Enlarging the Croatian Wordnet with WN-Toolkit and CroDeriV

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Krešimir Šojat
Sveučilište u Zagrebu
ksojat@ffzg.hr

Matea Srebačić
Sveučilište u Zagrebu
msrebaci@unizg.hr

Abstract

Wordnet is a standard semantic resource for several Natural Language Processing tasks and it is available for an increasing number of languages. The Croatian Wordnet (CroWN) was a relatively small resource with 10.026 synsets and 31.367 synset-variant pairs covering only 45.91% of the so-called Core WordNet. Comparing these figures with the size of the Princeton WordNet for English version 3.0, that has 117,659 synsets and 206,975 synset-variant pairs, it is clear that the CroWN should be expanded. First experiments for the expansion of the CroWN were performed using the WN-Toolkit, a set of Python programs for wordnet creation and expansion using dictionary, Babelnet and parallel-corpora based strategies. The WN-Toolkit was previously successfully applied to other languages as Spanish, Catalan and Galician. After this first expansion, CroWN reached 70.63% of the core wordnet. In the second step we used CroDeriv, a derivational database for Croatian and the manual creation of 1,457 synset-variant pairs until reaching 100% of the Core WordNet. After second step was completed, CroWN reached 23,137 synsets and 47,931 synset-lemma pairs.

1 Introduction

In this paper we explain the methodology and results of the experiments for the enlargement of the Croatian Wordnet using the WN-Toolkit and the derivational database CroDeriV. The paper is organised as follows: first, we will explain the development of the previous version of CroWN and we will present some figures about the size of this wordnet before and after the expansion. Then we will present the WN-Toolkit and its main features. After that, in section 4, we will present the CroDeriV, morphological database of Croatian verbs which was used in one of our experiments. Next, the experimental methodology is presented followed by the results of the experiments. After that the main sources of errors are presented and analyzed. Finally, the conclusions and future work are presented.

2 CroWN and wordnets for other languages

The Croatian Wordnet has been developed under the Central and South-East European Resources (CESAR) project, funded by the European Commission (50%) and the Ministry of Science, Education and Sports of the Republic of Croatia (50%). The first version of the Croatian Wordnet had 10.026 synsets and 31.252 synset-variant pairs. The synset ID's are those of the Princeton WordNet for English v 3.0.

The Princeton WordNet for English version 3.0 has 117,659 synsets and 206,975 synset-variant pairs. In table 1 we can observe the number of synset variant pairs both in old and new versions of CroWN. In table 2 the number of synsets in both versions is shown. The starting version of the Croatian Wordnet covered 45.91% of the so-called Core WordNet (Boyd-Graber et al., 2006), that is, approximately the 5,000 most frequently used word senses. After the automatic expansion described in this paper 100% of the core synsets were covered.

The Open Multilingual Wordnet[1] (OMW) (Bond and Paik, 2012) provides free access to several wordnets in a common format. The new CroWN is also distributed in the Open Multilingual Wordnet website. In table 3 we can observe all the wordnets in OML with the relative position

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

Proceedings of Recent Advances in Natural Language Processing, pages 480–487,
Hissar, Bulgaria, Sep 7–9 2015.

POS	Old version	New version
Overall	31,252	47,901
Nouns	16,726	27,001
Verbs	13,669	17,904
Adjectives	857	2,594
Adverbs	0	402

Table 1: Number of synset-variant pairs in old and new versions of CroWN

POS	Old version	New version
Overall	10,026	23,120
Nouns	7,373	16,178
Verbs	2,351	4,736
Adjectives	302	1,814
Adverbs	0	392

Table 2: Number of synsets in old and new versions of CroWN

regarding the number of synsets (on the left) and the % of the Core WordNet (on the right), both for the old version (hrv-o) and the new version (hrv-n) of the CroWN. As we can observe in the table, regarding the number of synsets, the old CroWN occupied the 21st position, and the new version reached the 17th. With respect to the % of the Core WordNet, the old version occupied the 24th position and the new one, as it reached the 100%, occupies the 4th position.

These figures indicate that the CroWN, after the enlargement described in this paper, is a much more valuable resource, although there is still a lot of work to be done.

