Chapter 2
Corpus Tools and Technology

ABSTRACT

This chapter will give insight to available corpus tools that are free for end users and have open language databases with different types of texts. The main corpus tool that will be presented is NoSketch Engine, an online interface for searching, creating, and analyzing corpus data. The chapter will provide information about available corpora in NoSketch Engine and additional corpus tools that are available to users for conducting language analyses, preparing teaching materials, exploring and learning the language.

CORPUS TOOLS AND TECHNOLOGY

To use and search computer corpora, corpus tools which make conducting language analyses a lot easier for lecturers and other language experts are used. To analyze corpus data, lecturers "need software that allows them to search, manipulate and store the data, precisely what corpus tools enable" (Kilgarriff et.al, 2014). Today, most lecturers and linguistics experts use online corpus tools that do not require a special process of installation of the software on a computer, which makes the first step in working with corpora and their tools much simpler and easier. The size of computer corpora is growing by the day, and they contain large databases with several million tokens (Grazib, 2009). In order to speed up and simplify language analysis, corpus tools that make the process of using and searching corpora of millions of tokens a lot simpler and more efficient were created.

DOI: 10.4018/978-1-7998-3680-3.ch002

Each computer corpus needs a certain software, a tool without which it would be impossible to search it. Every corpus uses a different type of software, and the features of those software types differ, but there are certain common characteristics which appear in almost all of them. One such characteristic and the most important function is "the option to search the corpora by request, after which the program most commonly provides information on the number of tokens in the corpus and their examples in the texts in which they appear, called concordances" (Nesselhauf, 2005). In the paper *Corpus Linguistics: A Practical Introduction,* several other common characteristics of corpus tools are enumerated, such as the ability to sort and narrow down search results according to relevance, the option to search for words and phrases which appear at a certain distance, the existence of options to show a larger chunk of context than the one shown in concordances, the option to store search results so that they can be used later on etc. Furthermore, the option to show the frequency of tokens and an overview of all the words the corpus contains together with the display of their frequency is also mentioned as a common characteristic of corpus tools. Apart from the main characteristics of corpus tools, a categorization of those tools which describes them more closely and states their main characteristics also exists.

Kilgarriff and Kosem (2012) give a detailed overview of corpus tools in their paper and categorize them. The first category includes computer-based tools and online tools. As examples of computer-based corpus tools, whose main features are the facts that they have to be installed on a computer and that they operate independently, WordSmith and MonoConcPro are mentioned, while examples of online tools include Sketch Engine developed by Adam Kilgarriff and Pavel Rychlý, the Corpus Retrieval System (CoREST) developed by the Department for Digital Dictionaries and Text Corpora at the Society for Danish Language and Literature, and the tools developed by the linguistics professor Mark Davies, available at http://corpus.byu.edu (Kilgarriff, Kosem, 2012).

The next category includes corpus-dependent and corpus-independent tools. Corpus-dependent tools can only be used for a specific corpus, and this corpus is in most cases created as a part of an institution's project. As examples, Kilgarriff and Kosem (2012) give SARA and BNCWeb user interfaces which were created to search the LPBritish National Corpus and Spanish Corpus de Referencia del Español Actual - CREA. The group of corpus-independent tools includes all the tools that allow the user to load and analyze any corpus.