Building Croatian WordNet

Ida Raffaelli*, Marko Tadić*, Božo Bekavac*, Željko Agić**

*Department of Linguistics
**Department of Information Sciences
Faculty of Humanities and Social Sciences
University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
{ida.raffelli, marko.tadic, bbekavac, zeljko.agic}@ffzg.hr

Abstract. This paper reports on the prototype Croatian WordNet (CroWN). The resource has been collected by translating BCS1 and 2 from English, but also by usage of machine readable dictionary of Croatian language which was used for automatical extraction of semantic relations and their inclusion into CroWN. The paper presents the results obtained, discusses some problems encountered along the way and points out some possibilities of automated acquisition and populating synsets and their refinement in the future. In the second part the paper discusses the lexical particularities of Croatian, which are also shared between other Slavic languages (verbal aspect and derivation patterns), and points out the possible problems during the process of their inclusion in CroWN.

Keywords: WordNet, Croatian language, lexical semantics.

1 Introduction

WordNet has become one of the most valuable resources in any language for which the language technologies are tried to be built. One could say that having in mind the state-of-the-art in LT, a WordNet for a particular language could be considered as one of the basic lexical resources for that language. Semantically organized lexicons like WordNets can have a number of applications such as semantic tagging, word-sense disambiguation, information extraction, information retrieval, document classification and retrieval, etc. In the same time carefully designed and created WordNet represents one of possible models of a lexical system of a certain language and this pure linguistic value is sometimes being neglected or forgotten.

Following, but also widening the original Princeton design of WordNet for English [7], since EuroWordNet [18], a multilingual approach in building WordNets has taken the ground resulting in number of coordinated efforts for more than one language such as BalkaNet [17], MultiWordNet [9]. A comprehensive list of WordNet building initiatives is available at Global WordNet Association web-site[1].

In spite of efforts to coordinate building of WordNets for Central European languages (Polish, Slovak, Slovenian, Croatian, Hungarian) since 2nd GWC in Brno

[1] http://www.globalwordnet.org/gwa/wordnet_table.htm.

from 2004, building WordNets for these particular languages have started separately by respective national teams. The Croatian WordNet (CroWN from now on) is being built at the Institute of Linguistics, Faculty of Humanities and Social Sciences at the University of Zagreb. This paper represents the first report on the work-in-progress and the results that it presents are by all means preliminary.

The second section of the paper deals with the method of creating CroWN, dictionaries and corpora used. The third section discusses some particularities of Croatian lexical system that have been observed and which has to be taken into consideration while building the CroWN. The paper ends with future plans and concluding remarks.

2 The Process of Building

2.1 Method

To build a WordNet for a language there are two methods to choose from: 1) expand model [19], which in essence takes the source WordNet (usually PWN) and translates the selected set of synsets into target language and then later expands it with its own lexical semantic additions; and 2) merge model [19], where different separate (sub-)WordNets are being built for specific domains and later merged into a single WordNet. Both approaches have pros and cons with the former being simpler, less time- and man-months (i.e. also financially) consuming, while the latter is usually quite the opposite. On the other hand the results of the former approach are WordNets that are to a large extent at the upper hierarchy levels isomorphous with the source WordNet thus possibly deviating from the real lexical structure of a language. This can be noted particularly in the case of typologically different languages when number of discrepancies starts to grow. The latter approach reflects the lexical semantic structure more realistically but it can be hard to connect it with other WordNets and to make this resource usable for multilingual applications as well.

Having no semantically organized lexicons for Croatian except the [13] which exists only on paper, for initial stages of building CroWN we were forced to use existing monolingual Croatian lexicons which we had in digital form i.e. [1]. Also having very limited human and financial resources we were also forced to opt for expand model but we wanted to keep in mind all the time that it should not be reduced to a mere “copy, paste and translate” operation and that one should always take care about the differences in lexical systems. The expand model has being successfully used in a number of multilingual WordNet projects so we believed that this direction could not be wrong if we also consider thorough manual checking as well.

