CORPUS-BASED STUDIES IN LINGUISTICS

Pasenko K.V.
Cherkasy Institute of Banking of the University of Banking of the National Bank of Ukraine (Kyiv),
Cherkasy, 164 V. Chornovil Street, tel. (+380472) 71-99-51,
e-mail: *k.pasenko@mail.ru*

In recent years a lot of investigations have been devoted to how computers can facilitate language learning. In this work we have made an attempt to examine the basic questions: what corpus is and what corpus linguistics is, how it can be applied to teaching English. Over the past twenty years, an approach to the study of language referred to as corpus linguistics has largely become accepted as an important and useful mode of linguistic enquiry [1]. Corpus linguistics is the study of language data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts [7].

The advent of computers led to the creation of what we consider to be modern-day corpora. The first computer-based corpus, the Brown corpus, was created in 1961 and comprised about 1 million words. Today, generalised corpora are hundreds of millions of words in size, and corpus linguistics is making outstanding contributions to the fields of second language research and teaching [2, p. 2].

Electronic (digital) language corpus is a new thing. It has a history of nearly half a century. Therefore, we are yet to come to a common consensus as to what counts as corpus, and how it should be designed, developed, classified, processed and utilised. The uniqueness corpus linguistics lies in its way of using modern computer technology in collection of language data, methods used in processing language databases, techniques used in language data and information retrieval, and strategies used in application of these in all kinds language-related research and development activities [4].

Different problems of corpus studies were investigated in great detail by G. R. Bennett, D. Biber, N. S. Dash, G. Leech, T. McEnery, N. Nesselhauf, T. Virtanen and A. Wilson.

The term "***corpus***" is derived from Latin corpus "body". At present it means representative collection of texts of a given language, dialect or other subset of a language to be used for linguistic analysis. Theoretically, corpus is

(**C**)apable
(**O**)f
(**R**)epresenting
(**P**)otentially
(**U**)nlimited
(**S**)elections of texts.

It is compatible to computer, operational in research and application, representative of the source language, processable by man and machine, unlimited in data, and systematic in formation and representation [3]. The term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings: sampling and representativeness; finite size; machine-readable form; a standard reference [6, p. 29].

There are a number of areas where language corpus is directly used as in *language description, study of syntax, phonetics and phonology, prosody, intonation, morphology, lexicology, semantics, lexicography, discourse, pragmatics, language teaching, language planning, sociolinguistics, psycholinguistics, semiotics, cognitive linguistics, computational linguistics*, etc.

In fact, there is hardly any area of linguistics where corpus has not found its utility. This has been possible due to great possibilities offered by computer in collecting, storing, and processing natural language databases. The availability of computers and machine-readable corpora has made it possible to get data quickly and easily and also to have this data presented in a format suitable for analysis corpus as knowledge resource:

1) corpus is used for developing multilingual libraries;
2) designing course books for language teaching;
3) compiling monolingual dictionaries and thesaurus (printed and electronic);
4) developing bilingual and multilingual dictionaries (printed and electronic);
5) various reference materials (printed and electronic version);
6) developing machine readable dictionaries (MRDs);
7) developing multilingual lexical resources, electronic dictionary [4].

According to some researchers, 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 concordance is a software program which analyses corpora and lists the results. The main focus of corpus linguistics is to discover patterns of authentic language use through analysis of actual usage [5].

Corpus linguistics is a comparatively new field of language research and application. Corpus linguistics is perhaps best described for the moment in simple terms as a study of language based on examples of real life language use. It has a long and interesting history. Yet the term corpus linguistics is a relatively modern term [6]. Corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics and so on. Corpus linguistics does, however, allow us to differentiate between approaches taken to the study of language and, in that respect, it does define an area of linguistics or, at least, a series of areas of linguistics. Hence we have corpus-based syntax as opposed to non-corpus-based syntax, corpus-based semantics as opposed to non-corpus-based semantics and so on. So, while corpus linguistics is not an area of linguistic enquiry in itself, it does, at least, allow us to discriminate between methodological approaches taken to the same area of enquiry by different groups, individuals or studies [6].

