Building Slovene WordNet

Tomaž Erjavec, Darja Fišer

Department of Knowledge Technologies,
Jožef Stefan Institute,
Jamova 39, Ljubljana, Slovenia,
tomaz.erjavec@ijs.si

Department of Translation,
Faculty of Arts, University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia,
darja.fiser1@guest.arnes.si

Abstract

A WordNet is a lexical database in which nouns, verbs, adjectives and adverbs are organized in a conceptual hierarchy, linking semantically and lexically related concepts. Such semantic lexicons have become one of the most valuable resources for a wide range of NLP research and applications, such as semantic tagging, automatic word-sense disambiguation, information retrieval and document summarisation. Following the WordNet design for the English language developed at Princeton, WordNets for a number of other languages have been developed in the past decade, taking the idea into the domain of multilingual processing. This paper reports on the prototype Slovene WordNet which currently contains about 5,000 top-level concepts. The resource has been automatically translated from the Serbian WordNet, with the help of a bilingual dictionary, synset literals ranked according to the frequency of corpus occurrence, and results manually corrected. The paper presents the results obtained, discusses some problems encountered along the way and points out some possibilities of automated acquisition and refinement of synsets in the future.

1. Introduction

WordNet (Fellbaum 1998) is an extensive lexical database in which words are divided by part of speech into nouns, verbs, adjectives and adverbs and are then organized into a hierarchy of nodes, where each node represents a concept. Words describing concepts are called literals and literals denoting the same concept are grouped into a synset. In WordNet, synsets are connected to one another with semantic and lexical relations, such as hiper-/hyponymy, meronymy, antonymy, derivative, etc.

As a semantic lexicon WordNet has become one of the most valuable resources for a wide range of NLP research and applications, such as semantic tagging, automatic word-sense disambiguation, information retrieval and document summarisation, thus representing a general trend in the field. Significant improvements in the overall performance of such systems by including WordNet have been reported in numerous publications, e.g. (Volk et al. 2002; Banerjee, Pedersen 2002; Gonzalo et al. 1998; Stevenson, Greenwood 2006).

Recent projects such as EuroWordNet[1] (Vossen 1998), BalkaNet[2] (Tufiş et al. 2004) and MultiWordNet[3] (Pianta et al. 2002) initiated the development of WordNets for many other languages, taking the idea into the domain of multilingual processing. Further WordNets are now being developed world-wide; the Global WordNet Association[4] maintains a list of existing WordNets which currently contains more than 30 languages.

One of such enterprises is the building of the prototype Slovene WordNet, which is presented in this paper. Our aim was to build a WordNet which would be self-contained but at the same time easily integratable with others. As far as coverage is concerned, we began with the production of a high quality core WordNet which would cover the basic lexical inventory of Slovene and could serve as a reliable starting point for further development and extensions. We wanted the WordNet to be useful for the widest possible range of applications in the later stages of the project, in both mono- and multilingual settings.

The paper is organized as follows: Section 2 presents the process of creating the Slovene WordNet, the resources used and the results. Section 3 discusses the limitations of the approach problems encountered along the way, and proposes some plans for further refinement and expansion of the WordNet. The paper ends with some Concluding remarks.

2. Creation of the Slovene WordNet

2.1. Approach

While several corpus resources exist for Slovene (FIDA[5], MULTEXT-East[6], SVEZ-IJS[7]), there is a general lack of semantic lexica. Therefore, much of the initial work on the Slovene WordNet had to be based on classical dictionaries and thus required extensive manual intervention. Being severely limited in the resources and manpower at our disposal, the expand model (Vossen 1998) seemed like the most suitable approach under the circumstances. In the expand model, which is much simpler to implement than the merge model, a fixed set of synsets is taken from an existing WordNet and these synsets are then translated into the target language. In the merge model, individual WordNets are developed independently and combined at the end of the process. The cost of the expand model is that the resulting WordNets are heavily biased by the original WordNet, which becomes even more disturbing when the linguistic systems differ considerably. Nevertheless, due to its greater simplicity, the expand model has already been adopted in previous multilingual WordNet development projects, such as the BalkaNet and MultiWordNet. And although former studies of the semi-automatic construction of expand WordNets are not entirely optimistic (cf. Rigau, Agirre 2002; Veronis, Ide 1994), we still believe it can be beneficent if we bear in mind the limitations of the approach and the reported suggestions

