Corpus Resources and Their Use in English Teaching

Qiang Cai, Jianping Zhang

Faculty of Foreign Studies

Jiangxi University of Science and Technology

Ganzhou City, the People's Republic of China

caiqiang628@163.com ; zping1992@163.com

Keywords: corpus resources; English teaching; various corpora; software tools

Abstract. Corpus linguistics is a newly developed subject with its specific characteristics and can be widely used in many aspects of language research and application. This paper briefly introduces the main features of corpus and analyzes its resources in details from corpora to software tools. Then it sums up their use in English teaching practice and inter-language study**.** The result of this research can help the English teachers and learners to get a better understanding of the corpus resources and use them more efficiently in language research and teaching practice.

Introduction

So far English teaching in China has made much progress, while such aspects like teaching methods, syllabus and teaching materials are still following some old modes, relying more on intuition and traditional knowledge of the language which affect negatively the efficiency of teaching. Therefore, English teachers have to explore more effective methods in their teaching and keep eyes open on some advanced research findings.

Since the 1960s, due to the rapid development of computer technology and its application in language study, more and more researchers are getting engaged in corpus establishment and relevant research. A corpus (plural corpora) refers to a collection of linguistic data, either compiled as written texts or as a transcription of recorded speech stored in an electronic database. Usually it is made of a batch of collected language materials for some purpose, following a certain sampling principle and certain classification method. After data getting input and annotated or coded, users can use retrieval software for language retrieving and statistics analyzing.

In China the study of corpus began in 1980s and the researchers have made great efforts to introduce corpus knowledge and have established various corpora. Their achievements help supply language research and teaching with efficient tools, enriching teaching resources and offering full reference for dictionary and textbook compiling. However, it is a pity that the application of corpus is still limited only in some researchers' study and not known and used widely enough to promote the development of English teaching. Considering such condition, this paper aims to deal with the following key issues: what resources does corpus have? How to use them into English teaching in China?

Features of Corpus

The reasons why corpus can be applied in English teaching lie in its many typical features, which are also of its advantages.

The first feature of corpus is its empirical approach to the description of authentic language. The collected corpus is from real life or discourse, not from subjective feelings or intuition. Therefore, the materials for language learning and corpus-based research are more realistic and reliable. The difference between intuition-based data, experimental-based data and corpus-based data lies in the fact that corpus-based data has "the richness of the evidence and the confidence we can have in the generality of that evidence, in its validity and reliability" (Kennedy 2000:8). The authenticity of

Published by Atlantis Press, Paris, France.

corpus can help learners to build real, diverse learning scenarios and get access to true language. And after learning about the laws and regulations of successful English users and constructing their own cognitive schemata, they can improve their language proficiency and solve the practical problems in their English learning.

Computability of data is the second feature of corpus or corpus linguistics. By means of computer, huge amounts of data can be processed with "incredible speed, total accountability, accurate reproducibility and statistical reliability" (Kennedy 2000:5). And at present, a variety of software has been developed for corpus retrieval.

In addition corpus has some other features like typicality of information, variety and timeliness of its data and in its development researchers have much autonomy etc.

Resources of Corpus

Since corpus has so many above-mentioned features or advantages, the researchers and users can explore and apply it in many needed aspects. Generally we can divide the resources of corpus into two parts, one is the various established corpora and those to be established. The other is all kinds of software programs or so called tools which can be used in corpus establishment and application.

Corpora as Resources

Different types of corpora

According to different criteria and purposes, corpora can be classified into different types. Hunston (2002:14-16) categorizes the corpus as following:

Specialized corpus is a corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essay written by students etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language.

General corpus is a corpus of texts of many types. It is unlikely to be representative of any particular "whole", but will include as wide a spread of texts as possible. It is also sometimes called a reference corpus because it is often used to produce reference materials for language learning or translation and it is often used as a baseline in comparison with more specialized corpora.

Comparable corpora are corpora which have two or more corpora in different languages or in different varieties of a language. Comparable corpora of different languages can be used by translators and by learners to identify differences and equivalences in each language. Comparable corpora of varieties of the same language can be used to compare varieties.

Parallel corpora are two or more corpora in different languages, each containing texts that have been translated from one language into the other, or texts that have been produced simultaneously in two or more languages.

Historical or diachronic corpus uses texts from different periods of time. The purpose of design such kind of corpus is to trace the development of aspects of a language over the time line.

Besides, there are other types of corpus like Monitor corpus, Learner corpus and Pedagogic corpora.

Some corpora abroad and in China

There are many existing corpora abroad and in China which the researchers and ordinary users can get access to for various purposes. Here are some of the well-known ones.

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the *BNC XML Edition*, released in 2007. Researchers or users can do the concordance or other study by getting access to it at http://corpus.byu.edu/bnc/.