3 The WN-Toolkit

The WN-Toolkit[2] (Oliver, 2014) is a set of programs developed in Python for the automatic creation of wordnets following the expand model (Vossen, 1998), that is, by translation of the variants (words) associated with the Princeton WordNet synsets. The toolkit also provides some free language resources. These resources are preprocessed so they can be easily used with the toolkit.

The WN-Toolkit implements the following strategies for WordNet creation:

- Dictionary based methodology: This strategy uses bilingual dictionaries to translate the English variants associated with each synset. This direct translation using dictionaries can be performed only on those English variants being monosemic, that is, variants associated to a single synset. About 82% of the English variants in the Princeton WordNet 3.0 are monosemic. These figures show us that a large percentage of a target wordnet can be implemented using this strategy, but we would not be able to extract the most frequent variants, as common words are usually polysemic.

- Babelnet based strategies: BabelNet (Navigli and Ponzetto, 2010) is a semantic network and a multilingual encyclopedic dictionary with lexicographic and encyclopedic coverage of terms. Entries are connected in a very large network of semantic relations. BabelNet covers 50 languages, Croatian among them. In this methodology we simply extract the data from the BabelNet file to get the target wordnet. This strategy can only be applied to old versions of Babelnet, as new versions have a use restriction not allowing the creation of wordnets from its data.

- Parallel corpus based methodologies: In or-

[2]The WN-Toolkit can be freely downloaded from `http://sourceforge.net/projects/wn-toolkit/`

P	lang	synsets	lang	% CORE
1	eng	117,659	eng	100
2	fin	116,763	fin	100
3	tha	73,350	cmn	100
4	fra	59,091	**hrv-n**	**100**
5	jpn	57,184	bul	100
6	ind	51,822	ind	99
7	cat	45,826	zsm	99
8	por	43,895	swe	99
9	zsm	42,679	jpn	95
10	slv	42,583	fra	92
11	cmn	42,312	slv	86
12	spa	38,512	por	84
13	ita	34,728	ita	83
14	eus	29,413	tha	81
15	pol	28,757	cat	81
16	**hrv-n**	**23,120**	dan	81
17	glg	19,312	nob	81
18	ell	18,049	spa	76
19	fas	17,759	eus	71
20	arb	10,165	nno	66
21	**hrv-o**	**10,026**	ell	57
22	swe	6,796	pol	49
23	heb	5,448	arb	48
24	bul	4,999	**hrv-o**	**45.91**
25	qcn	4,913	fas	41
26	als	4,676	glg	36
27	dan	4,476	als	31
28	nob	4,455	qcn	28
29	nno	3,671	heb	27

Table 3: Number of synsets in old (hrv-o) and new (hrv-n) versions of CroWN

der to extract wordnets from a parallel corpus we need this parallel corpus to be semantically tagged with Princeton WordNet synsets in the English part. As these corpora are not easily available, we use two strategies for the automatic construction of the required corpora:

- By machine translation of sense-tagged corpora.
- By automatic sense-tagging of English-Croatian parallel corpora.

The WN-Toolkit also provides some resources, as dictionaries and preprocessed bilingual corpora.

4 The CroDeriV database

CroDeriV (Šojat et al., 2014) is a database that contains information about the morphological structure and derivational relatedness of verbs in Croatian. Nowadays it contains 14,192 Croatian verbs that are morphologically analyzed, that is, segmented into lexical, derivational and inflectional morphemes. The structure of CroDeriV enables the detection of verbal derivational families in Croatian as well as the distribution and frequency of particular affixes and lexical morphemes. Derivational families consist of a verbal base form and all prefixed or suffixed derivatives detected in available Croatian dictionaries and corpora. Language data structured in this way was further used for the expansion of other language resources for Croatian, such as Croatian WordNet and the Croatian Morphological Lexicon (Šojat and Srebačić, 2014; Šojat et al., 2014). Matching the data from CroDeriV on one side, and Croatian WordNet and the Croatian Morphological Lexicon on the other, resulted in significant enrichment of Croatian WordNet and enlargement of the Croatian Morphological Lexicon.

In this paper we present the procedure for using CroDeriV to further expand the Croatian WordNet.

