Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool

Miljana Mladenović[1] and Jelena Mitrović[2]

[1] Faculty of Mathematics, University of Belgrade, Serbia
www.matf.bg.ac.rs
[2] Faculty of Philology, University of Belgrade, Serbia
www.fil.bg.ac.rs
{ml.miljana,jmitrovic}@gmail.com

Abstract. In this paper we present a set of new additions and functionalities to recently introduced software tools and techniques that will help researchers in the area of semantics and especially developers of wordnets. The motivation lies in our wish to get an on-line, fully comprehensive, modular, multiuser and safe system for further development of the Serbian WordNet (SWN). The most important functionality is the establishment of semantic relations between Princeton WordNet 3.0 (PWN) and the Serbian WordNet 3.0. Other functionalities of this set of tools are based on other semantic resources: SentiWordNet, a publicly available lexical resource for sentiment analysis, Suggested Upper Merged Ontology (SUMO) and Morphological electronic dictionary of Serbian (SrpMD). They provide sophisticated search possibilities and procedures for easier and more comfortable growth of WordNet. All of the functionalities were developed using publicly available resources: PWN mapping techniques, SUMO mapping procedures, SentiWordNet mapping files and by using the SrpMD.

Keywords: Serbian WordNet, Semantic Resources, Sentiment Tagging, SUMO

1 Introduction

WordNet is a lexical semantic tool that was first developed at Princeton University by George Miller and his team. The goal of this original WordNet — Princeton WordNet (PWN) was to serve as a mental lexicon that would help scientists working on psycholinguistic projects [4]. PWN is a set of approximately 117,000 concepts interconnected by semantic relations to form a semantic network, where a concept denotes an abstract set of members grouped on the basis of their common properties. Each concept in PWN is represented by a set of English synonyms which form the synset for this concept. WordNet is organized according to principles governing human lexical memory and as such, it is an online lexical database designed for use under program control. In WordNet, a *word form* is represented by a string of letters, and a *concept* is represented by the set of synonym words that have the same or similar meaning. WordNet respects the syntactic categories noun, verb, adjective, and adverb. Concepts can be interconnected by semantic relations and word forms by lexical relations. Synonymy is WordNet's basic relation, because WordNet uses sets of synonyms (words with similar meaning) to represent word senses; Antonymy (opposing-name) is a symmetric relation between two word forms with oposite meaning; Hyponymy (sub-name) and its inverse, hypernymy (super-name), are transitive relations between synsets and they organize the meanings of concepts into a hierarchical structure. Meronymy (part-name) and its inverse, holonymy (whole-name) distinguish component parts. Troponymy (manner-name) is for verbs what hyponymy is for nouns.

Many wordnets for languages other than English were developed based on the Princeton WordNet, using the expand model, that is to say, by translating synsets from PWN into a target language. Wordnets developed in this way are all connected through the Inter-Lingual-Index (ILI) that links similar concepts between languages, which is highly advantageous for various multilingual applications. ILI was first introduced in the scope of EuroWordNet [22], a project which introduced multilingualism into this semantic network. The common framework introduced by EuroWordNet proved to be useful, but each wordnet built on the basis of PWN has to regularly upgrade to new versions, in order to keep up with the upgrades of PWN.

Serbian WordNet (SWN) was built in the same fashion. But, except for basic connection to PWN, SWN is also connected with other important lexical and semantic resources as are SUMO ontology, Serbian Morphological Dictionaries (SrpMD) and SentiWordNet. Mapping of SWN to SUMO, creation of relations between SWN and Serbian Morphological Dictionaries (SrpMD) and sentiment tagging of synsets are very important parts of processes of building relations among SWN and these lexical and semantic resources. In this paper, we present improvements of software maintaining for SWN tools and techniques to manage these relations.

The remainder of this paper is organized as follows – in Section 2, we will give a short history of the development of SWN, the environment in which it was developed, along with the limitations of that environment. Section 3 gives an overview of connecting Serbian WordNet with other semantic resources. Section 4 is focusing on the properties and the phases of development of new functionalities added to the SWNE tool and Section 5 gives directions for future work.

