Prethodno priopćenje

UDK 811.163.42'364
81–118.6:811.163.42
81'322.3
Prethodno priopćenje
Prihvaćeno za tisak 14. 02. 2007.

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

Building the Croatian Dependency Treebank: the initial stages

The paper presents work–in–progress on the building of the Croatian Dependency Treebank. Its design principles, procedures and the pilot corpus used within are described. Perspectives for further development of the Croatian Dependency Treebank are presented at the end.

1. Introduction

Treebanks have become a widely used language resource for different purposes, starting from more theoretical ones such as the study of the general theory of syntax, to empirically based studies of language–specific syntactic features, and up to the training of syntactic parsers. Treebanks already exist in a variety of theoretical approaches and formats for a number of different languages.

Having already developed the basic language resources for Croatian, primarily Croatian National Corpus (Tadić 2004), Croatian Morphological Lexicon (Tadić & Fulgosi 2003) and Croatian Lemmatization Server (Tadić 2006), our intention was to take a step further and start the syntactic processing of Croatian by building a Croatian treebank. Its first purpose would be to function as a syntactically annotated linguistic resource. It was expected that its existence would be one of the preconditions for thorough research on parsing of Croatian, i. e. for building, training and testing parser(s).

As a member of the South–Slavic sub–family of languages, Croatian exemplifies all the features of Slavic morphosyntax: it has a rich morphology (7 cases, 2 grammatical numbers, 3 simple tenses, 3 compound tenses, 3 moods, 4

participles, elaborate and asymmetric aspectual system, etc.), a relatively free word order and other very interesting syntactic features such as clitic placement, long–distance dependencies etc. This makes its syntactic description even more interesting, as was reflected in a number of traditional and contemporary grammars, with Katičić (1986) probably the most prominent one. Although many individual syntactic phenomena have also been studied from the generativist point of view, there is no comprehensive Croatian grammar using that or any other formal framework.

There was only one attempt to build a Croatian parser (Seljan 2003) using LFG formalism, but it remained in the prototype stage and certainly could not be used for larger–scale treebank (pre–)processing and/or building since it is not robust enough, covers only a limited number of selected syntactic structures and cannot cope with all types of multiple–clausal sentences.

The Croatian Dependency Treebank[1] (*Hrvatska ovisnosna banka stabala*, HOBS from now on) has been initiated as one of the tasks at the very end of the project *Development of Croatian Language Resources* with prospective longer–term continuation (2007–2009) in the following computational linguistic projects. It started at the end of 2005 when the decision to adopt and follow the already existing and tested formalism had been made.

2. Choosing the approach

2.1. Slavic treebanks

Since several Slavic treebanks (Czech, PDT: Hajič 1998; Bulgarian, BulTreeBank: Simov et al. 2002; Russian: Boguslavsky et al. 2000; Polish HPSG treebank: Marciniak et al. 2000; Slovenian, SDT: Džeroski et al. 2006) already exist in different stages of production, we investigated them in order to opt for the annotation system which would be best suited for our task and available human resources. Finally we decided to adopt the PDT approach. There were several reasons for this decision.

While the first treebanking projects used the constituency annotation system, dependency annotation has become increasingly popular during the past few years as the number of treebanks for languages other than English has increased. The constituency based annotation schemes in treebanks are motivated by underlying generative formalisms describing the hierarchy and composition of the constituents (such as S $\rightarrow$ NP VP) in a sentence. The dependency based annotation schemes are motivated by underlying dependency formalisms trying to define dependency relations between parts of a sentence (such as hit(the_boy, the_ball)). Each approach has its pros and cons and the best solution would probably be to have both annotations present for each sentence in a treebank. Already existing constituency–annotated treebanks have

1 The web address of the project is *http://hobs.ffzg.hr*.

been enriched with dependency annotation layers (e. g. Penn Treebank). In this way we can talk about the union of syntactic annotations and deal with syntactic description on a higher, more universal level where we can compare and combine features from both approaches.

The primary reason for using dependency structures instead of more informative lexicalized phrase structures is that they are more efficient to learn and parse while still encoding much of the predicate–argument information needed in applications (McDonald et al. 2005).

