A Corpus-Based Study on the Core Vocabulary Growth Models of College English Intensive Reading Materials

YANG Tao
University of Shanghai for Science and Technology, Shanghai, China

GU Yan
Xuzhou Kindergarten Teachers College, Xuzhou, China

This paper explores the intra-textual core vocabulary occurrence rates and core vocabulary growth patterns in the four sets of College English coursebooks favorably evaluated by the Ministry of Education of China in 2004. And the CECR (College English Curriculum Requirements) word list is adopted as the reference word list. The research result reveals that the single text occurrence rates of the core vocabulary of the four sets of books vary substantially from one to another, with New Horizon College English (INHCE) covering the most. In general, the vocabulary size steadily grows within the four sets of course books; however, the unsmooth vocabulary growth pattern does exist. And core vocabulary growth patterns vary due to the differences in the single text occurrence rates of the core vocabulary of the four sets of books.

Keywords: College English coursebooks, core vocabulary, vocabulary growth pattern, corpus

Introduction

A textbook is an important medium for learners to learn English, and it is also the main resource and tool guide for teachers’ teaching (Liang & He, 2009). In fact, textbooks, as a system, are a combination of elements of interaction, thus the selection and design of vocabulary should integrate into the scientific system of textbooks (Hao, 2007). Especially in China, the vocabulary selection and distribution in College English coursebooks is of critical importance since coursebooks are the major sources of vocabulary acquisition for non-English major college students of China.

There are more than 20 sets of College English coursebooks available to choose from under the policy of “one syllabus, many coursebooks” advocated by the Ministry of Education of China. Among them, four sets of College English coursebooks passed the evaluation organized by the Ministry of Education and were recommended as trial coursebooks used in 180 universities of China in 2004. They are New College English (INCE), New Horizon College English (INHCE), Experiencing English (IEE), and New Era Interactive English (INEIE). Intensive reading materials are the most important in-class vocabulary acquisition channels for non-English majors. The design of internal vocabulary distribution will directly affect the effect of students’ vocabulary acquisition. Therefore, this study takes these four sets of intensive reading materials as an example,
and uses the College English curriculum vocabulary as the core vocabulary to describe and compare the selection and distribution of core vocabulary in College English textbooks, especially its growth mode.

**Literature Review**

**Definition of Core Vocabulary**

The so-called core vocabulary refers to the most common part of the vocabulary, also known as the basic vocabulary, which plays an important role in the process of transmitting information in the communication of a certain language (Zhao & Zheng, 2003). Nation and Newton (1997) point out that in any book, regardless of the subject to which it belongs, more than 85% of the vocabulary involved is among the most commonly used 2,000 basic vocabulary. Quirk (1982), and Sinclair and Renouf (1988) have pointed out the importance of core vocabulary for foreign language learners. This shows that these high frequency words should be the primary content of teaching. Therefore, the formulation of the syllabus and the preparation of teaching materials should reflect the attention to the core vocabulary.

However, in the teaching practice, it is tough work to decide what core vocabulary should be included and what should be taught because no single test will on its own be sufficiently systematic measure and core vocabulary itself has no unambiguously clear boundaries (Carter, 1998). Furthermore, according to Carter (1998), from a pedagogic point of view, it is obvious that purely linguistic determination of coreness in vocabulary, however systematic and detailed, will be insufficient for some purposes.

Nevertheless, for College English teaching, the national syllabus prescribes what should be taught and defines what goal the learners should reach. Thus, coursebooks should specify those to be taught and facilitate learner’s learning to reach the goal. The vocabulary recommended by CECR, therefore, is decided as the core vocabulary in this study. The maximization of the occurrence of syllabus vocabulary should be taken into great consideration for the coursebook compilers.