2 Motivation and Discussion

The development of Serbian WordNet (SWN) was initiated [12] in the scope of the BalkaNet project [19], which aimed at extending the model applied in the EuroWordNet. The aim of BalkaNet, a project financed by the European Commission from September 2001 until August 2004, was to develop aligned semantic networks for several Balkan languages, namely Bulgarian, Greek, Romanian, Serbian and Turkish, as well as to extend the existing network for Czech. At the end of the BalkaNet project, in 2004, the Serbian WordNet had 7,000 synsets aligned to the Princeton WordNet version 2.0. In the subsequent years approximately 14,000 synsets were added to it. The development of SWN following the end of BalkaNet greatly depended on joint efforts of many volunteers, specialists in different scientific areas and the WordNet editor who was overseeing the additions to SWN. These new additions [11] pertained to adding synsets in certain conceptual domains, e.g. scientific domains – biological species, biomedicine, religion, law, linguistics, literature, librarianship, computer science, the culinary domain, sentiment analysis domain, etc.

XML like representations of the EuroWordNet data were produced with a tool named VisDic [8]. This tool was used in the BalkaNet project also and therefore, SWN was built with its help. VisDic is the tool that was used for the development and maintenance of SWN until recently. The need for new work environment and new tools arose mainly due to the fact that VisDic was limited to the desktop surrounding. That meant that each contributor worked on their own XML file and at times, many XML files had to be merged into one which was often quite a tedious task for the main SWN editor. That is why we decided to build a web application that could be used by many contributors at once and which would keep all the useful functionalities of VisDic. Many functionalities of VisDic were kept, as they were quite useful, but the need for a web-based set of tools prevailed. VisDic made the depiction of several WordNet files possible as well as the synchronization of the synset display in the files presented at the same time. The function of copying synsets from one WordNet file into another in order to better adapt them, helped to speed up the process of expansion and ensured a high degree of precision and correlation between those WordNet files. Keeping in mind the high functionality and the ease of use of the VisDic software, we wanted to add some more features and functionalities. In that regard, a new web application, the Serbian WordNet Edition (SWNE) allowed for the following:

- Improved system of morphological tagging of synsets, that is to say, improved connecting of SWN and SrpMD [10] which was attained by better control over the existing connections, control over new connections made in the process of adding new synsets and limiting the list of possible values of the connecting synset tag;
- Improved system of synchronous display of synsets and their semantic relations in SWN and PWN, as well as synchronized search of synsets in SWN and PWN;
- Connecting SWN synsets with SUMO ontology concepts [16] – control over correctness of the existing SUMO tags and control over establishing links with SUMO concepts during adding new synsets and values of SUMO tags, depending on the POS (Part of Speech) tag of the synset;
- New semantic tags in SWN, related to emotions and based on SentiWordNet semantic resource;
- Improved control over XML documents, at any time, by establishing control of well-formedness and validity, based on the existing XSD Schema [15] .

In the short time interval after making the SWNE application[1] available for usage, partial functionality improvements were made. Also, we started redesigning the user interface of the application in order to provide comfortable conditions and more simple and intuitive usage of the application. The SWN resource itself was also improved – a new element which assigns sentiment orientation to a synset, as well as the strength of the assigned sentiment was added.

At the moment, the SWNE 2.0 is a valuable tool for further development of the Serbian WordNet through collaboration of users whose interests lie in different domains that are related to language technologies.

3 Improvement of Connecting SWN with Other Semantic Resources

WordNet is organized by semantic and lexical relations. Semantic relation is a relation between concepts which are designed by synsets. Therefore a semantic relation can be seen as a pointer between synsets. Unlike synonymy and antonymy, which are lexical relations between word forms, hyponymy/hypernymy is a semantic relation between concepts: e.g., oak is a hyponym of tree, and tree is a hyponym of plant. Much attention has been devoted to hyponymy/hypernymy (variously called subordination/superordination, subset/superset, or the ISA relation). This semantic organization of WordNet is useful for tasks such as word sense disambiguation [9], resolving semantic opposition [5], semantic tagging [1], semantics annotation [21], sentiment classification [7].

Still, WordNet is not a semantic network in which only its internal hierarchical structure between concepts is of importance. Due to its universality and its usage in many areas of natural language processing, there is a growing number of lexical resources that establish different types of relations with WordNet concepts. In the next parts of this paper, we will mention some of them that we feel are most important.

