HR4EU – Using Language Resources in Computer Aided Language Learning

Daša Farkaš
Faculty of Humanities and Social Sciences, Zagreb
dberovic@ffzg.hr

Matea Filko
Faculty of Humanities and Social Sciences, Zagreb
msrebaci@ffzg.hr

Marko Tadić
Faculty of Humanities and Social Sciences, Zagreb
mtadic@ffzg.hr

Abstract

In this paper we present the HR4EU – web portal for e-learning of Croatian language. The web portal offers a new method of computer aided language learning (CALL) by encouraging language learners to use different language resources available for Croatian: corpora, inflectional and derivational morphological lexicons, treebank, Wordnet, etc. Apart from the previously developed language resources, the new ones are created in order to further facilitate the learning of Croatian language.

We will focus on the usage of the treebank annotated at syntactic and semantic level in the CALL and describe the new HR4EU sub-corpus of the Croatian Dependency Treebank (HOBS). The HR4EU sub-corpus consists of approx. 550 sentences, which are manually annotated on syntactic and semantic role level according to the specifications used for the HOBS. The syntactic and the semantic structure of the sentence can be visualized as a dependency tree via the SynSem Visualizer. The visualization of the syntactic and the semantic structure of sentences will help users to produce syntactically and semantically correct sentences on their own.

1. Introduction

In this paper we present the HR4EU – web portal for e-learning of Croatian. The HR4EU is the first portal which offers Croatian language courses which are free-of-charge and developed by language professionals. Moreover, the HR4EU also integrates bidirectional interaction with some of the language resources for Croatian developed previously. For the purpose of this paper, we will focus on the interaction between the HR4EU and one of these language resources – the Croatian Dependency Treebank (HOBS) and show how language resources, developed primarily for NLP tasks, can be used as a valuable tool in the computer aided language learning.

The paper is structured as follows: in Chapter 2, we briefly present the HR4EU portal and its relation to Croatian language resources. In Chapter 3, we describe two layers of the Croatian Dependency Treebank: the syntactic and the semantic layer, as well as the SynSem Visualizer, a newly developed tool for visualization of dependency trees. Chapter 4 is dedicated to the HR4EU sub-corpus of the HOBS, which was developed to facilitate the understanding of Croatian syntax and semantic relations between a verb and its arguments to the HR4EU users. The paper ends with the concluding remarks.

2. HR4EU – web portal for e-learning of Croatian

Since Croatian is a language with relatively small number of speakers, its presence on the web is limited. A few e-learning sites which offer users the possibility to learn Croatian are expensive (e.g. E-learning course of Croatian as a second and foreign language - HiT-1[1]), developed by non-native speakers of Croatian (e.g. Surface languages[2]) or present the learning material in static manner avoiding the usage of existing language technologies (e.g. Easy Croatian[3], Basic Croatian[4]). With the HR4EU portal we aim to bridge this gap and develop a modern e-learning system which integrates bidirectional interaction with previously developed language resources (LRs). This e-learning system is developed by linguists, which are also native speakers and have experience in building LRs. Moreover, the great efforts were made to make this portal visually attractive to users.

The HR4EU portal is divided into four sections:

