The formalization of Serbian: lexical resources and tools

Duško Vitas
University of Belgrade

In this presentation we will present the processing of texts/corpora in Serbian that are based on lexical recognition methods that are supported by Unitex/GramLab, a corpus processing suite. These methods use various types of morphological dictionaries; in more details, we will present the principles of the classification of inflected words, the formalisms of the production of morphological dictionaries of Serbian, the inflection of multi-word units and its formalization, the processing of derivational phenomena, and treatment of unknown words. The interface between these dictionaries and syntax is established through local grammars. Their use will be illustrated by examples of homography disambiguation with so-called ELAG grammars and by the recognition of some classes of nested named entities.

We will also briefly present corpora of contemporary Serbian, more specifically the aligned Serbian-Croatian literary corpus that is based on the ASPAC text collection. This corpus is used to investigate the statistical relevance of often listed discriminators of these two usus.