It also seems that dependency annotation is more suited for languages which are typologically similar to Slavic ones due to specific (morpho)syntactic phenomena. To mention just one of them: the problem of long–distance dependencies (and the notorious problem of branch–crossing) could be modeled much easier with dependency featuring formalism instead of the constituency based approach (i. e. with non–projective dependencies like those described in McDonald et al. 2005).[2]

The more–or–less free word order in Croatian sentences is another expected feature which should not cause so many problems to dependency–based annotation, but would surely present a computational problem to constituency–based parsers. In fact, no statistics have ever been calculated regarding clausal and/or sub–clausal structures in Croatian in order to present in exact figures which word–order is the most common one. This is also one of the results we expect from our treebank.

2.2. The PDT approach

Since there is no formal syntactic description of Croatian yet, we were free to choose any approach we found appropriate. The PDT approach was a clear choice because of: 1) its theoretical foundations (Vuković 2007, an excellent PhD which presents a thorough description of Prague Two level valency syntactic theory and suggestions how to adapt it to Croatian as well); 2) experience in building PDT; 3) practical software support (i. e. TrEd tree editor, Pajas 2000).

The same choice made by the SDT team also gave us an opportunity to tackle some problems in the same manner or even in co–ordination when we come across the same/similar phenomena in Slovenian and Croatian. Genetic

2 Although there are also annotation schemes that enable annotation of discontinued constituents (e. g. TIGER annotation scheme with two output formats: Negra, a text–based format, and TIGER, an XML–based format, whose specifications can be found at the project web page http://www.ims.uni–stuttgart. de/projekte/TIGER/), they seemed too complicated for building the syntactic treebank from scratch. Actually, TIGER annotation scheme enables encoding of both approaches, constituency and dependency based and in that respect it allows the union of syntactic annotations. Once the dependency relations in the Croatian Dependency Treebank are explicitly tagged, it could be relatively easy to convert its format into TIGER XML–based interchange format and add the constituency level of annotation later.

closeness of Croatian and Slovenian can be of great help here for both projects but it could also exhibit interesting and subtle differences.

There are several layers of annotation in PDT: morphosyntactic level (disambiguated and MSD–tagged corpus), analytical level, tectogrammatical level and even inter–level information (see Razímová & Žabokrtský 2006 on annotating grammatemes). In the first phase of building HOBS we will try to deal with the analytical level which is placed above the morphosyntactic level.

Syntactic similarity to Czech enabled us to start using the publicly available and well elaborated PDT annotation manual (Hajič et al. 1997) directly off the shelf, but at the same time we were able to track the divergence of syntactic behavior in Croatian. This method was also used for SDT and it looks as if it could be well suited for building the dependency treebanks or adding a dependency annotation layer to the existing treebanks of other Slavic languages.[3] A close connection with SDT and the same proven software as that used by the Prague and Ljubljana teams (Džeroski et al. 2006) will certainly help us speed up our process of manual annotation.

3. The pilot treebank

To test the selected methods and tools we started with a small pilot corpus and manually annotated the sentences on the analytical level.

3.1. The corpus

The chosen corpus is a part of Croatian National Corpus, i. e. CW2000 subcorpus: a newspaper corpus covering different topics and fields, originally a Croatian side of the Croatian–English Parallel Corpus (Tadić 2000). Its size is around 100,000 tokens.

3.2. MSD annotation and lemmatization

The corpus was automatically MSD tagged and lemmatized using the Croatian Lemmatization Server (Tadić 2006) on the unigram level. The tagset used was MULTEXT East v3 guidelines (Erjavec 2004) i. e. their specification for Croatian.[4] The corpus was manually disambiguated for MSD and lemmas. The first 500 sentences were selected for the pilot HOBS corpus. It was divided into portions of 50 sentences in length each, then converted from XML (XCES) format to TrEd's native FS format and further manually annotated using TrEd.

---

3 The situation somewhat resembles the role of Princeton WordNet in building WordNets for other languages using the "translation approach" where PWN served as a 'theoretical and practical seed' from which other WNs developed.

4 For detailed description of MULTEXT East morphosyntactic guidelines and language resources (i. e. parallel corpus of translations of Orwell's *1984* to a number of languages) see its web page at http://nl.ijs.si/ME/V3.