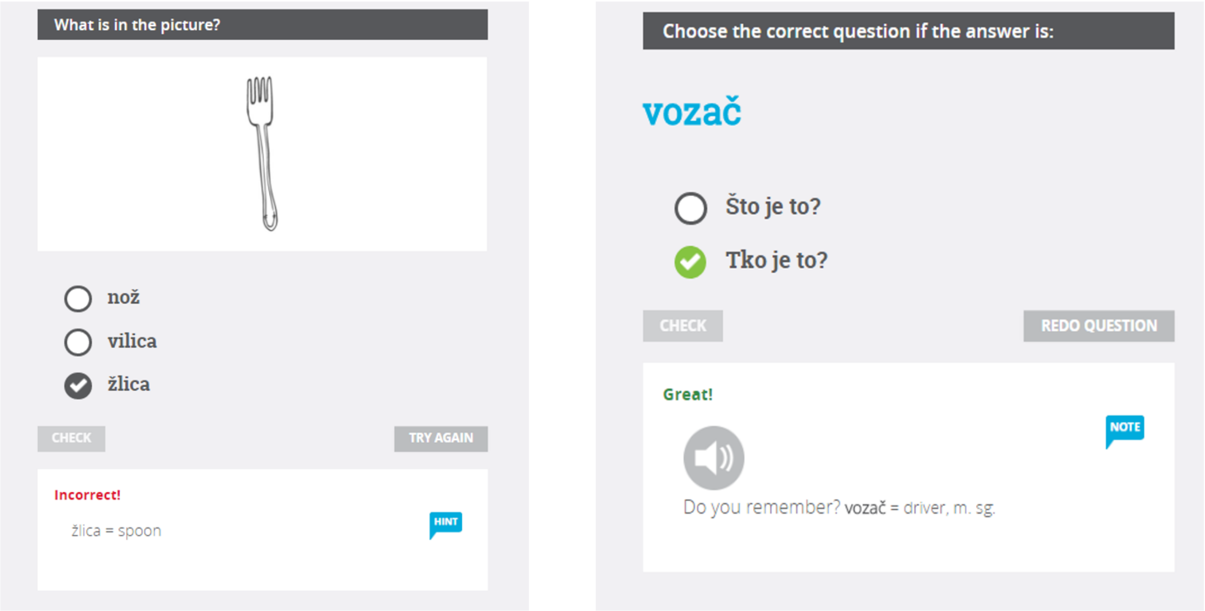
a) **Courses**, where users can find three general courses: beginner, intermediate and advanced, as well as two specialized courses: Croatian for students and Croatian for business users. Courses are equipped with interactive lessons, quizzes, dictionary, grammar books, tasks for practicing writing skills, etc. For the purpose of courses at the HR4EU portal we have recorded more than 1.600 audio tracks and approx. 200 illustrations, in order to obtain their interactivity and multimodality.

b) **Language Resources**, the section which includes description of LRs for Croatian language and a short video for each LR that is used as an additional learning tool throughout the courses. Short video tutorials provide users with the introduction to the particular resource (cf. 2.1.)

c) **About Croatia**, providing the cultural context for learning Croatian via nine interactive maps presenting most important cities, events, famous Croats, landscapes, cultural heritage, gastronomy and ethnology, etc.

d) **Living in Croatia**, offering useful information to foreigners in Croatia, e.g. the list of important state institutions.

The first section, Courses, is developed in Moodle, an open source e-learning platform, which provides teachers or course developers with numerous tools and activities that can be used in e-learning course (e.g. interactive lessons, quizzes with multiple question types, dictionary, books, and assignments). However, the Moodle itself is a "robust"[5] system, which was restructured and modified both visually and functionally in order to become interactive, attractive and effective e-learning tool. Several new plugins and possibilities were introduced, e.g. HINT and NOTE buttons (Figure 1), which provide users with help when they answer the question incorrectly, or with the additional information about words or grammar used in question if they answer the question correctly.



What is in the picture?

○ nož
○ vilica
● žlica

Incorrect!

žlica = spoon

Choose the correct question if the answer is:

vozač

○ Što je to?
✓ Tko je to?

Great!

Do you remember? vozač = driver, m. sg.

Figure 1: Hint and Note buttons

[1] http://www.unizg.hr/homepage/learn-croatian/e-learning-course-of-croatian/
[2] http://www.surfacelanguages.com/
[3] http://www.easy-croatian.com/
[4] http://basic-croatian.blogspot.hr/
[5] https://docs.moodle.org/31/en/About_Moodle

The other three sections are developed in WordPress, but with the identical visual theme as used in the Moodle part. The portal contains multimodal content: audio files, video tutorials, interactive maps, pictures, links to the various sites about Croatia, etc., since the quality approaches to the computer aided language learning have to use all the possibilities that are offered by multimodal technology. This is why the HR4EU portal introduces language learners with various language resources for Croatian and their usability when learning a new language.

2.1. Language resources at HR4EU

As stated before, apart from the interactive and multimodal content, the HR4EU portal introduces language learners to the usability of language resources in the (computer aided) language learning. Thus, the one part of the HR4EU portal is dedicated solely to Croatian language resources. There, the users can find out more about language resources which can be particularly useful for them and which are therefore used as a helping tool throughout the courses. Each resource, namely Croatian National Corpus[6] (216,8 million words; Tadić, 1996), Croatian Morphological Lexicon[7] (3,9 million word forms; Tadić and Fulgosi 2003, Tadić 2006), Croatian Wordnet[8] (23.122 synsets with 47.906 lexical units; Raffaelli et al. 2008; Oliver et al. 2015), CroDeriV[9] – a morphological database of Croatian verbs (14.491 verbs; Šojat et al. 2013), and Croatian Dependency Treebank[10] (4.000 sentences; Tadić 2007, Agić et al. 2014), is provided with a brief description, link to the respective search interface and a short video tutorial. Short video tutorials were made especially for the Croatian language learners, since the most of them have never seen or used Croatian LRs, or LRs in general, before.