1 http://www.illc.uva.nl/EuroWordNet/
2 http://www.ceid.upatras.gr/Balkanet/
3 http://multiwordnet.itc.it/english/home.php
4 http://www.globalwordnet.org/gwa/wordnet_table.htm
5 http://www.fida.net/slo/index.html
6 http://nl.ijs.si/ME/
7 http://nl.ijs.si/svez/

for improvement. The results can also be improved by complementing the WordNet refinement and enrichment by corpus-based techniques. Furthermore, the approach taken also fulfills one of the key project goals; the simple and unambiguous integration of the Slovene WordNet in a Princeton WordNet-centred multilingual WordNet infrastructure.

In our case, the notion proposed by Vossen (1998) that a relation holding between two synsets in the Princeton WordNet[8] (PWN) also holds between the corresponding synsets in the new language was taken a step further: we assumed that concepts and relations among them overlap across languages better if the languages are closely related. Instead of starting from the Princeton WordNet (PWN) we therefore used the Serbian WordNet (SWN) as the closest relative of Slovene in the WordNet family.

SWN's synsets were translated from English by hand and were validated against monolingual and bilingual dictionaries and corpora (Krstev et al. 2003; Krstev et al. 2004; Obradović et al. 2004), which is why it may be assumed that both synset equivalence across languages as well as Serbian synset contents are of high quality and representative of the actual language usage.

2.2. Resources used and WordNet creation procedure

The main resource for the automatic translation of literals was the Jurančič Slovene / Serbo-Croatian bilingual dictionary which was inverted to give pairs of Serbo-Croatian / Slovene lemmas. This lexicon was then used to automatically translate Serbian synset literals; the literals not found were retained in Serbian, and flagged for manual translation. Synset IDs and relations were preserved, while glosses, examples of use and sense numbers were omitted at this stage.

Given the expand approach, PWN 2.0 serves as the Interlingual Index, in the same way as in SWN and other BalkaNet WordNets. The SUMO/MILO[9] ontologies (Niles, Pease 2001) and DOMAINS[10] hierarchy (Bentivogli et al. 2004) have been aligned with PWN, which is why they automatically become available in each monolingual WordNet (Tufiş 2006), now including Slovene. The Suggested Upper Merged Ontology (SUMO) and the Mid-Level Ontology (MILO) with its domain ontologies form the largest formal public ontologies in existence today. They are being used for research and applications in search, linguistics and reasoning. WordNet Domains is an extension of PWN, where synsets have been annotated by domain labels, such as Medicine, Architecture and Sport. Semantic domains provide a natural way to establish semantic relations among word senses, which can be profitably used for word sense disambiguation: for each word we wish to disambiguate the domain of the context is estimated and compared to the domain of each word sense in the WordNet, finally the most similar one is selected.

In our project we adopted a top-down approach, concentrating on the top level ontology for the core WordNet, with a view of adding more specialised and language-specific concepts later. To this end, we retained only those synsets belonging to BalkaNet Base Concept sets. Synsets in BCS1 are essentially Base Concept Sets from EWN, while BCS2 and 3 have been selected on the basis of frequencies in the six languages involved in the BalkaNet project. BCS1, 2 and 3 are conceptually dense, which means that any concept has all its hypernyms up to the top of the hierarchies (Tufiş et al. 2004). So far, we concentrated on BCS1 and 2, which were first obtained through the procedure described above, giving us 4,688 synsets (1,219 from BCS1 and 3,469 from BCS2). This list was then checked for any missing hypernyms which were added to the WordNet, giving us a total of 4,841 top-level synsets. A surprisingly small number of the literals was left untranslated; only 676 out of 27,833. However, the automatically translated literals still required a substantial amount of manual clean-up which was carried out in VisDic, a tool for presentation and editing WordNet-like dictionary databases stored in XML format (Horak, Smrž 2004). A great advantage of VisDic is that it can be used to view and edit several dictionaries of various types (monolingual, translational, thesauri or generally linked WordNet lexicons) at a time in parallel dictionary panels.