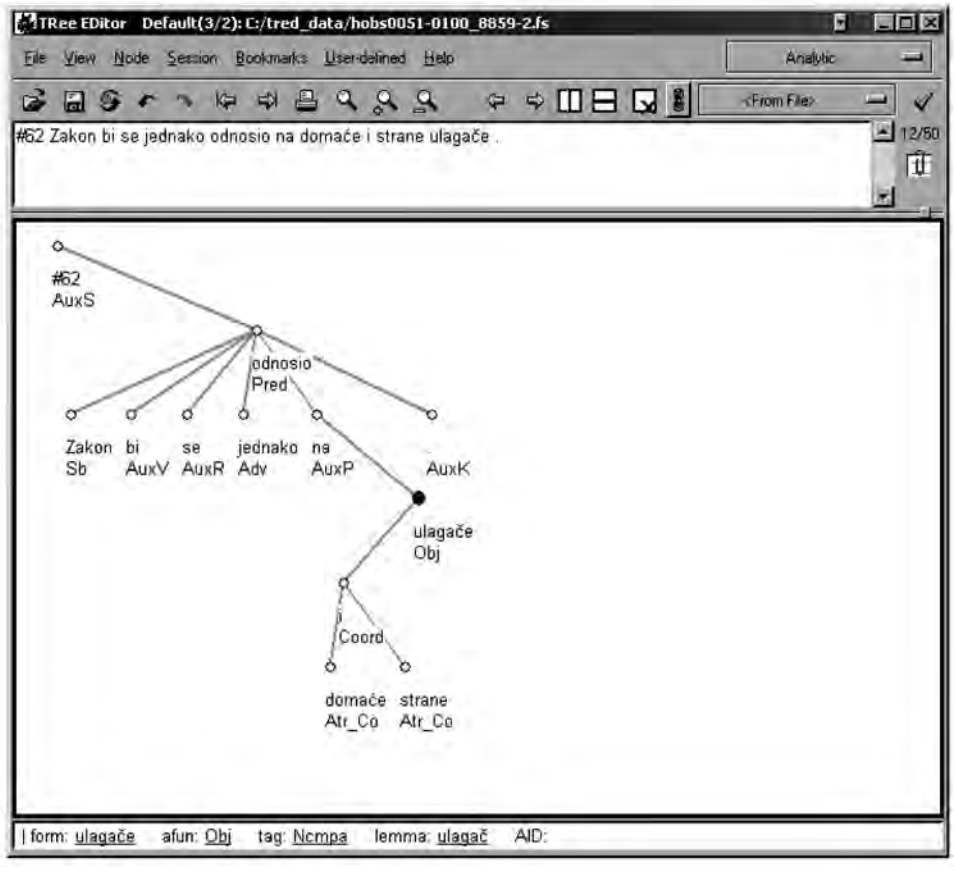

Figure 1. Example of annotated sentence from HOBS using TrEd

Since there is no parser and/or chunker available for Croatian, we could not perform the shallow parsing or chunking which would give us at least basic syntactic structures and preprocessed input data for TrEd. Instead, each sentence had to be manually annotated from scratch, which was tedious but rather instructive job. For the manual annotation of the pilot HOBS corpus four annotators were engaged without using the parallel annotating method. Although this method would certainly help in further consistency checking, the size of the corpus allowed us to simply exchange the 50–sentence portions between annotators and check them manually. Since the checking is still under way, we cannot come up with any serious statistics as yet, because it may change due to possible systematic changes in tag usage (e. g. two or more coordinated Atr tags should be changed to Atr_Co systematically since in the first run the correct Atr_Co tag was not applied consistently). At the moment we may say that the first 500 sentences contain 3717 nouns, 1840 verbs, 1475 adjectives, 532 adverbs, 828 pronouns, 239 numerals, 1237 prepositions, 898 conjunctions etc. The average length of a sentence is 25.10 tokens.

At the moment we are experimenting with preprocessing of chunks based on local regular grammars in an attempt to automatically annotate sub–clausal syntactic structures (see the "islands of certainty" by Abney 1996) such as adjective(s)+noun (A*N) or preposition+adjective(s)+noun (PA*N). The suggested chunks must have some identical MSD values (i. e. A*N should have the same number, gender and case; P and A*N should have the same case). The first experimental results suggest that these structures are quite common in Croatian and cover almost 15% of all tokens in corpus.