The lessons and quizzes are designed to encourage the users to use LRs, e.g., to find the appropriate word form or lemma in Croatian Morphological Lexicon or to learn semantically related words via lexical hierarchies or synsets in Croatian WordNet or derivationally related words in verbal derivational database CroDeriV.

Moreover, this system is designed in a way that language learners can also be helpful in improvement of existing LRs. Learners' activity will be used to enhance and enlarge existing LRs by tracking their activity yielding empty results, and adding them to the respective resources. Furthermore, users' answers in *Practice your writing skills* tasks will be used to build the new LR, the corpus of Croatian as a second language. This corpus will be particularly useful to language teaching specialists, because it will offer a possibility to extract morphological and syntactic errors of users.

However, some of the existing LRs weren't helpful for language users in their primary shape, because the language material they contain is too complex for language learners which have just begun to learn Croatian. Nevertheless, they served as a model for building a new LR on syntactic and semantic level which can be helpful even to the learners at a beginner level. In the following chapters we thus present the syntactic and semantic layer of the Croatian Dependency Treebank and the application of this model to the corpus of sentences from the HR4EU courses.

3. Croatian Dependency Treebank – HOBS

The Croatian Dependency Treebank (Tadić 2007, Agić et al. 2014) is a corpus of approx. 4.500 sentences extracted from the Croatia Weekly 100kw, the newspaper sub-corpus of the Croatian National Corpus. The sentences are manually tagged according to the modified Prague Dependency Treebank specification for annotation at the analytical level.[11] The part of the HOBS (approx. 3.500 sentences) is also manually tagged with semantic roles, according to the specification developed for the Croatian semantic role labelling. The SynSem Visualizer enables the queries across this 3.500 sentences which are annotated both on the syntactic and semantic level. Here we will briefly describe the two abovementioned layers and the queries enabled by the SynSem Visualizer.

[6] hnk.ffzg.hr

[7] hml.ffzg.hr

[8] crown.ffzg.hr

[9] croderiv.ffzg.hr

[10] hobs.ffzg.hr

[11] https://ufal.mff.cuni.cz/pdt2.0/

3.1. HOBS – syntactic layer

There are two slightly different manually annotated versions of HOBS at syntactic level. The first version is annotated in complete accordance to the Prague Dependency Treebank annotation guidelines for annotating at the analytic level (cf. Appendix 1, footnote 6). The second version is annotated with the modified PDT specification (cf. Appendix 1), which is adjusted to the syntactic structures of Croatian language. This specifically pertains to the different annotation of dependent clauses, which has also improved the parsing results. This second version is freely available for search via SynSem Visualizer (cf. 3.4.) and further annotated with semantic role labels.

3.2. HOBS – semantic layer

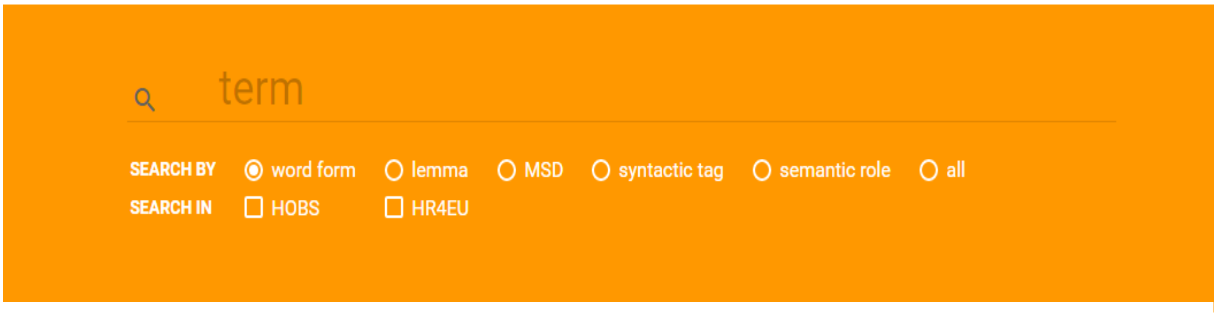
Semantic role labelling is essential for many NLP tasks, especially when it comes to information extraction. It is a logical step immediately after the resources on the syntactic level have been built.[12] Semantic layer of HOBS presents first steps towards automatic semantic role labelling in Croatian.

