Research Article

Data Analysis and Optimization of English Reading Corpus Based on Feature Extraction

Nengwei Fan

School of Foreign Languages, Yancheng Normal University, Yancheng, Jiangsu 224002, China

Correspondence should be addressed to Nengwei Fan; 0107009@yzpc.edu.cn

Received 18 July 2022; Accepted 10 August 2022; Published 31 August 2022

Academic Editor: Hengchang Jing

Copyright © 2022 Nengwei Fan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to solve the difficulties faced by English reading teaching, this paper proposes a feature extraction-oriented method for data analysis and optimization of English reading corpus. This method specifically includes the evaluation of vocabulary coverage, comparing the value and vocabulary distribution evaluation of the two sets of textbooks in helping students master English core vocabulary and adapt to the English reading environment as soon as possible, testing whether the teaching of the three sets of textbooks can enable learners to reach the level of free reading corresponding English materials, and testing whether the gradient of the textbooks can be in line with the average language level required for reading corresponding English materials. The experimental results show that, compared with the blog-1565 based on the four authoritative core word lists, the English core word coverage of the two tutorials has reached 91.8% and 93.7%, respectively, and can be improved to 96% and 97.4%, respectively, after optimization.

Conclusion. It is proved that the data analysis and optimization of English reading corpus based on feature extraction can play a directional role in the current English reading-related textbooks.

1. Introduction

With the development of the times, information technology changes with each passing day and penetrates into all aspects of social life. Information technology has a revolutionary impact on the development of education, which must be of great importance. We must "strengthen the development and application of high-quality educational resources" [1]. Reading is an important means for people to obtain information and improve their literacy. As far as foreign language learning is concerned, reading is not only the main goal of foreign language learning, but also an effective way to learn a foreign language. In non-English-speaking countries, reading has always been the main goal of English teaching. English reading can effectively promote the development of comprehensive language skills including listening and speaking ability, improve language level, enrich cultural knowledge, and promote the development of learners’ autonomous learning ability so as to achieve the goal of college English teaching [2]. Therefore, the integration and utilization of information resources and technical means to explore the construction of reasonable college English reading resources based on corpus has been paid more and more attention. Nowadays, English reading teaching is facing many difficulties, and the teaching effect is unsatisfactory, which restricts the development and improvement of learners’ comprehensive language skills. First of all, reading ability, as a language skill, needs a lot of effective practice, and it is difficult to improve significantly in a short time.

The development of reading ability is influenced not only by learners’ vocabulary, sentences, chapters, and other language knowledge, reading skills, and strategies, but also by learners’ personal experience, background knowledge, psychological cognition, and many other factors. Therefore, its development and improvement cannot be achieved overnight. In this environment, the data analysis and optimization of English reading corpus based on feature extraction will point out the path for the content of language learning. A large number of facts have proved that corpus research can present the most common language expressions and show the focus of language teaching [3].
Curriculum designers and teachers need to better reflect this information in the teaching materials and syllabus. Especially in the process of second language teaching, the distribution of information on these language factors can affect the selection of teaching materials, the teaching process, and the teaching focus, which is conducive to improving the effectiveness and efficiency of foreign language teaching and achieving the objectives of English teaching. Figure 1 shows the method of making an English reading corpus analysis system based on feature extraction.

