Further expansion of the Croatian WordNet

Krešimir Šojat, Matea Filko
University of Zagreb
Zagreb, Croatia
ksojat@ffzg.hr
msrebaci@ffzg.hr

Antoni Oliver
Universitat Oberta de Catalunya
Barcelona, Catalonia (Spain)
aoliverg@uoc.edu

Abstract

In this paper a semi-automatic procedure for the expansion of the Croatian Wordnet (CroWN) is presented. An English-Croatian dictionary was used in order to translate monosemous PWN 3.0 English variants. The precision values of the automatic process is low (about 30%), but the results proved valuable for the enlargment of CroWN. After manual validation, 10,884 new synset-variant pairs were added to CroWN, achieving a total of 62,075 synset-variant pairs.

1 Introduction

The building of the Croatian Wordnet has begun in 2004 at the Institute of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb. The Croatian WordNet is a lexical database built through the expand model (Vossen, 1998). The development of the Croatian Wordnet (CroWN) can be divided into two major phases (CroWN 1.0 vs. CroWN 2.0 / 3.0). Both versions are available for download and on-line queries. CroWN 1.0. (Raffaeli et al., 2008) was built completely manually. The main objective in this phase of the project was to translate and adapt the so-called basic concept sets extracted from the WN version 1.5 and used in the multilingual projects EuroWordNet (EWN) and BalkaNet (BN). For each synset a meaning definition was translated and adapted. Each synset in CroWN 1.0 is also accompanied by one or more examples of contextual usage. Synsets contain *literals* or *synset variant pairs* of the same part of speech. CroWN 1.0 comprises 10,000 synsets. 8500 of these are from the basic concept sets of EWN and BN. Approximately 1500 noun synsets were added using the same procedure. Although rich in information and data, CroWN 1.0 is a relatively small resource. In order to make it more useful in various NLP tasks, the second phase of the project was primarily oriented toward its enlargement. CroWN 2.0 and CroWN 3.0 (Oliver et al., 2015; Oliver et al., 2016) were built by using different automatic approaches. These versions of the lexicon are the result of joint work between two research teams from Zagreb and Barcelona. CroWN 2.0 and 3.0 contain only synset-variant pairs in Croatian, i.e. meaning definitions and examples of contextual usage have not been translated (yet). CroWN 2.0 and CroWN 3.0 are available at the Open Multilingual Wordnet website[1].

In this paper we present a semi-automatic method that was used for further expanding of CroWN, i.e. for the creation of its version 3.1.

The paper is structured as follows: in section 2 we describe the algorithms and procedures applied in the creation of versions 2.0 and 3.0 and provide some statistics regarding the number of synsets, POS distribution etc. Section 3 deals with the procedure and resources applied in the experiment presented in this paper. In section 4 results are discussed as well as advantages or potential disadvantages of the method applied here. Section 4 brings concluding remarks and the outline of future work.

2 Versions of the Croatian Wordnet

At this time, CroWN is the only resource for Croatian that deals with lexical semantics and also provides multilingual links to similar resources via The Open Multilingual Wordnet project. As mentioned, CroWN 2.0 and CroWN 3.0 are the result of joint work between two research teams from Zagreb and Barcelona. The 2.0 version of the CroWN was developed using the WN-Toolkit[2] (Oliver, 2014), a set of Python programs for the

[1] http://compling.hss.ntu.edu.sg/omw/

[2] http://sourceforge.net/projects/wn-toolkit

automatic creation of wordnets following the expand model. The WN-Toolkit implements 3 different strategies for wordnet creation:
1. Dictionary-based strategy - bilingual dictionaries are used to translate English variants associated with each synset. The strategy can deal only with monosemous English variants, i.e. variants associated with a single synset.
2. BabelNet-based strategy - the data from the BabelNet (Navigli and Ponzetto, 2010) file was extracted in order to obtain the data for CroWN.
3. Parallel-corpus-based strategy - in order to extract a target language wordnet, at least the English part of a parallel corpus should be sense tagged with PWN synsets. As such resources are rare and not easily available, two additional procedures were used for the creation of such a corpus: machine translation of sense-tagged corpora and automatic sense tagging of the English part of the parallel corpus.

Another line of work in CroWN 2.0 was oriented towards the enlargement of verbal synsets in CroWN. In CroWN 1.0 nouns make up almost 75 % of the whole lexicon (7391 noun synsets vs. 2318 verb synsets). The goal was to make CroWN a more balanced and representative resource for Croatian by enlarging the number of verbs. For this purpose we used CroDeriV (Šojat et al., 2013)[3], a large derivational database of Croatian verbs. The data was extracted and matched with PWN automatically. A more detailed account of the procedure and results is given in (Oliver et al., 2015). As in all other procedures described here, all candidates for synsets were manually checked and corrected if necessary. Taking into account that every automatic processing of data is followed by a manual revision, all procedures discussed here can be considered as semi-automatic.

