ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 8, No. 5, pp. 909-914, September 2017
DOI: http://dx.doi.org/10.17507/jltr.0805.09

The CEFR Stratification of English Productive Vocabulary of Chinese University Undergraduates Based on DIY Learner English Corpus

Dongyun Sun
College of Foreign Languages and Literatures, Fudan University, Shanghai, China

***Abstract*—This paper analyzes the productive vocabulary (PV) of non-English majors in a highly prestigious university in China through a DIY learner corpus of English compositions and the Productive Vocabulary Level Test. Based on the total PV and the average PV, this paper compares the corpus with the CEFR-aligned English Vocabulary Profile (EVP) of Cambridge University. The results show that some of the outstanding students can attain Level B2 of EVP while most students' PV is comparable to Level B1. The results of this study shed light on strengthening vocabulary teaching in College English teaching in China.**

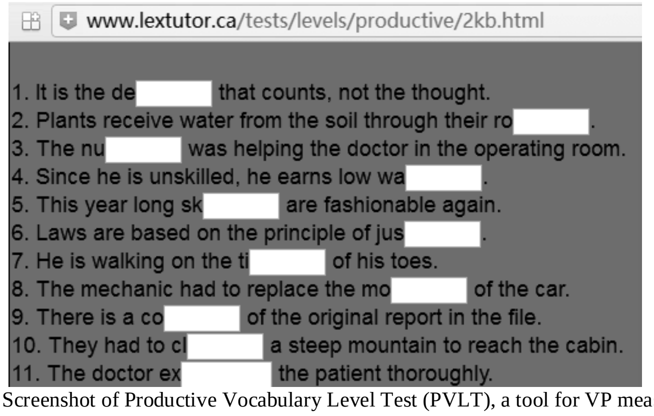
***Index Terms*—learner corpus, productive vocabulary, CEFR, English vocabulary profile**

I. Overview

A. *Introduction to Productive Vocabulary*

Productive vocabulary (PV), or active vocabulary, is an important indicator of language learning that gauges the amount and the level of learner vocabulary in actual use (Melka, 1997, p.84). In recent years, the English competence of leaners in China has generally improved steadily, and their verbal and written communicative competence has progressed significantly. However, many college students still take an examination-oriented approach to English learning, preparing for examinations by means of rote memorization of words. Although many people can pass the examination smoothly, such examinations cannot accurately indicate the learners' level of productive vocabulary. As a result, the phenomenon of Dumb English remains.

There are a number of approaches to the evaluation of productive vocabulary, the most famous of which being the Productive Vocabulary Level Test designed by Laufer & Nation (1999). The tool is roughly similar to a cloze test, offering the first letter of the word and requiring the learner to fill up the gap with needed words. Tom Cobb adapts its test tools to the online edition, as shown in the following figure:



www.lextutor.ca/tests/levels/productive/2kb.html

1. It is the de that counts, not the thought.
2. Plants receive water from the soil through their ro .
3. The nu was helping the doctor in the operating room.
4. Since he is unskilled, he earns low wa .
5. This year long sk are fashionable again.
6. Laws are based on the principle of jus .
7. He is walking on the ti of his toes.
8. The mechanic had to replace the mo of the car.
9. There is a co of the original report in the file.
10. They had to cl a steep mountain to reach the cabin.
11. The doctor ex the patient thoroughly.

Figure 1. Screenshot of Productive Vocabulary Level Test (PVLT), a tool for VP measurement

Although the evaluation tool has a wide range of international impact, it does not fully reflect the learners' oral and written language output in the practical application of the situation. In order to supplement PVLT and measure the actual practical English competence of learners, this paper establishes a DIY corpus of English compositions by non-English majors in a highly prestigious university (hereinafter referred to as "X University"), and makes a quantitative PV study of college students. The individual PV data was also obtained through the Productive Vocabulary Level Test (PVLT). Results of the study are compared against the benchmark of English Vocabulary Profile (EVP) to determine the CEFR level of PV of these learners. It is expected that this study can reflect the actual PV competence of high-level English learners in one of the top universities in China.