Up until now our top-down approach has been limited to the translation of BCS1, 2 and 3 from BalkaNet and additional data collecting from dictionary and corpora. The more specialized and more language-specific concepts will be added in further phases of creating CroWN. Table 1. shows basic statistics of POS in BCS1 and BCS2 of CroWN. The BCS3 is not included since it has not been completely adapted.

Table 1. Basic statistics on POS in BCS1 and BCS2 of CroWN.

	BCS1	BCS2	Total
Nouns	965	2245	3210
Verbs	254	1188	1442
Adjectives	0	36	36
Total	1219	3469	4688

2.2 Dictionary and its processing

The only dictionary resource we had available in machine readable form thus usable for populating the CroWN was [1]. Printed and CD-ROM edition of the dictionary contains approximately 70,000 dictionary entries. The right-side of lexicographic articles was divided into several subsections: part-of-speech and other grammatical information, domain abbreviations (e.g. anat. for anatomy), a number of entry definitions (containing various examples and synonyms), syntagms and phraseology, etymology and onomastics. Each of the subsections was labeled in original dictionary using a special symbol, making the dictionary easily processible. After extracting dictionary data and resolving some technical issues, we were left with 69,279 entries as candidates for the first phase of CroWN population. At this step, we omitted grammatical and lexicographic category information, phraseology, etymology and onomastics from articles but this information can be easily added later. In Figure 1 both original and simplified dictionary entries are shown:

pòstanak m
1. pojava, pojavljivanje, nastanak čega
2. prvi trenutak u razvoju čega; postanje
∆ Knjiga ~ka prva biblijska knjiga, govori o postanku svijeta

<ENTRY>
postanak
postanak **DEF** pojava, pojavljivanje, nastanak čega
postanak **DEF** prvi trenutak u razvoju čega; postanje
postanak **SINT** Knjiga ~ka prva biblijska knjiga, govori o postanku svijeta
</ENTRY>

Fig. 1. Original and reduced dictionary entry.

Each processed lexicographic element in reduced dictionary entry was tagged by the corresponding tag for definition, example and syntagm. Each headword was repeated before DEF and SINT tags, indicating that the definition and syntagm sections are linked to the entry. This redundant form was easily processed with regular patterns (local grammars) using NooJ environment [11]. The starting 69,279 entries now contained 88,352 different definition tags and 7,788 syntagm tags.[2]

[2] Note that the overall number of definitions is even bigger since we omitted as redundant the tags in single-line entries, i.e. those entries that contain only the headword and its right-side definition – their processing is trivial.

In this first extraction step we aimed at two things: 1) automatic linking of headwords to their definitions; and 2) creation of a set of well-defined lexical patterns which will be used to acquire additional knowledge from entries using information available in definitions and syntagms sections. We chose definitions and syntagms over all other lexicographic elements as definitions are more likely to contain well-formed word links than phraseology: e.g. the entry *crn* (en. *black*) has seven definitions in the dictionary and all of them are starting with *koji je* (en. *which is, that is*), providing a constant data extraction pattern. The same procedure is applicable to syntagms – *crni humor* (en. *black humor*), *crna lista* (en. *black list*), etc.

In dictionary filtering and pattern design, it was our intention to create correct and reliable set of WordNet entries containing basic information – their nearest hypo- and hyperonym classes, basic definitions and possible links to other entries.

In the preliminary test, which was used to determine whether the pattern method is feasible or not, we defined several lexical patterns and using NooJ tested them on our tagged and filtered dictionary. The simple patterns were defined in order to separate animate and inanimate nouns and also to try and link these nouns to other entry types similar in meaning. Some results are given in Table 2.

Table 2. Filtering definitions using lexical patterns.

