Pisco Med Publishing

Original Research Article

College English Teaching and Learning from the Perspective of Corpus Linguistics

Qianqian Cai[1] , Hui Di[2]

1.School of English, Tianjin Foreign Studies University (East China), Tianjin, 300204, People's Republic of China

2.Jiangsu Lidian Energy Group, Wuxi city, Jiangsu province, 214400, People's Republic of China

Abstract: Corpus linguistics is deemed as a methodology widely used in the field and theory of linguistics. The author will represent the points of general knowledge of corpus linguistics in this paper. In addition, the potential applications of it in English education are also included.

Keywords: Corpus linguistics; General knowledge; Potential applications; English education

1. Introduction

Bowker and Pearson (2002:230) proposed that a corpus could be captured as a group of authentic texts collected in electronic form based on a specific set of criteria."Authentic" means an example of real, naturally-occurring as Sinclair said and "live" language which can be used in everyday life. "Specific criteria" refers to the selection of texts from a corpus based on clear standards to serve as representative samples of a particular language or subset of that language (Sinclair, 1996; Bowker & Pearson, 2002:10). Corpus linguistics can be generally regarded as a methodology rather than an independent branch of linguistics. In this paper, Corpus linguistics is deemed as a methodology widely used in the field and theory of linguistics. The author will propose the points of basic items and tools of corpus linguistics. In addition, the potential applications of it in English education and statistics are also included.

2. Basic items and concepts

2.1 Markup

We usually have to construct information in the corpus first before extracting information from it. Two types of process for building information in a corpus are: makeup and annotation.

Markup characterizes a sequence of characters or other symbols inserted at exact points in a text or word processing file to show how the file will assume when printed or demonstrated, or to depict the logical structure of a document. Markup in the printed texts can be useful when we want to keep the original structural and appearance information of a text. We can simply design a coding system to mark up it.

2.2 Annotation

Garside, Leech and Mcenery (1997) posited that annotation was to add explanatory, linguistic information to an electronic corpus of linguistic data (spoken and/or written). An embodiment of corpus annotation is grammatical tagging (part-of-speech tagging or POS tagging), where a label or tag related to a word can demonstrate its syntactic type. For instance, in really_RR, the tag RR means really is an adverb.

When a corpus has been tagged, all of the implicit speech information is made explicit and accessible especially the grammatical role of every single word, thus more detailed searches become possible. A tagged corpus is without doubt a beneficial text resource, particularly when research interest is the study of grammatical features and patterns.

2.3 Tokens and types

A token is sometimes a word. A sentence with 20 words can include 20 tokens. But in that sentence, not all tokens are different. Maybe all the tokens can be classified into 17 different types of tokens.

Therefore, types are unique tokens, and one type should be different from another. The Standardized type/token ratio can help to get more information about a long text.

2.4 Mean word length

Mean word length can be beneficial in text categorization. "mean" is "average" and "word length" refers to how many characters a word embodies.

Word lengths of different texts can be compared so as to find which text has more long words. This belongs to the lexical variation which we can use to predict the category of a specific text. For example, when the mean word length of one text is longer than another, we may guess that the former can be materials more academic.

doi: 10.18282/l-e.v10i2.2279

3. Tools, applications and statistics

3.1 WordSmith

WordSmith and AntConc are two main retrieval tools in corpus linguistics. The former is much more sophisticated and powerful than the latter. The author will exclusively illustrate WordSmith which includes three main tools.

3.1.1 WordList

This tool visualizes a series of all the words or clusters in a text alphabetically or in frequency order. If in frequency order, to produce a frequency list, the WordList tool generates a list of all the different types (word forms) from the text based on their frequency of occurrence. This function can be beneficial in calculating the frequency of English words. According to the statement of a study, the first 700 high frequency words cover 70% of English usage, namely 70% of people's daily English listening, speaking, reading and writing originate from those 700 high frequency words. Therefore, teachers should attach importance to high frequency words in English teaching,.

