DOI: 10.23977/langl.2023.061317
ISSN 2523-5869 Vol. 6 Num. 13

A Study on the English Vocabulary Learning Based on the Application of Corpus Tools

Xiaomin Han

School of Foreign Language, Guangzhou Peizheng College, 53 Peizheng Road, Chini Town, Huadu District, Guangzhou, China
244591062@qq.com

Keywords: Corpus, AntConc, Range, vocabulary difficulty

Abstract: Skilled vocabulary use has long played an essential role in English reading, but vocabulary learning is also a hindrance for many learners. This paper suggests that corpus tools can be of great use in this regard. By storing large amounts of data and running it at a fast speed, corpus tools can illustrate to learners how vocabulary is used in different contexts, thus enhancing learning effect. The corpus tools can also be used to rate the vocabulary difficulty of a text, allowing learners to make specific target of their vocabulary learning. This paper will exemplify this point with an authentic IELTS reading text. The study of corpus tools will undoubtedly help to improve the effectiveness of English vocabulary learning, thereby developing learners' reading skills.

1. Introduction

Vocabulary learning has always been a vital part in improving English reading skills. Yet it is also a daunting task facing many English learners. Despite investing a large amount of time and effort in vocabulary memorization, many learners still find it hard to gain desirable results. Compared to grammar and syntax, vocabulary is more difficult to break through in a short period of time because of its sheer volume and variation, compounded by the fact that most people learn vocabulary by rote, which makes it even harder for them to make progress in vocabulary learning. As a result, learners often fail to understand the reading text in depth.

The development of computer technology has brought a new prospect to this dilemma. The corpus research has enabled language learners to make full use of real-life language resources to grasp the patterns of vocabulary use, which is in line with the data-driven learning (DDL) model proposed by Tim Johns, a concept highlighting an active mode of learning through the steps of "observe, describe, summarize and interpret" by offering students a large number of authentic language learning texts and teachers' assistance[1]. This highly proactive approach usually contributes to better learning outcomes. In this study, IELTS reading texts will serve as an example to illustrate how the corpus tools AntConc and Range can help with vocabulary learning.

2. The Main Features of IELTS Reading Texts

IELTS, known as the International English Language Testing System, is an international standardized English language test administered by the British Council, the University of Cambridge Examinations and the Australian Education International Development Agency (IDP). It is designed to test the English proficiency of people intended to study or migrate to English-speaking countries. The test system has gained worldwide popularity along with the acceleration of global integration, and the results achieved by candidates are accepted by more and more countries, making it an extremely influential international English test, together with TOEFL.

In a single sitting, IELTS Reading test contains a total of three essays, with a word count of around 700-1000 for each piece. These essays cover a wide range of content, covering a variety of science and humanities subjects. As a result, the vocabulary scope is also large and wide-ranging, forming an intimidating barrier to reading comprehension. The experience of many candidates has shown that the amount of vocabulary they command can make a huge difference to the outcome of the test. While language learners generally can understand the importance of vocabulary, the methods they often use, like rote memorization, fail to yield good results. Part of the reason is that vocabulary books are often arranged in alphabetical order rather than the frequency of use, which makes it difficult for learners to prioritize their vocabulary learning. In this study, common corpus tools are introduced to analyze the reading texts of IELTS, with the hope to facilitate vocabulary learning.

3. Corpus Collection and Corpus Tools

The word corpus (plural corpora) is derived from a Latin word, meaning “body”, and today corpus is more regarded as “electronic text collections”. These are the electronic collections of texts that represent a language or a variant of a language, collected according to certain sampling criteria[2]. Sinclair, a leading linguist, argues that corpus collections must be used primarily for generalizing language behaviors [3]. The corpus in essence is a large amount of authentic linguistic data, the analysis of which can provide learners with a more accurate picture of the main patterns of the concerning language, thus assisting learning.

This corpus featuring in this study consists of 132 reading essays in total selected from the IELTS Textbooks published by Cambridge University Press. In order to improve the accuracy of the research results, the author has removed the questions parts from the reading essays and kept the texts only. In addition, two commonly used corpus tools were applied in this research. The first one is AntConc developed by Professor Laurence Anthony of Waseda University. The tool is mainly used to analyze the frequency and distribution of words and phrases in the text. The other is Range, developed by Professors Paul Nation and Averil Coxhead in New Zealand, which is used to measure the difficulty of words.