Pattern	Extracted	Examples
onaj koji (en. *the one who*)	2138	*brojač* PATTERN *broji* (en. *counter* PATTERN *counts*) *psiholog* PATTERN *se bavi psihologijom* (en. *psychologist* PATTERN *does psychology*)
osoba koja (en. *the person that*)	90	*korisnik* PATTERN *se koristi računalom* (en. *user* PATTERN *uses a computer*)
osobina onoga koji (je) (en. *property of one who (is)*) *odlika onoga koji (je)* (en. *quality of one who (is)*)	170	*aktivnost* PATTERN *aktivan* (en. *activity* PATTERN *active*) *budnost* PATTERN *budan* (en. *awakeness* PATTERN *awake*)

We can make several conclusions from results of the type given in this table. The first one is that the pattern itself, if well-defined, can provide us with an insight on resulting entries; for example, *onaj koji* (en. *the one who*) clearly indicates a person, while *osobina onoga koji* (en. *a quality of one who*) indicates a property of an entity. Furthermore, although the [1] dictionary was written using a fairly controlled language subset, our patterns should still undergo parallel expansions in order to handle language variety that occurs in definitions (in Table 1: *property*, *quality* could be expanded with *feature*, *attribute*, etc.). Patterns should also be tuned with regard to article tokens occurring on its right sides; some of them could capture related nouns (*psychologist – psychology*) while others could link nouns to adjectives (*awakeness – awake*). Another possible enhancement to these patterns could be token sensitivity; if the dictionary were to be preprocessed with a PoS/MSD tagger or a morphological lexicon [16], pattern surroundings could be inspected and tokens collected with regard of their MSD and other properties (e.g. obligatory number, case and gender agreement in attribute constructions). Given these facts, we can come out with a conclusion to test: if carefully designed and paired with large, reliable dictionaries and MSD

tagging, pattern detection using local grammars could prove a good method for semi-automated construction of CroWN. Therefore, future dictionary processing and data acquiring tasks will include: enhancing all processing stages in order to collect even more definitions and syntagms that were left behind this first attempt in automatic CroWN population.

2.3 Corpora

We were aware that harvesting semantic relations encoded in the existing machine-readable dictionary, would still not be sufficient for building the exhaustive semantic net as WordNet should be. Therefore we also turned our attention to Croatian corpora and text collections in order to detect more examples and validate the existing ones.

As the treatment of compound words in WordNet from version 1.6. became more important, and since we had developed a system for detecting, collecting and processing compounds words (i.e. syntagms) [5], we decided to include them in CroWN right after completing the translation of BCS1-3. Overview of the compound words in the WordNet and their treatment is described in [10] so we will not go into details here.

When building an ontology from the scratch it is very useful to have a huge source of potential candidates for ontology population. For this task we used the downloaded Croatian edition of Wikipedia (http://hr.wikipedia.org) which at that time comprised 30,985 articles. For identification of distinctive compounds we extracted all explicitly tagged Wikipedia links, that undoubtfully point to a concept which was worded with at least two lower case words. The example can be seen in Figure 2.

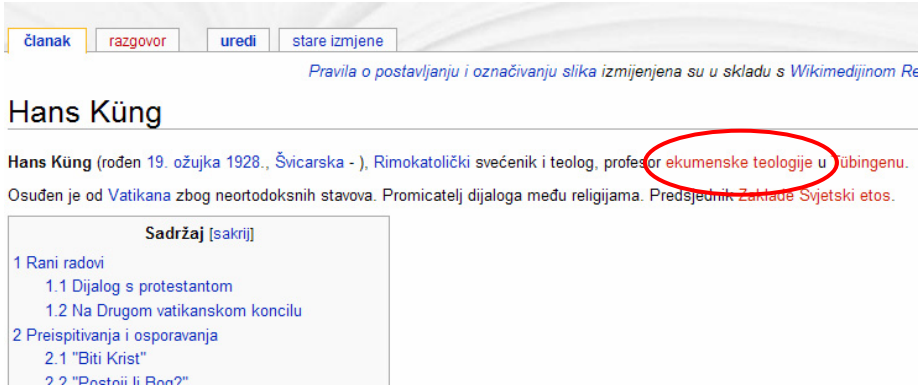

Fig. 2. Example of targeted compound from Wikipedia (circled text *ekumenske teologije*).

