On the Application of Corpus of Contemporary American English in Vocabulary Instruction

Xu Yusu[1]

[1] School of Foreign Languages, Zhejiang Gongshang University, Hangzhou, China

Correspondence: Xu Yusu, School of Foreign Languages, Zhejiang Gongshang University, Hangzhou, China. E-mail: xuyusu@126.com

Abstract

The development of corpus linguistics has laid theoretical foundation and provided technical support for breaking the bottleneck in traditional vocabulary instruction in China. Corpora allow access to authentic data and show frequency patterns of words and grammar construction. Such patterns can be used to improve language materials or to directly teach students. Therefore, this paper discusses how the Corpus of Contemporary American English (COCA) can be applied in vocabulary instruction in the following four different aspects: part of speech, collocation, morphology and word comparison. The above four aspects of application of COCA in vocabulary instruction and their examples have proved that corpora are robust in teaching.

Keywords: vocabulary instruction, COCA, mini-text, application

1. Introduction

Vocabulary has always been the top priority in English teaching and learning. As it is said, without grammar, one cannot express many things, but without words, one cannot express anything. However, the situation of vocabulary instruction among college students in China is far from being satisfactory.

The traditional vocabulary instruction in China mainly focuses on the presentation of Chinese meaning, part of speech, or at best several example sentences. Due to lack of language environment and sufficient input, students are often only aware of the meaning while ignorant of the word usage, grammatical construction, not to mention semantic and pragmatic patterns (Sun, 2004; Chu & Liu, 2007). In a word, students can recognize the meaning in context, but do not know how to put them into correct and active use in spoken or written English, which is why Chinese students make so much effort in learning English but the outcome is bitterly disappointing.

The development of corpus linguistics has brought breakthrough to such deadlock in vocabulary instruction. It thrives on data to analyze and discover what language speakers do. Large bodies of text reveal these patterns in words, grammar and discourse. When computer aid this process, those texts can be handled in seconds, especially if they are tagged for parts of speech or specific information. Romer (2009) claims that "corpus linguistics can make a difference for language learning and teaching and that it has an immense potential to improve pedagogical practice" (p. 84).

The advantages of corpora can be summarized as follows. First, corpora allow access to authentic data; they show frequency patterns of words and grammar construction. Such patterns can be used to improve language materials or to directly teach students. Second, the unique features of corpora, such as concordance and salience, will help students notice and process words in chunk, which can not only arouse their awareness of collocation, but also facilitate the lexical output. Third, the observation, analysis and interpretation of corpus data by students themselves can promote autonomy, which also gives opportunities for the development of their cognitive skills (Boulton, 2009). Therefore, data-driven instruction helps learners construct the metalinguistic awareness, improve lexical output and conduct autonomous study.

However, despite the above discussed benefits, the application of corpora in vocabulary teaching in China is still in its infancy because teachers are either lack of education concerning corpora or fearful of the unknown technique. Therefore, this paper intends to reveal how the Corpus of Contemporary American English (COCA) can be applied in vocabulary instruction.

2. A Brief Introduction of COCA and Mini-Texts

In order to provide a useful tool for the use of corpus in vocabulary teaching, an appropriate corpus must be selected, particularly one that enables study of the metalinguistic awareness and is user-friendly and accessible. While other free corpora exist, the Corpus of Contemporary American English (COCA), available online since 2008 (www.americancorpus.org), is the largest free English corpus and has significant advantages over other free corpora in terms of vocabulary study (Davies, 2009). First, the large size of COCA gives a sufficient patterning of English lexis and grammar, which will give an appropriate picture of word frequency in terms of how they are actually used. Second, the operation of COCA is so convenient that users do not need any special linguistic knowledge or computer technique to get access to all the resources. Meanwhile, COCA provides detailed instruction for each of its uses. Third, COCA has the benefit of being a balanced corpus in terms of register. It is balanced equally between its five registers of spoken, news, academic, fiction and magazine. Therefore, it gives users a more realistic picture of how and where words are used. Fourth, the texts are classified in terms of time, enabling users to observe the diachronic change of American English for every five years since 1990. What is more, COCA's unique interface allows for features of the metalinguistic awareness to easily be analyzed. The corpus is already tagged for part of speech, and offers easy search for collocates, synonyms, overall frequency and so on. Last but not the least, COCA has the ability to show example sentences simultaneously with frequency searches. These sentences, centering around one key word (or node word) as concordance lines or Key Word in Context (KWIC) lines, serve as ideal input to help students learn how the words fit in grammatically with other words as well as clues to meaning through surrounding words.