1. English Vocabulary Profile and CEFR

The project English Profile was initiated in 2007 by University of Cambridge ESOL Examinations, Cambridge University Press, British Council, Cambridge University, University of Bedfordshire, and English UK. It was also supported by the Council of Europe. At present the project is still underway. While the English Vocabulary Profile (EVP) has been completed, English Functions Profile and English Grammar Profile are still being researched. EVP is based on Cambridge Learner Corpus, which brings together hundreds of thousands of candidates from all over the world who participate in Cambridge English Exams. The EVP team elicits the vocabulary, concepts and phrases of learners at each level of English competence from the corpus composed of more than 45 million words (Good, 2010, p.114).

The stratification of levels ranging from A1 to C2 corresponds to the Common European Framework of Reference (CEFR) (Council of Europe, 2001), which divides language learners' language proficiency levels into three categories of A, B and C and six levels ranging from C2 (Mastery), C1 (Effective Operational Proficiency), B2 (Vantage), B1 (Threshold), A2 (Waystage) to A1 (Breakthrough). Among them, C2 and C1 are collectively referred to as the phase of proficient users; B2 and B1 are collectively referred to as the stage of independent users; A2 and A1 are collectively referred to as the stage of basic users. This standard has been widely implemented around the world, CEFR alignment issues are being extensively discussed and scrutinized. EVP offers a tool for analysis that is aligned with Cambridge English examinations, contributes to the development of glossary of learners for learning and research.

B. Construction and Analysis of Learners' Productive Vocabulary Corpus

Over the past two decades, the vigorous development of corpus linguistics has brought new tools and research paradigms for the research of learner vocabulary. In order to apply the empirical paradigm to understanding the vocabulary competence of learners, scholars have built a series of monolingual learner corpora. In the context of China, results show that Chinese EFL learners’ productive vocabulary is characterized by very limited in quantity, poor in collocation (Deng, 2005, p.9), over-reliance on high-frequency vocabulary (Deng, 2007, p.17), and a highly colloquial style of written language. With the improvement of English proficiency, the colloquial tendency in the written language of Chinese college EFL learners has not been satisfactorily balanced (Wen, Ding & Wang, 2003). Tan (2006) discussed the breadth and depth of PV knowledge of Chinese learners and established a development model for the PV of EFL learners. Zheng (2015) studied the diachronic development of free PV of English major freshmen. Lou & Ma (2012) compared the PV of Chinese and American Advanced English Learners’ Academic Writings through a corpus-driven approach to determine the vocabulary level by the embedded BNC word frequency list in Range BNC. So far none of the researches have aligned the PV level of learners with the CEFR level. Therefore this paper seeks to determine the CEFR level of PV of EFL learners in X University by comparing the corpus data against the EVP.

1. Principles for the construction of learners’ productive vocabulary corpus

To this end, this study takes the opportunity of giving learners assignments of English writing to collect the English compositions from undergraduate students of X University. These electronically submitted compositions then went through screening and tagging into Learners' Corpus of Productive Vocabulary (LCPV). The author has taken full account of the representativeness, balance of subject matter and capacity of the corpus, with the indicators described as follows:

1) Subject matter:

In order to diversify the subject as much as possible, the author required students to submit English writing assignments on 6 different subjects, each with about 400 words, with subject matters ranging from sociology, humanity, environmental protection, science and technology to psychology, covering some of the regular themes of college English textbooks so as to fully mobilize students to use their own mastery of the various fields of English vocabulary acquired in classroom learning. In this way, it is expected that this PV analysis can more accurately reflect the true vocabulary competence of EFL students of X University.

2) Source of corpus data:

This corpus is strictly limited to the third-grade undergraduate students of X University. The source of the corpus is limited to X University because X University is one of the top-notch universities in China. Undergraduates of this university are known for their good English competence, and the vocabulary profile of such students is also impressive. It is expected that results of this study will play a practical role in the future improvement of college English teaching. Originality of all written materials is strictly implemented to resolutely avoid plagiarism. The online submission system is equipped with a duplication-checking function, so if the student's English composition is similar to any composition in the system library, a system alarm is automatically prompted, thus ensuring the quality of the corpus. In addition, citations and quotes were manually deleted to ensure the cleanliness of data.