**Corpus-Based Studies on Vocabulary of English Textbooks**

There appeared many studies on the vocabulary design of English textbooks overseas. Simultaneously, with the fast development of corpus linguistics in China, the research on English textbooks by means of corpus is also increasing. He Anping (2001) uses corpus means to study topic distribution of middle school English textbooks. Zhao Yong and Zheng Shutang (2003) studied the core vocabulary coverage and recurrence rate of New Horizon College English textbooks; Zhang Wei and Ma Guanghui (2007) investigated junior high school English textbooks for the coverage of high-frequency words, and it is found that the number of high-frequency words and non-high-frequency words presented in the textbooks is very large; Liang Jianli and He Anping (2009) refer to CCEC (Collins, COBUILD English Corpus) as the reference for the new high school English textbooks. The vocabulary breadth and depth have been studied. It is found that this set of textbooks is basically consistent with CCEC in terms of grammar and vocabulary collocation of sub-cluster and most high-frequency verbs. In addition, Xie Jincheng (2010) investigated the overall presentation of vocabulary in Chinese primary and high school English textbooks and the study found due to the excessive concentration of presenting certain part of speech of vocabulary or semantic paradigm, some other vital parts and a large number of related paradigms are missing.

Most related research has established corresponding corpus of textbooks, or refers to the empirical information of large-scale corpus surveys. With the solid support of computer technology and corpus means,
these surveys cover different levels of textbooks and cover multiple dimensions of vocabulary, such as breadth, depth, paradigm, and collocation, etc. But these studies are mainly conducted from a certain angle to a certain extent. The vocabulary distribution characteristics of a textbook are described in a concentrated manner while comparative studies on multiple sets of textbooks are rare. In addition, most of the existing researches have made a general description of the vocabulary features of a set of textbooks through the frequency statistics of the target language phenomenon, paying attention to the rationality of the measurement, but neglecting the inner rationality of the lexical design structure. In fact, the overall vocabulary selection and measurement as well as the uniform internal structure arrangement are equally important factors to be considered. Therefore, while paying attention to the overall lexical features of the textbooks, this paper explores the structural distribution of the core vocabulary within each set of textbooks, that is, the growth model of core vocabulary.

**Research Methods**

**Research Objects and Means**

The main corpus of this study is the four sets of College English intensive reading materials approved by the Ministry of Education. Each set consists of one to four volumes. The textbook corpus only includes the core textbooks, i.e., the student’s books. The self-built College English textbook corpus has a capacity of 600,000 words. The reference corpus against the textbook corpus is the CECR (College English Curriculum Requirements) vocabulary which includes 2,323 positive vocabulary, 4,404 basic vocabulary, 5,465 medium requirement vocabulary, and 6,513 advanced requirement vocabulary.

The research tools mainly include computer database and programming software Visual FoxPro 6.0, corpus retrieval software Antconc 3.0, and SPSS software for statistical analysis of retrieved data.

**Research Steps**

Firstly, the overall distribution of the core vocabulary of four sets of intensive reading materials is described separately, and the selection of core vocabulary is examined from the perspective of quantity. Then, from the perspective of internal structure and design, the core vocabulary growth pattern of four sets of intensive reading materials is discussed. Many researchers have studied the relationship between texts and their internal vocabulary, and proposed corresponding internal vocabulary measurement models, such as Zipf, Carroll, Brunet, Guiraud, Tuldava, and Herdan. Professor Fan Fengxiang (2006) confirmed through empirical research that Brunet’s model is more suitable for vocabulary research in L2 or foreign language teaching (Fan, 2006). The following is Brunet’s vocabulary growth model formula:

\[ V(N) = \left(\log_w N\right)^{-\frac{1}{a}} \]  

(1) (Brunet, 1978)

\[ \log_w V(N) = \frac{1}{a} \log_w (\log_w N) \]  

(2) derived from (1)

\( \log_w V(N) \) is independent variable, \( \log_w N \) is dependent variable, \( W \) is base of logarithmic function, \( a \) is parameter for the expression.

This study is to use this formula to test the vocabulary growth pattern of four sets of intensive reading materials. Firstly, refer to the CECR vocabulary to calculate the core vocabulary occurrence rate of a single text, and then calculate the total vocabulary of each text and the total vocabulary of the four sets of intensive reading materials. Finally the overall growth curve of the core vocabulary is drawn upon that result.
Results and Discussion