3.1 Connection between WordNet and SUMO Ontology

Ontology is a conceptual representation of knowledge in a formal way. The Suggested Upper Merged Ontology (SUMO) was developed in the year 2000 thanks to Adam Pease [18], as the first formal ontology which established mapping with all synsets of the PWN resource. SUMO represents a relatively small number of concepts (approx. 1000), but a large number of assertions (approx. 4000) and rules (approx. 800) which enable processes of automated reasoning, information retrieval and can be used in many other NLP tasks. An important component of the SUMO ontology development tool is the Sigma Knowledge Engineering Environment[2] which provides integration of SUMO and user ontologies into a knowledge base and it is integrated with an automatic theorem prover Vampire.

[1] http://resursi.mmiljana.com/

[2] http://sigmakee.sourceforge.net/

In this environment, a user can also define their own assertions, which enriches the knowledge base and increases the possibilities of inferencing in first order logic.

Mapping enriches WordNet database through tagging of each synset with the corresponding SUMO concept. The WordNet synset may be declared: as equivalent to the SUMO concept, as subsumed by it, or as an instance of it. SUMO and WordNet both define conceptualizations (simplifications) of our world. WordNet with the chief purpose to map these conceptualizations into natural language terms. SUMO with the purpose to organize them into a logical structure. Thanks to publicly available files that contain unequivocal mapping pairs for PWN synsets and SUMO concepts, it was possible for us to generate the SUMO tag in SWN. SWN synsets acquired the SUMO concept tag through the process of horizontal parallelization (English synset – Serbian synset) and vertical inheritance (Serbian non-tagged synset inherited the SUMO tag of the parent synset, if one existed, otherwise, it was added manually). In this way, we created favourable conditions for the integration of RetFig, the ontology of rhetorical figures for Serbian [14] and SUMO ontology with a goal of enabling automatic reasoning in the process of rhetorical figures recognition.

3.2 Connection between WordNet and Morphological Dictionaries

Two very important linguistic resources for Serbian – SWN and SrpMD have been functioning in an integrated environment for a long time – WS4LR: A Workstation for Lexical Resources [13]. Each one of these resources uses a specific set of information and data that the other resource contains. Taking into account the experience in functioning of this tool, we made it possible for the SWNE web application to also use the data from morphological dictionaries for Serbian – in this way each SWN synset tagged with LNOTE tag can point out to a set of morphological rules of the lexeme defined by the tag LITERAL to which the LNOTE tag refers. Also, it is not possible to choose a morphosyntactic tag for LNOTE if it does not belong to the group of permitted tags. SrpMD contains many different morphological dictionaries in the LADL format [2], but for the purposes of connection with the SWNE application, the Simple lemmas DELAS dictionary was used. Its general entry is the following:

```
lemma.Knnn [+SinSem]
```

where Knnn is a unique label of the class of flective rules. In this phase of development of the SWNE, all Knnn labels from the DELAS set are available for correspondence with literals of synsets in the SWN.

3.3 Serbian WordNet Affective Tagging

In the recent years, with prompt development of e-commerce web sites, blogs, forums, discussion groups, social networks and short messaging systems, natural language processing is facing new challenges and the need to analyse emotionally charged messages that a text can contain arose. Sentiment analysis and opinion extraction are new tasks that call for a whole new range of semantic resources which will be able to support these new, sophisticated demands. Those resources are often integrated with WordNet, or they communicate with WordNet in various ways. Strapparava and Valitutti [20], starting from WordNet, precisely from its subset WordNet Affect, have built add-ons for WordNet-Affect in the form of affective labels (a-labels). Esuli and Sebastiani [3] created SentiWordNet, a publicly available lexical resource for sentiment analysis. In the work of Ohana and Turney [17], a method for sentiment classification of movie reviews using SentiWordNet was proposed.