3) Capacity:

Corpus must reach a certain scale to have practical significance, but if the corpus is not carefully designed and then the corpus is not representative. In that case, even millions of words cannot reach the desired effect of accuracy. Therefore, this study is based on the principle of convenience sampling, eliciting 6 essays from a class of 63 students in a span of two semesters. Several essays that were absent or delayed were excluded from the corpus, and one essay whose length was clearly inappropriate for the corpus was also excluded. The final word count of the corpus is 135,499 words. This data capacity is representative enough for a DIY specialized learner corpus.

C. *Corpus Analysis*

This study uses the corpus software Range developed by Paul Nation (2003) of Victoria University as an analytical tool. The reason why the author didn’t use Antconc or Wordsmith is the need of noise reduction. Even a lemmatized wordlist provided by Antconc is loaded with proper names as well as a number of spelling errors. Based on the somewhat controversial premise that some words in the learners’ productive vocabulary could be misspelled out of sheer carelessness or lack of proficiency, the author performed error correction of misspelled words prior to importing data into corpus tools, thereby improving the accuracy and reliability of corpus data. Since proper names such as names of persons or places in China are usually excluded from the PV, it is hardly possible to eliminate all these words manually. Even if these words are replaced by pronouns, this practice will disproportionately affect the word frequency results. By contrast, the benefit of using Range is that we can tokenize words efficiently and compare all words against frequency lists so that unfamiliar words such as proper names or misspelled words are excluded from the lists.

The reference list chosen for this study is based on BNC-COCA list by Mark Davies and revised for Range by Paul Nation.

TABLE 1:
RESULTS OF WORD STATISTICS OF LCPV USING RANGE29B

WORD LIST	**TOKENS/%**	**TYPES/%**	**FAMILIES**
1	108980/80.43	2487/29.21	**955**
2	12433/ 9.18	1769/20.78	**798**
3	6540/ 4.83	1411/16.57	**777**
4	2047/ 1.51	599/ 7.04	**443**
5	795/ 0.59	332/ 3.90	**276**
6	505/ 0.37	197/ 2.31	**171**
7	299/ 0.22	146/ 1.72	**131**
8	162/ 0.12	97/ 1.14	**91**
9	119/ 0.09	69/ 0.81	61
10	125/ 0.09	57/ 0.67	51
11	62/ 0.05	36/ 0.42	34
12	40/ 0.03	26/ 0.31	24
13	39/ 0.03	19/ 0.22	18
14	19/ 0.01	14/ 0.16	14
15	31/ 0.02	14/ 0.16	13
16	24/ 0.02	12/ 0.14	12
17	11/ 0.01	6/ 0.07	6
18	17/ 0.01	6/ 0.07	6
19	14/ 0.01	5/ 0.06	5
20	9/ 0.01	2/ 0.02	2
21	9/ 0.01	2/ 0.02	2
22	4/ 0.00	3/ 0.04	3
23	4/ 0.00	3/ 0.04	3
24	5/ 0.00	2/ 0.02	2
25	8/ 0.01	3/ 0.04	3
Not in the lists	3198/ 2.36	1195/14.04	?????
Total	**135499**	**8512**	**3901**

Analysis shows that the total number of words identified by the number of word families used by these 63 students amounts to 3,901. Then one question arises: the number of types is higher than expected before the experiment, so why do X University students have such a huge PV? A careful observation shows that the frequency of occurrence of these words varies drastically, with words in Leve1 1, 2 and 3 taking up the overwhelming majority. Whereas the actual use of words spanned all the 25 levels, the upper or uppermost levels contain few words that are statistically insignificant, as they do not represent the actual PV of learners. Therefore the author decides that only data from Level 1 to Level 8 are counted effective, because starting from Level 9, the ratio of word families in the corpus fell below 1%, which the author deems insignificant enough to be excluded. Therefore the total PV thus identified is 3,642 words.

It has to be noted that the author seeks to analyze PV by two indexes at the same time, that is, total productive vocabulary (TPV) and average productive vocabulary (APV). The former refers to the total number of types extracted from learner corpus; the latter refers to the average number of types used by individual learners. Obviously, the former is equivalent to the latter's aggregate value, so theoretically speaking, the greater number of students, the greater total PV. Therefore, the data in this corpus shows that 63 students have a total productive vocabulary of 3,642.