With all these strategies we reached the 70.63 % of the Core synsets ((Boyd-Graber et al., 2006)). Finally, we manually populated CroWN 2.0 with the remaining 1,456 synsets, thus reaching 100 % of the Core WordNet.

For the creation of the version 3.0 we used a new version of the WN-Toolkit. It implements several strategies for mapping lexical resources (Wikipedia, Wiktionary and Omegawiki). An extensive account of this procedure is given in (Oliver et al., to appear).

In table 1 the number of synsets and synset-variant pairs in each of the three versions is presented. More details will be given in the subsections below.

Version	Synsets	Synset-variants
V 1.0	10,026	31,367
V 2.0	23,137	47,931
V 3.0	25,658	51,168
V 3.1	31,614	62,075

Table 1: Number of synsets and synset-variant pairs in different versions of the CroWN

In the following section we explain the process of further extension of the CroWN V. 3.0 and the creation of the new V. 3.1.

3 Experimental part

3.1 Automatic creation of synset-variant candidate pairs

For the new extraction we have used the EH dictionary[4]. This is an on-line dictionary, and the source file is provided by the authors under request. The EH dictionary comprises 186,098 entries. The dictionary is a plain text file containing two columns: an English word and a Croatian word, with no POS information included, as in the following fragment:

```
mother majka
mother materinski
mother posiniti
```

However, correct information about the POS of each word is vital for the method applied here. We have therefore used the Croatian Morphological Lexicon ((Tadić and Fulgosi, 2003))[5] to automatically attach POS information to the dictionary entries. The data in this morphological lexicon is structured as follows (*majka* – mother; *materinski* – maternal; *posiniti* – to adopt as son).

```
majka majka Ncfsn
materinski materinski Afpmsny
posiniti posiniti Vmn
```

With such information we were able to attach the POS information to 79,608 dictionary entries:

```
mother majka n
mother materinski a
mother posiniti v
```

Dictionary entries with the POS information were used to translate monosemous English variants in PWN-3.0. A variant is regarded as monose-

[3] croderiv.ffzg.hr

[4] http://web.vip.hr/zcindori.vip/ehrjecnik/

[5] hml.ffzg.hr

mous, at least according to WordNet, if it is attached to a single synset. Table 2 shows the number of monosemous and polysemous variants in WordNet for each POS:

	Noun	Verb	Adj.	Adv.
All	117,798	11,529	21,479	4,481
Mono	101,863	6,277	16,503	3,748
Poli	15,935	5,252	4,976	733

Table 2: Monosemous and polysemous variants in PWN 3.0

The translation of the variants enabled the extraction of 62,353 Croatian synset-variant pairs. Table 3 displays the distribution by POS of the extracted data as well as the results of automatic evaluation. The evaluation was performed by comparing the extracted synset-variant pair with CroWN 3.0. In section 3.2.2 a more detailed evaluation is presented.

	Extract.	Eval.	Correct	%
All	62,353	30,123	9,357	31.06
Noun	33,451	17,829	5,803	32.55
Verb	14,230	8,754	2,695	30.79
Adj.	14,048	3,277	794	24.23
Adv.	624	263	65	24.71

Table 3: Extracted synset-variant pairs by POS and automatic evaluation figures

The automatically calculated precision values are low, about 31%. As the numbers indicate, there are 30,123 synset-variant pairs that were evaluated since they are present in the CroWN 3.0 versus 32.230 instances that could therefore not be evaluated. Further, 20,766 synset-variant pairs were evaluated as incorrect. A candidate is marked as incorrect if we have some variant for the given synset in the CroWN 3.0, but no the extracted variant. This extracted variant can be correct, but not present in the CroWN. The subset of pairs evaluated as incorrect can be also manually revised.

3.2 Manual revision and completion

In order to further evaluate the automatically extracted Croatian synset-variant pairs, all the results were revised by hand. During this time-consuming task we wanted to maximize our contribution and to expand CroWN as much as possible. Our revision was hence divided into several steps. First, non-evaluated candidates and candidates automatically evaluated as incorrect were set apart and evaluated in separate actions. Further, both sets of extracted Croatian synset-variant pairs were arranged according to PWN synset-IDs. Meaning definitions provided for PWN snysets were used as a criterion to evaluate candidates as correct or incorrect. In other words, each candidate was marked either as correct or incorrect on the basis of meaning definitions from PWN. During this process we were adding one or more Croatian variant pairs whenever it was possible. Finally, if none of the candidates for a particular synset was correct, we added new synset-variant pairs by hand as well.