4. Research Design and Result Analysis

4.1 The application of AntConc to identify vocabulary frequency

The latest version of AntConc has a number of useful features for vocabulary analysis, one of which is Word (also known as “word list generation”). This function allows users to rank the frequency of words in the text and to identify the frequency of the words used. Furthermore, if the words are re-ranked according to the range of distribution, a more complete picture of the use of high-frequency words can be obtained. This is because some words, though used frequently, may be restricted to specific texts. Only when the distribution of vocabulary use is also taken into account can the frequency of certain words be more accurately reflected.

After converting the 132 texts into the required format to form a self-built IELTS reading corpus, the author was able to use AntConc's Word function to obtain a word frequency table. However, it can be seen that many of the words in the front rows are common articles and prepositions like "a", "the", "and", "for", etc., which are not of much meaning. It is therefore necessary to continue to examine the table in order to obtain more valuable information. For example, in column 89 of the frequency table, the word is "used", with a frequency of 126 times in altogether 67 texts, which shows a wide range of word usage. Learners may wonder why "used" is so frequently applied in academic essays. To answer this question, we can double-click on the word, which would lead us to the KWIC (Keyword in Context) page, showing the original sentences in the texts as well as the more specific usage. In this case, it can be seen that the word "used" is most often followed by prepositions such as "in", "as", "for", "by" etc. A closer look at these sentences presented in KWIC also reveals further details. For example, the word "used" is mostly followed by the word "to", with 29 cases (see Table 1). But they are not of the same meaning. In most cases, "used to" indicates the passive voice, as in "it is used to add emotion and rhythm", which shows an infinitive structure to express purpose. Apart from this, it can also reflect the notion of regular past action (used to do sth), and a sort of habitual behaviors (be used to doing sth), but with much less frequency of 5 cases and 1 case respectively. Examples of these usages can be found in the following table.

Table 1: Part of the search result of "used" in AntConc

8	1142.txt	We are aware that it	**used**	to add emotion and rhythm.
9	1272.txt	knowledge regarding geography and methods	**used**	to analyze and interpret geographical
10	543.txt	and secondly, it is	**used**	to assist growth in plant.
11	843.txt	Baits can be	**used**	to attract and concentrate foragers
12	843.txt	is that they can be used	**used**	to collect over a period of time
13	911.txt	his purple solution could be	**used**	to colour fabric,
14	1422.txt	and epidemics for centuries, and were	**used**	to explain the spread of infection
15	1541.txt	can grow up quickly because it's	**used**	to exploiting water when it arrives?
16	623.txt	The ancient Gothic word for ten, tachund, is	**used**	to express the numbers 100 as
17	532.txt	hold back virtually all of the sediments that	**used**	to flow down the river.

As is seen, "used" is mostly applied in the passive case, and can be followed by a great variety of prepositions to express different meanings. On the contrary, it is used less often in the active form, showing past tense.

The KWIC interface presents some examples of sentences, but if learners do not understand them well enough and need to browse the whole sentence, they can click on the sentence and AntConc will jump to the File interface with the full text, making it easier for learners to see how the word is used in specific contexts.

Another example is the usage of "while". Drawing from my teaching experience, the word "while" is most often remembered by students as a conjunction indicating time, similar to "when", but a search of IELTS reading texts in AntConc shows that in academic essays, "while" is most often used to indicate a contrast. If students interpret "while" as an expression of time, their comprehension of the essay could be compromised.

Therefore, a combination usage of Word and KWIC in AntConc allows learners to more accurately identify high-frequency vocabulary and the different uses of these words, which can be a great help in improving reading skills. A major difficulty in reading is that students do not have a profound understanding of the different meanings of a word in various contexts, causing miscomprehension. Taking the advantage of fast data processing, corpus tools can generate word list based on frequency and categorize different usages into specific groups, allowing learners to compare and to distinguish.