In order to build a training set for the automatic semantic role labelling of Croatian texts, we first had to design a tagset for Croatian semantic role labelling. Since the manually tagged sentences will be used as a training set for the automatic semantic role labelling system, we had to be careful when it comes to our specification: the labels had to be verb-independent and of a limited number. The initial set of tags was revised during the manual annotation, i.e. the tags which proved to be very frequent and distinctive enough from the existing ones were added to the tagset. The final set consists of seventeen tags followed by the examples of sentences in which these tags should be used. (cf. Appendix 2 for the SRL specification for Croatian).

Tags can be divided in two groups: first group comprises verbal arguments, and second group adjuncts, mainly different types of adverbials and adverbial and attribute clauses. We have decided to include adjuncts into our SRL specification because they often give more specific and detailed information about the described event and can be very useful later in e.g. information extraction tasks.

3.3. SynSem Visualizer

The two above presented layers of the HOBS are encoded in the CONLL format. Although this format is useful to most of the professional linguists, it is not useful to the users of the HR4EU portal, and even to the non-computational linguists. This results in the lower usability and visibility of this language resource, so we decided to develop a visualizer which will enable the search and the hierarchical representation of the syntactic and semantic structure of Croatian sentences.



term

SEARCH BY ● word form ○ lemma ○ MSD ○ syntactic tag ○ semantic role ○ all

SEARCH IN □ HOBS □ HR4EU

Figure 2: SynSem Visualizer search interface

The SynSem Visualizer enables the queries across the Croatian Dependency Treebank on the syntactic and semantic level. It is developed as a database-driven web application, and written in Django, a widely used Python framework. It enables the graphical representation of the hierarchical sentence

[12] There are several LRs for Croatian at syntactic level: above mentioned Croatian Dependency Treebank, SETimes.HR Treebank (Agić and Ljubešić 2014), Universal Dependencies Treebank (http://universaldependencies.org/) and a dependency parser (Agić 2012).

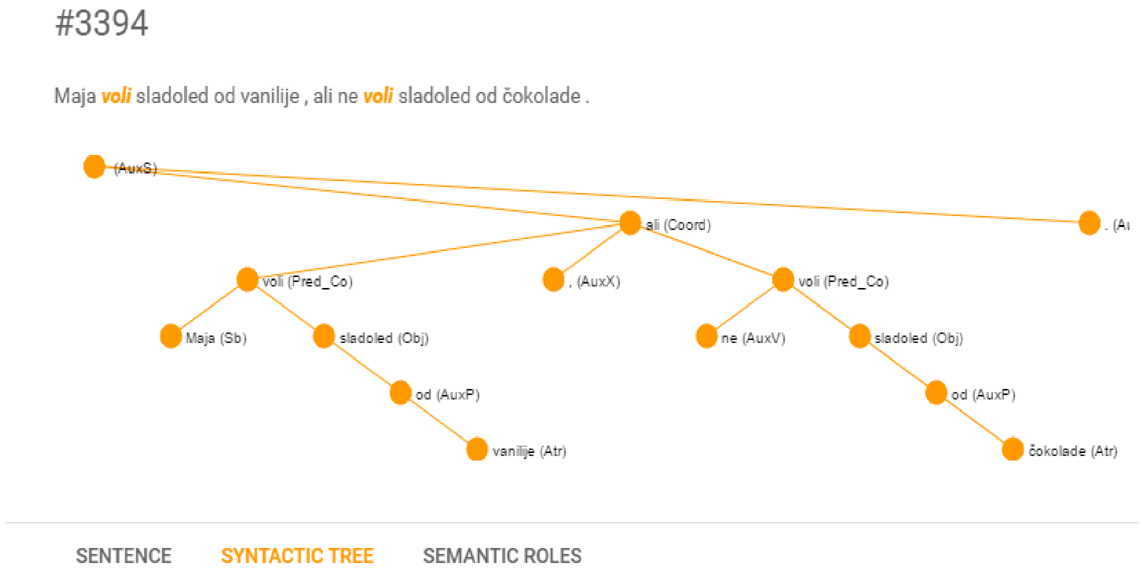
structure (cf. Figures 3, 4 for the representation of the hierarchical sentence structures via SynSem Visualizer).