Manual editing was performed in a series of steps, starting with a top-down approach, where an initial revision was carried out by chunks of related units (sets of related synsets, hyponymy trees, domains) rather than by individual units (synsets, senses, literals). The next step was manual translation of literals that were not translated via the dictionary followed by the revision of the translated synsets. The whole process was aided by various lexical resources (general and field-specific English-Slovene dictionaries, monolingual explanatory dictionary of Slovene) as well as by the FIDA reference corpus of the Slovene language (Erjavec et al. 1998). Manual revision was speeded up by classifying the literals into six bands according to their frequency in the lemmatised FIDA corpus. Band 0 – the lemmas that did not occur in the corpus (2,622 literals) – was examined with extra care in order to avoid unjustified exclusion of literals from the WordNet. These literals could not be automatically excluded from the WordNet since the corpus is only partially lemmatized, which is why some otherwise quite common literals fall into the lowest category. For example, words *era* (Eng. 'era') and *epoha* (Eng. 'epoch') fall into Band 0 but when the corpus is searched for *"er?"* and *"epoh?"* we receive 971 and 204 hits respectively. This suggests that automatic exclusion of such literals would be inappropriate.

2.3. Results

Table 1 shows the current top-level Slovene WordNet, which consists of 4,841 synonym sets and compares it with BCS1, 2, and 3 of the Serbian and Princeton WordNets. Out of 6,183 synsets in the Serbian WordNet, 73% have been included in the Slovene WordNet. As can be seen in the table, synsets from BCS1 and BCS2 are well-represented in the Slovene WordNet, while BCS3 is yet to be extended. Because this last stage is still on-going, the figures are likely to change in the future.

8 http://wordnet.princeton.edu/
9 http://www.ontologyportal.org/
10http://tcc.itc.it/research/textec/topics/disambiguation/index.html

	SloWN	SWN	PWN
BCS1			
nouns	965	965	964[11]
verbs	254	254	254
adjectives	0	0	0
adverbs	0	0	0
total	1219	1219	1218
BCS2			
nouns	2245	2245	2246
verbs	1188	1188	1188
adjectives	36	36	37
adverbs	0	0	0
total	3469	3469	3471
BCS3			
nouns	94	1187	2686
verbs	59	173	876
adjectives	0	135	265
adverbs	0	0	0
total	153	1495	3827
Grand total	**4841**	**6183**	**8516**

Table 1. Comparison of the number of synsets across POS in the three WordNets

As far as distribution of literals per synset is concerned (see Table 2), the average number of literals per synset is 2.13 for nouns and 2.35 for verbs. The longest Slovene synset among nouns in BCS1 is *{družina, rod, sorodstvo, pleme, klan, sorodniki, svojci, rodbina, žlahta}* (ENG20-07488154 {kin2}: group of people related by blood or marriage) with 9 literals and the longest Slovene synset among verbs in BCS1 is *{dodati, pridati, priložiti, navreči, primakniti, določiti, pridodati}* (ENG20-00176022 {add1}: make an addition (to); join or combine or unite with others; increase the quality, quantity, size or scope of) with 7 literals.

	SloWN	SWN	PWN
NOUNS			
synsets	965	965	964
literals	2056	1526	2135
avg. l/s	2.13	1.58	2.21
min l/s	1	1	1
max l/s	9	6	27
VERBS			
synsets	254	254	254
literals	607	481	729
avg. l/s	2.35	1.89	2.87
min l/s	2	1	1
max l/s	7	6	10

Table 2. Number of literals per synset in BCS1[12]

The relations used in the Slovene WordNet are summarized in Table 3. The hypernymy relation prevails with 4,727 occurrences, while the second most common relation (eng_derivative) is a lexical one, which is why it will be necessary to revise and replace it with Slovene data.

relation	no.	relation	no.
hypernym	4727	also_see	88
eng_derivative	2009	subevent	58
holo_part	299	causes	46
near_antonym	285	be_in_state	41
category_domain	170	holo_portion	34
verb_group	139	similar_to	8
holo_member	90		