For the purpose of enriching SWN with the data concerning sentiment measurement we explored SentiWordNet. It is a lexical resource for opinion mining based on the Princeton WordNet. It assigns three sentiment scores: positivity, negativity and objectivity to each record related to a WordNet synset. In that way SentiWordNet lexical resource determines positive and negative sentiment synset concepts. For each synset in PWN there is one record in SentiWordNet which has numerical scores Pos(s), Neg(s) and Obj(s) describing how positive, negative and objective the terms contained in the synset are. For example, the structure of a SentiWordNet record corresponding to a PWN synset defining adjective gladsome, is:

```
a   01361705 0.75 0 gladsome#1 “experiencing or expressing
    gladness or joy”; “a gladsome smile”; “a gladsome occasion”
```

where values 0.75 and 0 are numerical strength scores for a positive and negative affect, respectively, “a 01361705” means adjective in the PWN file with the ID=01361705, and # 1 is the number of a sense. We decided to introduce two tags (for positive and negative sentiment scores) for each SWN synset and to make a connection with the corresponding record in SentiWordNet. An application tool was created in order to change the current structure of SWN by adding a pair of tags as it is shown below:

```
<SRPWN>
 <SYNSET>
  <ID>ENG30-01361705-a</ID>
  <POS>a</POS>
  <SYNONYM>
   <LITERAL>drag<SENSE>1</SENSE><LNOTE></LNOTE></LITERAL>
        </SYNONYM>
        <DEF>koji oseća ili izražava radost i veselje</DEF>
        <BCS></BCS>
        <ILR>ENG30-01361414-a<TYPE>similar_to</TYPE></ILR>
        <NL>yes</NL>
        <USAGE></USAGE>
        <SNOTE></SNOTE>
        <STAMP>08/10/2012 00:00:00 jeca</STAMP>
        <SUMO>EmotionalState<TYPE>+</TYPE></SUMO>
        <SENTIMENT>
                <POSITIVE>0.75000</POSITIVE>
                <NEGATIVE>0.00000</NEGATIVE>
        </SENTIMENT>
 </SYNSET>
</SRPWN>
```

As it was the case with establishing relations with SUMO ontology, we followed the principle of horizontal parallelization (English synset – Serbian synset) and vertical inheritance (Serbian nontagged synset inherits the SENTIMENT tag of its parent synset, if one exists, otherwise it is added manually).

4 Improvement of the WordNet Editing Tool

Establishing relations between SWN and other semantic resources that was described in Section 3 of this paper was a two-phase process. In its first phase, relations between the existing SWN synsets and corresponding concepts of those resources were generated using an additional software tool, i.e. new tags in SWN synsets were generated, according to established relations, as was previously explained. In the second phase, the SWN tool itself enabled permanent integration with described resources, so for each new synset that would be added to SWN those relations are available.

The types of relations between SUMO ontology and WordNet 3.0 are defined with mapping methods set by the ontology authors themselves [16], which we are showing in Table 1.

Table 1: Types of relations between WordNet synsets and SUMO classes

Type of relation between WordNet synset and SUMO class	**Type of relation tag WordNet-SUMO**
WordNet synset is an instance of the SUMO concept	**@**
WordNet synset is subsumed by the SUMO concept	+
WordNet synset is equivalent to the SUMO concept	=
SUMO concept is an instance of the WordNet synset	**:**
SUMO concept is subsumed by the WordNet synset	**[**
SUMO concept is equivalent to the WordNet synset	**]**

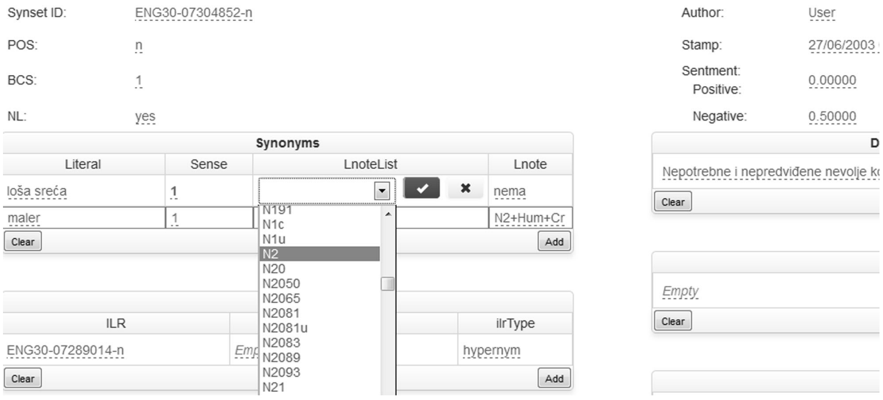
In Table 2, the results of integration of SWN with semantic resources after the first phase of development are given.