Definition of internal compound structures serves as filter for elimination of unwanted candidates. Examples of such patterns are combinations of MSDs like Adjective + Noun: e.g. *plava zastava* (en. *blue flag*); Noun + Noun-in-Genitive: e.g. *djeca branitelja* (en. *children of defenders*); Noun+Preposition+Noun-in-case-governed-by-

preposition: e.g. *hokej na ledu* (en. litteraly *hockey on ice*) etc. The compound dictionary collected in this way has also been included in lexical pattern processing of dictionary text described in the previous section.

Since we are still in the process of collecting and processing basic resources to create CroWN, we have not used Croatian National Corpus (HNK) for collecting literals. However it will be used in the process of corpus evidence and validation of literals within synsets used in CroWN.

Of course the last step before the inclusion of new items in CroWN is always the human checking and postprocessing of retrieved candidates where the final judgment about their inclusion and position in CroWN is taking place.

3. Particularities of Croatian

In this part of the paper we would like to discuss some underlying problems that we have detected while we were examining the structure of the Croatian lexical system which could, we believe, be relevant for building WordNets of other languages.

Except the necessity to be compatible with other WordNets, CroWN should preserve and maintain language specificity of Croatian lexical system in order to be a computational lexical database which reflects all semanti specifics of lexical structures in Croatian. Specifics of semantic and lexical structures in Croatian will especially become relevant in the construction of synsets at deeper hierarchical levels. Beside linking synsets with basic relations such as (near)synonyms, hypo/hypernyms, antonyms and meronyms, some of morphosemantic phenomena typical not only for Croatian, but also for other Slavic languages, should be taken into consideration and integrated in the construction of synsets and linking lexical entries within a synset.

Two of the most problematic language-specific phenomena of Croatian (which are shared with other Slavic languages) that should have inevitable impact on creating CroWN are: 1) verbal aspect and 2) derivation. Although these phenomena are traditionally considered as morphological processes, their impact to the semantic structure of a lexical unit should not be neglected in labeling lexical entries in CroWN. Moreover, as we will try to show all of these two morphological processes exhibit some regularity in patterns in Croatian derivation which could be exploited for automatic labeling of lexical entries. Regular derivational patterns characteristic for each morphological category should not be considered without close examination of their role in changing the semantic structure of a certain lexical entry in the CroWN. In other words, regularity of morphosemantic or derivational patterns could be useful for automatic labeling of senses in the CroWN, but at the same time there are many cases in the lexical system where some of these patterns considerably have motivated the change of the meaning from the basic lexical item.

3.1 Verbal Aspect

In one of the most recent Croatian grammars [12] aspect is defined as an instrument to express a difference between an ongoing action (imperfective aspect) and an action that has already been finished (perfective aspect). The category of aspect enables the

division of verbs in Croatian into perfective verbs and imperfective verbs which stand in binary opposition. Perfective verbs could be derived from the imperfective verbs and vice versa, imperfective verbs could be derived from the perfective verbs. Traditionally, aspectual verbal pairs are treated as separate lexical entries and in lexica they are sometimes listed as separate headwords and sometimes under the same headword (usually imperfective). Both practices can exist in the same dictionary in parallel. Some of the most prominent derivational patterns in the formation of both, perfective and imperfective verbs are the following:

1) Perfective verbs could be formed from imperfective verbs by substitution of the suffix of the verbal stem of an imperfective verb. The perfective verb e.g. *bac**i**ti* (en. *to throw*) is formed by substitution of the suffix ***-a*** of the verbal stem *bac**a**ti* (en. *to throw* as imperfective verb) with the suffix ***–i***. Similar substitutional patterns cover other suffixes.