Table 3. Relations used in SloWN

3. Discussion and future plans

This section lists some problematic aspects of employing the expand model over PWN. Some problems originate in the PWN itself, others stem from the inherent complexity of translating what is fundamentally an English language resource, still others are a consequence of our particular production method based on the translation of Serbian synsets into Slovene. The quality of the Slovene WordNet is thus heavily influenced by the quality and consistency of the resources used: the PWN, the Serbian WordNet and the Slovene / Serbo-Croatian bilingual dictionary. All in all, the automated translation of synsets resulted in high recall but very low precision: in BCS1, 1,108 of 1,219 of synsets were changed.

The first problem was that the automated translation of Serbian synsets into Slovene failed to translate any multi-word literals, which is why all collocations had to be extracted and added to the WordNet manually. Currently, the Slovene WordNet contains 1,344 multilingual literals.

A typical, and expected, translational error occurred in translations of polysemous literals where they were translated with equivalents that would be acceptable for some senses but not for this particular one (e.g. Eng.: {ending, conclusion, finish} (event whose occurrence ends something), SR: {*konac, kraj, svršetak, završetak*}, SI: {*izid, iztek, konec, končanje, kraj, krajnik, obrobje, nit, sklep, sukanec, zaključek, zatrep*}), requiring substantial manual clean-up. *Krajnik* is a relatively rare expression describing the end or edge of an object, not an event, and the word *obrobje* refers to the outer part of a place. Mistranslations *nit* and *sukanec* (Eng. 'thread') occurred because of the Serbian *konac*, a homonymous literal in the synset, which can mean either 'end' or 'thread'. Nevertheless, this problem would have been much worse if we started from PWN and an English-Slovene dictionary, as the weak relatedness of the two languages means that much more unresolved polysemy would have resulted from the automatic translation. So, for example, the English 'glass' would have been tranlsated into both *steklo* (Eng. {glass1}: a brittle transparent solid with irregular atomic structure) and *kozarec* (Eng. {glass2}: a glass container for holding liquids while drinking), while using the Serbian WordNet, this problem is avoided (Sr. *staklo* / Sl. *steklo*, Sr. *čaša* / Sl. *kozarec*).

As far as lexical discrepancies across languages are concerned, the literature suggests that syntactic and connotational divergences be disregarded and that expressions with the same denotation but different connotation be regarded as synonyms (Vossen 2005; Bentivogli, Pianta 2000). Nevertheless, this principle was sometimes difficult to follow as many cases of

11 The missing synset is ENG20-12509740 which is actually present in PWN but the tags for BCS1 are missing. However, two missing synsets have been identified in BCS2 in the Serbian and Slovene wordnets (synset ID: ENG20-00467580-n, synset ID: ENG20-01597253-a). The missing nominal synset {Go Fish} is a card game while the missing adjectival synset {little:4, small:4} describes someone or something not fully grown. The identified missing synsets will be subsequently added by hand to the Slovene wordnet.

12 Figures for BCS1 only are presented because editing of BCS2 and 3 has not been completed yet.

connotational inconsistencies originate in the PWN: in some cases, such literals are kept within the same synset (e.g. {grandma, grandmother, granny, grannie, gran}) while separated in others (e.g. {mother, female parent} -> [hypo] {ma, mama, mamma, mom, momma, mommy, mammy}).

The exceedingly fine granularity of senses in the PWN creates another difficulty for both Serbian and Slovene as there as many cases when such separation seems unjustified, e.g.:

{fluid:1} (a substance that is fluid at room temperature and pressure) -> [hypo] {liquid:1} (a substance that is liquid at room temperature and pressure)

{fluid:2} (a continuous amorphous substance that tends to flow and to conform to the outline of its container: a liquid or a gas) -> [hypo] {liquid:2} (a substance in the fluid state of matter having no fixed shape but a fixed volume)

In Serbian, fluid:1 is translated as *tećna supstanca*, fluid:2 as *fluid* and both liquid:1 and liquid:3 as *tećnost, tekućina*. In Slovene, there is no distinction between the senses fluid:1 and fluid:2, and between liquid:1 and liquid:3, nor has any evidence in Slovene resources been found to support the hyponymy relation as represented in the PWN. That is why having two synsets with identical contents and unjustified hyponyms seems redundant:

{*tekočina*:1, fluid:1} -> [hypo] {*tekočina*:1x}
{*tekočin*a:2, fluid:2} -> [hypo] {*tekočina*:2x}

The cross-PoS problem was already encountered during the construction of the Serbian WordNet; it occurs when a literal belonging to one PoS in English corresponds to a literal belonging to some other PoS in Serbian (Krstev et. al 2004). When translating the WordNet into Slovene, the problem was inherited: because literals were translated from Serbian, not from English, PoS of Slovene literals corresponds to their Serbian equivalents, regardless of the original PoS in English (e.g. Eng.: N {inverse, opposite}, SR: Adv {*obrnuto*, *suprotno*} , Sl.: Adv {*obratno*, *nasprotno*}). This means that across languages, synsets of different PoS are aligned and that the hypernym of such a synset will belong to a different PoS, which is in contradiction with the criteria for identification of relations between synsets (Vossen 3005).

While adopting solutions from Serbian was generally considered an advantage for easier construction of the Slovene WordNet, this was not the case with lexical gaps and denotational differences. According to Bentivogli and Pianta (2000), a lexical gap occurs when a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words. A denotational difference occurs when a concept is lexicalized in the target language but is more general or more specific than its source counterpart. In the Serbian WordNet, non-lexicalized concepts and generalizations are represented with their explanatory translation equivalents. As these multi-word literals could not be found in the bilingual dictionary, they were left untranslated, including the ones that are lexicalized in Slovene (e.g. Eng.: {great grandparent}, SR: {*roditelj babe i dede*}, SI: {*prastari starš*}), Eng.: {comestible, edible, eatable, pabulum, victual, victuals}, SR: {*jestive materije*}, SI: {*živilo*}).

So, while lexical gaps are represented with explanatory translation equivalents in the Serbian WordNet, the MultiWordNet approach (Pianta et al. 2002) suggests creating empty nodes whenever the lexical concept of one language has no correspondent in the other. This is in line with the EuroWordNet principle, saying that WordNets should avoid excessive dependency on the lexical and conceptual structure of the source language which can be achieved by allowing the new WordNet to diverge where necessary from the source (Vossen 1996). As our goal is to avoid developing an arbitrary lexical database, we will aim to follow this principle in further refinement of the Slovene WordNet. For example, {plant4} (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience) is not lexicalized in Slovene, that is why the concept was translated to Slovene with a free combination of words *"igralec iz publike"*, and a note "nonlexicalized" was included.

The existing database can now be further refined, augmented and updated. Our future plans involve the following: BCS3 needs to be added to the WordNet, the sense assignment conflict across synsets will need to be addressed and relations between synsets validated. Since Serbian is a closely related language to Slovene, we do not expect serious problems in this area but the Hierarchy Preservation Principle (Tufiş, Cristea 2002) still holds. According to it, semantic relations (e.g. hyponymy, holo-part) can be automatically imported but lexical relations (e.g. derivative, participle) are in general not valid across languages, and will thus have assigned anew.

More semantic information could be extracted from the monolingual explanatory dictionary of Slovene (SSKJ), especially by using dictionary definitions and terminological and phraseological fields. Several patterns in definitions could be exploited to extract hyponyms (e.g. 784 dictionary entries for "female form for") or lexical relations (e.g. 5,244 dictionary entries for "gerund of", 1,419 dictionary entries for "diminutive form of"). Later on, concept glosses for each synset and sample sentences for each literal will be added with the help of the explanatory dictionary of Slovene and the available corpora.

Furthermore, our plans for the future are to increasingly use automated means to refine the existing synsets and to acquire new ones with domain-specific vocabulary. We are considering strategies, such as extracting terms from existing available Slovene terminological lexica and other glossaries. We also believe we could benefit greatly from using multilingual parallel corpora, such as the EU ACQUIS corpus (Steinberger et al. 2006), to extract bilingual and multilingual lexica, and use those to find Slovene translation equivalents of, say, English literals. The obvious problem with this approach is the need to perform word-sense disambiguation on the English part of the corpus first; however, experiments show that this is easier to do on multilingual corpora than on the monolingual ones (Ide et al. 2002). Finally, we also need to consider formalised ways of evaluating the progress of the Slovene WordNet and to identify possible application areas.

