Identifying Syntactic Patterns in Croatian WordNet Synset Definitions

Božo Bekavac, Krešimir Šojat
University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics, Ivana Lučića 3, Zagreb, Croatia
E-mail(s): bbekavac@ffzg.hr, ksojat@ffzg.hr

Abstract: *The paper presents syntactic pattern definitions designed for Croatian WordNet in order to create unambiguous and consistent synset definitions. The rules are implemented in form of finite-state transducers and tested on already existing version of Croatian WordNet. Results are presented using standard evaluation measures.*

Keywords: Croatian WordNet, meaning definition, syntactic patterns, chunking, Intex.

1. Introduction

Croatian WordNet (further *CroWN*) is a lexical database for the Croatian language built according to the principles of the Princeton WordNet (further *PWN*) [4], a large lexical database for English, and similar multilingual projects done primarily for European languages [13, 14]. Wordnets are lexical databases that group words (*literals*) into sets of synonyms (*synsets*), accompanied by short definitions of the synset meaning (*glosses*). Synsets usually, but not always, comprise examples of contextual usage of literals in sentences. The structure of each wordnet is based on several major semantic relations such as synonymy (relevant for the members of a particular synset – *literals*) and relations as e.g. hyponymy/hyperonymy (relevant for the relations between whole synsets). Such a semantic lexicon based on a network of words can be used by humans as a dictionary and thesaurus as well as by machines as a source of various data used in natural language processing and artificial intelligence applications. The building strategy of the CroWN can be roughly divided into two major phases. The first one consisted of the translation and adaptation of the so called basic concept sets used in the projects EuroWordNet I and II (further *EWN*) (basic concept set 1 and 2, *BCS* 1 and 2) and BalkaNet (further *BN*) (basic concept set 3, *BCS* 3).

The first phase in the building of the CroWN consisted of manual translation of BCS 1-3 extracted from the WN version 1.5. and used in EWN and BN as the core of each national wordnet developed in these projects. In EWN and BN this set of synsets was chosen and agreed upon in order to ensure the compatibility between involved national wordnets. This approach was followed in the first phase of the CroWN building in order to establish the multilingual compatibility of the CroWN and wordnets build in EWN and BN as well. The second phase in the building of the CroWN consists of extension of the set of translated and partially adapted synsets from the BCS 1-3 and thereby established lexical hierarchies based on the semantic relation of hyponymy/hyperonymy according to the principles given in [8] and [15, 16]. The second phase of the project is due to begin in the autumn of 2010.

2. The structure of the CroWN

At the present time CroWN comprises 8510 synsets. Synsets consist of nouns, verbs and adjectives, i.e. autosemantic parts of speech or so called semantically full words except adverbs. The overall statistics is given in the Table 1.

Table 1. Number of synsets and POS division in BCS 1-3

	BCS1	BCS2	BCS3	total
synsets	1219	3469	3822	8510
nouns	965	2245	2681	5891
verbs	254	1188	876	2318
adjectives	0	36	265	310

For editing and browsing CroWN we use VisDic, a graphical application originally developed for viewing and editing wordnets but also extended to other dictionary databases stored in XML format [5]. An example of a CroWN synset and its structure is given in the Fig. 1. The structure of the synset contains the

information on part-of-speech, unique ID number of the synset and BCS to which the synset {glazba, muzika}:1 belongs. The digit following the semicolon indicates particular senses of polysemous words processed in the lexicon. Further lines refer to data from the Suggested Upper Merged Ontology (*SUMO*) and the MId-Level Ontology (*MILO*) as well as from their more specific domain ontologies mapped to PWN 2.0 used as a source of BCS 1-3 and to other wordnets involved in BN project [18]. In the following line the definition of the synset meaning is given. Lines below the definition line indicate various semantic relations between this synset and other synsets in the lexicon.

POS: n **ID**: ENG20-06591368-n **BCS**: 1
Synonyms: glazba:1, muzika:1
SUMO/MILO: = Music
Domain: music
Definition: umjetnost izražavanja tonovima, glasovima i šumovima
-->> [**hypernym**] *[n] auditorna komunikacija:1, slušna komunikacija:1
<<-- [**category_member**] *[n] glazba:3, muzika:3
<<-- [**category_member**] *[v] svirati:3
<<-- [**hyponym**] *[n] odlomak:2, pasus:1, glava:1, stavak:1
<<-- [**hyponym**] *[n] skladba:1, kompozicija:1, glazbena kompozicija:4, muzička kompozicija:4
<<-- [**category_member**] *[n] stil:1, stil izražavanja:3
<<-- [**category_member**] *[v] sniziti; snižavati:3
<<-- [**category_member**] *[n] glazbenik:1, muzičar:1

Figure 1. Sample synset from CroWN

In the following sections of the paper we focus on the methods of extraction of syntactic patterns of definitions used to illustrate synset meanings and their analysis in terms of predominant patterns and overall consistency of usage throughout the CroWN.