2. Literature Review

A corpus is a collection of real language samples. After scientific collection, classification, and annotation, a corpus of appropriate size can reflect and record the actual use of language and help people analyze and study the laws of the language system. Busby and others have established an academic English spoken corpus and designed academic English vocabulary and reading and writing textbooks using the corpus [4]; Lange and others explored ESP teaching and literature research by using the method of corpus linguistics [5]; Komolafe and others applied the corpus-driven approach to ESP teaching practice [6]. Looking at the current situation of ESP research in the last ten years, it is found that there is still room for the development of ESP corpus research in both theory and practice. Alotaibi and others used sketch engine system to discuss the specific application of corpus technology in ESP vocabulary data teaching from two aspects: corpus construction and teaching activity design [7]. Secondly, most studies focus on static or dynamic corpora, and there are few studies using corpus technology to serve ESP teaching. Xu and others tried to use the new senior high school English (NSEC) textbook corpus and corpus retrieval software to design senior high school English reading teaching. Through pre-reading, prediction, context guessing, and fast reading plot, they deeply discussed the teaching of reading micro skills [8]. Lesnevsky and others analyzed the pragmatic methods and teaching significance of corpus in assisting in understanding the main content of the text, retelling the text, learning anaphora, extracting the main line of story development, and guessing words in the context in combination with the case of corpus-assisted high school English reading teaching [9]. In order to obtain first-hand language data and explore the combination of corpus theory and textbook evaluation theory, in addition to the ready-made BNC corpus, the author has created two corpora [10, 11]. One is a 1 million word English annotation corpus of American and English newspapers and periodicals, and the other is a corpus of English reading textbooks. The corpus of the PCRC comes from three sets of reading materials used by English majors, namely, the American and English newspapers and periodicals reading course (universal, intermediate, and advanced) and the selected English newspapers and periodicals (volume 1–4) published by Peking University Press and the extensive reading course (volume 1–4) published by Shanghai Foreign Language Education Press. The course of reading American and British newspapers and periodicals (hereinafter referred to as American and British newspapers and periodicals) and selected readings of English newspapers and periodicals (hereinafter referred to as English newspapers and periodicals) are the only two sets of self-contained textbooks among all published textbooks for reading newspapers and periodicals, and they are for English majors. At present, some high-quality newspaper reading textbooks are also published, but they are not systematic. In the teaching of reading for English majors in colleges and universities across the country, extensive reading course is very popular and representative. Therefore, the source of the corpus of the PCRC is obtained from these three sets of reading materials [12]. All teaching materials and articles are scanned, identified, corrected, and marked into electronic text storage. Each corpus is attached with detailed annotation, including the source of the corpus, subject matter, genre, author, publication time, and other information. See Tables 1 and 2 for the composition of NEC (newspaper corpus) and PCRC (textbook corpus) corpora built in this research institute.

3. Method

3.1. Vocabulary Coverage Assessment. The vocabulary coverage assessment aims to test the ratio of a certain type of vocabulary that students should master as presented in the textbook. It is an important basis for assessing the value of the textbook in language learning. The teaching objects of “American and English newspapers” and “English newspapers” cover English majors in the upper and lower grades of undergraduate courses, so the two sets of textbooks are comparable [13]. It can be seen from Table 2 that the “extensive reading course” is only for junior English majors. It is not comparable in this assessment, so it is not used for the time being.

3.1.1. Evaluation Objectives and Investigation Contents. This evaluation has two objectives:

(1) By means of corpus, this paper tests the extent to which the two sets of textbooks “American and English newspapers” and “English newspapers” are consistent with the vocabulary learning requirements stipulated in the “Syllabus for English Majors in Colleges and universities (2000 EDITION)” (hereinafter referred to as the “Syllabus”).

(2) Compare the value of the two sets of textbooks in helping students master English core vocabulary and adapt to English reading environment as soon as possible.

3.1.2. Investigation Process

(1) The vocabulary requirements of the syllabus are most directly reflected in the vocabulary of CET-4 and CET-8. First, import the preprocessed vocabulary into wordsmith to obtain the reference vocabulary LEM-8. Then, the sublibraries of “American and English newspapers” and “English newspapers” are extracted from the PCRC, and the generated word
lists are truncated to obtain the respective word lists of the two sets of textbooks. Finally, using wordsmith’s word list comparison function, check the degree to which the word lists of the two teaching materials are consistent with LEM-8 reference word lists. The higher the degree of coincidence, the closer the teaching materials are to the vocabulary requirements specified in the syllabus.