4. Conclusions

The paper has presented the creation of the Slovene WordNet which uses the expand model and was based on the Serbian WordNet. The process was speeded up by automatic translation of Serbian synsets into Slovene with a bilingual dictionary. The results were examined and corrected by hand, resulting in about 5,000 synsets in BCS1 and 2. The paper also discussed the results obtained, limitations of the method used, as well as problems encountered along the way. The problems are similar to the ones encountered by other WordNet developers (cf. Cristea 2004; Pala 2004), which is why similar solution strategies have been adopted but, in certain cases, have not yet been fully implemented. In addition to that, a certain type of problems was caused by the chosen WordNet production method that is based on the translation of Serbian synsets into Slovene via a dictionary, such as mistranslations of polysemous literals and unsuccessful automated translations of multi-word literals. The quality of the Slovene WordNet is thus heavily influenced by the quality and consistency of the resources used: the PWN, the Serbian WordNet and the Slovene / Serbo-Croatian bilingual dictionary. We also discussed our future plans for further WordNet refinement and expansion.

As mentioned, the current Slovene WordNet is still being actively developed, so the numbers reported are expected to change soon. More up-to-date information is available on the home page of Slovene WordNet, at http://nl.ijs.si/slownet/. The WordNet is available without charge for research use; conditions on use and instructions for obtaining it are given on the Web page.

Acknowledgements

We would like to thank Duško Vitas and Cvetana Krstev from the University of Belgrade for permission to use the Serbian WordNet and all their assistance with producing the Slovene WordNet. The research reported in this paper was financed by the bilateral Slovene-Serbian project »Development of Slovene and Serbian Language Resources for Machine Translation«.

References

BANERJEE, S.; PEDERSEN, T. (2002): An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proc. of the Third CICLing Conf.*, Mexico City, Mexico, pp. 136-145.

BENTIVOGLI, L.; FORNER, P.; MAGNINI, B.; PIANTA, E. (2004): Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *Proc. of COLING 2004 Workshop on "Multilingual Linguistic Resources"*, Geneva, Switzerland, pp. 101-108.

CRISTEA, D., MIHAILA, C., FORASCU, C., TRANDABAT, D., HUSARCIUC, M., HAJA, G., and POSTOLACHE, O. (2004). Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets. In *Romanian Journal on Information Science and Technology*, Tufiş, D. (ed.), 7/ 1-2, pp. 125-147.

ERJAVEC, T.; GORJANC, V.; STABEJ, M.: Korpus FIDA. *In Proc. of the Intl. Multi-Conf. Intl. Society '98,* Jezikovne tehnologije za slovenski jezik, Ljubljana, Slovenia, pp. 124-127.

ERJAVEC, T. (2004) MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proc. of the Fourth Intl. Conferenc. on Language Resources and Evaluation, LREC '04*, pp. 1535 - 1538, ELRA, Paris, France.

ERJAVEC, T.; IGNAT, C.; POULIQUEN, B.; STEINBERGER, R. (2005): Massive multilingual corpus compilation: Acquis Communautaire and totale. In *Proc. of the 2nd Language & Technology Conf.*, Poznan, Poland.

FELLBAUM, C. (ed.) (1998): *WordNet: An Electronic Lexical Database*. MIT Press.

GONZALO, F. VERDEJO, I. CHUGUR, J. CIGARRÁN (1998): Indexing with WordNet synsets can improve Text Retrieval. In *Proc. of the COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing*, Montreal, Canada.

HORÁK, A.; SMRŽ, P. (2004): New Features of WordNet Editor VisDic. In *Romanian Journal of Information Science and Technology*. Dan Tufiş (ed.), 7/ 1-2, pp. 201-213.

IDE, N.; ERJAVEC, T.; TUFIŞ, D. (2002): Sense discrimination with parallel corpora. In *Word sense disambiguation: recent successes and future directions: Proc. of the workshop*, University of Pennsylvania, Philadelphia, USA, pp. 54-60.

