The Application of Corpora in the Compilation of English Textbooks

Taking COCA as the Example

Haiying Liu
School of Foreign Languages
Chongqing University of Arts and Sciences
Chongqing, China
e-mail: 731417532@qq.com

***Abstract*—This paper analyzes the problems existing in current compilation of English textbooks. And it puts forward that corpora play an important role in the material source, data analysis, the choice of important points in teaching, the design of the classroom activities, teaching material research and the assessment of English textbooks. Therefore, more attention should be paid to the application of corpora in English textbooks.**

Keywords—corpus; English; the compilation of English text books

I. INTRODUCTION

Corpus linguistics as a subject of the combination of quantitative and qualitative analysis is becoming more and more important. Corpus refers to written language and oral language materials processed and stored for the research of language by computers. In the late 1950's, corpus gradually developed under the impetus of the development in computer technology. Corpus linguistics research, because of its objectivity and verifiability, is gradually applied to language teaching. The effects have penetrated into every field of English teaching.

Let us take the Corpus of Contemporary American English (COCA, website:http://corpus.byu.edu/coca/) as the example. It is the largest freely-available corpus of English, and the large and balanced corpus of American English. The corpus was created by Mark Davies of Brigham Young University, and it is used by tens of thousands of users every month (linguists, teachers, translators, and other researchers). The corpus contains more than 450 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. It includes 20 million words each year from 1990-2012 and the corpus is also updated regularly (the most recent texts are from Summer 2012). Because of its design, it is perhaps the only corpus of English that is suitable for looking at current, ongoing changes in the language .

The interface allows you to search for exact words or phrases, wildcards, lemmas, part of speech, or any combinations of these. You can search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near faint, all adjectives near woman, or all verbs near feelings), which often gives you good insight into the meaning and use of a word.

The corpus also allows you to easily limit searches by frequency and compare the frequency of words, phrases, and grammatical constructions, in at least two main ways:

By genre: comparisons between spoken, fiction, popular magazines, newspapers, and academic, or even between sub-genres (or domains), such as movie scripts, sports magazines, newspaper editorial, or scientific journals

Over time: compare different years from 1990 to the present time

You can also easily carry out semantically-based queries of the corpus. For example, you can contrast and compare the collocates of two related words (little/small, democrats/republicans, men/women), to determine the difference in meaning or use between these words. You can find the frequency and distribution of synonyms for nearly 60,000 words and also compare their frequency in different genres, and also use these word lists as part of other queries. Finally, you can easily create your own lists of semantically-related words, and then use them directly as part of the query.

Text-books are the fundamental tool of learners. The study found that 98% classroom instruction get the source from text-books instead of the teachers, while 90% homework also gets guidance from text-books (Suarez, 2001). Take University English textbook editing as the example. With the development of corpus linguistics, the existing university English textbooks have shown the deficiency. The subjectivity in the editing University English textbooks has been questioned. At present domestic efforts in this respect is not enough. Corpus can provide resources which have not been fully utilized.

Fund project: GPA115070 National education science project of the 12th five-year plan in 2011, Ministry of Education (The Application of Corpora in the Compilation of English Textbooks); (Z2010WY21Corpus-based Pragmatic research on English Modal Verbs)

II. THE STATUS OF THE COMPILATION OF ENGLISH TEXTBOOKS

A. The Choice of the Contents of English Textbooks is not Reasonable

Although most of the English textbook writers claimed that the choice of textbook content is reasonable and scientific, but if you check carefully, you will find the choice of these materials, to a great extent, lacks the objectivity and the scientific basis. Compilers only based on subjective intuition and their experiences in teaching, therefore it is not enough in guiding teaching.

For example, if you search "cherish the time" (Book Two, New Horizon College English, 2010) in the Contemporary English Corpus (COCA), no results will be got. However, if you use "value time", you can get the result of the search, so"value time"can be used in American English as phrases, but "Cherish the time " can not be used in American English.

For another example, the search results appear in COCA show the frequency of the following phrases:

TABLE I. THE SEARCH RESULT OF THE FREQUENCY OF FOUR PHRASES IN VOCA

Phrases	Frequency
on...mind	2735
in...mind	24116
in university	1891
at university1	37256

The table shows that in America， majority people tend to use "in... Mind" and "at... University", rather than "on... mind" and "in... university" (Book Four, New Horizon College English, 2010) used in the textbooks.