It is generally accepted that the invention and advancement of computer technology in the last century has eventually added a new dimension to the field of linguistics. In recent times, as a result of this innovation, a comparatively new field, namely, the Computational Linguistics has evolved as an important area of Artificial

Intelligence, which aims at looking at language as an essential instrument of human communication directly linked with human cognition [9]. It is a method of carrying out linguistic analyses. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most wide-spread methods of linguistic investigation in recent years [8].

Corpus linguistics is a multidimensional area. It is an area with a wide spectrum for encompassing all diversities of language use in all domains of linguistic interaction, communication, and comprehension. The introduction of corpus in language study and application has incorporated a new dimension to linguistics.

Corpus linguistics is an approach that aims at investigating language and all its properties by analysing large collections of text samples. Corpora, and other (non-representative) collections of machine readable text, now mean that the lexicographer can sit at a computer terminal and call up all the examples of the usage of a word or phrase from many millions of words of text in a few seconds. This means not only that dictionaries can be produced and revised much more quickly than before − thus providing more up-to-date information about the language − but also that the definitions can (hopefully) be more complete and precise [6].

Corpus examples are important in language learning as they expose students at an early stage in the learning process to the kinds of sentences and vocabulary which they will encounter in reading genuine texts in the language or in using the language in real communicative situations.

Corpora have been used not only in language teaching but also in the teaching of linguistics. The link between findings of corpus-based research and (foreign) language teaching is that corpus evidence suggests which language items and processes are most likely to be encountered by language users (what is frequent and typical) and may thus deserve more time in classroom instruction. Corpora and corpus-data 1) help teachers and students make better informed decisions and improve teaching material to become more authentic, i.e. representative of contemporary usage; 2) help students to develop their own descriptive and analytical skills for improving language awareness.

By their nature, corpus activities are for more advanced levels, but they can be adapted for students at lower levels. By focusing on words which have a high frequency of occurrence and by concentrating on the usual rather than the exceptional, teachers can help learners acquire the language more efficiently, especially at elementary and intermediate levels.

To summarize, corpus linguistics is the foundation and an integral part of most linguistic studies. While it may be argued that corpus linguistics is not really a domain of research but only a methodological basis for studying language, one can in fact use corpora as the basis for an empirical approach to linguistics. Corpus linguistics adherents believe that reliable language analysis best occurs on field-collected samples, in natural contexts and with minimal experimental interference.

References

1. Baker, P. (2010) Sociolinguistics and Corpus Linguistics. Edinburgh: Edinburgh University Press. http://www.ling.lancs.ac.uk/staff/paulb/socioling.htm

2. Bennett, Gena R. (2010) Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. Ann Arbor, Michigan: University of Michigan. 144 p. http://www.press.umich.edu/pdf/9780472033850-part1.pdf

3. Dash, N. S. (2005) Corpus Linguistics and Language Technology: With Reference to Indian Languages. New Delhi: Mittal Publications.

4. Dash, N. S. (2010) Corpus Linguistics: A General Introduction. Proceedings of the Workshop on Corpus Normalization, LDCIL, CIIL, Mysore, 25 August, pp. 1-25. http://www.ldcil.org/download/Corpus%20Linguistics.pdf

5. Krieger, D. (2003) “Corpus Linguistics: What It Is and How It Can Be Applied to Teaching”. http://iteslj.org/Articles/Krieger-Corpus.html

6. McEnery T., Wilson A. (2004) Corpus Linguistics: An introduction (2nd ed.). Edinburgh: Edinburgh University Press, 247 p.

7. McEnery T., Hardie A. (2012) Corpus Linguistics: Method, Theory and Practice (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press, 312 p.

8. Nesselhauf, N. (2004) “Learner corpora and their potential for language teaching.” In John Sinclair, ed., How to Use Corpora in Language Teaching. Amsterdam & Philadelphia: Benjamins, pp. 125-152.

9. Nesselhauf, N. (2011) “Corpus Linguistics: A Practical Introduction”. http://www.as.uniheidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf