VERB VALENCY FRAME EXTRACTION USING MORPHOLOGICAL AND SYNTACTIC FEATURES OF CROATIAN

Krešimir Šojat*, Željko Agić, Marko Tadić***

*Department of Linguistics, **Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{ksojat, zeljko.agic, marko.tadic}@ffzg.hr

ABSTRACT

The paper presents an approach to valency frame extraction for Croatian verbs on basis of morphological and syntactic features of wordforms from syntactically annotated sentences. We have used a gold standard sample of approximately 1200 sentences and 30.000 tokens from the Croatian Dependency Treebank and a frame instance extraction algorithm. We extracted 936 verb frame instances for 424 different verbs – consisting of lemmas, morphosyntactic tags and syntactic functions of the encountered wordforms – and manually assigned tectogrammatical functors to their elements. Distributional properties are given in terms of co-occurrences for each of these features. The obtained results will serve for further development of valency frame extraction procedures.

1. Introduction

Recent enhancements of the Croatian Dependency Treebank both in size and annotation quality enabled the development of procedures for (semi-)automatic extraction of valency frames for Croatian verbs. The initial experiment, presented in (Agić et al. 2010), produced a rule-based procedure for the extraction of specific instances of verb valency frames from the treebank. On the basis of the results shown there, we present in this paper an extension of that specific line of research in terms of improvements of the algorithm to be used in further semi-automatic construction of verb valency frames. More specifically, in this paper we attempt to induce a set of statistically verified rules for the assignment of the most probable tectogrammatical functors to sentence elements on the basis of their morphosyntactic features and syntactic functions. In order to achieve the objective we extracted valency frame instances for verbs from a gold standard section of the treebank, manually annotated the extracted elements for tectogrammatical functors and established a set of relations between verbs, tectogrammatical functors (that roughly correspond to the notion of semantic or theta roles), syntactic functions and morphosyntactic features in the form of statistical distributions of their co-occurrences. We hope to use these distributional properties in the process of semi-automatic valency frame induction by applying the acquired rules on unseen portions of the treebank. To the best of our knowledge, other than (Agić et al. 2010) and (Šojat et al. 2010) – the latter implementing a rule-based approach to valency – no similar experiments in verb valency frame extraction were done on Croatian texts.

In the following section of the paper, we present the recent advancements in the development of Croatian Dependency Treebank in more detail. Sections 3 and 4 present the setup of the experiments and discuss the obtained results. We conclude the paper with an outline of future research, specifically emphasizing the utilization of the treebank and verb valency lexicons in stochastic dependency parsing.

2. Croatian Dependency Treebank

Croatian Dependency Treebank (hr. *Hrvatska ovisnosna banka stabala*, HOBS further in the text), as described in e.g. (Tadić 2007) and (Agić et al. 2010), is a dependency treebank built along the principles of Functional Generative Description (FGD) (Sgall et al. 1986), a multistratal model of dependency grammar developed for Czech. In a somewhat simplified version, the FGD formalism was further adapted in the Prague Dependency Treebank (PDT) (Hajič et al. 2000) project and applied for the sentence analysis and annotation on the levels of morphology, syntax – in the form of dependency trees with nodes labeled with syntactic functions – and tectogrammatics.

Annotation of a sentence at the morphological layer consists of attaching several attributes to the tokens such as morphological lemmas and morphosyntactic tags. At the analytical layer, the sentence is represented in the form of a tree with labeled nodes. In the syntactic analysis of a sentence a set of analytical functions such as subject or object are attached

to nodes of the tree as attributes. On the tectogrammatical layer, i.e. on the layer of the representation of sentence meaning and semantic relations among its elements sentences are also represented as rooted trees with labeled nodes. Unlike the analytical layer, not all the morphological tokens are represented at the tectogrammatical layer (e.g. there are no prepositions, nodes representing omitted subject are introduced, etc.). Similarly to the analytical layer, the edges of the tree represent relations between the nodes, the type of the relation being indicated by a set of labels. The total of 39 attributes can be assigned to every non-root node of the tectogrammatical tree. Every node representing a verb or a certain type of a noun has a valency frame assigned to it by means of a reference to a valency dictionary PDT-VALLEX (Hajič et al. 2003) (cf. http://ufal.mff.cuni.cz/pdt2.0/).