In addition, the author compared the description of the language in the English textbooks to the search results in the COCA, to examine how the actual users to express the content.

The results showed that the contents of the textbooks have the great differences with the American actual use of language.

For example, the author searched the usage of "advise" in COCA. In its that-clauses, the text books always teach us to use the original form of verbs, however the corpus shows that most Americans (68.30%) use the present tense rather than some grammar textbooks (Plain English grammar, Lou Guangqing, 1996; New-thinking English Grammar,Chen Xiaofeng, 2013) require. The following sentences are quoted from COCA.

1) In general, experts advise that for healthy adult women, about 55 grams of protein per day is sufficient.

2) Since the Main Street light rail opened, no major development has sprung up along its seven-mile corridor. Experts advise that patience is required; it takes five years or more for such growth.

B. The Principle of the Priority of High-frequency Language Phenomenon is not Embodied

Part of English textbooks ignore important language structures, but put too much emphasis on some less common language structures. The author investigated the most commonly used dozens of words in English textbooks, and found out these words less used in the real US society. In English textbooks, for example, "now" is used as a time reference of present continuous tense (Tao ran and Guo Xiaodan, 2008, English grammar).

In COCA survey, 10 sentences were randomly chosen, among them, only in one sentence, the verb form is the present tense. A sentence is in the present perfect tense, while eight sentences are in the present continuous tense. This suggests that not using the examples in real use to compile textbooks will be misleading, corpora should be used for textbooks compiling, especially to discriminate the most common language structures, and common language structures. At present, some Chinese English textbooks, have not yet to embody the principle of the priority of high-frequency language phenomenon.

C. The traditional teaching of grammar and vocabulary in English textbooks is separated

The separation of the traditional grammar and vocabulary in language textbooks has been challenged. Before the advent of corpus, to deeply and effectively study the relations of the combination of the words is not realistic. Now, the corpus evidence can be made full use in compilation of English grammar textbooks and it broke the isolating of vocabulary and grammar description. Appling corpora opens new ways of language description in English grammar textbooks.

III. ENLIGHTENMENT TO THE COMPILATION OF ENGLISH TEXTBOOKS BY USING CORPORA

The rapid development of corpus linguistics provides a solid foundation to the use of real language.

A. Large Sources of Corpora

Using large online corpora (COCA, for example), textbook of language can be selected from multiple domains, such as spoken, fiction, popular magazines, newspapers, and academic texts. As for textbook compilation, examples can be chosen from the corpus, reflecting the real language used in American society.

For example, in COCA corpus, "worth" and "worthy" were searched respectively in five domains: spoken, fiction, popular magazines, newspapers, and academic texts. Table II and III are the data of the frequency of the two words:

TABLE II. THE SEARCH RESULT OF THE FREQUENCY OF "WORTHY" IN THE FIVE DOMAINS IN COCA

domain	spoken	fiction	magazines	newspapers	academic
Frequency	684	1154	1722	1336	1764

TABLE III. THE SEARCH RESULT OF THE FREQUENCY OF "WORTH" IN THE FIVE DOMAINS IN COCA

domain	spoken	fiction	magazines	newspapers	academic
Frequency	8478	7034	13203	2015	5547

According to the tables, "worth" is used more, especially in the fiction texts, "worthy" is used less, especially in the spoken texts.

B. More Truthful Accurate Examples by Using Data from Corpus

On the frequency and distribution of various language phenomenon, comprehensive and reliable statistics can be conducted. When describing language phenomenon, what is found in corpus can be attached to quantitative description of the phenomenon. Discussion for the discovery from the corpus can also be presented in textbooks to discuss and analyze the data.

In COCA, for example, after the search, we got the result: "audience" is used 34395 times as a collective noun, accounting for 72.69%. As a countable noun, "audiences" are using 7708 times, accounting for 18.31%. Therefore, in America, 72.69% of the word is used as a collective noun. Authors can retrieve a lot of examples from the corpus, through the analysis of the examples, summarizing the differences in the specific context background and the difference of classification, qualitative conclusion can be made.