The HOBS and the HR4EU corpora can be searched independently, or they can be both searched at the same time. The SynSem Visualizer enables search by word form, lemma, morphosyntactic tag, syntactic tag or semantic role (cf. Figure 2). Users are provided with the MSD, syntactic and semantic specifications on the website, so they can easily manage their searches.

4. HR4EU sub-corpus

Although the language resources are mainly used in NLP tasks, they can also be used in the computer aided language learning. This is why they are an essential part of courses at the HR4EU portal. However, language learners often don't have linguistic background, so some of the resources have to be adjusted to their needs, or even new resources, which can be used both in NLP and CALL have to be built. For the purpose of the HR4EU portal we have developed a sub-corpus of the HOBS which consists of approximately 550 syntactically and semantically annotated sentences used in the Croatian language courses available at the HR4EU web portal. These sentences are manually annotated on both syntactic and semantic level according to the model and specification used for HOBS and presented in previous chapter.

The syntactic structure of sentences in the HR4EU sub-corpus is not as complicated as the syntactic structure of the newspaper sentences contained in HOBS. It is adjusted to the beginner level users of Croatian.[13] However, even the syntax of the simple sentences of the morphologically rich language as Croatian can be challenging to the speakers of other, especially non-Slavic languages. Thus, the graphical representation of syntactic structure of sentences used in courses could improve the users' understanding of different grammatical relations. The example of the syntactic tree from the HR4EU corpus is presented in Figure 3.



#3394

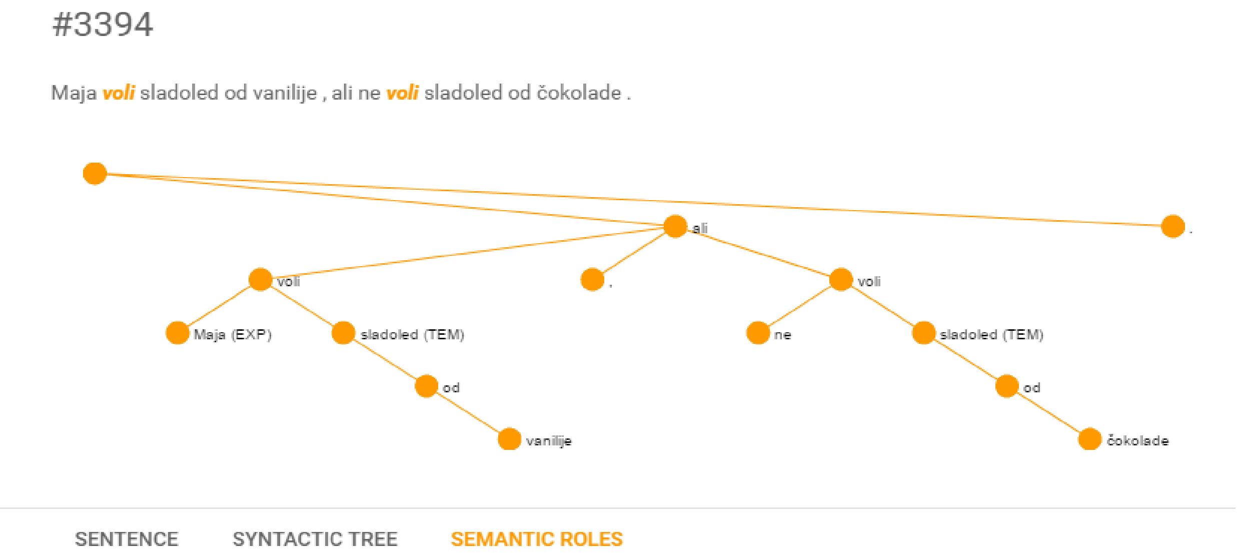
Maja **voli** sladoled od vanilije , ali ne **voli** sladoled od čokolade .