If in alphabetical order, we can easily find any word we want by tapping in the word needed. An alphabetically-ordered list can be helpful in many ways. Firstly, it conveys information about how often each word-form occurs, so that we can find its frequency information easily. Secondly, All types with the same initials are gathered up, making it easy to access a specific word form with similar initials. Thirdly, because related word-forms will often appear in the vicinity of each other, types such as strength, strengthen, strengthened, strengths are listed in close proximity. Therefore, it enables us to see whether all forms of a particular lemma are actually used in a corpus. Alphabetical lists can also be ordered by word endings with the reverse order facility, which helps to recognize word families including nouns, verbs, adverbs, adjectives) with the same suffix.

In addition, WordList could produce lists of chunks and clusters ranked by frequency, in alphabetical order or in reverse alphabetical order. If English learners have a good grasp of the chunks and clusters used by native speakers, their language production will be smooth and native like. A spoken corpus of successful users of English (SUE) could also be produced for students to learn the English Lingua Fr anca (ELF), which better meets the linguistic reality nowadays (Huang Ruoyu & He Gaoda, 2009).

Moreover, the text can add function "Stop list" to delete any words we do not want. For example, after making a word list of a current text, we can simply put all the function words into the Stop list, then start Wordsmith tools based on a specific order and a text without function words can be seen. In English writing teaching, teachers can harness this function to produce a list of conjunction words for use of students' writing structure.

3.1.2 Concord

Concord produces a word index, which helps visualize any word or phrase in context and appreciate the word collocates. Concord could also calculate and illustrate the collocation of the highest frequency. Collocations are words that usually appears near the search term. The word collocation patterns can be produced through the Concord to find the actual words in habitual company, which could forward the word teaching in English education. In that case, the Concord could also further the acquisition of all meanings of a particular word. However, that is dificult for the single word list because the words in it are all out of context.

3.1.3 KeyWords

This tool enables us to find the key words in a text (KWIC). When one text or corpus is compared with another, it will recognize words with highly frequent occurrence and can serve for discourse analysis or statistic research. For instance, a text about "business" may characterize keywords such as profit, shares, stock, and company. Here KeyWords can help us analyze and study these texts.

Moreover, teachers can use KeyWords to make up a writing fodder about the necessary words and structure for students to learn and practice. And KeyWords can help enrich the contents of students' writing in this way.

3.2 Statistics in corpus analysis

Collocation can be observed in any context informally, but a statistical measurement of it is more trustworthy. Statistics in corpus analysis can be used for the judgement on whether two words in a corpus can form collocations or if a word or a structure is used more frequently in one text than another.

The chi-squared test probably serve as the most commonly used significance test in corpus linguistics (Lee, 2006) and it has two advantages: (1) there is no assumption of normal distribution which the linguistic data often fail to meet; (2) it is easy to calculate, even without a computer statistics package. However, chi-square test does not work well with frequency less than 5 while Log likelihood does. The chi-square score changes drastically when two corpora differ greatly in size and its value usually goes extremely high.

4. Conclusions

Corpus linguistics exerts a great influence on English learning and teaching. English education based on corpus is a new subject in the field of language teaching research, which can provide students with authentic language materials and achieve the purpose of learning English.

English education based on corpus helps actualize student status as the subject and emphasizes teachers' instructing and facilitating role. Teachers should impart knowledge about how to utilize corpus and basic tools and competence educaion can be possible.

References:

[1] Bowker, L., & Pearson, J. (2002). Working with specialized language: a practical guide to using corpora. Routledge.

[2] Sinclair, J. (1996). EAGLES: Preliminary recommendations on corpus typology. Retrieved March 20, 2019 from.http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

[3] Garside, R., Leech, G. N., & Mcenery, A. M. (1997). Corpus annotation: linguistic information from computer text corpora. Taylor & Francis.

[4] Lee, H. (2006). Parallel Optimization in Case Systems: Evidence from Case Ellipsis in Korean*. Journal of East Asian Linguistics, 15, 69-96.

[5] Hu, C., & Yang, B. (2015). Using Sketch Engine to Investigate Synonymous Verbs. International Journal of English Linguistics, 5, 29.