All is based on the quantitative data, with the example analysis and with the aid of the specific use of language for them. Modern linguistics has one of the most essential difference with traditional grammar, which is the way of using description language, rather than with the method of rules. What a corpus collect is the actually-used language. Based on corpora, we can really compile grammar in textbooks by using descriptive language.

Grammar textbooks often introduce "some" and "any" , the two qualifiers in this way: "Some" is often used in affirmative sentences. "Any" is often used in negative sentences or questions, because "any" has the negative meaning. And it is often used in conditional clauses. In addition there are three kinds of usage of "some", but textbooks seldom mention that "any" used in the affirmative situation. So, whether in language use is not to use any positive? According to the search result of COCA, 10 sentences with "any" among 400026 were randomly selected. By observing the 10 sentences, "any" are found in affirmative sentences in 9 sentences, 90% of the total.

Again, for example, the collective noun should be looked as a whole or the individuals. In fact, it is quite complicated, which should be distinguished in classification. The first is two kinds of usages can be carried out, such as "youth". The frequency of "the youth is" in the search result of the frequency is 26. The frequency of "the youth are" in the search result of the frequency is 28. The second category is tend to regard the words as a whole and as a singular, such as "crew "and "family". For example, the search result of the frequency of "crew is" is 152 times, while the search result of "crew are" is 24 times. Similarly, the search result of the frequency of "family is" is 926, while the search result of the frequency of "family are"is 109.

The third kind tends to take it as a countable noun. For example, the search result of the frequency of "the police is" is 69, while the search result of the frequency of "the police are"is 770. The search result of the frequency of "The rich is" is 24, while the search result of the frequency of "The rich are" is 172. So, we should give up regulative description while we compile a textbook, we should base it on corpus, with the real language use for this, embodying a descriptive methods.

C. Applying Corpus to Select the Teaching Focus

Paying attention to the highest frequency of words is the key of teaching syllabus, which is one of the main principles need to be followed. This, to a certain extent, avoids the blindness of the syllabus formulation, choice of contents. Scott Thornbury, based on BNC (the British National Corpus, website: http://corpus.byu.edu/bnc/,The BNC was originally created by Oxford University Press in the 1980s - early 1990s, and now exists in various versions on the web.) retrieves from the corpus the highest-frequency words, and lists the first 200 high-frequency words. He compiles them known as the textbook, Natural Grammar. The book's purpose is to let the students master the grammar paradigm, and collocation of the high-frequency words

Research in corpus can find practical regulations, revealing the typical features of language, so as to reflect the reality of the language. In general, the word frequency in the discourse is very uneven, but regular. And corpora display the asymmetry distribution characteristics: 95% articles are written with 3000 to 4000 words, which are composed of 60% function words. This asymmetry distribution characteristic of language brings the teaching enlightenment, whether to differentiate language phenomenon, such as high-frequency words, low-frequency meaning of high-frequency words, high-frequency words, high-frequency meaning of low-frequency words, lexical collocation patterns of high-frequency words, high-frequency grammatical structure, etc. High-Frequency language points are in learners' most need to learn. From this perspective, corpus research has great potential, because its retrieval software can provide word frequency. Identification of most common words and high-frequency and lexical collocation can help to make these clear, such as the focus of English teaching, the contents and the order of the contents, which are very useful information.

For example, in the current English textbooks, "will","shall" and "be going to" are considered different only in some details, but all can express the time in the future. Learners will think three ways to express the future tense is the same, and use them randomly. In fact, the detection can be obtained in COCA statistics. We can see that native speakers use "will", "shall" to express time in the future, the retrieved data to show the future time is 912979 times for "will" and

17123 times for “shall” respectively, with the highest frequency (70.65%) in the present tense, followed by general present tense. And the frequency of other ways (such as, be going to, be to do, be about to structure), which show the future time, is only 9. 38%. This proved that if we do not focus on the typical language phenomenon, it is in the violation of the language communication, therefore it will reduce the effectiveness of teaching. Textbooks on the arrangement of these expressions can embody the frequency of use in the corpus, which can play a better guidance to the learners .