The ongoing construction of HOBS closely follows the guidelines set by the PDT, with their simultaneous adaptation to the specifics of the Croatian language. More detailed account of the HOBS project plan is given in (Tadić 2007). HOBS at this moment (2010-09) consists of approximately 2.870 sentences in the form of dependency trees that were manually annotated with syntactic functions using TrEd (Pajas 2000) as the annotation tool, whereas the manual annotation of sentences on the tectogrammatical layer is currently not conducted. These sentences, encompassing approximately 70.000 tokens, stem from the magazine Croatia Weekly, i.e. the Croatia Weekly 100 kw (CW100) corpus that is a part of the newspaper sub-corpus of the Croatian National Corpus (HNK) (Tadić 2000). The Croatia Weekly sub-corpus was previously XCES-encoded, sentence-delimited, tokenized, lemmatized and MSD-annotated by linguists using a semiautomatic procedure (cf. Tadić 2002). Thus, each of the analyzed sentences contains the manually checked information on part-of-speech, morphosyntactic category, lemma, dependency and analytical function for each of the wordforms. Such a course of action, i.e. the selection of the corpus, was taken in order to enable the training procedures of various state-of-the-art dependency parsers (Buchholz et al. 2006), (Nivre et al. 2007), to choose from a wide selection of different features in this and the upcoming experiments with stochastic dependency parsing of Croatian texts. Basic stats for HOBS and the experiment sets are given in Table 1 and will be further discussed in the following section. Sentences in HOBS are annotated according to the PDT annotation manual for the analytical level of annotation, with respect to differing properties of the Croatian language and consulting the Slovene Dependency Treebank (SDT) project (Džeroski et al. 2006). The utilized analytical functions are thus compatible with those of the Prague Dependency Treebank. Further work on HOBS includes, among other tasks: enlarging the treebank, cross-validating the treebank annotation, designing a manual for HOBS annotation and conducting a comparative analysis of HOBS, SDT and PDT.

3. Experiment setup

Two basic components were made available for conducting this experiment: the Croatian Dependency Treebank in CoNLL (cf. Buchholz et al. 2006) format and the algorithm for extracting verb valency frame instances from it, i.e. the algorithm presented in (Agić et al. 2010).

The treebank, i.e. its 2.870 manually annotated sentences, is stored in the native TrEd feature structure (FS) format. Using TrEd, we converted the treebank into the Czech sentence tree structure (CSTS) format and then easily translated this format into the CoNLL format by simple regular expressions. Further, we implemented a script for CoNLL token validation and filtered out sentences with invalid tokens. The results of this filtering are given in Table 1: token encoding issues invalidated 171 sentences and thus left a total of 66.930 tokens available for the experiment. The aforementioned token encoding issues were mainly caused by missing escape sequences for decimal numbers within FS-formatted sentences and are currently being corrected. However, out of the 2.699 valid sentences available in CoNLL format, at the moment of conducting this experiment, only 1242 were already double-checked by expert linguists dealing with adapting the PDT formalism to the specifics of Croatian syntax. Therefore, once again as indicated by Table 1, only 1.242 sentences and 29.892 tokens were used here.