The Corpus of Contemporary American English（COCA） is the first large corpus of American English, and it is freely available online at http://corpus.byu.edu/coca/. It contains more than 360 million words of text, including 20 million words each year from 1990-2007, and it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. The corpus will also be

Published by Atlantis Press, Paris, France.

updated at least twice each year from this point on, and will therefore serve as a unique record of linguistic changes in American English.

Chinese Learner English Corpus (CLEC) is a one million word corpus covering the texts from five kinds of English learners of high school students, English and non-English major college students at different grades. The CLEC has mistakes tagged. Its purpose is to make a precise description of Chinese Learners' English. It can observe the various characteristics and speech errors and provide useful feedback for English teaching with the help of quantitative and qualitative approaches. CLEC search online is at http://www.clal.org.cn/corpus/EngSearchEngine.aspx.

Spoken and Written English Corpus of Chinese Learners (SWECCL) consists of two corpora, one is SECCL, a spoken English corpus of Chinese college students with one million words of which the data are from oral English tests of English major college students; the other is WECCL, Written English Corpus of Chinese Learners with one million words of which the data are compositions of different topics of English major college students. SWECCL has published by Foreign Language Teaching and Research Press and the website for Colligation is at http://www.fleric.org.cn/corpora/.

Besides those above-mentioned corpora, there are many corpora of other categories too, such as The Brown Corpus, International Corpus of English (ICE), MICASE (Michigan Corpus of Academic Spoken English), The Bank of English etc. Many of those are available on line and some of them are for free while some others require the users to buy a site license in some limited time. Some other corpora come on CDs, either for institutional use or personal use.

To create a corpus on one's own

Although some corpora are well-designed and widely used with excellent results, they may have some limitations and are not suitable to meet some specific demands and for some purposes. Then it is necessary for the researchers to create a corpus on their own.

Usually the creation or compilation of a corpus contains three main stages: corpus design, text collection or capture, and text encoding. The creator has to decide what kind of corpus is to be created and what resources (time, money, knowledge) are at disposal. In corpus design the creator has to deal with such general issues like type, content, structure and size of the corpus. Once the design of a corpus is determined, the intended texts should be captured. The creator should make efforts to make his corpus as representative as possible of the language from which it is chosen and keep the balance in corpus collection. The last stage of creating a corpus is text encoding which includes corpus markup and annotation. "Corpus markup is a basic step in corpus construction", providing "textual (e.g. paragraph and sentence) and contextual information (e.g. text type, speaker gender and bibliographic source)" (Mc Enery et al. 2006). Annotation refers to all the additional information that is put on the texts in order to help the researcher to retrieve as much relevant information as possible. Corpus annotation can take many forms such as POS tagging, parsing, semantic annotation and so on.

Software Tools of Corpus as Resources

As corpus linguistics is developing so fast that more and more corpus theories and software have been explored for the study of language features and beyond.

Computer software used in corpus linguistics can be basically divided into three kinds: concordance programs, coders, and specifically made programs used to answer certain research questions. Concordance programs are search engines which give the result of the search as text samples. The search engines can offer the users text samples of a specific word, a collocation, a lemma or a syntactic construction, showing their location and frequency in exact context. Coders refer to the programs which are made particularly to tag raw text material for variables like morphological properties, word category, lemmas, or even syntactic functions.

Concordance programs and coders are available as either freeware or commercial software. Here we introduce some of them. Wordsmith software is one of the most commonly used tools in corpus study. It is designed by Mike Scott of University of Liverpool and published by Oxford University Press. Wordsmith is based on Microsoft Windows operating system with better features and more friendly after several upgrades. Like most corpus software, it reads plain text files and its three main functions are concordance, wordlist and keywords. Wordsmith software is a commercial one. The latest 5.0 version is available online for download at http://www.Lexically. Net / wordsmith /. With payment the users can get registration number and after registration they can obtain the genuine software. AntConc is a free and green tool developed by the Japanese scholar Laurence Anthony. It has the same functions of concordance, wordlist and keywords and as reliable as Wordsmith with some evidence of some comparison and investigation by Wang Chunyan (2009). And the mostly used speech tagging codes are Tree Tagger and CLAWS.

Of course if the research project is specific and concordances and coders can not be of any use for the exact purpose, then the researchers or users have to design their own computer program, or have it done by a computer expert.

The Use of Corpus Resources in English Teaching

Use of the English corpus resources for various language researches can further reveal the laws and regulations of the language and help English teaching and learning. To be exact, with the research on classroom language teachers can improve their awareness and sensitivity of English; and through the study of inter-language of learners it can help teachers understand the laws of learning and accept some scientific and rational teaching methods. Indeed the use of corpus resources in English teaching can bring along some new ideas to teaching practice and improve teaching methods which plays an important role in English teaching development.