2) Perfective verbs could be formed by adding the prefix (e.g. ***pre-***, ***na-***, ***u-***, ***pri-***, ***do-***, ***od-***, ***pro-***, etc.) to the verbal stem of an imperfective verb. Many perfective verbs are formed this way: *gledati* (en. *to look*) – *pregledati* (en. *to look over*, *to examine*), *hodati* (en. *to walk*) – *prehodati* (en. to walk a distance, used often in a metaphorical sense, meaning *to walk a flu over*), *pisati* (en. *to write*) – *prepisati* (en. *to copy in writing*) and many others.

As it could be observed from the previous examples, adding the prefix pre- to the verbal stem of an imperfective verb enables the formation of the perfective verb using regular and frequent derivational pattern, but it also triggers some of not negligible changes of the semantic structure of the basic verbal meaning. If we take the example of the aspect pair *pisati* (en. *to write*) – ***pre**pisati* (en. *to copy in writing*) the semantic change of the perfective verb ***pre**pisati* is quite significant with respect to the imperfective verb *pisati*. Though, there is another derivational pattern for the formation of a perfective verb from the imperfective pisati i.e. it is possible to add the prefix ***na-*** to the same verbal stem. The aspect pair *pisati* – ***na**pisati* does not exhibit a significant semantic shift of the derived verb toward a new meaning as in the previous case. Moreover, the derivational pattern has been introducing only the distinction between an ongoing and an already finished action

The same pattern exhibit the aspect pair *gledati* (en. *to look*) – ***pre**gledati* (en. *to examine*) pointing again to the significant semantic shift of the perfective verb, whereas the aspect pair *gledati* – ***po**gledati* (perfective verb formed by adding the prefix ***po-***) is exclusively related with respect to the differentiation of the type of action.

3) The most prominent pattern of the formation of the imperfective verbs from the perfective ones is the substitution of the suffixes of the verbal stem with derivational morphemes such as *-a-*, *-ava-*, and *-iva-* like in examples: *preporuč-i-ti* › *preporuč-a-ti*, *prouč-i-ti* › *prouč-ava-ti* and *uključ-i-ti* › *uključ-iva-ti*.

It is necessary to point out that this kind of formational pattern does not trigger significant semantic changes of the formed (imperfective) verb. The aspect pair is in binary opposition only with respect to the type of action (prefective or imperfective) they are referring to.

Basically, in Croatian grammars [3] and [12] verbs which differentiate with respect to the type of the action are considered as aspect pairs. However, aspect pairs could also differentiate with respect to the nature of the action or the way the action is

effected. This way of differentiating verbs which form an aspect pair is highly semantically motivated and should be taken into consideration when placing the lexical entries within a synset. For example the aspect pairs *kopati* (en. *to dig*) – ***ot****kopati* (en. *to dig up*), *kopati* – ***za****kopati* and *kopati* – ***po****kopati* semantically differentiate primarily with respect to the nature of the action. In the first aspect pair the perfective verb exhibits the beginning of the action (inchoative meaning), the two other pairs exhibit the end of the action (finitive meaning). It should also be pointed out that perfective verbs *zakopati* and *pokopati* do not have the same meaning. The verb *pokopati* means "to bury", whereas *zakopati* could mean "to bury" but also "to cover with something".

Grammar [12] distinguishes 11 different meanings of the aspect pairs with respect to the nature of the action and it is clear that this task will not be simple and without problems. The main issue is could we differentiate between these subtle senses using automatic techniques instead of tedious manual validation against the corpora.

3.2 Derivation

As Pala and Hlaváčková in [8] point out, derivational relations in highly inflectional languages represent a system of semantic relations that definitely reflects cognitive structures that may be related to language ontology. Derivational processes are deeply integrated in language knowledge of every speaker and represent a system which is morphologically and semantically highly structured. Therefore, as it is stressed in [8], derivational processes can not be neglected in building Czech Wordnet as well as any other Slavic language WordNet.