In the past two decades, Chinese scholars have carried out many studies on the relationship between the size of vocabulary and language competence, such as Gui (1985), Yu (1991), Zhou & Wen (2000), Deng (2001), Shao (2002) and so on. The vocabulary size of college students ranges from 1800~2200 (Huang, 2004), 2006 (Wang, 2001), 2404 (Deng, 2001) to 2574 (Shao, 2002). It has to be noted that findings of the investigations are all about the size of receptive vocabulary (RV), but PV is very different. Researchers agree that RV is greater than the PV, and that indicators of students in South China Agricultural University show that the percentage of PV in RV was 51% (Zhong, Adisa & Chonlada, 2005, p.134). Since the number of RV is much higher than that of PV, then why is the total PV of non-English majors in X University roughly comparable to the number of RV of college students in the above studies? The author hypothesizes that the causes of this gap may lie in the following three aspects:

1) The excellent student body of X University. As one of the top universities in China, X University has the ability to recruit the best students from all over China, and their English competence on average is also among the best in the country. Therefore it is no wonder that their PV is higher than the national average.

2) The popularity of electronic dictionaries. With the advancement of computer technology, access to electronic dictionaries and search engines becomes so easy that learners can find whatever they want to express during the process of writing. This is also a variable that is difficult to control in this study. If the essay is written in the examination context, it may more accurately reflect the true level of their PV.

3) The overall improvement of foreign language learning environment. Globalization has also brought tremendous opportunities for development in China, so that more people can open their eyes to see the world, study abroad, enjoy overseas travel opportunities. Students gain greater access to authentic English materials on the Internet, with a huge amount of US dramas, English news and listening materials. Such an ideal environment for foreign language learning was beyond one's imagination in the past. So naturally the PV of college students in the past decade has grown exponentially.

It should be noted that the PV calculation here is based on the compositions of 63 students in the class as a whole for corpus analysis, so the size of PV is the aggregate of all. The advantage of this algorithm is that we can see the overall trend, and it can also compensate for the deviation from the limited number of individual essays and data sparsity. However, it cannot fully reflect the differences between individuals.

While it is possible to calculate the individual PV of each learner in the corpus by using corpus tools, the relative sparsity of data makes invalidates such statistics. As an expedient, the author calculated the individual PV by asking the same group of students to take a PVLT test. The tests were administered in a simulated examination environment in a language laboratory where students have access to the Internet. As this test is pretty simple, results were soon obtained and analyzed in comparison with the TPV obtained through corpus analysis. These results manifest a great individual difference among the population. The individual PV of all students is shown in the following figure:

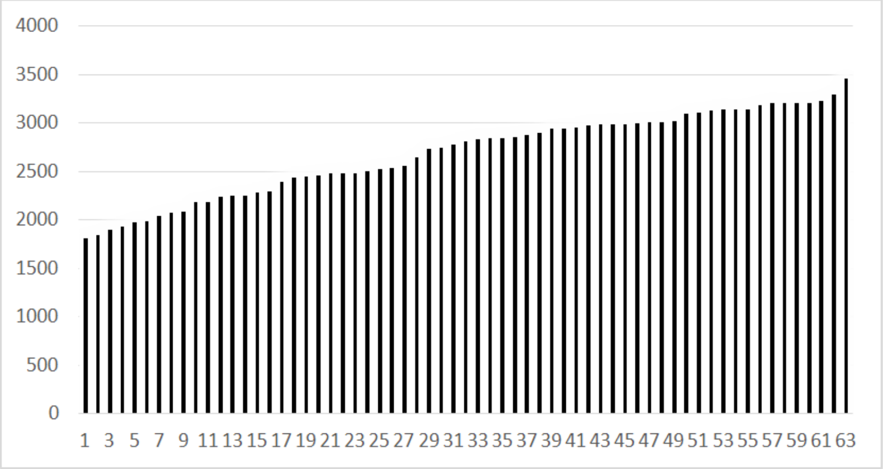


Figure 2. The individual PV of students acquired through PVLT

As shown in Figure 3, the PVs of all individual students were arranged in a histogram in a cascading manner. The chart shows a certain individual difference among the PVs of non-English majors in X University, ranging from around 1,900 to around 3,400, but the PVs of most students are between 2,500 and 3,000 (mean score, or APV is 2,685 while the median score is 2,830). This value does not differ significantly from the PV data from corpus analysis. In fact, the number of positive vocabularies is likely to increase significantly if different productive vocabulary measurements are used.

