



Journal Website:
<https://theamericanjournals.com/index.php/tajssei>

Copyright: Original content from this work may be used under the terms of the creative commons attributes 4.0 licence.

Research Article

READING ACTIVITIES BASED ON USING RESOURCES OF THE CORPUS LINGUISTICS

Submission Date: March 21, 2023, Accepted Date: March 26, 2023,

Published Date: March 31, 2023 |

Crossref doi: <https://doi.org/10.37547/tajssei/Volume05Issue03-09>

Khursantosh Anvar Kizi Tolmasova

Master Degree Student, Uzbekistan State University Of World Languages, Uzbekistan

ABSTRACT

In this article Corpora, in the sense “databases that contain texts”, are a widely used tool in language research. This tool can be a helpful resource for language teachers and learners as well, and many corpora are available online. In this chapter, Hilde Hasselgard discusses the use of corpora in the English classroom. She explains what a corpus is, and what answers a corpus can provide to questions about language use. Then she presents various possibilities for making use of corpora in language learning

KEYWORDS

Corpus, language teaching, linguistics, make claims, linguistics studies, each, databases that contain texts, particular profession.

INTRODUCTION

In recent years a lot of investigation has been devoted to how computers can facilitate language learning. One specific area on the computer frontier which still remains quite open to exploration is corpus linguistics. Having heard a declaration that corpora will revolutionize language teaching, I became very curious

to find out for myself what corpus studies have to offer the English language teacher and how feasible such an implementation would be. This article will address those questions by examining what corpus linguistics is, how it can be applied to teaching English, and some of the issues involved. Resources are also included

which will assist anyone who is interested in pursuing this line of study further.

LITERATURE REVIEW

The linguist M.A.K. Halliday (1991) argues that “the immense scope of a modern corpus, and the range of computing resources that are available for exploiting it, make up a powerful force for deepening our awareness and understanding of language” (p. 41). Both teachers and learners of English

can benefit greatly from learning how to use a corpus: they will have practically unlimited sources of information on English language use at their

disposal. The corpus can offer invaluable assistance where our intuitions

(even in our first language) are insufficient, for example when it comes to

studying collocations, making sense of ambiguous words and expressions,

and assessing how frequent an expression is compared to another. In fact,

research has shown that people’s intuitions about frequency are unreliable

because we notice what stands out rather than what seems normal. But

in order to benefit from the corpus, we must know how to use it. That

means knowing how to ask good questions, how to find appropriate ways

of searching in the corpus and, most importantly, how to interpret the

search results. Anyone who has acquired these skills will soon discover

that corpus use is addictive. And unlike many other addictions, corpus

linguistics is purely beneficial.

METHODOLOGY

With the purpose to ensure the “comparability” of the text samples under consideration, the project for the creation of the International Corpus of the English language was primarily subordinated. Continuing the traditions of British descriptive linguistics, project manager S. Greenbaum has significantly expanded the traditional practice of registering English speech use, including new materials reflecting regional and dialectal characteristics. In conditions where countries that officially recognize English as a second state language are striving for “linguistic independence” and are considering their own version of the language as a criterion for correct use, it is important that the options do not diverge too far. The creation of an international database was intended to help preserve the identity of at least the written form of the English language [6]. The presence of a large number of texts in electronic form greatly facilitated the task of creating large representative corpora of tens and hundreds of millions of words, but did not eliminate the problems: collecting thousands of texts, removing copyright problems, bringing all texts into a single form. Balancing the corpus by themes and genres take a lot of time.

Representative corpora exist or are being developed for German, Polish, Czech, Slovenian, Finnish, Modern Greek, Armenian, Chinese, Japanese, Bulgarian and other languages.

The first large computer corpus is considered the Brown Corpus (BC, Brown Corpus), which was created in 1960 at Brown University and contained 500 text fragments of 2 thousand words each, which were published in English in the United

States in 1961. As a result, it sets the standard of 1 million tokens for creating representative corpora in other languages. According to a model close to the BC, in the 1970s, a frequency dictionary of the Russian language by L. N. Zazorina built

on the basis of a corpus of texts with a volume of 1 million words was created. It included approximately equal proportions of socio-political texts, fiction, scientific and popular science texts from various fields and drama.

The Russian corpus, created in the 1980s at the Uppsala University, Sweden, was built on a similar model. Its volume numbering of one million words is sufficient for the lexicographic description of only the most frequent words, since words

and grammatical constructions of the average frequency occur several times per million words. From a statistical point of view, a language is a large set of rare events. The British Corpus of the English language Among the projects that are currently intensively developing, a special place is occupied by the British Corpus of the English language, numbering more than 100 million words. It possesses a carefully balanced material, including more than 4 thousand texts, which represent a variety of genres and varieties of language: from spoken English and newspaper articles to full-text novels. The corpus can be used primarily as a source of examples of live speech in teaching English and for research purposes to identify new trends in language development. So, each of such everyday words as polite or sunshine occurs in the bookmaker only 7

times, the expression polite letter only once, and such collocations as polite conversation, smile, request — never.

For these reasons, as well as in connection with the growth of computer power capable of working with large volumes of texts, in the 1980s several attempts were made around the world to create larger corpora.