However, the direct application of COCA in vocabulary instruction has its limitations. First of all, the demand of computer and internet connection poses challenge for traditional classrooms. Second, the query would generate so many entries that they would baffle students and the processing and analysis would take excessive time. Third, some of the entries would be exceedingly difficult with regard to the students' English level, thus causing unnecessary burden in cognition and impair their confidence and motivation.

Therefore, the result from the query needs to be modified into "mini-text"（Liang, 2009）before being used in the classroom. Otherwise the number of example sentences would be too heavy a burden and unnecessary for students to handle. Besides, the print of mini-texts makes it possible to bring them to traditional classrooms to use. One advantage of COCA is that is provides the tool to select the wanted entries from the query and save them to the list a user creates, which can be screenshot and put into use as mini-text.

The construction of mini-texts should meet the following requirements. First, the mini-texts should include example sentences which can reveal the most frequent uses of the queried words. Flowerdew (2009) points out "knowledge of … relative frequencies can be helpful to language practitioners in deciding what items to teach" (p. 330). Therefore, frequency in corpus helps teachers to decide which example sentences of the queried words to be included in mini-texts to create specialized word lists. Second, according to the Input Hypothesis (Krashen, 1985), learners progress in their knowledge of the language when they comprehend language input that is slightly more advanced than their current level. Thus, teachers should only select an appropriate amount of corpus data that fit students' level of English, avoiding too many new and difficult words. Third, according to different teaching purposes, data from different genres, for example, spoken, academic, news, etc. can be targeted to arouse students' awareness of different language features. Finally, teachers should also consider students' age, interest, and the time they live in, so as to choose data that can resonate with them to arouse their interest and enhance their motivation.

Therefore, in vocabulary instruction, teachers first select from the textbook the new words that need to be explained, and then use COCA to construct mini-texts for their classroom, which are informed by frequency, collocation and add variety in structure and context. The following part is to elaborate on the feasible usage of COCA in vocabulary instruction, with examples included.

3. The Application of COCA in Vocabulary Instruction

3.1 Part of Speech

Part of speech is an important concept in grammar, which can enable students to learn how to use new vocabulary words correctly. What is more, it is prevalent that English words have more than one part of speech for different definitions. Due to the tagging and user-friendly tools of COCA, the corpus can list example sentences around the searching word, which can give students an idea of how the word fits in grammatically with other words as well as clues to meaning through surrounding words. Therefore, COCA can help students identify part of speech knowledge dramatically.

Take the word "trigger" for example. To perform a part-of-speech search, simply choose the color-coded KWIC display button at the top of the screen, insert the word "trigger" into the WORD query box and press SEARCH. The search immediately yields all the example sentences with "trigger" as the node word, all of which are highlighted by two different colors to notify its part of speech, pink for verb and blue for noun. So we see that "trigger" can be used both as verb and noun.

To create the mini-text for "trigger", teachers should select an appropriate number of entries for both of its part of speech. At the same time, teachers should also consider frequency, context diversity, difficulty and appeal to students. First choose one entry, and then type in "trigger" in the box of CREATE NEW LIST, and then click SAVE LIST, the entry is automatically saved in the list of "trigger". Usually ten entries would be included in the mini-text to provide sufficient input for the new word.

3.2 Collocation

In addition to part of speech, students also need to learn collocation. As Firth puts it, "You shall know a word by the company it keeps" (1957, p. 11). And studies show that the recurrence of words in various contexts is essential for students to get to know the correct colligation, semantic prosody and pragmatic pattern. With COCA's ability to present the naturally occurring usage with frequency, students can discover the most authentic collocation of new words. For example, we already know that "trigger" can be used both as noun and verb. But what kind of nouns usually follow "trigger" when it is used as verb? What about its colligation, semantic and pragmatic information?

Therefore, to search for the frequent noun collocates of "trigger" as verb, simply type "trigger.[v*]" into the query box. Collocates of the part of speech "noun" can also be specified by choosing "noun. ALL" from the drop down menu POS LIST, which exempts users from memorizing part of speech tags. To adjust the window of words around "trigger", simply choose "0" and "3" in each box after the COLLOCATES query box. The first number represents the window of words before "trigger" and the second number refers to the window of words after "trigger". Eight appropriate entries from the list of example sentences are chosen to create the so-called mini-text as follows.

Figure 1. Mini-text of "trigger" as verb

Therefore, we can find that the frequent noun collocates of "trigger (v)" are illness and crisis etc., which have negative connotation, or defense and innovation etc., which have positive connotation. As a result, COCA can not only raise the students' metalinguistic knowledge, but also facilitate them to process and memorize collocation in chunk, which helps to develop intuitions and inferences to use the words correctly.

