Simon Krek
"Jožef Stefan" Institute
Artificial Intelligence Laboratory
Ljubljana, Slovenia

Kaja Dobrovoljc
Trojina, Institute for Applied Slovene Studies
Ljubljana, Slovenia

WG1: Lexicon-Grammar Interface

SKETCH GRAMMAR OR PARSER – A COMPARISON OF TWO MWE EXTRACTION METHODS

In the poster we will describe a comparison of multi-word expression extraction from the ssj500k and Kres corpora (Logar Berginc et al. 2012) based on the sketch grammar, in this case for Slovene, which is defined as a series of grammatical relations or gramrels using regular expressions over POS-tags in a tagged corpus (Kilgarriff and Tugwell 2002, Krek and Kilgarriff 2006, Krek 2012), and MWE extraction from the same sources parsed with the MSTParser adapted for Slovene (Rupnik, Dobrovoljc and Krek 2012).

A detailed sketch grammar for Slovene was developed for the extraction of lexical data from the Gigafida corpus (Logar Berginc et al. 2012) for the purposes of compiling Slovene Lexical Database (Kosem, Gantar and Krek 2013). The Sketch Engine tool enables the extraction of collocations (and corpus examples) based on grammatical patterns which were identified as relevant for the purposes of Slovene Lexical Database compilation. Lexical data based on the same or similar grammatical patterns can be extracted from the corpus using parsed data in the Sketch Engine tool as shown for Turkish in (Ambati, Reddy and Kilgarriff 2012).

MSTParser (McDonald, Lerman, Pereira 2006) was recently developed for Slovene and trained on the ssj550k corpus treebank (Erjavec, Fišer, Krek and Ledinek 2010). The Kres reference corpus (100 million words) was parsed with the new tool which enabled the comparison of the two methods of identifying relevant patterns in the same corpus. The comparison is made by adapting or "translating" gramrels in the sketch grammar to treebank dependencies and extracting lexical data. The Sketch Engine tool is using logDice score (Richly 2008) for identifying and ranking more relevant MWEs in specific grammatical patterns. The poster will describe the difference between the two "word sketches" by comparing the similarity of extracted collocates and their ranking in particular grammatical patterns.

References:

Dobrovoljc, K., Krek, S., Rupnik, J. (2012): Skladenjski razčlenjevalnik za slovenščino. Proceedings of the 8th Slovenian and 1st International Language Technologies Conference. Ljubljana, Slovenia. Available at: http://nl.ijs.si/isjt12/JezikovneTehnologije2012.pdf.

Erjavec, T., Fišer, D., Krek, S., Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10), Malta, 2010.

Kilgarriff, A., Tugwell, D. (2002). Sketching words. In H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Euralex, pp. 125-137.

Krek, S., Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec, J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*. Ljubljana, Slovenia. Available at: http://nl.ijs.si/is-ltc06/proc/12_Krek.pdf.

Krek, S. (2012). New Slovene sketch grammar for automatic extraction of lexical data. Presented at *SKEW3 workshop*, 21-22 March 2012, Brno, Czech Republic. Available at: http://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf?format=raw

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Kosem, I., Gantar, P., Krek, S. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In: *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut. Available at: http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf.

McDonald, R., Lerman, K. and Pereira, F. (2006): Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. *Tenth Conference on Computational Natural Language Learning (CoNLL-X)*.

Rychly, P. (2008) A lexicographer-friendly association score. In Sojka, P. & Horák, A. (eds.) *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing*, RASLAN 2008, 6-9. Brno: Masaryk University.