Apart from being able to provide us with various statistical data in regards to the set of WordNet tags and values of elements and attributes, the SWNE 2.0 can also give details on the distribution of relations established with the previously described semantic resources. For example, distribution of the most frequent SUMO classes in Serbian WordNet, acquired thanks to the SWNE 2.0 tool is given in Table 3.

Table 2: Distribution of assigning semantic tags to SWN synsets

Total number of SWN synsets	21234
Number of SWN synsets with POSITIVE sentiment tag	1977
Number of SWN synsets with NEGATIVE sentiment tag	2065
Number of SWN synsets with LNOTE tag	13798
Number of SUMO classes assigned to SWN synsets	3292
Number of SWN synsets with SUMO tag	21183
Number of SWN synsets with SUMO tag :	3
Number of SWN synsets with SUMO tag @	1921
Number of SWN synsets with SUMO tag [	23
Number of SWN synsets with SUMO tag +	16411
Number of SWN synsets with SUMO tag =	2825

In the second phase of SWNE development, we generated software filters that allow for easy search and establishing relations between SWN synset tags and corresponding semantic resources. Figure 1 shows filters that represent the connection of a SWN synset with the SrpMD resource in the form of a drop-down list. For each LITERAL tag from the LnoteList list we can choose a corresponding Knnn unique class tag of flective rules from the DELAS dictionary.



Synset ID: ENG30-07304852-n
POS: n
BCS: 1
NL: yes

Author: User
Stamp: 27/06/2003
Sentment: Positive: 0.00000
Negative: 0.50000

Synonyms

Literal	Sense	LnoteList	Lnote
loša sreća	1		nema
maler	1		N2+Hum+Cr

ILR		ilrType
ENG30-07289014-n	Emp	hypernym

Nepotrebne i nepredviđene nevolje k

Empty

Fig. 1: Connecting SWN synset with SrpMD DELAS dictionary

In order to establish relations with SUMO ontology, for each new synset, we can choose a SUMO class, i.e. a concept from the drop-down list called sumoList. The type of relation that is to be established between a SWN synset and a SUMO class can be chosen from the sumoTypeList drop-down list. SWNE 2.0 is

Table 3: Distribution of 10 most frequent SUMO classes in SWN

SUMO class and type of relation to a SWN synset	**Number of SWN synsets**	%
FloweringPlant +	1577	7.43
SubjectiveAssessmentAttribute +	2065	3.36
Bird +	13798	2.76
PreparedFood +	3292	2.75
Device +	3	1.41
Fish +	1921	1.39
Text +	23	1.32
FruitOrVegetable +	16411	1.14
Position +	2825	1.11
EmotionalState +	2825	0.91

a tool which provides insight into the speed of the development of SWN. It can provide us with the distribution of frequency of addition of new synsets and of all concepts of corresponding semantic resources. Figure 2 shows the representation of the relation between SWN and SUMO.

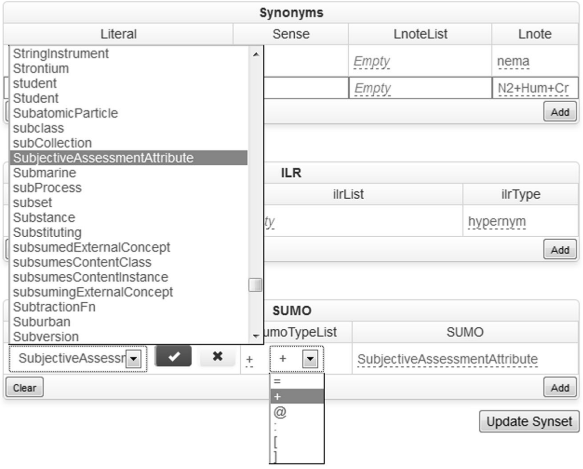


Fig. 2: Connecting of a SWN synset with SUMO ontology using a drop-down list

5 Future Work

The main advantage of the SWNE 2.0 software tool is enabling and strengthening collaborative work on the development of the Serbian WordNet. This tool allows supervising of the process of development through monitoring relevant statistics, such as the total number of synsets, the number of semantic relations, relations in regard to other semantic resources, quality and speed of tagging synsets (statistics listed by days and by authors). Further work on the SWNE tool, as well as on the development of new tools for semantic related research in Serbian will be leaning on the overall development of morphological dictionaries, the SWN and its relations with other semantic resources, above all, with geolocational network GeoWordNet [6] which is a semantic resource built around the connection between the WordNet and the GeoNames geographic database; connecting SWN with an Ontology of Rhetorical Figures RetFig is also planned in the near future, in order to increase the size of the database of rhetorical figures in Serbian, using SUMO ontology and ontological reasoning.