Comparing the APV data from PVLT and the TPV data obtained from LCPV with the EVP glossary to determine the CEFR levels of PV, we can arrive at the correspondence of the PVs of non-English majors in X University to the CEFR levels. As mentioned above, the EVP project was based on the Cambridge English Proficiency Test papers to determine their alignment with the CEFR. According to Capel (2010, p.5), the vocabulary of A1 to B2 is as follows:

A1 grade 601
New Words
B1 new words 1,429
B2 new words 1,711

Based on the above values, we can calculate that learners at CEFR B1 Level have a PV of 2,955, while the B2 Level vocabulary is 4,666. According to this standard, the TPV of non-English majors in X University is 3,642, and the APV is 2,685. Taking into account the pros and cons of the two methods, and the individual differences between learners, the author speculates that the PV of some of the outstanding students reaches the B1 level of CEFR and even close to the CEFR B2 Level, while most ordinary learners' PV level is around CEFR B1 Level.

II. Unique Word Analysis

In order to further understand the specific differences between the PV of non-English majors in X University and the corresponding level of CEFR, the author extracts the unique words from both LCPV and the glossary of EVP using the method proposed by Feng (2010). The program Concordance 3.2 was used to extract two lists of unique words which were then compared and analyzed.

First of all, the TPV was roughly proportionate in number to the B2 level of CEFR and was thus compared against each other. It is found that a total of 3,542 words belong both to B2 Level of EVP and the TPV; the number of unique words in B2 vocabulary is 1,592, while that of the unique words of the TPV is 1,174. This difference is worthy of attention. After observation, the causes of the differences are mainly reflected in the following three aspects:

1) Genre and scope of knowledge.

As the EVP project is based on Cambridge English Examination corpus, with its huge data spanning multiple years, a variety of genres, and involving a wide range of knowledge. Thus the basic vocabulary coverage is more comprehensive, with such words as “zoo”, “zoology”, “vet” and other words involved, which are absent in the compositions of non-English majors of X University. Therefore, this does not mean that they do not grasp these words, but in the composition does not involve these genres.

2) The influence of Chinese and English.

Unique word analysis shows that some of the more common English vocabulary did not appear in the writings by Chinese students, such as “access”, “accessible” and so on. Given the lexical gaps between two languages, this is understandable, but the language competence of students has yet to be strengthened.

3) The washback effect of language proficiency tests.

Some college students demonstrate a strong vocabulary competence, with their individual PV amounting up to 3,400, and some has even attained the C1 or even C2 Level of CEFR. After communicating with some of these students, the author learned that they have begun to prepare for the Graduate Record Examination (GRE) and other language proficiency tests, and thus consciously or unconsciously they used the vocabulary that they acquired during this stage of learning.

III. Summary and Discussion

A. Summary

This study is an attempt to align college students' PV with its corresponding CEFR level according to EVP. The preliminary results show that the average PV of the second grade non-English majors of X University is close to B1 Level of CEFR, and their TPV is close to the B2 Level of CEFR. Considering the limited samples and the deviation of the experimental method, it can be argued that there is a clear gap between the PV of the vast majority of students and the C Level of CEFR. This shows that even in one of the top universities in China, students still have to improve their English competence, and vocabulary learning should still play a key role in college English teaching.

B. Innovation

The innovation of this research is to use the DIY learner corpus to measure the productive vocabulary, and two indicators of TPV and APV were analyzed. This method effectively complements Laufer & Nation (1999)'s Productive Vocabulary Level Test. Results show that the vocabulary level basically corresponds to the CEFR level by students' self-evaluation.

C. Limitations and Perspectives

The study also has some limitations, including those in the amount of data, the range of topics, and the interference of access to the dictionary or reference materials. These are to be further addressed in the follow-up study.

The follow-up study can follow the following approaches:

First, build a student English portfolio by the CEFR level so that students can perform self-assessment, which is then incorporated into an electronic portfolio (e-portfolio) to establish the alignment between self-assessment and PV of CEFR.