(2) The survey of coverage assessment is to compare the two textbooks’ vocabulary generated in the above survey with the English core vocabulary to see the extent to which they are consistent with the English core vocabulary. The higher the degree of coincidence, the greater the help of the textbooks for students to master the English core vocabulary, and the higher the value of enabling students to adapt to English reading activities as soon as possible. The study of core vocabulary should be the focus of vocabulary teaching and the focus of teaching materials. This survey involves the selection and definition of English core vocabulary. It is found that 87% of the vocabulary in English is concentrated on 2000 high-frequency words. These core high-frequency words should be the focus of English vocabulary teaching. On the basis of these results, this study extracted the intersection 1565 words of the four lists (the shared words of the four core lists), and the integrated English core list (blog-1565) is more scientific and representative.

\[
dm(t) \times \frac{1}{w(t)} = r(t)[N - m(t)].
\]

Here \( m(t) \) represents the expected mean function of the number of detected defects within the time interval \( [0, t] \); \( w(t) \) represents the test workload consumption rate function, which is the test workload shown in formula (1) is very helpful for calculating coverage.

![Figure 1: The method of making an English reading corpus analysis system based on feature extraction.](image)

| Table 1: Structure and distribution of NEC (newspaper library) (unit: word order). |
|---------------------------------|-----------------|-----------------|-----------------|
| NEC                            | British straight news report (BSNR) | 112 995         |
| English sublibrary of British newspapers | British opinion writing (BOPW) | 389 584         |
|                                | Subtotal         | 502 579         |
| American newspaper English sublibrary | American straight news report (ASNR) | 143644         |
|                                | American opinion writing (AOPW) | 360 125         |
|                                | Subtotal         | 503769         |
|                                | Total            | 1006 348        |

| Table 2: Structure and distribution of PCRC (textbook library) (unit: word order). |
|---------------------------------|-----------------|-----------------|-----------------|
| PCRC                            | Primary         | Intermediate   | Token           |
| American and British newspapers | Shape symbol    | 22 211          | 122 811         | 145 022         |
|                                 | Classifier      | 4 751           | 13 560          | 14747           |
| English newspapers and periodicals | Shape symbol    | 59 211          | 102 017         | 161 228         |
|                                 | Classifier      | 9 434           | 12412           | 16 379          |
| Extensive reading tutorial      | Shape symbol    | 174 866         | 174 866         | 174 866         |
|                                 | Classifier      | 14 742          | 14 742          | 14 742          |
|                                 | Total            | 481 116         | 481 116         |
workload function (recorded as the derivative of \( w(t) \) to the test time \( t \), \( w(t) = \frac{dW(t)}{dt} \). When the boundary condition is \( m(0) = 0, W(0) = 0 \), different forms of test workload function \( w(t) \) are substituted into the above differential equation and solved to obtain various NHPP Software reliability models.

3.2. Vocabulary Distribution Assessment

3.2.1. Research Purpose and Tools. In this study, three sets of teaching materials from the whole PCRC library are selected, because these three sets of teaching materials are systematic and constitute a language level gradient in reading content. This study has two purposes:

(1) Using real data to verify whether the vocabulary distribution among the three sets of textbooks is scientific, whether there is a level of gap in language level, and whether it meets the requirements of textbook compilation step-by-step.