Feature	Treebank	This experiment
Sentences	2699	1242
Tokens	66930	29892
Lemmas	8995	5501
MSD tags	798	649
Analytical functions	80	65

Table 1. Treebank stats

The previously mentioned extraction algorithm – described in more detail by (Agić et al. 2010) – was also modified for purposes of this experiment. Its previous version was designed to detect only the verbs annotated with analytical functions *Pred*, *Pred_Co* and *Pred_Pa* and descend one level down the dependency tree to retrieve subjects (*Sub*), objects (*Obj*), adverbs (*Adv*) and nominal predicates (*Pnom*) or two levels down to retrieve the same tokens (annotated as *Sub*, *Obj*, *Adv*, *Pnom*) introduced by using subordinate conjunctions (*AuxC*) and prepositions (*AuxP*). Here, we adapted the algorithm to retrieve any verbs found in the dependency structure, regardless of their respective analytical functions and position within the dependency trees. The adaptation itself is implemented in order to raise the recall of the algorithm (while still maintaining its precision by not changing the simple set of descending rules), i.e. to retrieve as much verbs as possible given the limited size of the treebank sample used in the experiment.

```
              biti  [ dovršiti dovršena Vmps-sfp Pnom ]     [ studija studija Ncfsn Sb ]
       (biti Obj)  [ dovršiti dovršena Vmps-sfp Pnom PAT ] [ studija studija Ncfsn Sb ACT ]

 djelovati(djeluje  [ neozbiljno Neozbiljno Rnp Adv ]      [ odustajanje odustajanje Ncnsn Sb ]
            Pred)  [ neozbiljno Neozbiljno Rnp Adv MANN ] [ odustajanje odustajanje Ncnsn Sb ACT ]

       oslobodit i  [ nikada Nikada Rt Adv ]     [ zloduh zloduha Ncmsg Obj ]
 (oslobodili Pred)  [ nikada Nikada Rt Adv THL ] [ zloduh zloduha Ncmsg Obj PAT ]

          postati  [studij studiji Ncmpn Sb]     [fakultet fakultet Ncmsn Obj]
   (postali Pred)  [studij studiji Ncmpn Sb ACT] [fakultet fakultet Ncmsn Obj PAT]

        postojati  [ objektivno Objektivno Rnp Adv ]      [ problem problem Ncmsn Sb ]
 (postoji Pred_Co)  [ objektivno Objektivno Rnp Adv MANN ] [ problem problem Ncmsn Sb ACT ]

         prerasti  [ šuma u->šumu Spsa->Ncfsa AuxP->Adv ]
(prerastao ExD_Co)  [ šuma u->šumu Spsa->Ncfsa AuxP->Adv EFF ]

          započeti  [ proces Proces Ncmsn Sb ]     [ već već Rt Adv ]
 (započeo Pred_Co)  [ proces Proces Ncmsn Sb ACT ] [ već već Rt Adv MANN ]

       zaustaviti  [ oni ih Pp3-pa--y-n-- Obj ]     [ dolina u->dolini Spsl->Ncfsl AuxP->Adv ]
  (zaustavio Atr)  [ oni ih Pp3-pa--y-n-- Obj PAT ] [ dolina u->dolini Spsl->Ncfsl AuxP->Adv LOC ]
```

Figure 1. An example verb valency frame instance and its annotation

The algorithm was run on the treebank sample, extracting 2930 valency frame instances. Tectogrammatical functors were afterwards manually assigned to the extracted wordforms, as illustrated in Figure 1. A total of 936 frame instances were annotated for 424 different verbs. The following section presents the results obtained by counting co-occurrences of tectogrammatical functors and valency frames on the one side and verbs, morphosyntactic tags and analytical functions on the other.