As already mentioned derivations in Slavic languages are highly regular and are suitable for automatic processing. In the paper [8] 14 (+2) derivational patterns have been adopted as a starting point for the organization of so-called derivational nests of the Czech Wordnet. They are aware of the main problem considering the derivational patterns and relations. Although there exists a significant number of cases where affixes preserve their meaning in Czech as well as in Croatian, it should be taken into consideration that there is also many cases where affixes do not preserve their prototypical meaning and become semantically opaque. This certainly poses a problem for automatic processing of derivational patterns and relations.

If we consider perfixation as one of possible derivational processes in Croatian as in Czech [8] as well, it is without any doubt that prefixes denote different relations such as time, place, course of action, and other circumstances of the main action. There are many cases where prefixes preserve their prototypical meaning, often related to its prototypical meaning as prepositions since most of them developed from prepositions. For example Croatian prefix *na-* has been developed from the preposition *na* with a prototypical meaning referring to the process of directing an object X on the surface of an object Y. There are many verbs in Croatian formed with the prefix *na-* where the prefix has preserved this meaning: *baciti* (en. *to throw*) – *nabaciti* (en. *to thow on sth/smb*), *lijepiti* (en. *to stick*) – *nalijepiti* (en. *to stick sth. on sth.*), *skočiti* (en. *to jump*) – *naskočiti* (en. *to jump on sth.*).

Unfortunately, this is not the only meaning of the prefix *na-* in Croatian. It also serves for derivation of a large number of verbs meaning *to do sth. to a large extent*.

For example: *krasti* (en. *to steal*) – *nakrasti* (en. *to steal heavily*), *kuhati* (en. *to cook*) – *nakuhati* (en. *to cook lot of food*, or *to cook for a long time*). In [2] there are three more meanings of the prefix *na-* and they should be integrated in any kind of automatic processing of prefixation in CroWN. Though in our opinion the greatest problem would represent some cases where the same verb, as a result of a prefixation, changes a meaning towards a completely new domain but still preserving some of possible meanings of the prefix.

Such an example is the verb ***na****pustiti*. The verb *pustiti* means *to drop*, *to let go/loose* while napustiti has two semantic cores or two basic meanings. One is related to the first meaning of the prefix *na-* (*to put X on the surface of Y*) and it is *to drop X on the surface of Y*. The other meaning is related to another possible meaning of *na-*; *lead to a result*. So *napustiti* could also mean *to abandon, to quit, to give up*. The connection between two semantic cores is hard to grasp for an average speaker of Croatian, but it could be explained with respect to different meanings of the prefix *na-*. In the CroWN the verb *napustiti* should be linked to the verb *pustiti* and its (near)synonyms, as well to verbs such as *ostaviti*, *odustati* which are both (near)synonyms of *napustiti*. What co-textual patterns will be detected in the corpus and will there be any explicit means to univocally differentiate between these senses remains to be seen.

As shown from the previous examples, derivational patterns such as suffixation and prefixation could not be considered as formal processes using affixes with simple and unique semantic value. Moreover, in highly grammatically motivated languages such as Croatian, as well as in any other Slavic language, suffixation and prefixation should not be regarded as grammatical processes which always result in same transparent and regular semantic changes of the basic lexical item. In many cases affixes used in derivational patterns lose their prototypical meaning enabling significant changes of the semantic structure of the basic lexical item thus influencing the organisation of highly structured morphosemantic relations

4. Future Plans and Concluding Remarks

Being at the very beginning of creating CroWN, this section could be expected to be quite extensive. In order to keep things moderate, we will list only the most imminent future plans to develop CroWN.

The first step would be the digitalization of [13] dictionary and its preprocessing for later usage. Being a lexicographically well-formed dictionary of synonyms in Croatian, this resource would provide us with huge amount of reliable data for direct CroWN synset acquisition and refinement.

The next step is refining and elaborating patterns for extraction of semantic relations from the dictionaries and corpora. This does not only include more complex lexical patterns but also additional dictionaries and corpora including mono- and multilingual such as Croatian-English Parallel Corpus [14] etc.