(2) To test whether the teaching of these three sets of textbooks can enable learners to freely read the corresponding English materials, that is, to test whether the gradient of the textbooks can be in line with the average language level required for reading the corresponding English materials. This study analyzes the samples from the three sets of textbooks one by one with the help of range analysis software. The basic function of range is to calculate the word frequency range and relevant statistical data of the input text with reference to the three-level benchmark vocabulary [14]. Based on this characteristic, the range analysis software can be combined with the reference corpus to conduct quantitative research on the whole set of teaching materials. The analysis results can provide a scientific basis for the compilation of teaching materials and the selection of teaching materials in teaching. In order to ensure the objectivity and scientificity of the study, the author sampled the three sets of textbooks by layers according to the three standards of volume, subject matter, and genre, and took random sampling among the layers. After preprocessing, the three sets of textbook samples become range recognizable texts, and then the 440 word frequency breadth data calculated by the range and the 360 data calculated from the reference corpus samples are input into SPSS to test one by one [15]. With the help of SPSS, it is necessary to call an independent sample \( t \)-test to calculate whether texts from two different sources have the same difficulty level distribution. This test can compare the difficulty between each textbook and the reference corpus [16]. Since the independent sample \( t \)-test can detect whether the two independent samples come from the same population and whether the data distribution of the two samples is the same, that is, it can test whether there is no difference between the textbook text and the reference corpus text because they both come from the same text population with the same difficulty level. Therefore, we can compare the difficulty of the textbook with the average language level required for reading English materials. Since all the texts of the “American and English newspapers” and “English newspapers” subdatabases of PCRC are from English newspapers, the evaluation of these two sets of textbooks is based on the NEC corpus (newspaper library).

4. Results and Discussion

As can be seen from Figure 2, there is no substantive difference in the vocabulary coverage of the two sets of textbooks, reaching 41.3% and 44.8%, respectively. The coverage of level 8 vocabulary is low, and the number of exclusive words exceeds the number of shared words. Exclusive words are supersylabic words, which are not included in the learning requirements of LEM-8. From the perspective of material sources, the newspaper articles included in the two sets of textbooks are articles read by the British and American public every day, and most of the vocabulary in the textbooks is daily commonly used words. Through word by word analysis, it is found that another reason why the number of exclusive words in the vocabulary of the two sets of teaching materials exceeds the number of shared words is that there are a large number of proper nouns such as people’s names, place names, and abbreviations of institutions and companies such as NRA and Boeing [17]. The “Outline” has no high requirements for mastering proper nouns and abbreviations. After retrieving LEM-8, it is found that there are only 29 proper nouns and 3 abbreviations. In the language practice of students after graduation, a large number of working vocabulary involve these two categories of words. The above analysis shows that LEM-8 deviates from the modern vocabulary actually used by Britain and the United States to a certain extent. The reason lies in the timeliness of language [18].

The calculation results show that, compared with the blog-1565 generated based on the four authoritative core word lists, the English core word coverage of the two tutorials is quite high, reaching 91.8% and 93.7%, respectively, and there is no qualitative difference between the two. Through further analysis of the missing words, it is found that although the prototypes of these word families are missing in the vocabulary of the two sets of textbooks, many of their derivatives have been reflected in the textbooks. If these derivatives are also included in the category of shared words, the vocabulary coverage will be increased to 96% and 97.4%, respectively. This shows that the two sets of textbooks are of great help in mastering English core words [19]. The comparison of the above two findings shows a contradiction. The two sets of textbooks have a high coverage of English core vocabulary but a low coverage of vocabulary in the syllabus. This just shows that when the syllabus contains the vocabulary that students need to master, the frequency of vocabulary is not the only selection standard [20].

From the data in Table 3, we draw the following conclusions through analysis.
(1) There is a significant difference between the first-class vocabulary and rare vocabulary of the “universal edition” and the average level of British and American newspaper English at 95% confidence. Since the “popularization book” sets the minimum language level for learners to start reading British and American newspapers, the most commonly used words that learners have mastered will appear in the articles as much as possible, so that the vocabulary of level 1 is more than the average level, and rare words are used as little as possible, resulting in the vocabulary beyond level 3 being less than the average level.

(2) There is no significant difference between the vocabulary at all levels of the “intermediate edition” and the average English level of British and American newspapers and periodicals. It can be seen that the “intermediate edition” has improved the language level of readers to the level of being able to read most British and American newspapers and periodicals normally.

(3) Compared with the average language level of British and American newspaper English, the first-class words and rare words in the advanced edition have significant differences within the 95% confidence interval. As the compiling principle of “Advanced Edition” is the highest level of British and American newspaper English, its difficulty is higher than that of ordinary British and American newspapers.

