Adding MWEs to Serbian Lexical Resources Using Crowdsourcing

Jelena Mitrović, Miljana Mladenović, Cvetana Krstev

University of Belgrade

Working Group: WG1: Lexicon/Grammar Interface

In this poster, we will present a project aiming at enriching Serbian WordNet with MWEs using the power of crowdsourcing. Crowdsourcing has become a popular and quite effective tool for human computation in various scientific fields. This model is mostly used for tasks which can easily be carried out by people but tend to be difficult for computers. Many NLP projects have been successfully carried out via crowdsourcing recently (Mitrović, 2013).

We started this project in the hopes of acquiring valid linguistic data that we can use to enrich Serbian WordNet (SWN) and the Ontology of Rhetorical Figures for Serbian (RetFig) (Mladenović and Mitrović, 2013) with multi-word expressions that are used the most in everyday language as we will get the most use out of this kind of lexical data. Mechanical Turk, Amazon's platform, has proved to be very effective in various NLP related crowdsourcing tasks. Unfortunately, we haven't been able to use it in Serbia, due to territorial restrictions of this tool, so we had to find another way to collect linguistic data. Google Forms proved to be a good solution.

In the first part of the project, we extracted Similes (figures of speech that directly compare two things through the explicit use of connecting words), which are always MWEs, from the Corpus of Contemporary Serbian Language (Vitas and Krstev, 2012; Krstev, 2008). We made a selection of the Similes that were most likely to be used in everyday speech and incorporated them into Google Forms questionnaires. We circulated these forms via social networks and the potential participants were asked to mark the expressions according to their own perception of what constitutes an expression that is used in everyday speech. One of the questions in each form (*Crn kao ulica* "As black as a street") was the control question, used for checking whether forms were being filled in truthfully or just automatically. The expected answer to that question was "No". If a participant's answer to the control question was "Yes", his answers were not taken into consideration at all. We used only one control question in order to keep the task of filling in the form interesting, as well as because the forms were quite short. At first, we wanted to add one more control question whose expected answer would be "Yes", but it was very difficult to find a construct everyone uses, so we decided against adding another control question.

For the purpose of enriching SWN and RetFig we developed new software, consisting of two independent, complementary parts. The first part of the software used the results of the crowdsourcing technique to generate adjective-noun pairs which were later used for expansion of SWN with the new relation named "Characteristic Attribute". In that regard, we first measured the contribution of participants and determined the relevant set of participants whose results were taken into account as their results proved to be statistically significant and were used to establish the final set of pairs. In order to measure the participants' contribution we generated sets of answers where each set had <=30 expressions in the form <adjective>as<noun> to which each participant could give one of two given answers, "Yes" or "No". Sets of answers were then converted into

matrices where each row presented answers of each participant and each column presented one expressions in the form <adjective> as <noun>.

Content of each cell of the matrix had the value 1 if the participant marked a certain expression with "Yes" and the value 0 if the participant marked that expression with "No". Rows of the matrix were compared against each other with a paired t-test in order to determine that there was no substantial difference between arithmetic means of participants' answers. First, we assumed that the first participant was reliable and we compared contributions of all other participants against his contribution. If the group he belonged to was bigger, we assumed that the first participant was reliable, and that entire group was considered relevant, on the contrary, the other, opposite group of participants' answers was considered relevant. Statistical significance was determined at $p<0.05$.

The second part of our software was used to generate new relations – "Characteristic Attribute" – between appropriate synsets to which the noun part of the <adjective> as <noun> expressions belongs to and the synsets to which the adjective part of the same expressions belongs to. In that regard, the SWN web application (Mladenović et.al, 2014) was enhanced with the possibility to facilitate the choice of an appropriate synset. This new addition enabled automatic creation of the new relation for those adjective-noun pairs whose relation was 1:1, i.e. both the noun and the adjective had one sense and one-to-one mapping was used. If a noun was polysemous, or an adjective had more than one sense (although that case was not so common), the software gave us an opportunity to choose an appropriate synset, with the right meaning conveyed by the adjective-noun pair. The noun *miš* "mouse" is one of those polysemous nouns, as it can represent both a computer device and an animal, e.g. *Mokar kao miš* "As wet as a mouse" – 0.8% of participants marked this expression as an expression used in everyday speech, and *Mali kao miš* "As little as a mouse" – 0.6 % of participants also marked this expression positively and they were both added to SWN. We also added: *Crn kao ugalj* "As black as coal"; *Crn kao noć* "As black as night"; *Mekan kao duša* "As soft as a soul"; *Mekan kao svila* "As soft as silk". We added 70 new expressions in total. Table 1 gives an overview of the project.

	Total expressions	Total participants	Participants whose contribution was stat. significant	Expressions marked Positively
1	30	46	37	13
2	42	138	112	19
3	41	150	132	17
4	41	100	83	21
Total	154	434	364	70

Table 1. Project in numbers

We plan on continuing with this project as it is cost-effective and gives good, valid results. As we do not have a referent collection of Simile in Serbian, we will collect the ones Serbian people use the most by asking them to give us their suggestions, through another crowdsourcing project. We will evaluate the new data and use new constructs to further enrich our lexical resources via crowdsourcing. SWN enriched in this way will be used for building software for figurative speech recognition and automatic tagging, and will be used to produce features for sentiment analysis tasks, as well as to enrich RetFig with new examples of the rhetorical figure Simile.

References

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

Mitrović, Jelena. 2013. Crowdsourcing and its Application. INFOtheca, Volume XIV, No 1, pages 37a-46a

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.

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