3. Definition of synsets in CroWN

As stated above, the work in the first phase of the building of the CroWN primarily consisted of the translation of literals, but also of the translation of meaning definitions and examples of contextual usage. This work was first done manually by several persons and afterwards edited by another team. Very soon during the editing it became obvious that the translated definitions can be problematic in several aspects.

First, some of the definitions in PWN are logically circular in terms that the definition comprises the terms being defined, i.e. literals. Besides, some of the definitions can hardly be translated into Croatian. The notorious example is the English literal *something*, which is defined as *a thing of some kind* (fortunately, this literal is not a member of BCS 1-3). More severe problems that editors faced belong to the second and much larger group. In this paper we shall focus on one of the major problems from this group, namely the inconsistency in terms of syntactic patterns used in definitions by different translators as well as on a possible solution to this problem. Having in mind usefulness and applicability of CroWN in various NLP tasks such as terminology extraction, automatic creation of glossaries, question answering, machine learning of lexical semantics relations, automatic construction of ontologies etc. a certain uniformity of syntactic and lexical features in meaning definitions is a necessary precondition. In order to enable these tasks and to achieve the overall consistency of the lexicon, at least two general principles should be followed as much as possible: (1) definitions should comprise members of the same lexical hierarchy as *genus proximum* (preferably the one on the first level above) and (2) *differetia specifica*, i.e. distinctive semantic features of the literal defined should be stated consistently in terms of their syntactic features. Since the VisDic enables browsing through the lexical hierarchies of CroWN, the first principle can be more or less acceptably fulfilled in the process of editing. On the basis of experience so far, the second principles cannot be fulfilled without automatic or semi-automatic processing of existing definitions. In order to determine which syntactic patterns were used and which syntactic patterns should be used in the process of editing and further extending of CroWN we decided to conduct an experiment on the BCS1 focusing on noun synsets. The description of the applied method and results are given in the following sections.

4. Experiment setup

In order to improve definition consistency and uniformity in CroWN, 965 definitions of nouns from the BCS1 were analyzed in terms of their basic syntactic features. This analysis was

preceded by two procedures. The first one consisted of writing rules for syntactic patterns to be used in editing and future work. The rules consist of elements defined as VP (verb phrase that can consists of a verb), NP (noun phrase that can consists of a noun or a adjective(s) + noun) and PP (preposition + (adjective(s)) + noun) etc. In such a way we constructed 10 different local grammars. The grammars range from simple ones like NP NPg (*g* stands for the genitive case in Croatian) designed for detecting simple syntactic patterns of only one NP as *brzina kretanja* (*the speed of movement*) or NP PP to more complex ones, that also included terminals such as *koji* (which), *što* (what), *čije* (whose) etc. In writing these rules we tried to obey the aforementioned principles of *genus proximum* and *differentia specifica* as well as to incorporate the experience gained so far in the process of editing. The second procedure consisted of the application of these rules to the definitions for noun synsets in BCS1. This procedure was applied in order to test the design and the applicability of the rules. On the basis of our direct insight in existing definitions, the design of the rules aims at capturing syntactic patterns that are or should be most frequently used in definitions. In other words, the aim of these local grammars (syntactic patterns) is to provide unambiguous detection of elements used in definitions, i.e. to provide an important step towards the automatic "knowledge extraction".

5. Grammar construction

In order to obtain unambiguous definition patterns, we established another two principles in the phase of grammar construction: (1) although grammars allow only limited flexibility, at the same time the rules should be flexible enough to allow defining of all noun synsets, (2) structures should be kept as simple as possible. This implies that inserted structures (e.g. inserting a new clause in another, already existing, clause) which cause discontinuity between coherent parts of a sentence (chunks) are not allowed. These principles can be illustrated with the following synset definitions taken from the BCS1:

1. osoba koja upravlja (*a person who rules*)
2. osoba koja stvara umjetnička djela (*a person who creates works of art*)
3. pojava koja uključuje postupnu smjenu različitih stanja (*a phenomenon marked by gradual changes through a series of states*)

First step in the process of the structure description is the identification of chunks in Croatian definitions. Chunks are the non-recursive cores of "major" phrases [1]. Keeping structures as simple as possible in our approach reflects itself in the possibility of structure description only using chunks, i.e. without a further need to implement full (context free) parsing. A simple representation of the sentences above in the explained manner could look like:

1. NP koji (*which*) VP
2. NP koji (*which*) VP NP
3. NP koji (*which*) VP NP NPg

It can easily be observed that the structures above are regular. First three elements are obligatory (NP, koji and VP), the following two elements (NP and NPg) are optional. Such a structure can be rewritten in the form of the following regular expression:

NP koji (*which*) VP NP*_NPg*

Such a regular expression is the representation of one rule or one of allowed definition structures. Since the researchers involved in the building of the CroWN are not and presumably will not be formally educated computational linguists, for the sake of simplicity rules are displayed in a more intuitive way:

Figure 2. Graphical description of the rule

Full color rectangles are obligatory parts of the rule while the broken line indicates the optional ones. The biggest rectangle, which surrounds all smaller rectangles, states that all the elements are parts of the same rule. The horizontal line between NP and NPg means that if NPg appears in definition, previous NP is obligatory as well.

