f ws" refers to "Hepax Legomena", which means the word which just appears only one time in each corpus.

B. Text Coverage Compared between English and Chinese Vocabulary

The following Table 4 is a comparison of text coverage of vocabulary between English and Chinese (Luo Jian-ping, 2013). In that table, "Ws or chs" means English words or Chinese characters; "En txt coverage" means English text coverage, and "Ch txt coverage" is Chinese text coverage. "First 50" refers to the first 50 most often used words which are sorted according to the frequency, and so it is with the other "First 100", "First 1000" and so on.

TABLE 4.
TEXT COVERAGE OF VOCABULARY OF ENGLISH AND CHINESE

Ws or chs	En txt coverage	Ch txt coverage
First 50	43.3%	29.1%
First 100	50.9%	40.5%
First 1000	74.1%	90.3%
First 1500	78.4%	95.4%
First 2000	81.4%	97.8%
First 3000	85.5%	99.5%
First 4000	88.1%	99.9%
First 5000	90.0%	——
First 10000	94.9%	——

By the way, the data from Table 4 are from the same corpora as Table 3, also with more than one million tokens respectively for English and Chinese.

IV. DISCUSSION

A. English Has a Much Lower Word Frequency and Text Coverage

As Table 3 reveals, the total tokens of both English and Chinese are nearly the same in numbers, but the numbers of types are quite different. English has far more types than Chinese, and means to have to use more words, nearly 8 times as Chinese, which clearly shows that English has a much larger vocabulary than Chinese. For word frequency, the rate is much lower in English than that in Chinese. It is seen clearly that nearly four fifths of English words have to enjoy a much lower frequency, lower than 10 times. According to the inference from the statistics above, each word in such a large vocabulary size would certainly take up a much low word frequency. Averagely, the frequency of every English word is 28.6 times, while every Chinese character has a frequency of 227.2 times. For IVA approach, it could also be seen that only 20.8 percent of words show up more than 10 times, which means nearly 80 percent of words cannot enjoy enough frequency to be picked up from a natural reading. But on the contrary, more than 60 percent of Chinese characters share their frequency over 10 times.

For the text coverage, the table tells that, for the first 100 words, English has a higher text coverage than Chinese, but soon it lags behind. Wholly, English text coverage is much lower than Chinese does. As Table 4 shows, to the 95 percent text coverage, which mean the least required for reading comprehension (Schmitt & McCarthy, 1997), English has to use up more than ten thousand words, while Chinese only use one thousand and five hundred characters.

From the tables, seemingly there are more than 7300 English words with a frequency of above 10 times, which makes it possible to pick up them from just reading, but they only make up so small a part of the vocabulary size (20.8%). The number is far away to meet the requirement of 95 percent text coverage, the least required ability of reading. On the other hand, there is one third of English words that only appear one time (36.9%) so that it would be impossible to be picked up just from reading.

So, it could be concluded that English has a much large vocabulary, but a much low word frequency and text coverage, which would certainly decide an unfavorable condition for the College English learners to follow IVA approach.

B. How Much Reading Needs for College English Learning by Means of IVA?

According to the College English Requirement, there are about 4700 English words which all the College English learners should learn and master (The Chinese Minister of Education, 2007). In concrete terms, New Horizon College English, a College English course book that is widely used in many universities and colleges in China, there are about 5000 new English words for the learners to learn. But in the course book, there are only about 800 words which appear more than ten times in frequency, which means there are more than 4000 words that the learners could not pick up and accumulate from just reading the book. So, what could be done? Adding more reading outside the class is the only choice by those who follow the approach of IVA. But how much reading should they add?

According to this author's study, another group of statistic data from the same English corpus shows that, in order to give every one of the required 5000 words a chance to show up ten times at least, the reading quantity out of classrooms should be added up to more than 660,000 words, eleven times as much as the in-class reading for the College English learners. In another word, these learners must read 800 more texts with 800 words each in their after-class free time. Or, when they finish reading one text in class, there are still another ten texts waiting for them to read after class.

Obviously, is it possible for these learners to do that? The answer is NO. Even though they are interested in reading so much, they could not have so much time for that job. As non-English majors, these English learners have many other subjects waiting for them, and that is much impractical to force them to put so much time on this non-major course (Luo Jian-ping, 2013).