Figure 3: The example of the syntactic tree - HR4EU sub-corpus

Maja voli sladoled od vanilije, ali ne voli sladoled od čokolade.
Maja-NOMsg like-PRES3sg ice-cream-ACCsg from vanilla-GENsg,
but no like-PRES3sg ice-cream-ACCsg from chocolate-GENsg
'Maja likes vanilla ice-cream, but she doesn't like chocolate ice-cream.'

[13] The vocabulary used in the HR4EU courses is mostly based upon the corpora frequency lists, and the syntactic structure of sentences follows the grammar content which is presented in the lesson.

The challenges for the decoding of the syntactic structure stated above can be expanded to the understanding of the role of the verb arguments as well. Graphical representation of the semantic structure of the sentence can, therefore, improve the learners' accurate interpretation of semantic roles of the verb arguments, i.e. they can easily see "Who did What to Whom" (Palmer, 2010). The understanding of these basic relations in the sentence is crucial for the foreign language learners, and along with the understanding of the syntactic structure helps them to build correct sentences on their own. The example of the semantic tree of the same sentence from the HR4EU corpus is presented in Figure 4.



#3394

Maja ***voli*** sladoled od vanilije , ali ne ***voli*** sladoled od čokolade .

Figure 4: The example of the semantic tree - HR4EU sub-corpus (cf. Figure 3 for glosses and translation)

5. Conclusion

In this paper we have presented the HR4EU – web portal for e-learning of Croatian and its bidirectional relation to language resources for Croatian. The HR4EU is the first completely free-of charge portal with e-courses of Croatian language developed by language professionals. Moreover, it is the first portal which takes advantages of the language technologies in the computer aided language learning. The interrelation between the HR4EU and the one of the existing LRs for Croatian – the Croatian Dependency Treebank – is described in this paper.

The resources like HOBS are most commonly used in NLP tasks, e.g. parsing (syntactic layer) and automatic semantic role labelling (semantic layer). However, they can be extremely useful in the CALL as well, but they have to be modified to serve the language learners' purposes. We have applied the same model used for the HOBS to less complex sentences used in the HR4EU courses to help our users to understand the syntactic and semantic structure of Croatian sentences. They can, moreover, use this resource if they are not sure of the verbal frame, e.g. if they don't know which preposition they should use with the particular verb to express the particular argument. The other language resources, especially different morphological lexica, can also be helpful in the CALL, and further stress the importance of language technologies in the computer aided language learning. The application of LRs to other domains, along with NLP, extends their visibility and usability.

Acknowledgement

This paper and hereby presented project are fully supported by European Union, European Social Fund, under the project grant HR.3.2.01.-0037 Mrežni portal za online učenje hrvatskoga jezika HR4EU.

References

Agić, Ž. (2012). *Pristupi ovisnosnom parsanju hrvatskih tekstova.* PhD thesis, University of Zagreb, Faculty of Humanities and Social Sciences.

Agić, Ž., Berović, D., Merkler, D., Tadić, M. (2014). Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2313-2319.

Agić, Ž., Ljubešić, N. (2014). The SETimes.HR Linguistically Annotated Corpus of Croatian. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 1724-1727.

Oliver, A., Šojat, K., Srebačić, M. (2015). Enlarging the Croatian Wordnet with WN-Toolkit and CroDeriV. In Angelova, G., Bontcheva, K., Mitkov, R. (eds.) Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria: BAS, pp. 480-487.

Palmer, M., Gildea, D., D. Nianwen, X. (2010). *Semantic Role Labelling*. Morgan and Claypool Publishers.

Raffaelli, I., Bekavac, B., Agić, Ž., Tadić, M. (2008). Building Croatian Wordnet. In: Tanács, Attila; Csendes, Dóra; Vincze, Veronika; Fellbaum, Christianne; Vossen, Piek (eds.) Proceedings of the Fourth Global WordNet Conference 2008, Szeged: GWC, pp. 349-359.

Šojat, K., Srebačić, M. and Štefanec, V. (2013). CroDeriV and the Morphological Analysis of Croatian Verb, *Suvremena lingvistika* 39/(75): 75-96.

Tadić, Marko (1996). Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika* 41-42.

Tadić, M., Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In: *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* (Budapest 2003), ACL, pp. 41-46.