Formal representation of our rules is implemented in Intex, a development environment for making formal descriptions of natural languages using finite-state transducers (FSTs) and their immediate application on large corpora in real-time [10]. All constructed rules of definitions are presented in Fig. 3.

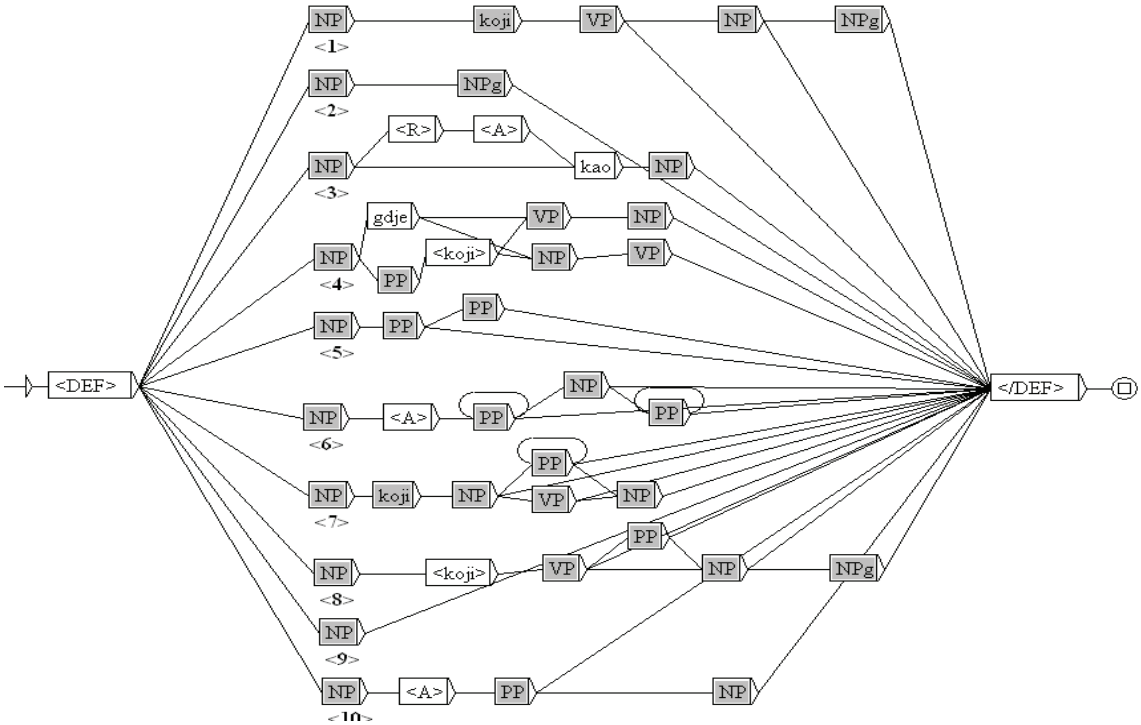

Figure 3. Presentation of Intex FST syntactic patterns

The rules presented in the Fig. 3 operate over POS annotated definitions between the opening and closing definition <DEF> tags. POS and lexical information is taken from the Croatian Morphological Lexicon [17]. Grey colored states in transducer represent coherent structures (mostly chunks). White colored states refer to POS tags or lemmas. The whole set of rules is applied simultaneously on the text implementing the longest match principle. Such an approach provides a dynamic disambiguation in cases when one token can belong to two or more chunks, i.e. the longest chunk is taken as accurate. The longest match principle and the dynamic disambiguation are in more detail described in [3]. Since our rules are implemented in the form of FSTs, the definition processing produces annotated chunks. Above mentioned definitions are recognized by rule number 1 and automatically marked as:

1. <NP>osoba</NP> koja <VP>upravlja</VP>
2. <NP>osoba</NP> koja <VP>stvara</VP> <NP>umjetnička djela</NP>
3. <NP>pojava</NP> koja <VP>uključuje</VP> <NP>postupnu smjenu</NP> <NPg>različitih stanja</NPg>

6. Results and evaluation

We created ten syntactic patterns for the definitions of noun synsets according to the principles explained in sections 3, 4 and 5. CroWN BCS1 containing 965 definitions of nouns was divided into two parts: a training part consisting of 482 noun synset definitions, and a testing part consisting on unseen 483 noun synset definitions. The rules applied on the CroWN BCS1 testing part containing 483 unseen synset definitions recognized 190 definitions. Using standard evaluation measures, our rules achieved precision of 97,3 % and recall of 39,4 %.