C. Negative Transfer of College English Vocabulary Acquisition

As the statistics shows, the number of the often-used Chinese characters is relative much small, compared with that of English words. In fact, as Chinese dictionaries show, in *New Chinese Dictionary*, there are only about ten thousand

Chinese characters in it, and in *Modern Chinese Dictionary of Words and Expressions*, there just collects fifty-six thousand Chinese words or expressions. Thus, different from English, Chinese characters would have a much more high frequency and text coverage, and in fact, only 1500 often-used characters have covered the text to the rate of 95%, seen from the above tables. Then, the certain number of characters would have chance to show up much more frequently in a limited reading material, and therefore it would be possible for a Chinese beginner to pick up them from his or her relative small daily reading. So, in history, Chinese has a learning tradition of vocabulary acquisition called "dushu shizi", which means learning words just through reading. That might be the early IVA of Chinese style.

Unfortunately, quite a few scholars of College English teaching failed to notice the difference between Chinese and English, and took it for granted that the Chinese students could follow the same way to acquire English vocabulary just through not much reading quantity to learn College English. Their empirical researches or experiments focused more on the so-called effectiveness of IVA with just one or two reading passages or novels and then they strongly claimed that it was good to College English teaching and learning, even though they found the efficiency was too limited. Only a few researches once noticed the restriction of word frequency on the efficiency of IVA which was warned in their researches (Luo Weihua & Deng, Yaochen, 2009), but their further probing is not seen.

Under such a wrong influence of IVA over College English teaching and learning, the result is unavoidably a sad story that the students' vocabulary is too small to do well in their College English tests for a long time.

D. Necessary of ILL for College English Learning

One more problem not in favor of the approach of IVA as seen from the above statistics in Table 3 and Table 4 is that the English word frequency is not only much low, but also highly discrete. The frequency of English word is distributed highly among the first one thousand words, but scatters quickly away and drops sharply from the second and the third and the other thousand words, which agrees with the findings of Gui Shichun (2010) and means they have fewer and fewer frequency times encountering with their readers.

Also, there is still another restriction. Before using the approach of IVA, the English learners should first have a basic vocabulary. Li & Tian (2005) hold that the application of IVA requires the student at least to learn 2000 words first of all; Gai (2003) claims that "2000 to 3000 words are wanted first, and for College English learning, 5000 to 6000 words have to be the base for IVA". Some other scholars suggest that the EFL learners could not have a good understanding of what they have read nor have acquired any new words from reading as the English native speakers do until they have a good command of at least 5000 word families first (Coady, & Huckin, 1997; Nation, 2001). In all, mastering the first several thousand words is the first most crucial thing to do before using IVA" (Nagy, Hermann & Anderson, 1985; Nation, 2001).

Accordingly, it is firmly believed that ILL approach is undoubtedly more feasible than IVA approach for College English learners. A foreign language would not really be learned incidentally (Luo Jian-ping, 2013). So in reading, to consult a dictionary or combine IVA with ILL has proved to be necessary (Gao Xinhua, 2010), and in class teaching, modes of new words presentation have turned out to be much effective in class teaching (Zeng Jian-xiang, 2007). One scholar from Guangdong University of Petrochemical Technology strongly hold in her research paper that teachers should use strategies to help learners to memorize new English words intentionally rather than incidentally, which could actually help to arouse their interest more in learning (Huang Ling-yan, 2013).

V. Conclusion

This paper has focused on the discussion of the feasibility of incidental vocabulary acquisition for College English teaching and learning with the help of the comparative study based on the English and Chinese corpora of more than one million tokens collected by the author himself. This study finds that English enjoys a large vocabulary but a far low word frequency and text coverage relatively, compared with Chinese. In the English corpus of more than one million tokens, nearly eighty percent types appear fewer than ten times, and the number of the types with a frequency above ten times is too small to reach the 95% text coverage, which is generally seen as the least required for reading comprehension. Then, this paper suggests that, for College English learning in China, if the learners follow the approach of incidental vocabulary acquisition to pick up their new words from reading, they will have to add up their reading outside classroom to a quantity of more than 660,000 words, eleven times as much as that reading in class. That is, they will have to read 800 more texts with 800 words each after class, or have to read another ten texts in their free time after they finish learning one text in class. Undoubtedly, this is a reading load too heavy for them to bear. As a conclusion, this paper strongly argues and proves that the approach of intentional vocabulary learning is rather feasible for College English teaching and learning than that of incidental vocabulary acquisition.

References

[1] Coady, J. & Huckin, T. (1997). Second Language Vocabulary Acquisition A Rationale for Pedagogy. Cambridge: Cambridge University Press, 225-237.

[2] Duan, Shiping & Yan, Chensong. (2004). Multiple Choice Glossing on Incidental English Vocabulary Acquisition. *Foreign Language Teaching and Research.* 3, 213-218.

[3] Fu, Yonghe. (1985). The New Results of The modern Chinese Character frequency Statistics. *Language Planning.* 3, 44-45.

[4] Gai, Shuhua. (2003). A Review of Incidental Vocabulary Acquisition. *Journal of PLA University of Foreign Languages*. 2, 73-76.
[5] Gao, Xinhua. (2010). An Empirical Study of the Dual Unity between Conscious and Incidental Vocabulary Acquisition. *Shandong Foreign Language Teaching Journal*. 5, 55-59.
[6] Gui, Shichun. (2010). A Corpus-based Analysis of the Register of English Linguistics. Beijing: Foreign Language Teaching and Research Press. 32.
[7] Horst, M., Cobb, T., & Meara, P.. (1998). Beyond a clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11, 207-223.
[8] Huang, Ling-yan. (2013). Strategies of Memorizing College English Words From Lists and Cards. *Journal of Jiamusi Education Institute*. 8, 339+344.
[9] Laufer, Batia. & Hulstijn, J.. (2001). Incidental vocabulary acquisition in a second language: The construct of task induced involvement. *Applied Linguistics*. 22, 12- 26.
[10] Laufer, Batia. (1998). The development of passive and active vocabulary in second language: Same or different? *Applied Linguistics*. 19/ 2, 255- 271.
[11] Li, Guoying & Zhou, Xiaowen. (2011). Improvement in Statistic Method to Chinese Character Frequency Study. *Journal of Beijing Normal University* (*Social Sciences*), 6, 45-50.
[12] Li, Hong & Tian, Qiuxiang. (2005). A Study of Second Language Incidental Vocabulary Acquisition. *Foreign Language Education*. 3, 52-56.
[13] Luo, Jian-ping. (2013). Is Incidental Vocabulary Acquisition Feasible to EFL Learning?. *English Language Teaching*. 10, 245-251.
[14] Luo, Weihua & Deng, Yaochen. (2009). A Study of English Lexical Repetition Pattern Based on BNC Texts. *Foreign Language Teaching and Research*. 3, 224-229 .
[15] Nagy, W., Heman, P., & Anderson, R.. (1985). Learning words from context. *Reading Research Quarterly*. 20, 233-253.
[16] Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press, 2001
[17] Saragi, T., Nation, P. & Meister, G. (1978). Vocabulary Learning and Reading. *System*. 6, 72- 78.
[18] Schmitt, Norbert & McCarthy, Michael. (1997). *Vocabulary: Desxription, Acquisition and Pedagogy.* Cambridge. Cambridge University Press. 12
[19] The Chinese Minister of Education. (2007). College English Curriculum Requirements. Shanghai: Shanghai Foreign Language Education Press.
[20] Waring, R. & Takaki, M. (2003). At What Rate do Learners Learn and Retain New Vocabulary from Reading a graded reader? *Reading in a Foreign Language*. 15, 130- 163.
[21] Wu, Wei & Xu, Hong. (2006). Effects of Frequency on Incidental Vocabulary Learning through Reading. *Journal of Chongqing University* (*Social Science Edition*). 4, 116-121.
[22] Xu, Tongqiang. (2005). Chinese Character-centered Theory and Language Study. *Language Teaching and Linguistic Studies*. 6, 1-11.
[23] Zahar, R., Cobb, T. & Spada, N. (2001). Acquiring Vocabulary Through Reading: Effects of Frequency and Contextual Richness. *The Canadian Modern Language Review*. 57, 541- 572.
[24] Zeng, Jiang-xiang. (2007). The Influence of Modes of Words Presentation on Intentional Vocabulary Learning. *Foreign Language Research*. 4, 131-136.

Jianping Luo was born in Guangzhou, China in 1954. He received his bachelor degree in linguistics from Guangzhou Institute of Foreign Languages, China in 1980s.

He is currently an associate professor in the School of Foreign Languages, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China. He was the dean of the Foreign Language Department of Maoming University from 2002 to 2007. His research interests include linguistics and College English learning and vocabulary acquisition.