Chapter 1
Types of Computer Corpora

ABSTRACT

This chapter will give an overview of different types of corpora and explain their differences. It will provide readers with different examples of computer corpora that are available to users and for which language analysis can be used. Based on differences of existing corpora, the chapter will show in which way users (language teachers) can use these corpora for creating teaching materials for their students. It will explain why certain types of corpora can be and are open educational resources that provide teachers with a large number of language examples. Additionally, the chapter will examine open source technology (e.g., available tools and programs for creating computer corpora).

INTRODUCTION

Definitions

One of the methods by which a language can be analysed and studied is corpus linguistics, which "in its broadest sense encompasses corpus-based language research" (Utvić, 2013), i.e. use of corpus to analyze language. A corpus is a collection of texts, written or spoken, stored in a computer (O'Keefee, McCarthy, Carter, 2007), online in the cloud[1], on web or in books, so it can be defined as a systematized collection of a natural language (Nesselhauf, 2005; Reppen, 2011). The Glossary of Language Technologies[2] defines a corpus as a "written or spoken language resource collected and annotated with

DOI: 10.4018/978-1-7998-3680-3.ch001

the purpose of: analyzing a language to determine its properties, analyzing human behavior (in the sphere of language use) in certain situations, training the system to adapt its behavior to specific language circumstances, empirical testing of a language theory, creating a test for a language-engineering technique or an application of the technique to see how it functions in practice." The same glossary also gives a definition of computer corpora for which it claims that they are "encoded in a standard and consistent way with the intention of keeping them open for computer searches." Information and communication technology plays a key role in development of corpora because "the creation of computer technology has made corpus research possible, as the computer is able to store, code, categorize, and retrieve massive amounts of information" (Durand, 2018, p. 132).

The term originates from the Latin word "*corpus, corporis* meaning a body, a whole, totality, a set, an almanac" (Utvić, 2013, p. 1). The term "corpus linguistics" was created in the 1980s (Leech, 1992) and it is was first mentioned by Aarts and Mejia in 1982 in the paper "Grammars and Intuitions in Corpus Linguistics", and it appeared again in 1984 in the book "Corpus Linguistics I: Recent Developments in the Use of Computer Corpora". In 1991, on an international symposium of British, Dutch, Swedish and Norwegian linguists, a new group of researchers under the name corpus linguists was formed. From 1996, they have also been publishing their own journal - *The International Journal of Corpus Linguistics* (Leech, 1992; Utvić, 2013). Utvić (2013) thinks that the authors of the definitions of corpus linguistics are stating their opinions about corpus linguistics and based on that treat it as a tool, method, methodology, methodological approach, discipline, theory, theoretical approach, theoretical or methodological paradigm or a combination of all of the above. He also claims that corpus linguistics is today considered to be primarily a methodology or a group of methodologies, not a separate theoretical discipline in the field of linguistics and also wonders whether or not corpus linguistics represents something more than mere methodology (Utvić, 2013). More detailed overview of definition and discussion on what corpus linguistics is can be found in Taylor (2008) "What is corpus linguistics? What the data says" where author states that "corpus linguistics is a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm (theoretical or methodological), or a combination of these" (Taylor, 2008, p. 180).

Corpus linguistics, apart from analyzing language with the use of corpora in its studies, wants to answer two important questions: "Which specific