5 Experimental methodology

In order to automatically evaluate the results, we compare the obtained wordnet with the existing Croatian Wordnet. If we get some variant for a synset, we compare if in the Croatian WordNet there is a variant for this synset, and if this variant is the same as the extracted one. If we got one of the variants in the reference wordnet, the result is evaluated as correct. If there are some variants in the reference wordnet, but not the one we extracted, this is evaluated as incorrect. If we don't have any variant in the reference wordnet for the particular synset, the result remains unevaluated, that is, we don't take into account this obtained variant in the evaluation results. The automatic precision values obtained in this way tend to be lower than the real values. Sometimes we obtain a variant that is correct, but we have other correct variants for the same synset in the reference wordnet. In these cases we evaluate our result as incorrect. On the other hand, as the reference Croatian Wordnet is not very big, we leave a lot of obtained variants without evaluation.

As we stated in the previous section, automatically evaluated values of precision tend to be lower than the real values. For this reason, for each experiment we have manually evaluated a subset of the non-evaluated and incorrect results in order to calculate a corrected value of precision.

We offer two values of corrected precision values:

- strict: we have also considered small errors (as capitalization, plural forms, etc.) as errors
- non-strict: we have considered small errors as correct

6 Experimental results

6.1 Dictionary-based strategy

6.1.1 Resources

In the table 4 we can observe the dictionaries (English-Croatian) we have used for the experiments along with the number of entries. As can be seen in the table, only freely available resources have been used.

Dictionary	Website	Entries
OmegaWiki	http://www.omegawiki.org/	1,692
Wiktionary	http://www.wiktionary.org/	29,216
Wikipedia	http://www.wikipedia.org/	70,387
Geonames	http://www.geonames.org/	1,353
Wikispecies	http://species.wikimedia.org/	1,785

Table 4: Dictionaries used for the dictionary-based strategy

The Wiktionary dictionary contains words in Croatian, Bosnian and Serbian, some of them written in Cyrillic. We have filtered the dictionary with the Croatian Morphological Dictionary (Tadić and

	Omegawiki	Wiktionary	Wikipedia	Geonames	Wikispecies	Combination
Total	646	1,905	4,196	429	772	7,247
Evaluated	176	522	409	0	183	1,156
Precision	83.52	79.31	57.95	-	75.96	70.33
Precision N	57.14	80.65	57.95	-	75.96	70.49
Precision V	-	67.86	-	-	-	66.10
Precision A	-	83.33	-	-	-	83.33

Table 5: Results for the dictionary-based strategy using automatic evaluation

Fulgosi, 2003; Oliver and Tadić, 2004) in order to get a list of Croatian words, so words in the Wiktionary dictionary not being in the Croatian Morphological Dictionary are deleted from the dictionary. After the filtering 7.437 entries remained. Entries from the Wikipedia are all with the first letter in uppercase. Once we have extracted the wordnet from Wikipedia we had to normalize the capitalization of the results. We have done this in an automatic way by comparing capitalization of entries from the Wikipedia with the capitalization of the variants of the same synset in the Princeton English WordNet. Entries in the Wikispecies dictionary are with the first letter in uppercase. In this case we have simply changed all to lowercase.

6.1.2 Results and evaluation

In the table 5 we present the results of the automatic evaluation for all the dictionaries and for the combination of all of them:

The value in the row Total shows the number of synset-variants pairs extracted using the given dictionary or the combination of all dictionaries. The value in Evaluated indicates the number of synset-variant pairs that could be automatically evaluated, that is, the number of synset-variant pairs already present in the Croatian WordNet. In the case of Geonames no single synset-variant pair was present, so we couldn't calculate figures of precision. We show the overall precision along with the precision for nouns, verbs, adjectives and adverbs. Note that we couldn't evaluate the precision for any adverb, as no adverbs were present in the previous version of the Croatian WordNet.

As we can see, an overall automatic calculated precision of 70.33% is achieved. We have manually evaluated 10% of the non-evaluated synset-variant pairs and 10% of the evaluated as incorrect. For strict precision we have achieved 84.49% (more than 14 points higher than automatic evaluation) and for non-strict 90.72% (more than 20 points higher).

The dictionary-based strategy has allowed to extract 6,091 new synset-variant pairs.

6.2 BabelNet based strategy

6.2.1 Resources

For our experiments we have used BabelNet version 2. In this strategy we simply extract the information for Croatian from the BabelNet file.

6.2.2 Results and evaluation

In table the results of automatic evaluation are presented.