RESEARCH INSTRUMENTS

As explained above, a corpus is a digital collection of texts. A more precise definition is that a corpus is a structured database of natural texts prepared for use in linguistic research. That is, the texts are not produced in order to be included in a corpus, but in order to communicate in authentic settings. However, the selection of texts and the way in which they are organized have been planned with linguistic research in mind. This means that not all databases of texts are corpora, because they may have been created for other purposes. Examples are archives of newspapers and public documents. When linguists compile a corpus, they typically aim to make the corpus representative: that is, it should give a fair image of the language of a particular period, region or genre, for example. Users of a corpus need to be aware of what the corpus contains and thereby what kind of language it is meant to represent. For example, we cannot make claims about language in literature or in classroom settings based on a corpus that consists exclusively of newspaper text.

In order to conduct a study of language which is corpus based, it is necessary to gain access to a corpus and a concordancing program. A corpus consists of a databank of natural texts, compiled from writing and/or a transcription of recorded speech. A concordancer is a software program which analyzes corpora and lists the results. The main focus of corpus

linguistics is to discover patterns of authentic language use through analysis of actual usage. The aim of a corpus based analysis is not to generate theories of what is possible in the language, such as Chomsky's phrase structure grammar which can generate an infinite number of sentences but which does not account for the probable choices that speakers actually make. Corpus linguistics only concern is the usage patterns of the empirical data and what that reveals to us about language behavior.

One frequently overlooked aspect of language use which is difficult to keep track of without corpus analysis is register. Register consists of varieties of language which are used for different situations. Language can be divided into many registers, which range from the general to the highly specific, depending upon the degree of specificity that is sought. A general register could include fiction, academic prose, newspapers, or casual conversation, whereas a specific register would be sub-registers within academic prose, such as scientific texts, literary criticism, and linguistics studies, each with their own field specific characteristics. Corpus analysis reveals that language often behaves differently according to the register, each with some unique patterns and rules.

RESULTS AND DISCUSSIONS

Even without using corpora ourselves, we can benefit from teaching materials and reference tools that are based on corpus investigations. For example, most major English-language dictionaries are corpus-based. As a result, the dictionaries can give reliable information on the kinds of contexts a word or phrase occurs in; for example whether it is academic or colloquial, rare or frequent. Besides being an excellent source of authentic examples that can show the usage patterns of words and phrases, corpora can be used to define a core vocabulary. Studies have shown that the

2000 most frequent words in a ten-million-word corpus of spoken and written English account for 83% of the text (O'Keefe et al., 2017). This means that a learner who has chapter 23-8 acquired these words will get by in most situations. In addition, it may be helpful to work with specialized corpora to identify and teach slightly less frequent words that are nevertheless useful in the kind of situations and genres that the learners are likely to encounter. One use of specialized corpora is to extract word lists to tailor the vocabulary teaching to specific learner groups and learning purposes. For example, young learners may focus on vocabulary that is frequent in corpora of children's books and everyday conversations in family settings, while those learning for a particular profession (such as cooking, carpentry, engineering, business) could learn vocabulary drawn from corpora of books, magazines, journals and transcribed lessons within their field, or, say, from recorded service encounters if the goal of the learning is to be able to serve customers in English.

CONCLUSION

The linguist M.A.K. Halliday (1991) argues that "the immense scope of a modern corpus, and the range of computing resources that are available for exploiting it, make up a powerful force for deepening our awareness and understanding of language" (p. 41). Both teachers and learners of English can benefit greatly from learning how to use a corpus: they will have practically unlimited sources of information on English language use at their disposal. The corpus can offer invaluable assistance where our intuitions (even in our first language) are insufficient, for example when it comes to studying collocations, making sense of ambiguous words and expressions, and assessing how frequent an expression is compared to another. In fact, research has shown that people's intuitions about frequency are unreliable because we notice what

stands out rather than what seems normal. But in order to benefit from the corpus, we must know how to use it. That means knowing how to ask good questions, how to find appropriate ways of searching in the corpus and, most importantly, how to interpret the search results. Anyone who has acquired these skills will soon discover that corpus use is addictive. And unlike many other addictions, corpus linguistics is purely beneficial.

REFERENCES

1. Biber, D., Conrad, S., & Leech, G. (2002). Longman student grammar of spoken and written English. Longman. Halliday, M.A.K. (1991).
2. Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), English corpus linguistics: Studies in honour of Jan Svartvik (pp. 30–82). Longman.
3. Hasselgård, H., & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), A taste for corpora. In honour of Sylviane Granger (pp. 33–82).
4. Benjamins. Hidalgo, E., Quereda, L., & Santana, J. (Eds.). (2017). Corpora in the foreign language classroom. Rodopi. Johansson, S. (2011). A multilingual outlook of corpora studies. In V. Viana, S. Zyngier, & G. Barnbrook (Eds.) Perspectives on corpus linguistics.
5. Benjamins. Mukherjee, J. (2016). Korpuslinguistik und Englischunterricht. Eine Einführung. Peter Lang. O’Keeffe, A., McCarthy, M. & Carter. R. (2017). From corpus to classroom. Language use and language teaching. Cambridge University Press. Römer, U. (2018). Corpus research applications in second language teaching. Annual Review of Applied Linguistics, 31, 205–25.
6. Sinclair, J. M. (Ed.) (2004). How to use corpora in language teaching. Benjamins.