Electronic Tools and Resources for Multi-Word Unit detection and research in Serbian

Working Group: WG1: Lexicon/Grammar Interface

Jelena Mitrovic, Univesity of Belgrade, Faculty of Philology

The aim of this poster is to present recent developments related to new tools and resources for Natural Language Processing in Serbian pertinent to the research of Multi-word units. The idea is to enable a connection between lexical resources for Serbian, such as e-dictionaries, which now have a sizeable number of MWUs, the Serbian WordNet (SWN) and domain ontologies, such as the recently developed Ontology of rhetorical figures for Serbian, in order to facilitate research of the complex linguistic phenomenon that are MWUs. As far as the research of Multi-word units goes, Serbian electronic dictionaries were used in that regard the most, so far. The system of Serbian e-dictionaries covers both general lexica and proper names and all inflected forms were generated from 130,600 simple forms and 11,324 MWU lemmas (Krstev, 2008). The MWUs e-dictionary for Serbian is a morphological dictionary whose construction is complex due to the very rich inflectional nature of Serbian. This dictionary is being built systematically according to the inflection description supported by the Multiflex system (Savary, 2009). Apart from complex prepositions, adjectives, conjuctions and interjections, this e-dictionary also contains complex adjectives e.g. *mrtav pijan* 'dead drunk' and complex nouns e.g. *nemasno mleko u prahu* 'fat free powdered milk' (Krstev et al., 2010).

The new set of tools that has been developed for Serbian WordNet recently (Mladenović et.al, 2014) will help developers of wordnets not only to increase the number of synsets but also to ensure their quality, thus preventing it to become obsolete too soon. These new tools will help with upgrade, cleaning and validation and facilitate the production of a clean, up-to-date WordNet, while the new Web application will enable search, development and maintenance of a WordNet. All of the mentioned upgrades will make MWUs research and detection more straightforward. Serbian WordNet, now having 21,212 synsets will be an important source for new MWU entries in e-dictionaries since the percentage of MWUs that appear in it is approximately 32.5%,

RetFig (Mladenović and Mitrović, 2013) is a formal domain ontology of rhetorical figures for Serbian and it represents a knowledge base of rhetorical figures in Serbian, which can also be used for other languages, with minor changes. The RetFig ontology was developed taking into account a plethora of rhetorical figures in the morphologically rich Serbian language, as well as in regard to various classifications of rhetorical figures that exist. The ontology can be accessed at http://resursi.mmiljana.com/MemberZone/RetFig.aspx after simple authentication. As some rhetorical figures can be considered as Multi-word units, this ontology can bring a new dimension to research related to MWUs which are very frequent in

everyday language, as well as in languages for special purposes, just like many rhetorical figures are. Some examples of MWUs in regard to rhetorical figures are:

1) Oxymoron – a rhetorical figure in which apparently contradictory terms appear in conjunction e.g. *topli led* 'warm ice', *živi mrtvac* 'living dead', *glasna tišina* 'loud silence', etc.
2) Periphrasis – the use of indirect and circumlocutory speech or writing, e.g. *vrh sveta* 'Top of the World' to describe the Himalayas, *Velika jabuka* 'The Big Apple' for New York, *Grad svetlosti* 'The City of Lights' for Paris, etc.

A new development that will be possible thanks to these new tools is adding sentiment-related adjective and noun pairs to the Serbian WordNet, using the newly developed tools and according to the lists of rhetorical figures acquired with the help of the Ontology for rhetorical figures for Serbian – Retfig, all the while using the help of e-dictionaries and the already existent MWU entries, thus enriching the SWN with MWUs. This project will be rolled-out through a crowdsourcing system in which participants will make noun-adjective and adjective-noun pairs they feel are natural in the Serbian language. Also, new MWUs can be detected using the combination of a recently formed culinary corpus (Vujičić-Stanković et.al, 2014), the Corpus of Contemporary Serbian (Vitas and Krstev, 2012), the Serbian WordNet and the RetFig ontology.

References

Krstev Cvetana, Stanković Ranka, Obradović Ivan, Vitas, Duško, Utvić Miloš. 2010. Automatic Construction of a Morphological Dictionary of Multi-Word Units. LNCS 6233 Springer Berlin Heidelberg, pages 226-237.

Krstev, Cvetana. 2008. Processing of Serbian – Automata, Texts and Electronic Dictionaries. Faculty of Philology, University of Belgrade.

Mladenović, Miljana, Mitrović, Jelena, Krstev, Cvetana. 2014. Developing and Maintaining a WordNet: Procedures and Tools. Proceedings of 7th Global WordNet Conference, Tartu, Estonia, pages 55-62.

Mladenović, Miljana, Mitrović, Jelena. 2013. Ontology of Rhetorical Figures for Serbian. LNAI 8082, Springer Berlin Heidelberg, pages 386-393.

Savary Agata. 2009. Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. Implementation and Application of Automata. Springer Berlin Heidelberg, pages 237-240.

Vitas, Duško and Krstev, Cvetana. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. In Prace Filologiczne, vol. LXIII, Warszawa, pages 279-292.

Vujičić-Stanković, Staša, Krstev, Cvetana, Vitas, Duško. 2014. Enriching SerbianWordNet and Electronic Dictionaries with Terms from the Culinary Domain. Proceedings of 7th Global WordNet Conference, Tartu, Estonia, pages 127-132.