Total	12949
Evaluated	1,934
Precision	66.65
Precision N	66.65

Table 6: Results for the BabelNet-based strategy using automatic evaluation

Note that with this strategy we have only been able to extract synset-variant pairs for nouns. 10% of the non-evaluated synset-variant pairs, as well as 20% of the evaluated as incorrect have been manually evaluated. A strict value of 88.96% and a non-strict value of 96.8% have been calculated.

6.3 Machine translation of sense-disambiguated corpora

6.3.1 Resources

In order to extract wordnets from a parallel corpus we need this parallel corpus to be semantically tagged with wordnet synsets in the English part. As these corpora are not easily available, we use two strategies for the automatic construction of the required corpora:

- By machine translation of sense-tagged corpora. We use manually sense tagged English corpora (as Semcor, for example) and we automatically translate the English text into the target language. We are using Google Translate, as it is a statistical system capable to perform a quite good lexical selection task when translating, that is, in some cases is capable to

	Semcor	PWGC	Senseval 2	Senseval 3	Combination
Total	3,111	4,916	144	135	7,123
Evaluated	1,616	2,066	73	82	2,853
Precision	83.17	79.77	83.56	81.71	80.41
Precision N	84.8	80.77	77.08	78.23	81.29
Precision V	78.71	74.25	95	90.63	75.82
Precision A	80.11	85.26	100	50	89.12

Table 7: Results for the parallel corpus strategy using automatic evaluation and machine translation of sense-tagged corpora

select the correct translation of a polysemic word.

- By automatic sense-tagging of English-Croatian parallel corpora. To perform the sense-tagging we have used Freeling and UKB (Padró et al., 2010) (Agirre and Soroa, 2009). The tagging has been performed sentence by sentence.

In both cases, we need to POS tag the Croatian text, getting both the lemma and the POS information. We have used Hunpos with a model for Croatian, and we have developed a program to get the associated lemma from the Croatian Morphological Lexicon. Once we have these corpora, the task of extracting a wordnet is equal to word-alignment task. We have used GIZA++ to align the lemmatized parallel corpora and we have developed a script (that will be included in the WN-Toolkit) to extract the wordnets from the aligned files. In the table 8 we can see the information about the sense-tagged corpora for machine translation strategy.

Corpus	Sentences	Tokens eng	Tokens hrv
Semcor	37,176	794,748	721,282
PWGC	113,404	1.529,105	1,303,386
Senseval 2	238	5,493	5,129
Senseval 3	300	5,530	5,022

Table 8: English sense-tagged corpora used in the experiments

The algorithm for wordnet creation from parallel corpora allows to adjust two parameters:

- Minimum frequency: the minimum value of frequency of the synset in the corpus.
- Minimum percent: The relation between the frequency of the first candidate and the second candidate.

In our experiments for Croatian we have fixed these values to:

- minimum frequency: 5 (except for very small corpora, as for example Senseval 2 and Senseval 3)
- minimum percent: 50.

These values have been fixed after performing several extraction experiments using the Croatian-English parallel corpus.

6.3.2 Results and evaluation

In table 7 we can observe the values of precision, calculated in an automatic way, for the strategy of machine translation of sense-tagged corpora. No distinction between monosemic and polysemic variants is done here, offering an overall value. As expected, for bigger corpora we are obtaining more synset-variant pairs. We are again not obtaining precision values for adverbs, as no adverbs were found in the previous version of CroWN.

We have manually evaluated 10% of the non-evaluated synset-variant pairs, as well as 20% of the evaluated as incorrect. This allowed us to calculate a corrected strict precision of 87.76% (7 points higher than automatic precision) and a non-strict precision of 94.26% (more than 13 points higher than automatic precision).

6.4 Automatic sense tagging of parallel corpora

6.4.1 Resources

In table 9 we can observe the information for the corpus used in the automatic sense-tagging strategy.

Corpus	Sentences	Tokens eng	Tokens hrv
cro-eng p.c.	62,566	1,790,041	1,590,637
EUBookshop	6,104	131,217	126,607
hrenWaC	47,475	1,282,007	1,152,552
SETIMES 2	205,910	4,629,877	4,662,863

Table 9: English sense-tagged corpora used in the experiments

6.4.2 Results and evaluation

In table 10 the results for automatic sense-tagging of English-Croatian parallel corpora are shown. Here again, no distinction between monosemic and polysemic variants has been made.