The study found that 60% in-phrase in COCA is composed of about 150 words. In most cases, the phrase in the textbooks arrangement are in random sequence without the connection with the frequency. In face of a large number of phrases, learners often find it difficult to memorize them. In fact, according to frequency, it is easy to divide phrases into different levels. For example, the most commonly used phrase can be marked with "4", the less commonly used phrase with "3" and so on. In learner's mind, the concept of relative frequency, importance can be established. So students can arrange reasonably. There is no doubt that this consciousness is important in teaching plan. Published COBUILD grammar textbook, is compiled according to the results of the study in COBUILD corpus (COBUILD, an acronym for Collins Birmingham University International Language Database, is a British research facility set up at the University of Birmingham in 1980 and funded by Collins publishers.Website:http://www.mycobuild.com/homepage.asp x.) What is written in the book is examples from the corpus, the real language material, with the order of definitions of terms determined by the word frequency obtained from the corpus. This kind of interpretation method reflects the language situation more objectively, originally, also more conducive to language teaching. Compilers can get word frequency statistics to provide the scientific basis to formulate the syllabus. And the corpus also puts forward the frequency of syntax structure, which can be base for the formulation of the syllabus.

The development of corpus linguistics brings a new revelation to the compilation of English textbook.Writers should properly consider the typical features of language structure and teaching requirements, and arrange teaching content accordingly. After knowing textbook's main frame and language knowledge framework, the information structure of the external frame system, which is speech, grammar, reading, writing, language culture, etc., we can decide the arrangement of teaching contents, referring to the arrangement of Longman Contemporary Dictionary in the new version, which are compiled based on the large-scale corpus, with the provided word frequency.

D. Corpus and the Design of the Classroom Activities

English teaching is to solve two questions: what to teach and how to teach. What about teaching? The principle is: most teachers follow textbooks arrangement because of the authority of textbooks. And foreign scholars Pasch and Norsworthy proposed that "critical thinking" (2001)should be integrated into the language classroom teaching, and its characterization can be summarized as the following: to encourage students to actively participate in the process of constructing knowledge. Questions should be open, with a variety of answers. Students should be given enough time to think about the questions. Classification exercise and inductive activities should be designed. So Corpora can be applied to a variety of teaching activities, such as vocabulary teaching, grammar teaching, writing teaching, reading teaching and translation teaching. The following is the example in vocabulary teaching.

Teachers can put forward requirements for students through the search in COCA to distinguish the use of “goods” and “cargo”. Students easily get sentences from COCA:

1) Well, I believe that there are economic goods that we can achieve by doing the right things in space.

2) We must ration the minimum requirements of life and give our people ration cards, and keep prices fixed for these goods.

3) Japan, the biggest source of those imports, shows fewer goods going out.

4) He can't leave the cargo bay.

5)This compartment restricts the boat's ability to carry large cargo items.

6)Transcontinental mail was indispensable cargo.

The above examples can be summed up: “goods” is a collective noun, and is usually used as a plural concept, and “ cargo” is also a collective noun, is usually used as a singular concept, and often as noun qualifier placed before a noun.

For another example, students can also be required to use COCA corpus to distinguish the phrases, such as “near-sighted” and “short-sighted”, and to provide typical examples for them. Students can quickly select the following from the corpus, e.g:

1) Can we use your glasses? Are you far-sighted or near-sighted?

2) A lot of people thought that was a short-sighted move, that quitting would end her career as an elected politician.

From these examples, we can distinguish that “ near-sighted ” refers to “ unable to see distant objects clearly ”, and “ short - sighted ” refers to “ lacking foresight or scope ”.

Therefore, by using corpora, we can design inquiry-based classroom activities.

For example, we can require students to search the different usage of “ ear ”and “ ears ”, and to generalize and discuss their meanings. By searching in COCA, we can find they not only have the differences in number, through the examples from corpus, it is easy to find that the singular form of “ ear ” is used in metaphors, such as, keep/have one's ear to the ground, goes in one ear and out the other, turn a deaf ear, play by ear, have itching ears...

Corpus of auxiliary classroom teaching can stimulate students' curiosity and motivation. Secondly, it can present students abundant corpus examples. Students don’t carry out simple memory activity, but focus out the analysis and

inductive activities for cultivating their ability, so as to improve the learning effect.

As the compilers' paying more attention to the application of corpus linguistics in the classroom activities, corpus-aid classroom teaching practice will be more abundant.

E. Corpus and the Research on the Compilation of Textbooks

Currently corpora can be used in the study of textbooks, which mainly include: teaching materials research, classroom discourse research, learners' interlanguage research.

With the ais of corpora, textbooks are now analyzed and researched in the arrangement of topics, of vocabulary, and of grammar points. In corpora, classroom discourse between teachers and students can be qualitatively and quantitatively studied. Our country already has multiple learners' corpora: "Chinese learner corpus", "College English learners' spoken English corpus", "Chinese students oral English corpus ". College of English at South China normal university has built corpus base with the English-textbook corpus and the English classroom discourse corpus. With the aid of learners' corpora and the target language corpora, we can operate on interlanguage analysis, such as the analysis of the learners' grammar, vocabulary use.

Research based on corpus approach gives English teachers the new methods: teachers base on batch of materials from corpora to analyze, reason, classify, and summarize for conclusions, with less subjectivity and one-sided viewpoints.

The purpose of English teaching is to make the learners' language as close as possible to the native language, while corpus research can provide more representative, authentic native language. Therefore, with corpora's great capacity, advanced technology, compiling suitable Chinese textbooks is imperative. The use of new textbooks must have changed English teaching by combining teaching and corpora together in practice.

F. Corpus can be Used as Test Means for English Textbooks

Using corpus' position retrieval software for word frequency statistics, based on the distribution of certain words in the discourse situation and frequency statistics, we can be more objective to judge the subject of the discourse and the difficulty, and make the trade-offs at quantitative standards. According to the statistical results from corpora, the distribution and arrangement of the high frequency words are relatively stable. If word frequency statistics in a text are in the exception condition, it shows that the discourse lacks representativeness, and it is more special, whether it should be selected to the textbooks should be reconsidered.

Therefore, textbooks without the basis of the corpus is uncompetitive, which eventually will be eliminated by history.

IV. CONCLUSION

With the rapid development of corpus linguistics, related theories and technology are constantly updated. In the theory and technology background, based on corpora, compared with the previous practice, textbooks can be compiled in a more comprehensive, accurate, reliable, descriptive way to describe the real language use.

Because corpora can reveal the most typical features of language, finding the regulation in a language, actually reflecting the reality of the language use. The analysis of the linguistic features of statistical results can, in a large extent, reduce the blindness of English teaching.

Large corpora's capacity and advanced technology are bound to change the traditional English textbooks by combining teaching with the use of real language to adapt to the teaching practice in China.

ACKNOWLEDGMENT

First of all, I'd like to extend my sincere thanks to my colleagues who offer me great help to the research. Moreover, I'd like to extend my sincere respect to all authors whose research I cited in the paper.

REFERENCES

[1] HolmesJ, " Doubt and certainty in ESL textbooks, " Applied Linguistics, vol. 468, pp. 182-195, September 1988.

[2] Pan Fan, " Corpus linguistics and transformation of language teaching thought ," Foreign language journal, vol. 238, pp. 82-85, August 2000.

[3] Sun Qibiao, " Evaluation of the college English integrated course (new edition) - a corpus assisted college English teaching material evaluation, " Journal of Hefei university of technology (Social science edition), , vol. 86, pp. 137-140, October 2009.

[4] Tang Liling, " KWIC function of *New horizon college English* textbooks corpus and data driven learning model," Electric power education in China, vol. 147, pp.206-270, October 2009.

[5] Wang Jianxin, "Several important stages in the history of the development of corpus linguistics," Foreign language teaching and research, vol 116, pp. 52-59, October 1998.

[6] Xue Xiaojuan, "COCA-based study on two synonyms 'Confused' and 'puzzled' ," overseas English, vol.120, pp. 263-266, September 2013.

[7] Yang Anwen, "New college English teaching corpus construction research ," Journal of southwest university for nationalities (humanities and social science edition), vol. 104, pp. 180-188, October 2011.

[8] Zhai GongHua, "Corpus linguistics and its application," Journal of Shandong university of science and technology (Social science edition), vol 24, pp. 100-106, December 2004.