3.2.1 Problems for the automatic approach

Manual evaluation of candidates revealed several problematic cases for the automatic method of expansion applied here. Problems that we faced regard to several aspects:
1. Problems that result from linguistic features of Croatian and American English as well as cultural differences that are reflected in conceptualization and lexicalization. One of the problems that we faced is related to the processing of multi-word expressions. For example, one of the senses of the noun *wall* in PWN is defined as *"a difficult or awkward situation"*. This candidate was translated with Croatian *zid*, a wall (as in *brick wall*). The problem for this and similar examples is that the Croatian noun is normally used in this sense only in idioms, e.g. *naići na zid*, *naći se pred zidom*. In other words, English synsets list literals that are used only as parts of idioms or phrasemes in Croatian.
2. Besides, several problems resulted from the fact that Croatian collocations composed of adjectives and nouns, e.g. *genska ekspresija*, generally act as a single semantic unit, whereas in English synsets only a noun is listed as a literal. Unlike in English, in many cases Croatian candidates were obligatory multi word expressions.
3. Further, we came across numerous cases in which PWN literals cannot be lexicalized in Croatian due to its morphological properties. Although derivation of nouns from verbs is common in Croatian, it is not possible for numerous PWN literals (e.g. there are no derivatives for *skidder, slider, slipper* defined as *"a person who slips or slides because of loss of traction"* and *chew, chaw, cud, quid, plug, wad* defined as *"a wad of*

something chewable as tobacco").

4. We also found several examples when concepts represented by PWN literals are lexicalized with completely other lexical means. For example, the closest relatives of the PWN literal *near miss* defined as *"an accidental collision that is narrowly avoided"* are various Croatian verbal idioms, e.g. *promašiti za dlaku, izbjeći za malo, "miss by a hair's breadth"* etc.

5. Some concepts from PWN do not exist at all in Croatian, e.g. *dictator*, as *"a speaker who dictates to a secretary or a recording machine"*, or *show-stopper, showstopper, stopper* as *"an act so striking or impressive that the show must be delayed until the audience quiets down"*. Since we could not come up with a better solution, in CroWN 3.1 we marked such examples with the tag GAP. The same mark was used for numerous expressions denoting concepts from various domains characteristic almost exclusively for the US. Problems that result from cultural differences pertain to specific terms used in stock market, the US legal system, sports as baseball and American football, cuisine etc. For example, PWN literals *bomber, submarine, torpedo* denote the same type of sandwich eaten in the US. The meaning definition for this synset points out that different names are used in different sections of the United States. Such words are almost impossible to translate or adapt without additional explanations. Candidates from this group exclusively belong to the non-evaluated part of the obtained candidates.

The second group of problems pertains to differences between Croatian and English:

6. An issue that poses a challenge to the adopted expand model pertains to cases when PWN literals can be translated only with Croatian words of different POS. For example, adjectival synset containing the adjective *several* should be translated with the adverb *nekoliko*. Similarly, but not so often, PWN literals can be translated only with Croatian suffixoids, i.e. units that are neither words nor morphemes, e.g. –ology, -ism etc. E.g., the most accurate translation of the PWN's *stasis "an abnormal state in which the normal flow of a liquid (such as blood) is slowed or stopped"* is the Croatian suffixoid *-staza*, although word *zastoj* can be used. Further, parts of English compounds are also sometimes listed as literals, e.g. *wort* is defined as: *"usually used in combination: 'liverwort'; 'milkwort'"*, which makes the processing almost impossible.

7. PWN verbal literals referring to both causative and reflexive senses of English verbs are also higly problematic. In Croatian, as it is common in Slavic, these are different verbs and consequently different lemmas. Lemmas for reflexive verbs include the reflexive pronoun *se* (e.g. *otopiti se* 'to become melted'), whereas causatives do not co-occur with *se* (e.g. *otopiti* 'to melt'). Such cases pose a challenge for the construction of verbal synsets in CroWN. On top of that, there is group of reflexive verbs that co-occur with the so-called reflexive particle *se* (e.g. *smijati se* 'to laugh'). As far as the discussed method of expansion is concerned, there were numerous cases when only infinitives were recognized, while reflexive pronouns or particles were missing.

8. Although phrasal verbs do not exist as a separate category according to Croatian grammars, based on the examples from CroWN, (Katunar et al., 2012) argue that they should be recognized and treated as such. In some cases, the meaning of verbs is altered by co-occurring prepositions, e.g. verb *držati* 'to think' vs. *držati do* 'to value'. The applied automatic approach can account only for infinitives, thus yielding incorrect candidates.