Particularly important for quality checking of CroWN will be proving the frequency data of literals and their meanings with Croatian reference corpus, namely Croatian National Corpus [15].

We expect to gain some insight also from checking correspondence with WordNets of genetically close languages (Slovenian, Serbian) [6,17] as well as culturally close languages (Slovenian, Czech, Hungarian, German, Italian), particularly at the level of culturally motivated concepts.

In this paper we have presented the first steps in creating Croatian WordNet which consist of translating BCS1, 2 and 3 from English into Croatian. Also we have described procedures for additional synset population from a machie-readable monolingual Croatian dictionary using lexical patterns and regular expressions. Similar procedure has been applied for compound words collection from a semistructured corpus of Croatian Wikipedia articles. Particularities of Croatian and possible problematic issues for defining synset structures are being discussed at the end of the paper with the hope that their solving will lead to a more thorough and precise semantic network of Croatian language.

Acknowledgments. This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1002, 130-1300646-1776 and 036-1300646-1986.

References

1. Anić, V.: Veliki rječnik hrvatskoga jezika. Novi liber, Zagreb (2003)
2. Babić, S.: Tvorba riječi u hrvatskome književnome jeziku. Croatian Academy of Sciences and Arts-Globus, Zagreb (2002)
3. Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., Znika, M.: Priručna gramatika hrvatskoga književnog jezika. Školska knjiga, Zagreb (1979)
4. Bekavac, B., Šojat, K., Tadić, M.: Zašto nam treba Hrvatski WordNet? In: Granić, J. (ed.) Semantika prirodnog jezika i metajezik semantike: Proceedings of annual conference of Croatian Applied Linguistics Society, pp 733--743. CALS, Zagreb-Split (2004)
5. Bekavac, B., Vučković, K., Tadić, M.: Croatian resources for NooJ (in press)
6. Erjavec, T., Fišer, D.: Building Slovene WordNet. In: Proceedings of the 5th LREC (CD). Genoa (2006)
7. Fellbaum, M. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
8. Pala, K., Hlaváčková, D.: Derivational Relations in Czech WordNet. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, pp 75--81. ACL, Prague (2007)
9. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: developing an aligned multilingual database. In: Proceedings of the First Global WordNet Conference, pp. 293--302. Mysore, India (2002)
10. Sharada, B.A., Girish, P.M.: WordNet Has No 'Recycle Bin'. In: Proceedings of the Second Global WordNet Conference, pp. 311--319. Brno, Czech Republic (2004)
11. Silberztein, M.: NooJ Manual (2006), http://www.nooj4nlp.net
12. Silić, J., Pranjković, I.: Gramatika hrvatskoga jezika. Školska knjiga, Zagreb (2005)
13. Šarić, L., Wittschen, W.: Rječnik sinonima. Neretva-Universitätsverlag Aschenbeck und Isensee (2003)
14. Tadić, M.: Building Croatian-English Parallel Corpus. In: Proceedings of the 2nd LREC, pp. 523--530. Athens (2000)

15. Tadić, M.: Building Croatian National Corpus. In: Proceedings of the 4th LREC, pp. 441--446. Las Palmas (2002)
16. Tadić, M.: Croatian Lemmatization Server. In: Vulchanova, M., Koeva, S. (eds.) Proceedings of the 5th Formal Approaches to South Slavic and Balkan Languages Conference, pp. 140--146, Bulgarian Academy of Sciences, Sofia (2006)
17. Tufiş, D. (ed.): Special Issue on the BalkaNet Project. Romanian Journal of Information Science and Technology. 7 (1--2), 1--248 (2004)
18. Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998)
19. Vossen, P. (ed.): EuroWordNet: General Doccument, Final, Version 3. University of Amsterdam (2002), http://www.illc.uva.nl/EuroWordNet/ docs/GeneralDocPS.zip