Adding MWEs to Serbian Lexical Resources Using Crowdsourcing

WG1
Lexicon/Grammar Interface

Jelena Mitrović
Miljana Mladenović
Cvetana Krstev
University of Belgrade
5th PARSEME Meeting, Iași, Romania, 23-24 September, 2015

The power of crowdsourcing in NLP projects (Mitrović, 2013) used to enrich Serbian lexical resources with MWEs related to the rhetorical figure Simile, e.g. Hladan kao led "As cold as ice" for the purpose of evaluating an automatic method of adding new relations to Serbian WordNet (SWN) and enriching the Ontology of Rhetorical Figures for Serbian (Mladenović and Mitrović, 2013).

Similes (Adjective-Noun constructs) extracted from the Corpus of Contemporary Serbian Language (Vitas and Krstev, 2012; Krstev, 2008) and used to automatically determine relevant Adjective-Noun constructs, according to the algorithm given in pseudocode in Figure 1.

```
Input: Adjective As Noun text file
Output: a pair of WordNet mutually inverse
semantic relations (specificOf/specifiedBy)
for each input adjective-noun pair
for each adjective-noun pair in adjective-noun pairs
if ((adjective exists in Wordnet.adjective.literals)
    and (noun exists in Wordnet.noun.literals)) {
        if ((Wordnet.senses(adjective).Count==1)
            and (Wordnet.senses(noun).Count==1)
            and (Wordnet.sense(adjective).FirstSense)
            and (Wordnet.sense(noun).FirstSense) ) {
              Create Relation(specificOf,adjective,noun);
              Create Relation(specifiedBy,noun,adjective);
        }
        else
          for each (sense in Wordnet.senses(adjective)) {
            add to adjective senses(adjective,sense,synsetId)}
          for each (sense in Wordnet.senses(noun)) {
            add to noun senses(noun,sense,synsetId)}
        }
}
```

FIGURE 1. Algorithm for WordNet expansion

372 candidates that can be connected by the relation SpecificOf/SpecifiedBy were produced. For the rest of the possible ADJ-NOUN pairs present in SWN, a web page in the SWNE2 application (Mladenović et al., 2014) was created for semi-automatic input.
Google Forms survey advertised via Facebook was used for finding Adjective-Noun constructs in everyday language. Table 1 shows the number of questions and participants per each form.

GOOGLE FORM	NUMBER OF QUESTIONS PER FORM	PARTICIPANTS PER FORM
1	30	46
2	42	138
3	41	150
4	41	100
TOTAL	**154**	**434**

TABLE 1. Distribution of questions and participants per form

Inter-annotator (participant) agreement in this survey was determined in 4 steps:

1) If there is no substantial difference between arithmetic means of the participants' answers according to a paired t-test, go to step 2.
2) As it is shown in Table 2, 7 subsets of questions and answers were thus created.
3) All 7 units were converted into matrices: each row – answers of each participant, each column – one question in the form <adjective>as<noun> -- value 1 for "Yes" and value 0 for "No" answers.
4) From each set, five participants whose difference in the paired t-test was the slightest were chosen.
5) Inter-annotator agreement was finally evaluated using the Krippendorff α coefficient (Kalpha) (Lombard et al., 2012) – results given in Table 2.

FORM SET	NO. OF PARTICIPANTS	NO. OF QUESTIONS	KALPHA VALUE	NO. OF QUESTIONS ANNOTATED WITH YES
1	5	30	α = 0,7575*	16
2a	5	21	α = 0,713*	17
2b	5	21	α = 0,698*	15
3a	5	21	α = 0,688*	5
3b	5	20	α = 0,484	
4a	5	21	α = 0,434	
4b	5	19	α = 0,375	
TOTAL		154		53

TABLE 2. Inter-annotator agreement over Google Forms and number of items which belong to reliable forms and were annotated with "Yes"

In Table 3. Some results are presented according to the number of votes by the reliable participants. The first column shows constructs that were added to SWN, and the second column shows the ones that were not assessed as used in everyday language, therefore, not added to SWN.

5 OUT OF 5 VOTES	2 OR LESS OUT OF 5 VOTES
Tačan kao sat "Like clockwork"	Brz kao misao* "Quick as a thought"
Hladan kao led "Cold as ice"	Lak kao ptica* "Light as a bird"
Hladan kao špricer "Cool as spritzer"	Beo kao kreda "White as chalk"
Tvrdoglav kao mazga "Stubborn as a mule"	Debeo kao bure "Fat as a barrel"
Lagan kao pero "Light as a feather"	Blistav kao zvezda "Shiny a s a star"

TABLE 3. Adjective-Noun constructs as evaluated by online participants
* Frequency of occurrence in the Corpus k ⩾ 4, but were not selected in the survey.

How much the change of the frequency of occurrence in the Corpus threshold (k) influenced the relevance of automatically selected ADJ-NOUN pairs, measured based on the results obtained through the surveys?

FREQUENCY THRESHOLD	BY ALGORITHM	BY HUMANS	HUMANS/ALGORITHM
K = 1	93	53	57%
K = 2	44	32	73%
K = 3	32	27	84%
K = 4	23	19	83%

TABLE 4. Relation of manually and automatically selected pairs depending on the frequency threshold

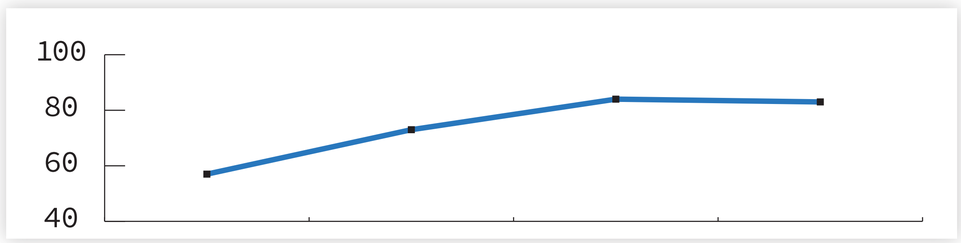

FIGURE 2. Relationship of selected pairs obtained with the survey method compared to the ones obtained with the method of the most frequent occurrence for different frequency thresholds.

References

Krstev, Cvetana. 2008. Processing of Serbian – Automata, Texts and Electronic Dictionaries. Faculty of Philology, University of Belgrade.
Lombard, Matthew Lombard, Snyder-Duch, Jennifer and Campanella Bracken, Cheryl. 2002. Content analysis in mass communication: Assessment and reporting of inter-coder reliability. Human Communication Research, 28(4):587–604.
Mitrovic, Jelena. 2013. Crowdsourcing and its Application. INFOtheca, Volume XIV, No 1, pages 37a-46a.
Mladenović, Miljana, Mitrović, Jelena. 2013. Ontology of Rhetorical Figures for Serbian. LNAI 8082, Springer Berlin Heidelberg, pages 386-393.
Mladenovic, Miljana, Mitrovic, Jelena, Krstev, Cvetana. 2014. Developing and Maintaining a WordNet: Procedures and Tools. Proceedings of 7th Global WordNet Conference, Tartu, Estonia, pages 55-62.
Vitas, Duško and Krstev, Cvetana. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. In Prace Filologiczne, vol. LXIII, Warszawa, pages 279-292.

Acknowledgements

Group for language resources and technology