6.5 Use of CroDeriV

6.5.1 Resources

In our experiments we have used CroDeriV to expand the verbal subset of the CroWN. Using this derivational database we have created a list of 13,781 verb lemmata. Once we have created the verb list we have tried to find their translation in a free Croatian-English on-line dictionary[3]. We have used a script to automatically query this on-line dictionary in case the verb is not already in the CroWN. In this way we have done queries and obtained a list of 10,463 Croatian verbs with translations into English. For each Croatian verb we have assigned the synsets of the English verb, which we obtained as a translation variant.

6.5.2 Results and evaluation

Candidates	10463
New verbal synset-variant pairs	2921
New verbal synsets	2271

Table 11: Number of candidates, synset-variant pairs and synsets for verbal expansion using CroDeriV

In table 11 we can observe the number of candidates, synset-variant pairs and synsets obtained by using CroDeriV for the expansion of the verbal part of the CroWN. The obtained precision is very low, only 27.91%, due to the fact that verbs are highly polysemous units and all of the synsets in which the translation of the Croatian verbal lemma occurs were listed among the candidates, which resulted in an average of 6,8 candidate synsets per verbal lemma. However, in the majority of cases at least one of the candidate synsets was correct. Moreover, numerous candidates were not completely incorrect, since only the reflexivity of the Croatian verb in question had to be corrected in order to correspond to the offered PWN synset. All of these cases were manually corrected. Finally, the results show that in some cases more than one synset-variant pair per synset was found, and in the final step the synset-variant pairs corresponding to the same PWN synset were grouped into same synset in CroWN as well.

Although the overall precision of this procedure is not as high as with monosemous units, it yielded a rather satisfactory number of both new synset-variant pairs and new synsets. However, this method significantly contributed to the improvement of the CroWN's coverage of lemmas from various Croatian corpora.

6.6 Manual creation until reaching 100% of Core WordNet

After applying WN-Toolkit strategies, CroWN encompassed 70.63% of the Core synsets. We decided to add the remaining part of this set, namely 1,456 synsets. The majority of these synsets comprise senses of polysemous units. The following procedure was applied to polysemous units:

1. the literals from these synsets were automatically translated into Croatian;

2. the obtained results were manually checked and corrected.

A manual evaluation and correction of the remaining 1,456 Core synsets was performed. The results of this procedure can be divided into following groups as far as:

1. only one of the translation candidates was correct,

2. two or more translation candidates were correct,

3. none of the translation candidates was correct.

For the first and the second group additional synset-variants in synsets with at least one automatically obtained correct translation was provided. For the last group at least one correct translation for all synsets was provided. The result of these procedure is 100% of Core WordNet synsets represented in CroWN 2.0.

7 Main source of errors

The manual revision of the results has allowed us to devise the main source of errors. We can highlight the following:

[3] http://www.rjecnik.net/

	cro-eng p.c.	EUBookshop	hrenWaC	SETIMES 2	Combination
Total	2,209	673	3,834	5,583	7,395
Evaluated	866	344	1,560	1,908	2,569
Precision	79.56	75.29	78.46	71.96	70.07
Precision N	78.81	74.73	78.64	71.83	69.73
Precision V	77.39	72.34	70.97	68.42	66.67
Precision A	91.94	87.5	90.63	85.05	86.47

Table 10: Results for the parallel corpus strategy using automatic evaluation and automatic sense-tagging of English-Croatian parallel corpora

- For dictionary-based and Babelnet-based strategies one important source of errors is the capitalization of the entries. In some of the used dictionaries (for example Wikipedia and Wikispecies), all the entries begin with a capital letter, regardless they are proper or common names.

- For dictionary-based and Babelnet-based strategies other important source of errors are some entries in forms other than nominative singular. Some of the dictionary entries are in nominative plural.

- For strategies based on parallel corpora (both machine translation of sense-tagged corpora and automatic sense-tagging of parallel corpora) numerous errors are produced by the Croatian tagger. As stated earlier, we have used a simple Hunpos tagger with a model for Croatian and a simple script for adding the lemmata. This tagger is not able to cope with multiword expressions and is not able to attach the reflexive particle *se* of reflexive verbs to the lemma.

- For the strategy based on parallel corpora using machine translation, another important source of errors is the quality of the machine translation system. We have used Google Translate, a state-of-the-art machine translation system, so we don't expect to make any improvement in this aspect.