The next step could be simple shallow parser covering predicates and their arguments (the commonest among them being the verb finite word–form, subject in the nominative, direct object in the accusative and indirect object in the oblique case).

4. Perspectives

Further work on HOBS will continue in several directions. First we will try to adapt the PDT analytical annotation manual to Croatian and in coordination with the SDT team.

We also plan to include more sentences in the annotated corpus and cover the whole CW2000 corpus until the total number of 4626 sentences is reached in the future.

It would also be interesting to syntactically annotate the Croatian translation of Orwell's *1984*, Part I. This would make possible a whole range of different experiments with other parallel translations of the same text, providing an opportunity to comparatively investigate syntactic phenomena in typologically and genetically similar and/or distant languages. We need not emphasize the use that such a resource would have in e. g. parallel grammar induction, machine translation, etc.

One of the most promising experiments (Barbu–Mititelu & Ion 2005) has shown that using sentence–aligned parallel corpora enables the syntactic annotation transfer even between typologically different languages. Having both sides in a parallel corpus syntactically annotated makes possible the automation of the evaluation process. We would like to test the Croatian translation of Orwell against other genetically close (Slovenian, Czech, Serbian, Bulgarian) and more distant languages (Romanian, English etc.) using this kind of evaluation.

Another experiment (Tufiş et al. 2006) included the transfer of verbal valency information but, in addition to the parallel corpus, it also needed wordnets developed for respective languages. This investigation has shown that the direct transfer of verbal valency information between verbs from Czech and their Romanian translation equivalents yielded almost 80% of correct verbal valencies in the target language. This experiment could be conducted on languages that are even closer, such as Czech and Croatian, where an even higher percentage could be expected. Having the Croatian translation of Orwell also syntactically annotated would also enable the automation of the evaluation

process. The second prerequisite for this investigation, i. e. the Croatian WordNet which is under construction, unfortunately does not yet exist in a size applicable to this type of experiment.

We would also like to build post–annotation tests for Croatian in order to check the consistency and quality of manual annotation procedure like in (Hladka & Pajas 2001).

Finaly, HOBS will be used as a test bed for dependency parsers applied to Croatian, whether being adapted from the already existing ones, or written on our own.

Acknowledgments

The work presented in this paper was partially funded by the Croatian Ministry of Science, Education and Sports within the project 130–1300646–0645

References

(Abney 1996) S. Abney, 1996, *Partial Parsing via Finite–State Cascades*, In *Journal of Natural Language Engineering* 2 (4), pp. 337–344.

(Barbu–Mititelu & Ion 2005) V. Barbu–Mititelu, R. Ion, 2005, *Cross–language Transfer of Syntactic Relations Using Parallel Corpora*, In *Proceedings of the Workshop on Cross–Language Knowledge Induction*, EUROLAN2005, Cluj–Napoca, Romania, pp. 46–51.

(Boguslavsky et al. 2000) I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin, N. Frid, Nadezhda, 2000, *Treebank for Russian: Concept, tools, types of information*, In *Proceedings of COLING 2000*, Saarbrücken, Germany, pp. 83–89.

(Böhmová et al. 2003) A. Böhmová, J. Hajič, E. Hajičova, B. Hladká, 2003, *The Prague Dependency Treebank: A Three–Level Annotation Scenario*, In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*, Kluwer, 2003, pp. 103–127.

(CoNLL–X 2006) *CoNLL–X Shared Task: Multi–lingual Dependency Parsing*, In P*roceedings of the Tenth Conference on Computational Natural Language Learning*, New York, USA, 2006. (URL:http://nextens.uvt.nl/čconll/)

(Čmejrek et al. 2004) M. Čmejrek, J. Cuoín, J. Havelka, J. Hajič, V. Kuboň, 2004, *Prague Czech–English Dependency Treebank: Syntactically Annotated Resources for Machine Translation*, In *Proceedings of the Fourth Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal, ELRA, 2004, pp. 1597–1600.

(Džeroski et al. 2006) S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky, A. Žele, 2006 *Towards a Slovene Dependency Treebank*, In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy, ELRA, 2006.