In order to annotate verbal frames we used a set of functors used to describe verb valency, namely 5 argument functors and functors for 32 free modification functors. This is the list of free modification we used:

(1) Argument functors: *ACT* (actor), *PAT* (patient), *ADDR* (addressee), *ORIG* (origin), *EFF* (effect)

(2) Temporal functors: *TWHEN* (when), *TFHL* (for how long), *TFRWH* (from when), *THL* (how long), *THO* (how often), *TOWH* (to when), *TPAR* (temporal parallel), *TSIN* (since when), *TTILL* (till)

(3) Locative and directional functors: *DIR1* (where from), *DIR2* (which way), *DIR3* (where to), *LOC* (where)

(4) Functors for causal relations: *AIM* (purpose), *CAUS* (cause), *CNCS* (concession), *COND* (condition), *INTT* (intention)

(5) Functors for expressing manner: *ACMP* (accompaniment), *CPR* (comparison), *CRIT* (criterion), *DIFF* (difference), *EXT* (extent), *MANN* (manner), *MEANS* (means), *REG* (regard), *RESL* (result), *RESTR* (restriction)

(6) Functors for specific modifications: *BEN* (benefactor), *CONTRD* (contradiction), *HER* (heritage), *SUBS* (substitution)

This set of functors was chosen because we believe that they are sufficient to capture and represent main syntactic and semantic relations within sentences covering major morphosyntactic functions such as subject, object and various types of adverbials. On the other hand, a similar set of functors is used in lexica dealing exclusively with verb valency, such as CROVALLEX (Mikelić Preradović et al. 2009), developed for Croatian.

4. Results and discussion

Seven distributional properties were obtained by analyzing the previously presented manual annotation of valency frame instances within our testing framework:

(1) frequency of applied tectogrammatical functors,

(2) frequency of verb lemmas,

(3) frequency of functor n-grams, i.e. valency frames,

(4) distribution of valency frames from the previous distribution according to the verb they represent,

(5) distribution of morphosyntactic tags across functors,

(6) distribution of syntactic, i.e. analytical functions across functors and

(7) the previous two distributions combined, i.e. the distribution of pairs of analytical functions and morphosyntactic tags across tectogrammatical functors.

These results are presented in a somewhat compressed form in tables 2, 3 and 4 and brief interpretation of the presented data is given further in the text.

Table 2 provides the frequency of functors used in annotation and appears to be rather straightforward and expected. Namely, the most frequent functors are *PAT* (Patient), *ACT* (Actor) and *LOC* (Location), accounting for more than 70% of all the assigned functors[1]. The counts for the Actor functor should therefore be incremented by the number of occurrences of the Patient functor in Table 2. Additionaly, due to the FGD formalism, every argument following the Actor in two- or three-argument frames is implied to be labeled as Patient regardless of its cognitive content.

Functor	Count	Percent
PAT[1]	773	36.07
ACT	637	29.72
LOC	128	5.97
TWHEN	115	5.37
MANN	114	5.32
ADDR	43	2.01
CAUS	35	1.63
MEANS	26	1.21
DIR3	24	1.12
CRIT	23	1.07
AIM	22	1.03
THO	22	1.03
Other	181	8.45

Table 2. Functor frequency

[1] It should be noted that the overall number of wordforms annotated as Patient (PAT) should not in any case be larger than the number for Actor (ACT); the Actor is thus implied by the Patient within all the frames, even though it may not explicitly occur.

In Table 3, the actual frames – sequences of tectogrammatical functors occurring with a verb – are counted. In this presentation form, we do not display the frames as attached to specific verbs, as e.g. in Figure 3. Rather, we simply display the frequencies of the frame types independently. The table indicates that the Actor-Patient (*ACT PAT*) frame is the most frequent one, once again taking into account the emphasized note regarding the Patient functor and frame (*PAT*) from the previous table[1].

Table 4 represents a key point of our experiment. It is extracted from an obtained distribution of pairs of analytical functions and morphosyntactic tags across the tectogrammatical functors. Basically, for each functor, occurrences of specific ordered pairs (analytical function, morphosyntactic tag) were counted. These occurrence maps were assigned to the functors. The distribution, as illustrated by the table, can be used directly in writing down simple rules for the inference of tectogrammatical functors from wordforms in unseen (but morphosyntactically annotated and dependency-parsed) text. In the table, for purposes of illustration, the distributions are given just for the six most frequent tectogrammatical functors (Actor, Patient and Locative) and ten most frequent pairs of morphosyntactic tags and analytical functions.