- For strategy based on parallel corpora using automatic word sense-disambiguation of the English part, one important source of errors is the word sense disambiguation, as it is a very difficult task. We have used a state-of-the-art word sense algorithm (Freeling+UKB), so we don't expect to make any improvement in the tagger. In these experiments the corpora were sense tagged sentence by sentence, thus reducing the context information available for the UKB algorithm. In future experiments we plant to sense tag the corpora grouping several sentences of the same document.

8 Conclusions and future work

In this paper we have described the procedures applied for the automatic acquisition of new CroWN synsets based on various dictionaries and parallel corpora. The results were both automatically and manually evaluated, and approximately 5,000 new synsets were detected as candidates for CroWN. As it has been stated above, the procedures proved valuable for the detection of monosemous vocabulary. However, it became obvious that the detection of correct senses of polysemous words is a highly challenging task. This especially pertains to procedures relying on sense-tagged parallel corpora, previously lemmatized and POS tagged. The main problem is non-availability of sense-tagged corpora for Croatian that could be used for more comprehensive approach. Further problems arise from not completely satisfactory results of lemmatization and POS tagging. One of our future goals is thus to create a Freeling module (including lemmatizer and POS tagger) for Croatian. In order to make the CroWN a more representative resource for Croatian, we plan to compare the list of words from CroWN and frequency list of lemmas from Croatian corpora. This procedure should enable the detection of gaps in the coverage of Croatian vocabulary and should result in a more balanced and usable wordnet for Croatian. Moreover, since CroDeriV is currently being expanded with other POS, we will use it for further expansion of other lexical hierarchies in CroWN. All these steps should also result in a sense-tagged corpus of Croatian that could be used for various NLP tasks.

As a future work we also plan to improve the WN-Toolkit. One of the improvements will be the inclusion of a methodology allowing to deal

with polysemous English variants. This methodology will make use of the definitions and the semantic relations in the dictionary and will try to match them with the definitions and relations in the Princeton English WordNet. This will allow us to match the correct target language translation to a given meaning. With the new version of the toolkit we plan to create wordnets for as much languages as possible and to contribute to the extension of the Extended Open Multilingual Wordnet[4] (Bond and Foster, 2013).

Acknowledgments

This research has been partially conducted thanks to a Networks Grants for Short visits, references 6500, 7008 and 7009, from the ESF Research Networking Programmes.

This research has been carried out thanks to the Project SKATER (TIN2012-38584-C06-01 and TIN2012-38584-C06-06) supported by the Ministry of Economy and Competitiveness of the Spanish Government.

The automatic translation of sense-tagged corpora have been performed thanks to an academic agreement with Google .

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, Sofia, Bulgaria. 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japan. 64–71.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In Petr Sojka, Key-Sun Choi, Christine Fellbaum, and Piek Vossen, editors, *Proceedings of the 3rd International WordNet Conference*, pages 29–36. Global Wordnet Association.

[4]http://compling.hss.ntu.edu.sg/omw/summx.html

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

Antoni Oliver and Marko Tadić. 2004. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In Maria Teresa Lino Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Language Resources and Evaluation - LREC'04*, pages 3366 – 3370, Lisbon, Portugal. European Language Resources Association, ELRA.

Antoni Oliver. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. In Heili Orav, Christianne Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, pages 7–15, Tartu, Estonia. Global Wordnet Association.

Lluís Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India, February. Global Wordnet Conference 2010, Narosa Publishing House.

Marko Tadić and Sanja Fulgosi. 2003. Building the croatian morphological lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, pages 41 – 46.

Piek Vossen. 1998. Introduction to eurowordnet. In Piek Vossen, editor, *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer Netherlands.

Krešimir Šojat and Matea Srebačić. 2014. Morphosemantic relations between verbs in Croatian WordNet. In Heili Orav, Christianne Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th Global WordNet Conference*, pages 262–267, Tartu, Estonia. Global Wordnet Association.

Krešimir Šojat, Matea Srebačić, Tin Pavelić, and Marko Tadić. 2014. CroDeriV: a New Resource for Processing Croatian Morphology. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, Maegaard B., Mariani, J., A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Language Resources and Evaluation - LREC'14*, pages 3366 – 3370, Reykjavik, Iceland. European Language Resources Association, ELRA.