Simon Krek, »Jožef Stefan« Institute / University of Ljubljana

Slovene and the South Slavic Language Family – the Obligation of Expressing Duality and Other Stories from the South

Slovene is one of the rare languages exhibiting the category of the dual grammatical number, and the only Indo-European one among the official languages of European Union (the other being Maltese from the Semitic family). Being a morphologically rich language, which is a common characteristic of Slavic languages, this unique feature contributes no less than 613 tags to the tagset of 1,902 possible combinations commonly used for PoS-tagging of Slovene (Erjavec and Krek 2008). In the tagset, the dual can be attributed to verbs, nouns, pronouns, adjectives and numerals. In the standard variant of the language Slovene speakers are obliged to use the dual in all cases where two objects, people or other entities are referred to. They are, therefore, confronted both with possibilities and obligations that are not available to speakers of other languages. We will explore some of the more interesting ones.

We will also place Slovene in the wider context of South Slavic languages and address the issue of natural language processing of very similar languages (e.g. Ljubešić and Kranjčić). Since the breakup of the former Federal Republic of Yugoslavia where Serbo-Croatian, Slovene and Macedonian were defined as official languages in the constitution, four official standards originating from the former Serbo-Croatian (Bosnian, Croatian, Montenegrin and Serbian) are now used in the newly-formed states. Discriminating between these standards is not an easy task and poses some interesting challenges to the natural language processing community, taking into account also rather sensitive socio-linguistic aspects of the situation.

References:

Tomaž Erjavec, Simon Krek: The JOS Morphosyntactically Tagged Corpus of Slovene. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech 2008.

Nikola Ljubešić and Denis Kranjčić. Discriminating between VERY similar languages among Twitter users. *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*. Ljubljana.