(4) Level 2 and level 3 vocabulary belong to secondary and unusual words. Each volume in the course has the same level as the average vocabulary level of British and American newspapers and periodicals on these two levels of vocabulary, and there is no significant difference in the test, which reflects the good combination between the course and the actual language application of British and American newspapers and periodicals.

In this study, the results of Mann–whitney U test in the nonparametric test are consistent with the results of t-test for independent samples. This study confirms that the vocabulary distribution among the “American and British newspapers” has a level gap in language level. The compilation of the whole set of textbooks meets the requirements of step-by-step, and its gradient is in line with the average language level required for reading British and American newspapers. In an ideal state, learners who can complete the whole set of textbooks will likely exceed the average level required by the British and American newspapers, and their language literacy will reach a new level as shown in Table 4:

The SPSS test results of “English newspapers” show that there is no significant difference between the four levels of vocabulary in each volume and the average level of English in British and American newspapers. It can be seen that the full set of “English newspapers” sets the language level of potential readers at the level that they can normally read in most British and American newspapers. This shows that the language level of “English newspapers and periodicals” has been in line with the average level required for reading British and American newspapers and periodicals, but there is no gradient in the distribution of vocabulary among the subvolumes, there is no level gap in the language level, and the compilation of the whole set of textbooks does not reflect the requirements of step-by-step as shown in Table 5:

In general, the reading content of the fourth volume of the “extensive reading course” prepared for the second semester of grade 2 basically reached the reading level of Britain and the United States, while the first three volumes reduced the vocabulary difficulty of the reading materials in order to adapt to the level of lower grade students. Since the fourth volume of the “extensive reading course” is still used by junior English majors, there has not been a significant decrease in the number of first-class words and a significant increase in the number of rare words in the highest volume of the “American and British newspapers.”

![Figure 2: Survey results of coverage assessment.](image-url)
gradient in the distribution of vocabulary among the sub-
American newspapers and periodicals. However, there is no
equal to the average level required for reading British and
respectively. There is no qualitative difference between the two.
word lists, the coverage of English core words in the two
pared with BLOG-1565 based on the four authoritative core
method plays an important role in realizing the scientization
pragmatics, and other aspects, this paper proves that this
popularity of language, language rhetoric, latent semantics,
and effectiveness of English textbooks. Specifically, com-

5. Conclusion
This paper presents a method for data analysis and opti-
mization of English reading corpus based on feature ex-
traction and systematically evaluates three sets of textbooks
collected in the PCCC corpus with the real language data of
the corpus. By observing the theme of textbook selection, the
popularity of language, language rhetoric, latent semantics,
pragmatics, and other aspects, this paper proves that this
method plays an important role in realizing the scientization
and effectiveness of English textbooks. Specifically, com-
pared with BLOG-1565 based on the four authoritative core
word lists, the coverage of English core words in the two
tutorials is quite high, reaching 91.8% and 93.7%, respec-
tively. There is no qualitative difference between the two. The
language level of “English newspapers and periodicals” is
equal to the average level required for reading British and
American newspapers and periodicals. However, there is no
gradient in the distribution of vocabulary among the sub-