Frame	Count	Percent
ACT PAT	250	26.71
PAT[1]	157	16.77
ACT PAT TWHEN	30	3.21
ACT MANN PAT	23	2.46
ACT ADDR PAT	20	2.14
ACT LOC	20	2.14
ACT LOC PAT	20	2.14
MANN PAT	17	1.82
ACT CAUS PAT	16	1.71
ACT MANN	13	1.39
LOC PAT	12	1.28
ADDR PAT	11	1.18
Other	347	37.07

Table 3. Frame frequency

ACT (Actor)			PAT (Patient)			LOC (Locative)		
A-fun	**MSD**	**%**	**A-fun**	**MSD**	**%**	**A-fun**	**MSD**	**%**
Sb	Ncmsn	14.91	Obj	Ncfsa	11.25	(AuxP) Adv	(Spsl) Ncfsl	21.88
Sb	Np-sn	13.50	Obj	Ncmsa	9.18	(AuxP) Adv	(Spsl) Ncmsl	16.41
Sb	Ncfsn	12.87	Pnom	Ncmsn	5.69	(AuxP) Adv	(Spsl) Npmsl	10.16
Sb	Ncmpn	9.89	Obj	Ncmpa	4.53	(AuxP) Adv	(Spsl) Ncnsl	8.59
Sb	Npfsn	5.65	Obj	Vmn*	4.40	(AuxP) Adv	(Spsl) Npfsl	8.59
Sb	Pi-mpn--n-a--	4.71	Obj	Ncnsa	3.75	(AuxP) Adv	(Spsl) Ncmpl	5.47
Sb	Ncfpn	3.30	Obj	Ncfpa	3.49	(AuxP) Adv	(Spsl) Ncfpl	3.91
Sb	Ncnsn	2.98	Pnom	Ncfsn	2.72	Adv	Rl	3.13
Sb	Pi-msn--n-a--	2.51	(AuxC) Obj	(Css) Vmip3s	2.07	Adv	Css	1.56
Sb	Pi-fsn--n-a--	1.88	Obj	Ncmsn	1.81	(AuxP) Adv	(Spsg)Ncmsg	1.56
TWHEN (Temporal when)			**MANN (Manner)**			**ADDR (Addressee)**		
A-fun	**MSD**	**%**	**A-fun**	**MSD**	**%**	**A-fun**	**MSD**	**%**
Adv	Rt	30.43	Adv	Rnp	40.35	Obj	Ncfsd	13.95
(AuxP) Adv	(Spsl) Ncmsl	12.17	Adv	Rn	20.18	Obj	Ncmpd	9.30
Adv	Ncfsg	5.22	Adv	Css	5.26	Obj	Pp3msd--y-n--	9.30
(AuxP) Adv	(Spsg) Ncmsg	5.22	Adv	Rt	3.51	Obj	Ncmsd	6.98
Adv	Ncmpg	4.35	(AuxP) Adv	(Spsi) Ncnsi	3.51	Obj	Ncnsd	6.98
Adv	Ncfsi	3.48	(AuxP) Adv	(Spsl) Ncfsl	3.51	(AuxP) Adv	(Spsi) Ncnsi	4.65
Adv	Ncmsi	3.48	Adv	Rk	1.75	Obj	Np-sd	4.65
(AuxP) Adv	(Spsl) Ncfsl	3.48	Adv	Rnc	1.75	(AuxP) Adv	(Spsa) Ncmsa--n	2.33
Adv	Ncmsg	2.61	(AuxP) Adv	(Spsi) Ncmsi	1.75	(AuxP) Adv	(Spsa) Ncmsg	2.33
(AuxP)Adv	(Spsl) Ncnsl	2.61	Adv	Afpmsn-	0.88	(AuxP) Adv	(Spsa) Px--sa--npn--	2.33