KRSTEV, C.; PAVLOVIĆ-LAŽETIĆ, G.; OBRADOVIĆ, I.; VITAS, D. (2003): Corpora Issues in Validation of Serbian Wordnet, In *Lecture Notes in Artificial Intelligence*, LNAI 2807, Springer, pp. 132-137.

KRSTEV, C.; PAVLOVIĆ-LAŽETIĆ, G.; VITAS, D.; OBRADOVIĆ, I. (2004): Using textual resources in developing Serbian WordNet. In *Romanian Journal of Information Science and Technology*. Dan Tufiş (ed.), 7/ 1-2, pp. 147-161.

NILES, I.; PEASE, A. (2001): Towards a Standard Upper Ontology. In *Proc. of the 2nd Intl. Conf. on Formal Ontology in Information Systems*, FOIS-2001, Ogunquit, Maine.

OBRADOVIĆ, I.; KRSTEV, C.; PAVLOVIĆ-LAŽETIĆ, G.; VITAS, D. (2004): Corpus Based Validation of WordNet Using Frequency Parameters. In *Proc. of the Second Intl. WordNet Conf.* Brno, Czech Republic, pp. 181-186.

PALA, K.; SMRŽ, P. (2004): Building Czech WordNet. In *Romanian Journal of Information Science and Technology*. Dan Tufiş (ed.), 7/1-2, pp. 79-88.

PIANTA, E.; BENTIVOGLI, L.; GIRARDU, C. (2002): MultiWordNet: developing an aligned multilingual database. In *Proc. of the First Intl. Conf. on Global WordNet*, Mysore, India, pp. 293-302.

RIGAU, G.; AGIRRE, E. (2002): Merge and Expand Approaches to building wordnets. Tutorial2: Technologies to build Wordnets at the *First Intl. Conf. on Global WordNet*, Mysore, India.

STEINBERGER, R.; POULIQUEN, B.; WIDIGER, A.; IGNAT, C.; ERJAVEC, T.; TUFIŞ, D. (2006) The JRC Collection of the Acquis Communautaire – A multilingual parallel corpus with 20+ languages. In *Proc. of the Fifth Intl. Conf. on Language Resources and Evaluation, LREC '06*, ELRA, Paris, France.

STEVENSON, M.; GREENWOOD, M. A. (2006): Learning Information Extraction Patterns Using WordNet. In *Proc. of the Third Intl. WordNet Conf.*, Jeju Island, Korea, 2006, pp. 95-102.

TUFIŞ, D.; CRISTEA, D.; STAMOU, S. (2004): BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal of Information Science and Technology*. Dan Tufiş (ed), 7/ 1-2, pp. 9-43.

TUFIŞ, D.; CRISTEA, D. (2002). Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. In *Proc. of the Workshop on Wordnet Structures and Standardization, workshop in conjunction with The Third Intl. Conf. on Language Resources and Evaluation, LREC-'02*, Las Palmas, Spain, pp. 35-41.

TUFIŞ, D.; BARBU MITITELU E.; ION R.; BOZIANU L. (2004): The Romanian WordNet. In *Romanian Journal of Information Science and Technology*. Dan Tufiş (ed), 7/ 1-2, pp. 107-124.

TUFIŞ, D.; BARBU MITITELU, V.; BOZIANU, L.; MIHAILA, C. (2006): Romanian WordNet: New Developments and Applications. In *Proc. of the Third Intl. WordNet Conf., Jeju Island, Korea, 2006*, pp. 337-347.

IDE, N.; J. VERONIS (1994): Machine Readable Dictionaries: What have we learned, Where do we go? In *Proc. of the post-COLING '94 intl. workshop on directions of lexical research*, Beijing, pp. 137-146.

VOLK, M.; RIPPLINGER, B., VINTAR, Š.; BUITELAAR, P.; RAILEANU; D.; SACALEANU, B. (2002): Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. In: *Intl. Journal of Medical Informatics*. 67/1-3, pp. 97-112.

VOSSEN, P. (ed.) (1998): *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Press.