References
[1] N. A. Khan and S. M. Shafi, “Open educational resources repositories: current status and emerging trends,” International Journal of Digital Literacy and Digital Competence, vol. 12, no. 1, pp. 30–44, 2021.
[2] X. Gu and Y. Zhao, “Integration of teaching and assessing intercultural communicative competence,” Chinese Journal of Applied Linguistics, vol. 44, no. 2, pp. 241–258, 2021.
[3] E. Baik and H. Joo, “The effect of dramatic play using multimedia on physical expression, language expression and peer play behavior of young children,” Journal of Children’s Media & Education, vol. 20, no. 1, pp. 1–26, 2021.
[4] N. L. Busby and A. Dahl, “Reading rate of academic English texts: comparing L1 and advanced L2 users in different language environments,” Nordic Journal of English Studies, vol. 20, no. 1, pp. 36–61, 2021.
[5] C. Lange, “Review of esimajie, gut & antia (2019): corpus linguistics and african languages,” English World-Wide, vol. 42, no. 1, pp. 111–115, 2019.
[6] B. F. Komolafe, M. O. Ogundiran, F. Y. Zhang, and X. S. Qian, “A comparative perspective of teaching skill acquisition in pre-service physics teacher (pspt) training program in China and Nigeria,” Journal of Baltic Science Education, vol. 19, no. 3, pp. 356–373, 2020.
[7] S. S. Alotaib, T. Kumar, and A. Benyo, “Significance and teaching strategies of mathematics and its skills to the students of other departments at prince sattam bin abdulaziz university, ksa,” International Journal of Advanced Science and Technology, vol. 29, no. 7, pp. 13768–13783, 2020.
[8] J. Xu and T. Li, “The application of artificial intelligence and virtual reality in the auxiliary teaching of american science fiction literature,” Journal of Intelligent and Fuzzy Systems, vol. 1, no. 6, pp. 1–10, 2021.
[9] Y. Y. Lesnevsky, “Assistive technologies as a search tool for the demanded profile of special library,” Bibliotekovedenie [Russian Journal of Library Science], vol. 70, no. 2, pp. 135–147, 2021.
[10] D. Selva, B. Nagaraj, D. Pelusi, R. Arunkumar, and A. Nair, “Intelligent network intrusion prevention feature collection and classification algorithms,” Algorithms, vol. 14, no. 8, pp. 224, 2021.
[11] G. Pujadas and C. Muñoz, “Examining adolescent efl learners’ tv viewing comprehension through captions and subtitles,” Studies in Second Language Acquisition, vol. 42, no. 3, pp. 551–575, 2020.
[12] K. JuusoEsko, “Intelligent temporal analysis of coronavirus statistical data,” Open Engineering, vol. 11, no. 1, pp. 1223–1232, 2021.
[13] S. Zhu, Y. Han, X. Yang, and Yang, “An automatic analysis approach toward indistinguishability of sampling on the lwe
problem,” *Tsinghua Science and Technology*, vol. 25, no. 5, pp. 553–563, 2020.

[14] Q. Zhang, “Relay vibration protection simulation experimental platform based on signal reconstruction of MATLAB software,” *Nonlinear Engineering*, vol. 10, no. 1, pp. 461–468, 2021.

[15] R. Huang, S. Zhang, W. Zhang, and X. Yang, “Progress of zinc oxide-based nanocomposites in the textile industry,” *IET Collaborative Intelligent Manufacturing*, vol. 3, no. 3, pp. 281–289, 2021.

[16] A. Ji, B. Ji, L. R. Xin, A. Wg, Z. Jing, and C. Fza, “Degradation of Toluene in Surface Dielectric Barrier Discharge (SDBD) Reactor with Mesh Electrode: Synergistic Effect of UV and TiO2 Deposited on Electrode,” *Chemosphere*, vol. 288, 2021.

[17] M. S. Pradeep Raj, P. Manimegalai, P. Ajay, and J. Amose, “Lipid data acquisition for devices treatment of coronary diseases health stuff on the internet of medical things,” *Journal of Physics: Conference Series*, vol. 1937, no. 1, Article ID 012038, 2021.

[18] A. Sharma, G. Rathee, R. Kumar et al., “A secure, energy- and sla-efficient (sesse) e-healthcare framework for quickest data transmission using cyber-physical system,” *Sensors*, vol. 19, no. 9, p. 2119, 2019.

[19] L. Yan, K. Cengiz, and A. Sharma, “An improved image processing algorithm for automatic defect inspection in TFT-LCD TCON,” *Nonlinear Engineering*, vol. 10, no. 1, pp. 293–303, 2021.

[20] X. Zhan, Z. H. Mu, R. Kumar, and M. Shabaz, “Research on speed sensor fusion of urban rail transit train speed ranging based on deep learning,” *Nonlinear Engineering*, vol. 10, no. 1, pp. 363–373, 2021.