Electronic Tools and Resources for Detection and Research of Multi-Word Units in Serbian

WG1

Jelena Mitrović

advisor: Prof. Cvetana Krstev, PhD

University of Belgrade, Faculty of Philology
Department of Library Science and Information Science

Introduction

The aim of this poster is to present recent developments related to new tools and resources for NLP 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, the Serbian WordNet (SWN) and an Ontology of rhetorical figures for Serbian, in order to facilitate research of the complex linguistic phenomenon that are MWUs.

Serbian electronic dictionaries

Their system covers both general lexica and proper names. 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 (its building is supported by the Multiflex system (Savary, 2009)). Apart from complex prepositions, adjectives, conjunctions 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).

FIGURE 1. New SWN Web Interface

Serbian WordNet

The new set of tools that has been developed for Serbian WordNet recently (Mladenović at. al, 2014) are tools for upgrade, cleaning and validation that will facilitate the production of a clean, up-to-date WordNet, while the new Web application enables search, development and maintenance of a WordNet (Fig. 1). New tools can be accessed at http://resursi.mmiljana.com/WordNetS.aspx

All of the mentioned upgrades will also make MWUs research and detection more straightforward. Serbian WordNet, now having 21,212 synsets (Fig. 3) 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% (Fig. 2).

FIGURE 2. Distribution of MWUs in SWN

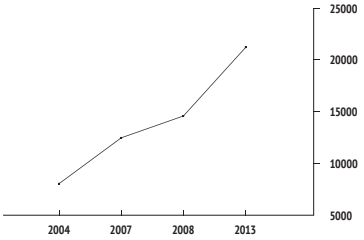


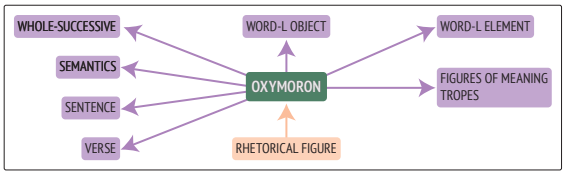
FIGURE 3. Number of SWN synsets

Ontology of Rhetorical Figures for Serbian

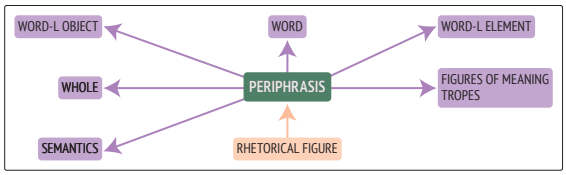
RetFig (Mladenović and Mitrović, 2013) is a formal domain ontology of rhetorical figures for Serbian – a knowledge base of rhetorical figures in Serbian (can also be used for other languages). As some rhetorical figures can be considered as MWUs, this ontology can bring a new dimension to their research.

RetFig ontology gives an unambiguous formal description of 98 rhetorical figures in Serbian, such as:

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.



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.



The ontology can be accessed at
http://resursi.mmiljana.com/MemberZone/RetFig.aspx after simple authentication.

Future work

1. Enriching the SWN with MWUs by adding sentiment-related adjective and noun pairs to the SWN using:

- the newly developed tools for SWN
- the lists of rhetorical figures acquired with the help of the RetFig Ontology
- e-dictionaries and the already existent MWU entries.
- through a crowdsourcing system in which participants will make noun-adjective and adjective-noun pairs they feel are natural for the Serbian language.

2. MWUs detection 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

[1] Krstev C., Stanković R., Obradović I., Vitas D., Utvić M. 2010. Automatic Construction of a Morphological Dictionary of Multi-Word Units. LNCS 6233 Springer Berlin Heidelberg, pages 226-237.
[2] Krstev, C. 2008. Processing of Serbian – Automata, Texts and Electronic Dictionaries. Faculty of Philology, University of Belgrade.
[3] Mladenović, M., Mitrović, J., Krstev, C. 2014. Developing and Maintaining a WordNet: Procedures and Tools. GWC 2014, Tartu, Estonia. Pages 55-62.
[4] Mladenović, M. and Mitrović, J. 2013. Ontology of Rhetorical Figures for Serbian. LNAI 8082, Springer Berlin Heidelberg, pages 386-393.
[5] Savary, A. 2009. Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. Implementation and Application of Automata. Springer Berlin Heidelberg, pages 237-240.
[6] Vitas, D. and Krstev, C. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. In Prace Filologiczne, vol. LXIII, Warszawa, pages 279-292.
[7] Vujičić-Stanković, S., Krstev, C., Vitas, D. 2014. Enriching SerbianWordNet and Electronic Dictionaries with Terms from the Culinary Domain. GWC 2014, Tartu, Estonia, pages 127-133.

Acknowledgements

Prof. Cvetana Krstev, PhD; Miljana Mladenović, PhDc; The Group for Language Technologies of the Faculty of Mathematics, University of Belgrade.