(Erjavec 2004) T. Erjavec, 2004, *MULTEXT–East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*, In *Proceedings of the Fourth Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal, ELRA, 2004, pp. 1535–1538. (URL for MULTEXT–East language resources and morphosyntactic specifications: http://nl.ijs.si/ME/V3/)

(Hajič 1998) J. Hajič, 1998, *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*, In *Issues of Valency and Meaning*, Prague, pp. 106–132.

(Hladka & Pajas 2001) B. Hladka, P. Pajas, 2001 Post–annotation Checking of the Treebank, Technical report, Charles University, Prague, 2001 (URL http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/AvsM. pdf).

(Katičić 1986) R. Katičić, 1986, *Sintaksa hrvatskoga književnog jezika*, Zagreb, HAZU.

(Marciniak et al. 2000) M. Marciniak, A. Mykowiecka, A. Przepiorkowski, A. Kupsc, 2000, *An HPSG–Annotated test Suite for Polish*, In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, ELRA, 2000, pp. 1671–1677.

(McDonald et al. 2005) R. McDonald, F. Pereira, K. Ribarov, J. Hajič, 2005, *Non–projective Dependency Parsing using Spanning Tree Algorithms*, In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*, Vancouver, BC, Canada, 2005, pp. 523–530.

(Pajas 2000) P. Pajas, 2000, *Tree Editor TrEd*, PDT, Charles University, Prague, 2000 (URL: http://ufal.mff.cuni.cz/~pajas/tred/).

(Razímová & Žabokrtský 2006) M. Razímová, Z. Žabokrtský, 2006, *Annotation of Grammatemes in the Prague Dependency Treebank 2.0*, In *Proceedings of the Workshop "Annotation Science" at the Fifth Conference on Language Resources and Evaluation (LREC2006)*, Genua, Italy, ELRA, pp. 12–19.

(Seljan 2003) S. Seljan, 2003, *Leksičko–funkcionalna gramatika hrvatskoga jezika: teorijski i praktični modeli*, PhD dissertation, University of Zagreb, Zagreb, 2003.

(Simov et al. 2002) K. Simov, P. Osenova, M. Slavcheva, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, Krasimira, A. Simov, M. Kouylekov, 2002, *Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank*, In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain, ELRA, 2002, pp. 1729–1736.

(Tadić 2000) M. Tadić, 2000, *Building the Croatian–English Parallel Corpus*, In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece, ELRA, 2000, pp. 523–530.

(Tadić 2002) M. Tadić, 2004, *Building the Croatian National Corpus*, In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain, ELRA, 2002, pp. 441–446.

(Tadić 2006) M. Tadić, 2006, *Croatian Lemmatization Server*, In *Proceedings of the Formal Approaches to South Slavic and Balkan Languages (FASSBL2006)*, Sofia, Bulgaria, Bulgarian Academy of Sciences, 2006, pp. 140–146.

(Tadić & Fulgosi 2003) M. Tadić & S. Fulgosi, 2003. Building the Croatian Morphological Lexicon, In T. Erjavec, D. Vitas (eds.) Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL2003, Budapest, 2003, pp. 41–46.

(Tufiş et al. 2006) D. Tufiş, V. Barbu–Mititelu, L. Bozianu, C. Mihăilă, 2006, Romanian WordNet: New Developments and Applications, In Proceedings of the 3rd Conference of the Global WordNet Association, Seogwipo, Jeju, Republic of Korea, 2006, pp. 337–344.

(Vuković 2007) P. Vuković, 2007, *Prednosti dvorazinske valencijske sintakse u sintaktičkom opisu slavenskih jezika – na primjeru češkoga i hrvatskoga jezika*, PhD dissertation, University of Zagreb, Zagreb, 2007.

Sastavljanje Hrvatske ovisnosne banke stabala: početne etape

Članak donosi međurezultate sastavljanja Hrvatske ovisnosne banke stabala koje je istraživanje u tijeku. Opisuju se njezina načela oblikovanja, postupci i uporabljeni pilot korpus. Na kraju se članka predstavljaju perspektive za daljnji razvitak Hrvatske ovisnosne banke stabala.

Key words: dependency treebank (linguistics), dependency grammar, corpus linguistics, parsing, Croatian

Ključne riječi: stabla ovisnosti (lingvistika), gramatika ovisnosti, korpusna lingvistika, parsanje, hrvatski jezik