Table 4. Distribution of (analytical function, MSD) pairs for the most frequent functors

A simple example of utilizing the data in Table 4 for the inference of functors in unseen, but preprocessed text would be the one for assigning the Actor (*ACT*) functor. Namely, if a wordform (1) annotated as a noun in the nominative case (*N…n*) and (2) with an assigned syntactic function of subject (*Sb*) is encountered, the Actor functor should also be assigned to it. Such rules could also assign confidence measures to the outputted functors; these measures could be based e.g. on the occurrence percentages given in Table 4. Once again taking the Actor functor as an example, the confidence of assigning this functor to a {subject, nominative} noun would be at least 60 percent, a number derived by adding the percentages of {subject, nominative} entries in the table.

5. Conclusions and future work

In this experiment, we have designed and implemented one possible approach to semi-automatic extraction of a valency frame lexicon for Croatian verbs and also to the refinement of existing lexicons by using the Croatian Dependency Treebank as an underlying resource. We have automatically extracted 2930 verb valency frame instances and annotated 936 frames with tectogrammatical functors selected from the FGD formalism. We analyzed these annotations and provided two important results: (1) the distribution of valency frames for each of the encountered verbs and (2) the distribution of analytical functions and morphosyntactic tags for each of the tectogrammatical functors. The first result directly enables the enrichment of existing valency lexicons, such as CROVALLEX (Mikelić Preradović et al. 2009), while the second result enables the implementation of a rule-based system for automatic assignment of tectogrammatical functors to morphosyntactically tagged and dependency-parsed unseen text.

We divide our future research plans in the track of valency frame extraction into several directions. In the first one we will try (a) to implement and evaluate the previously mentioned rule-based system for assigning semantic roles to wordforms in unseen text and (b) to investigate the possibilities of semi-automatic enrichment of CROVALLEX with the verb valency frames extracted in this experiment. In the second one information on verb valency could also be utilized for the enrichment of Croatian WordNet (Raffaelli et al. 2008), namely by adding the valency frames to the verbs it encodes (cf. Pala & Sedlácek 2005)). Also, as implied in the previous sections, the treebank itself requires both enlargement and enhancements. Extensive efforts are currently underway with respect to these goals.

This procedure of automatic detection of valency frames will be used also in several other projects dealing with factored SMT (e.g. ACCURAT) where valency information will represent one of the layers of additional linguistic annotation that will be taken into account when developing translation models.

We also consider various approaches to dependency parsing of Croatian. Future research plans for this line of research are rather extensive. Regarding dependency parsing of Croatian by using the Croatian Dependency Treebank, we shall undergo various research directions in order to increase overall parsing accuracy.

In the first run we should investigate the performance of all freely available state-of-the-art data-driven dependency parsers. For example, in Table 5, the baseline scores obtained by using the linear-time algorithms of the MaltParser system (Nivre et al. 2007) are presented as an illustration of improvement possibilities with respect to the rather poor accuracy scores obtained in the trial run.

In the second run fine-tuning of all the available parameters for these should be investigated with respect to the specific properties of Croatian. Experiment with combining parsers and different parsing settings along the lines of experiments with the Index Thomisticus treebank (Passarotti & Dell'Orletta 2010) should also be conducted. Specifically, we would like to look into the possibilities of hybridization of the before-mentioned state-of-the-art data-driven parsers by linking them with language specific resources such as valency lexicons, following e.g. (Zeman 2002), being that a valency lexicon of Croatian verbs (CROVALLEX) already exists and the basic idea of verb valency (and valency in general) actually implies and constrains the dependency relations within a sentence. These research paths will be accompanied by a more elaborate investigation into all the different variables, i.e. treebank-encoded properties of Croatian language influencing the various aspects of dependency parsing accuracy.