3.3 Morphology

COCA can be further used to raise the students' metalinguistic awareness of morphology, the branch of linguistics that studies word forms. For example, the word "press" means "to act upon with steadily applied weight or force". Adding various prefixes, we can derive new words like "compress, depress, impress, repress", which all contain the meaning of the root word "press". Therefore, if students know how to break down words into parts to find meaning or create new words by attaching affixes, they can activate and optimize what they already know.

For example, "out" as a common prefix to verb, often means "overtake". Through COCA, teachers can get access to all verbs starting with "out" as well as the frequency and example sentences. So we choose KWIC,

insert "out*.[v*]" in the query box, and search for all the wanted verbs. The mini-text is designed as follows, which includes the most frequent words and their complete context. The same way can also be adopted to study words starting with "sub-, trans-, audi-" or ending with "-ology, -tive, -ful".
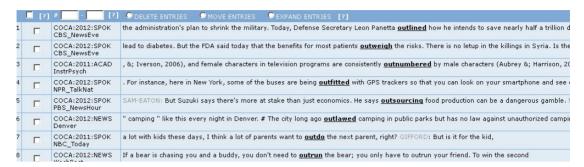

Figure 2. Mini-text of words with prefix "out"

3.4 Word Comparison

A lot of English words are of similar meanings and always cause confusion. Especially after their definitions are translated into Chinese, they tend to be misused by students. For example, "stable" and "steady" almost have the same definition in Chinese. So if students only remember the Chinese meaning without getting to know their uses, mistakes often emerge.

In COCA, users can compare two words or phrases and their differences in meaning by comparing their collocates. First we choose COMPARE, insert "stable" and "steady" in the two query boxes, and input "[nn*]" in COLLOCATES and select "0" and "3". It means that we will compare the three nouns around "stable" and "steady" on the night. To sharpen the contrast between them, the first value of MINIMUN FREQUENCY is set at "10", the second at "0", which means the collocation frequency with "stable" should be above 10 while that with "steady" only 0. And the result is sorted by REVELENCE. The result below shows that the collocates of "stable" and "steady" are apparently different.
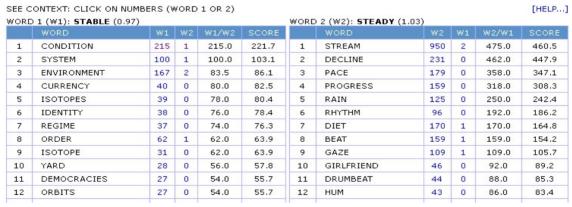

Figure 3. Result of "stable" and "steady" comparison

Obviously, the corpora search can present the collocations of the node word, together with its frequency pattern and context. The respective mini-texts for "stable" and "steady" are designed as follows.

Figure 4. Mini-text of "stable"

Figure 5. Mini-text of "steady"

We can find that "stable", followed by "currency, community, situation, relation", is static and means immobile and unchangeable; while "steady", together with "stream, decline, pace, gaze, rain", is dynamic and emphasizes continuance.

4. Conclusion

The above four dimensions of application of COCA in vocabulary instruction and their examples have proved that corpora are robust in teaching. With authentic data, various context and word frequency, students get access to the most desired learning materials in an instant. Once students have adequate understanding of vocabulary principles and how to use corpora, they can be used for autonomous and individualized study. However, teachers should first guide them through the process, and give them examples to follow. Class projects, homework, and individual tutoring can be used to teach students to gradually explore on their own.

Acknowledgments

References

Boulton, A. (2009). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, *35*(1), 81-106.

Chu, X. Q., & Liu, C. M. (2007). Self-access to advanced English vocabulary acquisition–Corpus-based approach. *Media in Foreign Language Instruction*, (5), 12-17.

Davies, M. (2009). The 385+million word corpus of contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159-190. http://dx.doi.org/10.1075/ijcl.14.2.02dav

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-1955 in Studies in Linguistic Analysis*. Philological Society, Oxford; reprinted In F. Palmer (Ed., 1968), *Selected Papers of J. R. Firth*. Longman, Harlow.

Flowerdew, L. (2009) Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, *14*, 393-417. http://dx.doi.org/10.1075/ijcl.14.3.05flo

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. New York: Longman.

Liang, M. C. (2009). Mini-text and its application in foreign language teaching. *Media in Foreign Language Instruction*, (3), 8-12.

Romer, U. (2009). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and Language Teaching*. Philadelphia: John Benjamins. http://dx.doi.org/10.1075/scl.33.09rom

Sun, H. Y. (2004). A corpus-based study of semantic characteristics of adjective collocations in CLEC. *Modern Foreign Languages*, *27*(4), 410-418.

Copyrights