Second, the learner corpus will be consistently expanded with new students enrolled in each new academic year so that the data acquired will be more precise, reflecting a wider range of topics and individuals.

In conclusion, this study is a useful attempt to measure the productive vocabulary of college students. Results show that the PV of non-English majors of X University is generally higher than previous assessments, but most still revolve around B1 Level of CEFR. Clearly there is still great room for improvement. This needs to be addressed in the future college English teaching.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the Ministry of Education of China (15YJC740071).

REFERENCES

[1] Capel, A. (2010). A1-B2 vocabulary: insights and issues/03 from the English Profile Wordlists project. *English Profile Journal*, 1 (1), 1-11.

[2] Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

[3] Deng, Y., T Wang (2005). Statistics in Collocation Extraction and Computer Implementation. *Technology Enhanced Foreign Language Education*, 5, 26-29.

[4] Deng, Y. (2007). Review of Learner Corpora and SLA Research. *Foreign Language World*, 1, 16-21.

[5] Deng, Z. (2001). An inquiry into the survey of English vocabulary—comments on a national vocabulary survey. *Foreign Language Teaching and Research*, 33(1), 57-62.

[6] Feng, Q. (2010). The Practical Translation Course (English-Chinese Translation) (Third Edition). Shanghai: Shanghai Foreign Language Education Press.

[7] Good, M. (2010). Meet the English Profile Wordlists: describing what learners Can Do. *Dictionaries: Journal of the Dictionary Society of North America*, 31 (1), 113-117.

[8] Gui, S. (1985). Investigation and Analysis of Vocabulary in English Majors in China, *Modern Foreign Languages*, 1, 1-6.

[9] Huang, J. (2004). On the vocabulary size of the College English Syllabus. *Foreign Language World*, 1, 1-5.

[10] Laufer, B. and Nation, P. (1999). A vocabulary size test of capacities productive ability. *Language Testing*, 16 (1), 33-51.

[11] Lou, X., & Ma, G. (2012). A Comparison of Productive Vocabulary in Chinese and American Advanced English Learners' Academic Writings. *Theory and Practice in Language Studies*, *2*(6), 1153-1159.

[12] Melka, F. (1997). Receptive vs. productive aspects of vocabulary. *Vocabulary: Description, acquisition and pedagogy*, *33*(2), 84-102.

[13] Nation, P. (2003). Range29b. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

[14] Shao, H. (2002). An empirical study of College English vocabulary proficiency of Chinese normal college students during Band 1-4 stage. *Foreign Language Teaching and Research*. 34(6): 421-425.

[15] Tan, X. (2006). A study of Chinese English learners' productive vocabulary development. *Foreign Language Teaching and Research*, 38(3), 202-207.

[16] Wang Q. (1998). On the First Discussion on English Vocabulary of College Students in China, *Foreign Languages*, 2, 24-28.

[17] Watt, R. J. C. (2004). Concordance 3.20. http://www.concordancesoftware.co.uk/ (Accessed 11 November 2012)

[18] Wen, Q., Ding Y. & Wang W. (2003). Chinese Students' English Writing Colloquial trend—a high level of English Learners Corpus analysis. *Foreign Language Teaching and Research*, 35, 268-274.

[19] Yu, A. (1991). A Survey and Analysis of English Vocabulary of Trainees. *Foreign Language Teaching and Research*, 1, 42-47.

[20] Zheng, Y. (2015). A longitudinal study on free productive vocabulary development from the Dynamic Systems Theory perspective. *Foreign Language Teaching and Research*, (47)2, 276-288.

[21] Zhong, Z., Adisa T, & Chonlada L. (2005). Testing Vocabulary Size, Depth and Strength. *Journal of South China Agricultural University (Social Science Edition)*, 4, 131-137.

[22] Zhou, D. and Wen B. (2000). A track investigation of English vocabulary of Chinese college students. *Foreign Language Teaching and Research,* 32(5), 356-361.

Dongyun Sun was born in Anhui Province, China. She received her PhD from Fudan University in 2015. She is currently a lecturer in the College of Foreign Languages and Literatures, Fudan University. Her research interests include educational linguistics and translation teaching.