As expected, precision is very high because of relatively predictive definition structures. A few mistakes that occurred are the result of incorrect POS tags. Therefore, precision could be raised by improving POS tagger. As far as recall is concerned, its value in comparison to various efforts in definition extraction [6, 11, 12] substantially higher. On the other hand, in terms of consistency of tested definitions recall is unexpectedly high since there were no prescribed syntactic patterns to be used in the process of BCS 1 definition creation.

To the best of our knowledge, there is no closely related work done in this field for any other Slavic language. Our results could be

compared with automatic definition extraction experiments [6, 11, 12] conducted for Slavic languages, where our F-measure of 55,9 % by far outperforms other reported results. These experiments are mostly conducted on less structured texts [cf. 7].

7. Conclusion and future work

The construction of Croatian WordNet BCS1 started without any syntactic structures constrains. In the process of synset definition editing a possibility of bringing definitions to relatively uniform and regular structures in terms of syntactic patterns was spotted. The explication and the standardization of rules for syntactic patterns set in this work provide consistency checking of definitions for future synset definitions.

From the perspective of automatic meaning extraction we provided a framework for detection of the first parent node in a lexical hierarchy (*genus proximum*) and specific semantic components (*differentia specifica*) of defined term. Since contemporary Q&A systems use less structured content sources, our work could also provide a means for achieving better results in this field.

As far as the editing and the extension of CroWN, our work provides means to formalize syntactic patterns in definitions and to avoid the usage of "innovative" structure definitions. Preferably, future synset definitions should be completely structured in accordance with designed rules. These rules, as well as the ones that will be designed for other POS categories (verbs and adjectives), in future work will be applied on CroWN BCS 2-3.

8. Acknowledgements

This work has completed within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants 130-1300646-1002 and 130-1300646-0645.

9. References

[1] Abney S, Partial Parsing via Finite-State Cascades, Journal of Natural Language Engineering 2 (4); 1996. p. 337–344.

[2] Barnbrook G, Defining Language: A local grammar of definition sentences, John Benjamins Publishing Company, Amsterdam /Philadelphia; 2002.

[3] Bekavac B, Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima. PhD Thesis, Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb; 2005.

[4] Fellbaum, Ch. Editor. Wordnet: an electronic lexical database; The MIT Press; Cambridge/ London; 1998.

[5] Horak, A, Smrz, P. VisDic - Wordnet Browsing and Editing Tool. Proceedings of the Second International WordNet Conference - GWC 2004. Brno, Czech Republic : Masaryk University; 2003. p. 136-141.

[6] Przepiórkowski A, Degórski L, and Wójtowicz B. On the evaluation of Polish definition extraction grammars. In Zygmunt Vetulani, editor, Proceedings of the 3rd Language & Technology Conference, Poznan, Poland; 2007, p. 473–477.

[7] Przepiórkowski A, Spousta M, Simov K, Osenova P, Lemnitzer L, Kuboň V, Wójtowicz B, Towards the automatic extraction of definitions in Slavic. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, Prague, Czech Republic; 2007. p. 43-50.

[8] Raffaelli I, Tadić M, Bekavac B, Agić Ž. Building Croatian WordNet. Proceedings of the 4th Global WordNet Conference; 2008, p. 349-359.

[9] Silberztein M. INTEX: a Finite State Transducer toolbox, Theoretical Computer Science #231:1, Elsevier Science; 1999. p. 33-46.

[10] Silberztein M. INTEX Manual. ASSTRIL, Paris; 2000.

[11] Simov K, Osenova P, BulQA: Bulgarian-Bulgarian Question Answering at CLEF 2005. In CLEF; 2005.

[12] Tanev H, Socrates: A question answering prototype for Bulgarian. Recent Advances in Natural Language Processing; 2004.

[13] Tufiş D, editor. Special Issue on the BalkaNet Project. J. Romanian Journal of Information, Science and Technology 7 (1–2); 2004, p. 1–248.

[14] Vossen P, editor. EUROWORDNET: A multilingual database with lexical semantic networks, Kluwer Academic Publishers, Dordrecht / Boston / London; 1998.

[15] Šojat K, Sintaktički i semantički opis glagolskih valencija u hrvatskom. PhD thesis, Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, 2008.
[16] Šojat K, Verbs in the Croatian WordNet; in press
[17] Tadić M, Fulgosi S, Building the Croatian Morphological Lexicon, Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, Budapest, ACL; 2003. p. 41-46
[18] http://www.ontologyportal.org