Chapter 3
Searching Methods for Corpora in NoSketch Engine

ABSTRACT

In this chapter, readers will get insight into a free online system for creating and searching corpora, NoSketch Engine, and its abilities. The chapter will provide valuable information about all corpus tools and functionality that users can use for conducting linguistic research, preparing teaching materials, exams, or exercise for their students. The basic syntax for querying corpus, a corpus query language (CQL), will be explored. Along with that, the chapter will provide examples of CQL syntax so that users can easier comprehend the basics for searching and exploring large language databases in different languages.

INTRODUCTION

Searching Methods for Corpora in NoSketch Engine

With NoSketch Engine, users can search corpora based on different kind of query. There are six types of query: simple, lemma, phrase, word, character and CQL. Additional options which corpus offers to its users is view of context which provides additional analysis of lemmas and words. Also, there is an option of analysis text type in which searched tag is shown. In this chapter author will show all query types and additional possibilities of query analysis which are available in NoSketch Engine. In category query

DOI: 10.4018/978-1-7998-3680-3.ch003

type user can select which type of search he or she wants to conduct. Query type simple refers to most simple searching method of selected corpora. In this search user can type word or a phrase and based on it user can get all language forms.

Query type lemma[1] refers to searching lemmas in corpus. For example, if user wants to find all tags of the headword[2] *žena* (eng. woman) result of this search would be exclusively for nominative of singular in all other forms in singular and plural like *žene, ženama*, resume and so on.

With query type word user can search exact word in her exact morphosyntactic form without lemmatization[3]. For example, if user wants to search corpus for word the *dog* he will get results only for genitive singular but he will not get all the rest forms. This query type offers to the user additional possibility where user can specify which form or word he or she wants to find. In drop down menu PoS (Part of Speech) user can choose from different available language forms. Tadić (2009) states that PoS tagging is associating language categories to each headword in text, this is also called morphosyntactic tagging. This process enables user to select which type of word he or she wants to find in corpus, e.g. noun, verb, adverb, preposition, etc. For example, if user wants to search for word *pila* (eng. chainsaw) as noun than in field Word Form he or she will type *pila* and from drop down menu at PoS will select *noun common.* Search result will show just headword in this searched form.

Query type character is used for searching certain string of characters, e.g. if user wants to find string of characters (letters) which occur together in certain word. Figure 1 shows search results for string *potr* in Croatian hrWaC corpus. This search doesn't have any restrictions and it will find all word in corpus which contains string of mentioned characters no matter on which position in word they appear (beginning, middle or end of the word). On the Figure 1 is example for string of characters *hyper* in English ukWaC corpus.

Last query type which is available in NoSketch Enigine is CQL (Corpus Query Language) a query language which is used for creating advance search queries with morphosyntactic tags. This is advanced query method for searching corpora with which user can narrow down his/hers results. Possibilities of CQL query which are available for users will be in more detailed explained in following subchapter.