Use in Teaching Practice

Corpus resources can be used in many aspects in English teaching practice. With the reference of English language description based on Corpus Linguistics research, People can formulate and revise curricula in a more scientific way, compile teaching materials more rationally and establish the wordlist more accurately. An important example is the data-driven learning (DDL) promoted by Johns (1991). DDL is a kind of "discovery learning" approach by which students should bring their questions to the corpus to discover their own answers. This new mode of learning and teaching will stimulate the students’ motivation and encourage them actively involved in English learning.

Teachers can use corpus resources in teaching of writing. Through the analysis of student errors in writing or with a comparison between a students’ composition corpus and a English native speakers’ corpus, teachers can get a better understanding of the influence which Chinese has on the students’ writing and they can get much advice and reference for the teaching practice. The teachers can also use corpus resources to improve the writing assessment model. For example, Lou Baocui (2001) studied the phenomenon of coinage in Chinese students’ composition by using the sub-corpus of Chinese Learner English Corpus and put forward some advice on the attitude and teaching measures which the teachers should take in their teaching.

The use of spoken English corpus also includes many aspects like small word research, prosody research, students’ communication strategies research, some sentence structure research etc. Extensive studies of the oral English can guide the teachers to a better teaching and improve the students’ speaking ability. For example, He Lianzhen et al (2004) studied the communication strategies used in the Oral English Exam of the College English Test, by using the relevant corpus and found the significant effects of oral English level on the ideas of communication strategies and the use of them. The results of the study can bring along much inspiration in developing the students’ communicating abilities.

Use in Inter-language Study

Another important aspect of the use of the Corpus resources in English Teaching is the corpus-based English learners’ inter-language study. Through the establishment of English learners’ written and spoken corpora, using the appropriate search tools, teachers can find out some common intermediate language problems, then with such information feed-back they can adjust and improve the curriculum development, teaching materials compilation and teaching practice, thus making English teaching more targeted and more effective.

At present, the English learners’ inter-language study has made considerable progress and established some large English learners’ corpora such as Long-man Learners' Corpus with 10 million words by Longman Publishing Group and the Cambridge Learners' Corpus with15 million words by

Cambridge University Press. In recent years, many inter- language research findings based on English learners’ corpora have been published. They include word frequency of English learners of different native language background, (Ringbom, 1998), learners’ discourse features (Granger & Rayson, 1998) and the value of Learners’ corpus data in the syllabus development (Meunier, 2002) and so on.

Conclusion

Corpus is a huge data base with real life language information which is stored in an electronic form for computer retrieval and research use. With such unique advantages as large capacity, the real life data and its faster and more accurate retrieval, the use of corpus resources is playing an increasingly important role in modern English education. In conclusion, corpus is not only a new research method, but also a new way of thinking, which can greatly enrich the language learning resources, reflect the new learning concept, highlight the dominant position of students and enhance their language awareness and self-learning ability. Therefore, we should explore and develop the corpus resources to increase the corpus application and use the advanced technology tools to do greater contribution for China's English teaching by investing more human and material resources into it.

References

[1] Granger, S. & P. Rayson. Automatic profiling of learner texts[A]. Learner English on Computer [C]. Ed. Granger. London and New York: Addison Wesley Long (1998)

[2] He Lianzhen. Corpus-based Investigation into Communication Strategies in CET – SET. Foreign Languages Research(1) (2004):60-65

[3] Hunston, S. Corpora in Applied Linguistics. Oxford: Oxford University Press (2002)

[4] Johns, T. Should you be persuaded-two examples of data-driven learning materials. English Language Research Journal.(4) (1991): 1-16.

[5] Kennedy, G. An Introduction to Corpus Linguistics.Beijing: Foreign Language Teaching and Research Press. (2000).

[6] Lou Baocui An analysis of coinages in Chinese learners composition. Foreign Language Teaching and Research(1) (2001): 63-68

[7] McEnery T. R. Xiao and Y. Tono. Corpus-Based Language Studies. Abingdon: Routledge (2006)

[8] Meunier, F. The pedagogical value of native and learner corpora in EFL grammar teaching .Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching[C].Ed. Granger et al. Amsterdam /Philadelphia: John Benjamins Publishing Company (2002)

[9] Ringbom, H. Vocabulary frequencies in advanced learner English: a cross-linguistic approach. Learner English on Computer. London and New York: Addison Wesley Longman (1998)

[10] Wang Chunyan. Applications of AntConc in Foreign Language Teaching and Research Computer-assisted Foreign Language Education (1) (2009):45-48

Published by Atlantis Press, Paris, France.