Metric	Nivre eager	Nivre standard	Stack projective
Labeled attachment (LAS)	58.29±0.67	55.07±0.84	57.58±0.68
Unlabeled attachment (UAS)	67.91±0.59	67.31±0.77	67.49±0.64
Attachment of labels (LA)	70.85±0.45	64.73±0.69	72.36±0.54

Table 5. Baseline dependency parsing scores (MaltParser)

Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347. This work was also supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1002, 130-1300646-1776 and 130-1300646-0645.

References

Agić Ž, Tadić M, Dovedan Z. (2008). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica, 32:4, pp. 445-451.

Agić Ž, Šojat K, Tadić M. (2010). An Experiment in Verb Valency Frame Extraction from Croatian Dependency Treebank. Proceedings of the 32nd International Conference on Information Technology Interfaces, Zagreb, SRCE University Computer Centre, University of Zagreb, 2010. pp. 55-60.

Buchholz S, Marsi E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL), New York, NY, pp. 149-164.

Džeroski S, Erjavec T, Ledinek N, Pajas P, Žabokrtský Z, Žele A. (2006). Towards a Slovene Dependency Treebank. Proceedings of Fifth International Conference on Language Resources and Evaluation, LREC'06, 24-26 May 2006. Genoa.

Erjavec T. (2004). Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Fourth International Conference on Language Resources and Evaluation. ELRA, Lisbon-Paris 2004, pp. 1535-1538.

Hajič J, Böhmová A, Hajičová E, Vidová Hladká B. (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. Treebanks: Building and Using Parsed Corpora, Amsterdam, Kluwer, 2000. See also URL http://ufal.mff.cuni.cz/pdt2.0/

Hajič J, Panevová J, Urešová Z, Bémová A, Pajas P. (2003). PDTVALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. Proceedings of the Second Workshop on Treebanks and Linguistic Theories, Vaxjo University Press, 2003, pp. 57-68.

Kübler S, McDonald R, Nivre J. (2009). Dependency Parsing. Synthesis Lectures on Human Language Technologies, Morgan\&Claypool Publishers, 2009.

Mikelić Preradović N, Boras D, Kišiček S. (2009). CROVALLEX: Croatian Verb Valence Lexicon. Proceedings of the 31st International Conference on Information Technology Interfaces, pp. 533-538. See URL http://cal.ffzg.hr/crovallex/index.html.

Nivre J, Hall J, Nilsson J, Chanev A, Eryigit G, Kübler S, Marinov S, Marsi E. (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. Natural Language Engineering, 13(2), 95-135.

Nivre J, Hall J, Kübler S, McDonald R, Nilsson J, Riedel S, Yuret D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007, Prague, Czech Republic, pp. 915-932.

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

Pala K, Sedlácek R. (2005). Enriching WordNet with Derivational Subnets. Proceedings of CICLing 2005, pp. 305-311.

Passarotti M, Dell'Orletta F. (2010). Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin. Proceedings of the Seventh conference on International Language Resources and Evaluation, ELRA, 2010.

Raffaelli I, Tadić M, Bekavac B, Agić Ž. (2008). Building Croatian WordNet. Proceedings of the 4th Global WordNet Conference, Szeged, Global WordNet Association, 2008, pp. 349-359.

Sgall P, Hajičová E, Panevová J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht, D. Reidel Publishing Company.

Šojat K, Vučković K, Tadić M. (2010). Extracting verb valency frames with Nooj. Finite State Language Engineering: NooJ 2009 International Conference and Workshop, Touzeur, Centre de Publication Universitaire, 2010. pp. 231-241.

Tadić M. (2002). Building the Croatian National Corpus. Proceedings of the 3rd International Conference on Language Resources and Evaluation, ELRA.

Tadić M. (2007). Building the Croatian Dependency Treebank: the initial stages. Suvremena lingvistika, 63, pp. 85-92.

Zeman D. (2002). Can Subcategorization Help a Statistical Dependency Parser? Proceedings of the 19th International Conference on Computational Linguistics.