9. Finally, the problem with the automatic approach is that it relies on one-to-one translation and therefore offers all translation equivalents from the dictionary in all their senses. This usually results in one or more correct and one or more incorrect candidates per synset if the word in case is highly polysemous.

However, in many cases new candidates for the already existing synsets were offered, i.e. candidates omitted in previous versions of CroWN. The result is a more diversified language resource.

3.2.2 Evaluation of the methodology

The manual revision of the candidates facilitated the calculation of precision values for two subsets: the non-evaluated candidates and the candidates automatically evaluated as incorrect. Table 4 presents these values. They are similar (in the region of 30 %) as the values shown in table 3 for the automatic evaluation of the non-evaluated subset. The precision values for the incorrect subset are lower, as expected, but in this subset there are still about 15 % of correct synset-variant pairs.

In table 5 the number of synset-variant pairs for each POS for versions 3.0 and 3.1 are shown.

	P	P_N	P_V	P_A	P_R
non-eval.	30.06	29.6	18.98	39.53	-
incorrect	14.11	16.54	11.21	22.92	-

Table 4: Precision figures for the manually evaluated subsets.

	3.0	3.1
Nouns	30,240	38,951
Verbs	17,913	18,645
Adjectives	2,623	4,064
Adverbs	415	415
Total	51,191	62,075

Table 5: Number of synset-variant pairs in version CroWN 3.0 and 3.1.

Once all the new synset-variant pairs had been manually validated and corrected, we could calculate final values of precision for the applied methodology. In table 6, we present these figures, which are in fact very similar to the precision figures of the automatic evaluation in table 3.

	Extract.	Eval.	Correct	P.
All	62,353	46,774	14,682	31.39
Noun	33,451	30,802	9,880	32.08
Verb	14,230	10,111	2,969	29.36
Adj.	14,048	5,598	1,768	31,52
Adv.	624	263	65	24.71

Table 6: Extracted synset-variant pairs by POS and automatic evaluation figures

4 Conclusions and future work

The main goal of the experiment procedure described in this paper was to expand the CroWN 3.0 with a) new synsets, and b) new literals in the existing synsets. The development of CroWN is not financially supported on a regular basis, therefore automatic and semi-automatic procedures for its further expansion are particularly valuable. When dealing with large amount of data, it is easier to manually edit the results of the automatic extraction of candidates than to work from scratch.

The use of the EH dictionary has allowed us to further expand the Croatian Wordnet. In previous works we have used other free lexical resources (namely Omegawiki, Wiktionary and Wikipedia) and a similar methodology. The precision values obtained with EH are much lower that those obtained with other resources. The main reason is the size of the EH dictionary, which is much larger and provides a lot of translation equivalents for each English word. Some of these translation provide similar meaning that are not suitable for the construction of a wordnet.

Acknowledgments

This research has been carried thanks to the project TUNER TIN2015-65308-C5-1-R (MINECO/FEDER, UE) and the short-term research support of the University of Zagreb.

References

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the 3rd International WordNet Conference*, pages 29–36. GWC.

Daniela Katunar, Matea Srebačić, Ida Raffaelli, and Krešimir Šojat. 2012. Arguments for Phrasal Verbs in Croatian and Their Influence on Semantic Relations in Croatian WordNet. In *Proceedings of the LREC 2012*, pages 33–39. ELRA, Istanbul.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the ACL*, ACL '10, pages 216–225, Stroudsburg, PA, USA. ACL. ACM ID: 1858704.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2015. Enlarging the Croatian WordNet with WN-Toolkit and Cro-Deriv. In *RANLP*, pages 480–487.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2016. Automatic expansion of Croatian Wordnet. In Sanda Lucija Udier and Kristina Cergol Kovačević, editors, *Metodologija i primjena lingvističkih istraživanja*, pages 171–185. Zagreb.

Antoni Oliver, Krešimir Šojat, and Matea Filko. to appear. The Croatian WordNet: CroWN 3.0. *Linguistic Issues in Language Technology - LILT. Special Issue on Linking, Integrating and Extending Wordnets*, 10.

Antoni Oliver. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. In *Proceedings of the 7th GWC*, Tartu, Estonia.

Ida Raffaeli, Bekavac Božo, Željko Agić, and Marko Tadić. 2008. Building Croatian WordNet. In *Proceedings of the 4th GWC*, Szeged, Hungary.

Krešimir Šojat, Matea Srebačić, and Vanja Štefanec. 2013. CroDeriV and the morphological analysis of Croatian verb. *Suvremena lingvistika*, 75:75–96.

Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian Morphological Lexicon. In *Proceedings of the EACL Workshop on Morphological Processing of Slavic Languages*, pages 41–46. ACL, Budapest.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.