4.2 The use of Range to analyze vocabulary difficulty and identify academic vocabulary

Range is a handy corpus tool for analyzing vocabulary difficulty. The principle of its operation is not complicated. It is based on comparing the words in a given text with a ready-made glossary and then analyzing the use of words in the text by looking at which words appear or do not appear in the glossary and the percentage of words that appear in the glossary[4].

The tool comes with it three levels of word lists, in ascending order of difficulty, Basewrd1.txt, Basewrd2.txt and Basewrd3.txt. The primary level contains 998 of the most commonly used word families in English (4119 words). The second level contains 988 common word families, totaling 3708 words. These two word lists cover the most frequently-used English words. The upper level of vocabulary (level 3) is mainly the academic vocabulary, with 570 families of words, totaling 3107 words.

Words that do not appear in the glossary are listed in a separate group. Generally speaking, it is the level 3 word list and the words outside the three word lists that can reflect the difficulty of the text. When the author loaded the first text from the IELTS reading corpus into Range for analysis, the following results were obtained in Table 2.

Table 2: Difficulty classification of IELTS reading text words retrieved by Range.

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	606/75.56	225/60.48	191
two	56/ 6.98	38/10.22	35
three	41/5.11	36/9.68	30
not in the lists	99/ 12.34	73/19.62	?????

This table illustrates the level of vocabulary difficulty in this text. As is seen, the number of word forms in the primary and secondary combined exceeds 80%. The percentages of academic words and words not collected in the word list are 5.11% and 12.34% respectively. The overall figure is close to 20%. Independent analysis of other texts in this corpus were also conducted, with relatively consistent results shown. Obviously, this two categories of the vocabulary tend to be the obstacles blocking students' reading comprehension and could be the highlight of teaching later on.

The advantages of Range are not confined to being an indicator of vocabulary difficulties, but also the sorting of vocabulary information, providing the frequency of word use. Usually the learners' focus should be on the third group of academic vocabulary. The words that are not categorized are often less commonly seen. Some may be the really high-level words; some may be the technical words or proper names including the names of people or places, and it is up to the learner to decide which words are worth studying in depth.

5. Conclusion

The acquisition of vocabulary is a crucial part of learning English. More than affecting reading comprehension, it also exerts an influence over listening, writing and speaking. If students do not have a good command of vocabulary, it will be difficult for them to make a substantial breakthrough in any of their English studies. Grammar and sentence structure are like the framework of a building, while vocabulary is the stacks of bricks of it. A building without bricks is just an empty frame, unable to convey an effective message. In reality, many students do pay full attention to vocabulary learning, but it is difficult for them to achieve good results simply by relying on rote memorization using the common vocabulary books that show the words in an alphabetical order.

The advent of corpus tools can help to solve this problem to some extent. Valuable data can be retrieved quickly through the process of running computing power, allowing learners to immediately understand the characteristics of the use of a particular word in a particular domain (e.g. academic), including its frequency and range of usage. The abundance of contrasting examples also reinforces the learner's impressions on the different usages of the same word, which is much more efficient than traditional dictionary searches. At the same time, the use of corpus tools to categorize vocabulary difficulty gives learners a clearer idea of which words are the focus for breakthroughs and which are less frequently used and thus less important. This makes it easier to improve the focus and efficiency of vocabulary learning.

In conclusion, vocabulary learning is something that will stay with learners for the rest of their lives, and a skillful use of corpus tools will make this process easier and more productive. This is a skill that both teachers and students should be able to acquire and apply.

References

[1] Johns, Tim. Should You be Persuaded: Two Examples of Data-driven Learning [J]. English Language Research Journal, 1991(4).

[2] Liang Maocheng, Li Wenzhong, Xu Jiajin. The Course of Applied Linguistic [M]. Forigen Language Teaching and Research Press, 2010:3

[3] Sinclair, J. Trust the text: Language, Corpus and Discourse [M]. London: Routledge. 2004

[4] Chen Shi. The Corpus Tool Range in Stylistic Research - An Example of the U.S. Presidential Inaugural Speech [J]. Journal of Language and Literature - Foreign Language Education and Teaching, 2009(9): 42-46.