CORPUS LINGUISTICS: LEXICAL RESOURCES

Kateryna Riabova
Lecturer,
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

Artificial intelligence was created, so software or a robot works for a person. Computer technologies will absorb not only society but also science as a whole. Linguistics as a science is also no exception. Information technology has offered linguistics its technical possibilities: automatic translation of a Web page, auto-recognition of language (spoken and written), analysis and calculation of data, automation of texts, and much more. However, for this to work, it is necessary to use not only computer technologies but also lexical resources: machine-readable text, dictionaries, thesauri, and tools for processing them.

Using computers and special software has improved the development of linguistics and language research. Linguists use corpora to solve several questions. Corpora provide empirical support, frequency of use information, and extralinguistic information or meta-information that allows comparisons between different text types. It has become much easier to create a practical base. The information obtained is widely used in lexicography, stylistics, linguistic variantology, translation studies, sociolinguistics, and in many linguistic studies (Baker, Hardie, McEnery, 2006).

What is corpus analysis, and what are its primary tasks? Characteristic features of such research are an empirical approach to studying language data, using voluminous texts (corpus) as a basis for analysis, using computer technologies for research, etc.

Corpus linguistics is a heterogeneous field of language research, within which sub-trends are distinguished that differ from each other in their approach to the analysis and processing of corpus data (McEney & Hardie, 2012).

Mode of communication - corpora of oral speech, corpora of written discourse, and corpora of mixed type are distinguished. Differences in the presentation of texts in the corpus determine particular approaches to the selection and processing of linguistic material and reveal significant linguistic differences between the obtained data. *Corpus-based & corpus-driven,* these corpora are used to prove, disprove or clarify a specific theory or hypothesis. *Data collection regimes* distinguish two broad approaches to the data collection regime in the corpus - the monitoring corpus and the static corpus approach. *Annotated/unannotated corpus*. The main difference is the presence of annotation, unique labels assigned to words in corpus texts to indicate diverse linguistic categories. *Total accountability & data selection* - vary depending on how the case is operated. *Multilingual and monolingual corpus.* Corpus types also

differ in the number of languages represented in the corpus. Multilingual corpora are built on material from two or more languages.

Let's consider the most famous corpora. *The Brown* corpus is probably the best-known and most balanced corpus that was started at Brown University in the 60-70s (Manning & Schutze, 1999). It contains about 1 million words, namely 500 text fragments of 2 thousand words each. This body is not updated, it is static. However, some studies try to detect historical changes in words or other syntactic structures.

The Susanne corpus counts a 130,000-word subset of the Brown corpus. It is annotated with information about the sentence's syntactic structure, unlike the Brown corpus, where it is only word for word. The most extensive corpus of syntactic sentence structure is the *Penn Treebank*, which collects texts from the Wall Street Journal. *The Canadian Hansards* is a bilingual corpus with parallel texts in two or more languages with translation. Such a corpus is significant for machine translation and other cross-lingual work. In addition to the corpus, do not forget about dictionaries. *WordNet* is an electronic English language dictionary in which words are placed in a hierarchy, and each set of words contains identical meanings (Manning & Schutze, 1999).

Thus, corpus studies have gone through a difficult path in their formation in the modern English language. There is no unequivocal attitude to such studies, but they are still gaining popularity.

References

Baker, P., Hardie, A., McEnery, T. (2006). *Glossary of Corpus Linguistics.* Edinburgh University Press.

MacEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge University Press.

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* Massachusetts Institute of Technology.

Skobnikova, O. (2019). Analysis of the concept *Family* on the basis of American national corpora. *Linguistic., Volume XXXII,* 115-120.

Zhukovska, V. (2013). Vstup do korpusnoi linhvistyky: navchalnyi posibnyk. [Introduction to corpus linguistics: a study guide]. *ZhDU im. I. Franka*. Zhytomyr Ivan Franko State University.