Occurrence Rates of the Core Vocabulary in the Intensive Reading Coursebooks

The overall distribution of the core vocabulary of four sets of intensive reading materials is counted. The occurrence rates of positive vocabulary in IEE, INCE, INEIE, and INHCE was 74%, 84%, 77%, and 92%, respectively. And the basic requirements for vocabulary occurrence rates are 64%, 78%, 70%, and 95%; nearly one-third of the vocabulary in the basic requirements vocabulary does not appear in IEE and INEIE. The medium-required vocabulary occurrence rates of the four sets of intensive reading materials were 55%, 68%, 62%, and 78%, respectively; the overall vocabulary rate was slightly lower than the basic requirement. The occurrence rates of advanced requirements vocabulary are 48%, 60%, 55%, and 67%, respectively. Among the four levels of vocabulary, advanced requirement vocabulary has the lowest occurrence rate on average, which is less than 58%. The tendency is more obviously seen in IEE and INEIE. Generally speaking, the core vocabulary arrangement of INHCE in the four sets of textbooks is the most ideal one. The final result is listed in Table 1.

Table 1
Occurrence Rates of the Core Vocabulary in the Intensive Reading Materials

| Name of books       | Name of word list | IEE      | INCE     | INEIE    | INHCE    |
|---------------------|-------------------|----------|----------|----------|----------|
|                     | Active words      |          |          |          |          |
|                     | Occurring         | 1,724    | 1,948    | 1,790    | 2,143    |
|                     | Non-occurring     | 599      | 375      | 533      | 180      |
|                     | Basic requirement | 2,817    | 3,434    | 3,075    | 4,192    |
|                     | Occurring         |          |          |          |          |
|                     | Non-occurring     | 1,587    | 970      | 1,329    | 212      |
|                     | Intermediate      | 3,015    | 3,722    | 3,360    | 4,270    |
|                     | Occurring         |          |          |          |          |
|                     | Non-occurring     | 2,450    | 1,743    | 2,105    | 1,195    |
|                     | Higher requirement| 3,132    | 3,922    | 3,572    | 4,336    |
|                     | Occurring         |          |          |          |          |
|                     | Non-occurring     | 3,381    | 2,591    | 2,941    | 2,177    |

Notes. N: number; OR: occurrence rate.

Intensive Reading Textbook Core Vocabulary Growth Model

First of all, the basic information such as the number and length of the texts of the four sets of intensive reading materials are presented in Table 2:

Table 2
Basic Information of Four Sets of Intensive Reading Materials

| Name of teaching materials | Number of text | Range     | Mean   |
|----------------------------|----------------|-----------|--------|
| IEE                        | 64             | 332-932   | 686    |
| INHCE                      | 120            | 571-998   | 818    |
| INCE                       | 64             | 642-2,015 | 1,120  |
| INEIE                      | 80             | 322-1,851 | 741    |

Then the occurrence rate of the core vocabulary in the independent text, that is, the ratio of the number of occurrences of the core vocabulary to the total number of words in each text is calculated. The results are shown in Table 3.
Table 3

The Core Vocabulary Occurrence Rate of a Single Text in Four Sets of Textbooks

| Name of materials | Vocabulary in single text Range | Mean | Number of core vocabulary in single text Range | Mean | Occurrence rates of the core vocabulary in single texts Range | Mean | SD |
|-------------------|-------------------------------|------|-----------------------------------------------|------|------------------------------------------------------------|------|----|
| IEE               | 163-409                       | 289  | 154-334                                      | 251  | 76.5%-95.1%                                                | 87.1%| 0.0368 |
| INHCE             | 228-420                       | 333  | 206-385                                      | 299  | 81.3%-96.1%                                                | 90.1%| 0.0237 |
| INCE              | 287-557                       | 419  | 256-487                                      | 362  | 76%-90.8%                                                  | 86.6%| 0.0290 |
| INEIE             | 151-574                       | 290  | 135-486                                      | 249  | 74%-94.2%                                                  | 86.4%| 0.0417 |

As is presented in Table 2 and Table 3, in IEE, the text is the shortest and the size of vocabulary is the smallest among the four sets of coursebooks, but the single text occurrence rate of core vocabulary is second high, which indicates that small vocabulary in each text leads to the low occurrence rate of core vocabulary. In INHCE, the text is not the longest and the average vocabulary of text is not the largest, but the single text occurrence rate of core vocabulary is the highest among the four sets, which accounts for the highest occurrence rate of core vocabulary. In INCE, the text is the longest and the average vocabulary of text is the largest, but the single text occurrence rate of core vocabulary is not the highest, which shows that although the vocabulary load is much heavier, core vocabulary are not necessarily covered effectively. The text in INEIE is longer than the one in IEE, but the single text occurrence rate of core vocabulary is the lowest, which shows that many words in each text are outside national syllabus vocabulary.

Furthermore, post hoc analysis using Tamhane is conducted and the result is shown in Table 4.

Table 4

ANOVA Statistics for the Single Text Occurrence Rates of the Core Vocabulary

| Book (I) | Book (J) | Mean difference (I-J) | Std. error | Sig. | 95% confidence interval |
|----------|----------|------------------------|------------|------|------------------------|
|          |          | Lower bound            | Upper bound|
| 2        | 1        | 0.007567               | 0.0067039  | 0.837| -0.010326 -0.025460    |
| 1        | 3        | -0.029365 (*)          | 0.0052840  | 0.000| -0.043583 -0.015146    |
| 4        | 3        | 0.005540               | 0.0060279  | 0.931| -0.010592 -0.021673    |
| 1        | 2        | -0.007567              | 0.0067039  | 0.837| -0.025460 -0.010326    |
| 3        | 4        | 0.036931 (*)           | 0.0051374  | 0.000| -0.023174 -0.050688    |
| 4        | 3        | -0.002026              | 0.0058998  | 1.000| -0.017773 -0.013720    |
| 1        | 4        | 0.029365 (*)           | 0.0052840  | 0.000| -0.015146 -0.043583    |
| 2        | 3        | 0.036931 (*)           | 0.0051374  | 0.000| -0.050688 -0.023174    |
| 4        | 2        | 0.034905 (*)           | 0.0042173  | 0.000| -0.023603 -0.046206    |

Notes. *: The mean difference is significant at the 0.05 level. Since the variances are not equal, the Tamhane result is adopted.

Book 1 = IEE; Book 2 = INEIE; Book 3 = INHCE; Book 4 = INCE.

The post hoc analysis of the core vocabulary occurrence rate of a single text indicates that there is a significant difference between the mean value of the core vocabulary in INHCE and the average of the other three sets of textbooks.

After obtaining the core vocabulary occurrence rate of a single text, we calculate the ratio of the core vocabulary of a single text to the total vocabulary of the text, and then the overall growth pattern of the core
vocabulary of the four sets of intensive reading materials can be obtained, as shown in Figure 1:

Figure 1. Core vocabulary growth pattern and vocabulary growth pattern.

Note: The core vocabulary growth pattern and the vocabulary growth pattern of IEE, INEIE, INHCE, and INCE are described respectively in Panel A, Panel B, Panel C, and Panel D. The X axis represents the number of words in the accumulated text, and the Y axis represents the corresponding new vocabulary. The solid line is vocabulary growth curve and the dotted line represents the core vocabulary growth. The curves are measured against text.

It is not difficult to see from the figure that the general trend of all growth curves is similar; namely, all experienced both the initial rapid growth phase and the subsequent steady growth phase; the vocabulary growth rate entering the steady growth period is slow, but it has not stopped. However, when we take a close look, the differences of growth pattern in both core vocabulary and vocabulary among the four sets of coursebooks can be detected. The growth curve of IEE (see Panel A) is not steep and the deviation of the core vocabulary curve from the vocabulary curve is the smallest, which means the input number of new core vocabulary in each text is gradually growing and the proportion of core vocabulary to vocabulary is appropriate. But the end-up vocabulary is too small, which accounts for the low occurrence rate of the core vocabulary. The growth curve of vocabulary of INEIE (see Panel B) is not very smooth and the deviation of the core vocabulary curve from the vocabulary is larger than IEE and INHCE. This shows that the vocabulary in each text is in uneven distribution, which leads to the uneven vocabulary growth. Moreover, the core vocabulary in each text is too small compared to its larger vocabulary. The growth curve of vocabulary of INHCE (see Panel C) is very
smooth and the deviation of the core vocabulary curve from the vocabulary is smaller, which indicates that the vocabulary is increasing gradually with more textual lexical input and the core vocabulary is rising steadily together with the vocabulary growth. The growth curve of vocabulary of INCE (see Panel D) is smooth but steep and the deviation of the core vocabulary curve from the vocabulary curve is the largest. This reveals that vocabulary is very large in each text but the core vocabulary is smaller compared to its larger vocabulary.

In sum, the core vocabulary in INHCE gradually grows with the vocabulary increasing and it covers the most core vocabulary. The core vocabulary in IEE increases gradually too but covers the least core vocabulary due to the smallest vocabulary in each text. The core vocabulary in INEIE does not grow smoothly and has the lower occurrence rate compared to its vocabulary. The core vocabulary in INCE grows slowly considering its larger vocabulary.

Combining the analysis of the core vocabulary distribution with the analysis of the vocabulary inner structural arrangement can more effectively test whether the design and arrangement of a set of textbook vocabulary is reasonable. For example, the investigation of the overall distribution of core vocabulary can only reflect the low occurrence rate of core vocabulary of IEE and INEIE, while the test of its growth model more clearly reflects the structural differences of the core vocabulary coverage of the two sets of textbooks. The gradually rising vocabulary presentation, the appropriate vocabulary control, as well as the overall difficulty control of vocabulary learning are the prerequisites for students to learn English well and maintain interest in English learning (Xiong, 2010). Textbooks without vocabulary control are not in accordance with the rules of teaching, and will cause learners to be at a loss (Zhao & Zheng, 2003). In short, the vocabulary arrangement of the textbook does not need to be confined to syllabus, nor can it deviate too much from the scope of the syllabus. Otherwise, it will result in waste of resources in both teaching and learning.

Conclusions

Strictly speaking, there is no vocabulary that has no meaning, but the limited time of the English class objectively requires the representativeness, conciseness, and applicability of the content. The core vocabulary control in College English textbooks should be paid attention to by the textbook editors, so that teachers and students can know which words should be prioritized in the primary stage, trying to avoid untargeted vocabulary teaching (Dai, 2000). To sum up, on the one hand, the coursebooks compilers should attach great importance to the recycling of the core vocabulary so as to help the learners acquire sufficient core vocabulary in an efficient way and meet the requirements of vocabulary stipulated by CECR. On the other hand, teachers can select different coursebooks based on the different degree of difficulty of coursebooks to meet the students’ needs.

References

Brunet, E. (1978). Le vocabulaire de Jean Giraudoux, Structure et évolution. Genève: Slatkine.
Carter, V. (1998). Vocabulary, applied linguistic perspectives. London: Routledge.
Dai, M. C. (2000). On second language vocabulary acquisition. Foreign Language Teaching and Research, 32(2), 138-144.
Fan, F. X. (2006). Quantitative characteristics of marine English vocabulary. Journal of Dalian Maritime University (Social Science Edition), 5(2), 128-132.
Guo, B. X. (2005). Vocabulary selection in English course content. Educational Theory and Practice, 25(2), 44-46.
Hao, Y. N. (2007). Compilation and design of vocabulary for College English textbooks. Crazy English Teacher Edition, 3, 68-70.
He, A. P. (2001). Foreign Language Syllabus Teaching Materials - Design and Evaluation. Guangzhou: Guangzhou Education Press.
Liang, J. L., & He, A. P. (2009). A corpus-based study on the breadth and depth of vocabulary in NSS textbooks. *Journal of School of Foreign Languages, Shandong Normal University, 11*(2), 9-16.

Nation, P., & Newton, J. (1997). Teaching vocabulary. In J. Coady and Th. Huckin, *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 238-254). Cambridge: Cambridge University Press.

Quirk, R. (1982). *International communication and the concept of nuclear English, style and communication in the English Language*. London: Arnold.

Sinclair, J., & Renouf, A. A. (1988). Lexical syllabus for language learning. In J. Coady and T. Huckin, *Second language vocabulary acquisition* (pp. 140-158). UK: Cambridge University Press.

Xie, J. C. (2010). Lexical corpus analysis of middle school English teaching materials. *Foreign Language Teaching Theory and Practice, 1*, 55-61.

Xiong, J. X. (2010). Core vocabulary and textbook compilation. *Course Materials Teach Methods, 10*, 62-68.

Zhao, Y., & Zheng, S. T. (2003). Focus on core vocabulary in College English textbooks. *Foreign Language and Foreign Language Teaching, 6*, 21-24.

Zhang, W., & Ma, G. H. (2007). Vocabulary analysis of English new target. *Foreign Language Teaching in Primary and Secondary Schools, 1*, 9-13.