References

1. Michael J. Cole and Jacek Gwizdka. Tagging Semantics: Investigations with Wordnet. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 446–446, New York, NY, USA, 2008. ACM.
2. Blandine Courtois and Max Silberztein, editors. *Dictionnaires électroniques du français*. Langue Française, 1990.
3. Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422, 2006.
4. Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.
5. Sandiway Fong. Semantic Opposition and WordNet. *Journal of Logic, Language and Information*, 13(2):159–171, 2004.
6. Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, and Biswanath Dutta. Geowordnet: A resource for geo-spatial applications. In *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part I*, ESWC'10, pages 121–136, Berlin, Heidelberg, 2010. Springer-Verlag.
7. Yoan Gutiérrez, Sonia Vázquez, and Andrés Montoyo. Sentiment Classification Using Semantic Features Extracted from WordNet-based Resources. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 139–145, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
8. Aleš Horák and Pavel Smrž. VisDic - WordNet browsing and editing tool. In *Proceedings of the 2nd International Global WordNet Conference*, pages 12–14, 2004.
9. S. G. Kolte and S. G. Bhirud. Exploiting Links in WordNet Hierarchy for Word Sense Disambiguation of Nouns. In *Proceedings of the International Conference on Advances in Computing, Communication and Control*, ICAC3 '09, pages 20–25, New York, NY, USA, 2009. ACM.

10. Cvetana Krstev. *Processing of Serbian – Automata, Text and Electronic Dictionaries.* University of Belgrade, Faculty of Philology, Belgrade, 2008.
11. Cvetana Krstev, Bojana Djordjević, Sanja Antonić, Nevena Ivković-Berček, Zorica Zorica, Vesna Crnogorac, and Ljiljana Macura. Cooperative Work in Further Development of Serbian WordNet. *INFOteka*, IX(1–2):59–78, 2008.
12. Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. Using Textual and Lexical Resources in Developing Serbian Wordnet, 7:147–161. Romanian Journal of Information Science and Technology, 2004.
13. Cvetana Krstev, Ranka Stanković, Duško Vitas, and Ivan Obradović. WS4LR: A Workstation for Lexical Resources. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 1692–1697, Genoa, ELRA, 2006.
14. Miljana Mladenović and Jelena Mitrović. Ontology of Rhetorical Figures for Serbian. In *Text, Speech, and Dialogue - Lecture Notes in Computer Science*, 8082:386–393, 2013.
15. Miljana Mladenović and Jelena Mitrović. Developing and maintaining a WordNet: Procedures and Tools. In *Proceedings of the 7nd International Global WordNet Conference*, pages 55–62, Tartu, Estonia, 2014.
16. Ian Niles and Adam Pease. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416, 2003.
17. Bruno Ohana and Brendan Tierney. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *9th. IT & T Conference*, Dublin Institute of Technology, Dublin, Ireland, 2009.
18. Adam Pease. *Ontology: A Practical Guide.* Articulate Software Press, Angwin, CA, 2011.
19. Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christodoulakis, Dan Christodoulakis, Svetla Koeva, George Totkov, Dominique Totkov, and Maria Grigoriadou. Balkanet: A Multilingual Semantic Network for Balkan Languages. In *Proceedings of th 1st International Global WordNet Conference*, pages 12–14, 2002.
20. Carlo Strapparava and Alessandro Valitutti. WordNet–Affect: an affective extension of WordNet. In *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC'04)*, pages 1083–1086, 2004.
21. Dan Tufiş and Dan Ştefănescu. Experiments with a differential semantics annotation for wordnet 3.0. *Decision Support Systems*, 53(4):695–703, 2012.
22. Piek Vossen. Introduction to EuroWordNet. *Computers and the Humanities*, 32(2–3):73–89, 1998.