Tadić, M. (2006). Croatian Lemmatization Server. In*:* Vulchanova, M. D., Koeva, S., Krapova, I., Vulchanov, V. (Eds.). *Formal Approaches to South Slavic and Balkan Languages*. Sofia: Bulgarian Academy of Sciences, pp. 140-146.

Tadić, M (2007). Building the Croatian Dependency Treebank: The Initial Stages. *Suvremena lingvistika* 63: 85-92.

Appendix 1 – Syntactic specification used in HOBS – syntactic level

Table 1: List of analytical functions from PDT

afun	Description
Adv	adverbial
Apos	apposition
Atr	attribute
Atv	complement hung on a non-verb. element
AtvV	complement hung on a verb
AuxC	subordinating conjunction
AuxG	other graphic symbols, not terminal
AuxK	terminal punctuation of a sentence
AuxO	emotional, rhythmic particles
AuxP	preposition
AuxR	reflexive passive
AuxT	reflexivum tantum
AuxV	auxiliary verb
AuxX	comma
AuxY	some particles
AuxZ	emphasizing words
Coord	coord. node
ExD	ellipsis
Obj	object
Pnom	nominal predicate
Pred	predicate
Sb	subject

Table 2: Syntactic sub-tagset for subordinate clause annotation in HOBS

tag	Description
Sub_Atr	attribute
Sub_Adv	adverbial
Sub_Obj	object
Sub_Pred	predicate
Sub_Sb	subject

Table 3: Syntactic sub-tags for adverbial clause subclassification

Sub Adv	Adverbial clause
Sub_Adv_loc	local
Sub_Adv_temp	temporal
Sub_Adv_mod	modal
Sub_Adv_caus	causative
Sub_Adv_cons	consequential
Sub_Adv_fin	final
Sub_Adv_cond	conditional

Appendix 2 – SRL specification used in HOBS – semantic level

Table 1. SRL Tagset for Croatian

Tag	Role	Example
AGT_anim	agent_animate	**Marko** je zapjevao. ***Marko*** *started to sing.*
AGT_inanim	agent_inanimate	**SDP** je izjavio... ***SDP*** *has declared...(political party)* Atomski **udar** je uništio grad. *Nuclear* ***attack*** *has destroyed the city.*
PAT	patient	Oduzeli su **im** ljudska prava. *They took* ***them*** *the human rights.* Marko je udario **loptu**. ***Marko*** *hit the ball.*
EXP	experiencer	**Krešo** uživa u hrenu i šunki. ***Krešo*** *enjoys eating ham and horseradish.*
BEN	beneficiary	Krešo uči **studente** matematiku. *Krešo teaches* ***students*** *math.*
RES	result	On je postao **Španjolac**. *He became* ***Spanish****.*
PART	participant	**Hrvatska** je potpisala ugovor s **Rusijom**. ***Croatia*** *signed the contract with* ***Russia****.*
TEM	theme	On je naučio **španjolski**. *He learned* ***Spanish****.* **Lopta** se odbila od zida. *The* ***ball*** *bounced from the wall.*
INS	instrument	**Ključem** je otvorio vrata. *He opened the door with the* ***key****.*
SRC	source	Svjetlost dolazi **od Sunca**. *The light comes from the* ***Sun****.*
QUAN	quantity	Povećali su trošarine **od 10 do 20 posto**. *They increased the exice duties* ***from 10 to 20 percent****.*
TMP	time	Putovanje je trajalo **od 6 ujutro do 7 navečer**. *The trip lasted* ***from 6 a.m. until 7 p.m.*** **Prošle godine** donijeli su odluku o ukidanju ropstva. *They decided to abolish slavery* ***last year****.*
LOC	location	Delegacija je otputovala **u Sarajevo**. *The delegation has departed* ***to Sarajevo****.*
LOC_ap	abstract location	Hrvatska se prima **u Partnerstvo za mir**. *Croatia is becoming a member of* ***Partnership for Peace****.*
CAU/FIN	cause, intention	Krešo je umoran **od naporna rada**. *Krešo is tired* ***of hard work****.* Povisili su poreze **radi smanjenja deficita**. *They increased taxes* ***to reduce deficite****.*
MNR	manner	**Veselo** su potpisali sporazum. *They* ***cheerfully*** *signed an agreement.* Igrao je **nepošteno**. *He played* ***unfairly****.*
ATR	attribute clause	Vidio sam dijete **koje** se igralo. *I saw a child* ***that*** *was playing.*