An Experiment in Verb Valency Frame Extraction from Croatian Dependency Treebank

Željko Agić[1], Krešimir Šojat[2], Marko Tadić[2]
[1]*Department of Information Sciences*
[2]*Department of Linguistics*
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{zeljko.agic, ksojat, marko.tadic}@ffzg.hr

Abstract. *The paper presents an approach to semi-automatic verb valency frame extraction from the Croatian Dependency Treebank. Our algorithm extracted 1923 verb valency frames for 594 different verbs. We discuss applicability of our method to semi-automatic verb valency lexicon creation and refinement, along with possibilities of utilizing it in the task of parsing Croatian texts.*

Keywords. valency frame extraction, valency lexicon, Croatian dependency treebank

1. Introduction

A verb valency lexicon or sub-categorization lexicon is a collection of linguistically annotated data on syntactic and semantic properties of verbs and their arguments in a given language. More specifically, such a lexicon encodes the number and properties of arguments controlled by a verb in a syntactic structure of a sentence. As metaphorically described by Tesnière [16], the concept of verb valency can be compared to the notion of valency numbers in chemistry: as valency in chemistry defines the number of chemical bonds formed by atoms of a given element, verb valency in linguistics determines the number and properties of words (*actants*) that co-occur with a particular verb in a sentence and that appear on the first hierarchical level below the verb in the dependency tree (*stemma*). Information on verb valency finds its usefulness in basic computational linguistic tasks such as chunking and (shallow and deep) parsing, but also in other natural language processing tasks (cf. [4]).

There are two basic approaches to the development of verb valency lexica: (1) manual construction conducted by expert linguists and (2) automatic induction by utilizing existing syntactically annotated corpora, i.e. treebanks, implementing rules for valency frame extraction, possibly followed by manual verification and refinement. The first approach comprises projects like VALLEX [18], a valency lexicon of Czech verbs, based on the Functional Generative Description (FGD) [11] and closely related to the Prague Dependency Treebank (PDT) [3] (although thorough experiments with automatic induction of valency frames on the PDT were also conducted [1, 10]). The emerge-and-refine approach to valency lexicon construction is illustrated by e.g. the TüBa-D/Z dependency treebank and valency lexicon [4] and the Index Thomisticus dependency treebank and valency lexicon of Latin [5]. It should also be noted that, when large quantities of syntactically annotated text in treebanks are available in parallel with manually assembled valency lexicons (e.g. resources available for Czech), emerging valency frames from the treebank may be used as a mechanism for continuing validation and further enrichment of existing lexicons.

As far as the state-of-the art in approaches to valency lexicon creation to Croatian language is concerned, the situation is as follows. There is one existing and publicly available, manually assembled valency lexicon, the CROVALLEX [6]. It encompasses 1739 verbs distributed into 5118 valency frames and 173 syntactic-semantic classes. CROVALLEX is constructed with respect to the FGD in terms of an underlying linguistic formalism and to VALLEX in terms of its structure. To our knowledge, except for the preliminary investigation described in [12], dealing with using local grammars to describe verb valency, no experiments with automatic verb valency frame acquisition from corpora of the Croatian language have been conducted.

Further in the text, we describe results of a preliminary experiment in extracting valency frames for Croatian verbs from the Croatian Dependency Treebank (hr. *Hrvatska ovisnosna*

banka stabala, HOBS further in the text) [15]. Firstly, we present the experiment setup, the current state of development of HOBS and the implemented rule-based strategy of valency frame extraction. Secondly, we analyze the verb valency frames acquired automatically from the treebank, independently and with respect to enriching the existing valency lexicon, i.e. CROVALLEX. Concluding sections of the paper discuss future perspectives of utilizing induced valency information for Croatian verbs with respect to development and refinement of other language resources [9] and natural language processing tools for Croatian.

2. Experiment setup

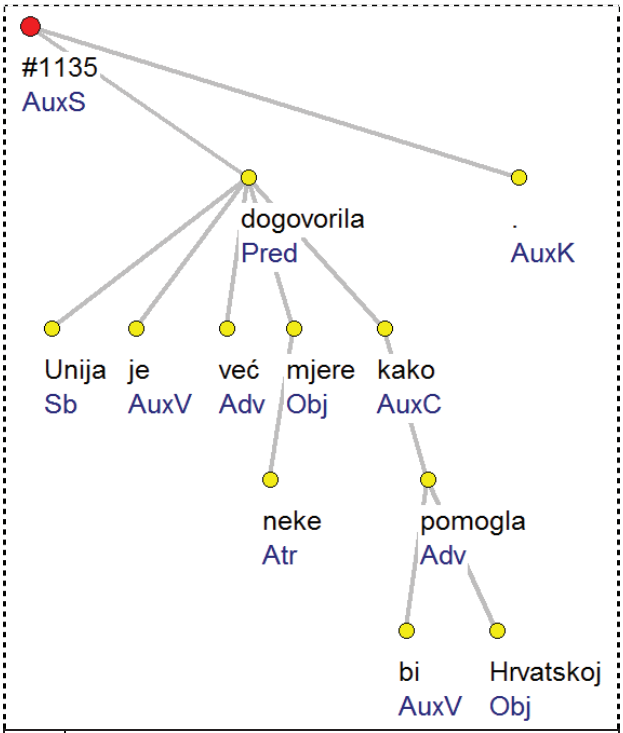
The primary goal of the experiment was to extract a preliminary list of valency frame candidates for Croatian verbs from HOBS and discuss its overall quality and applicability. The secondary goal was to check its appropriateness in validating and extending CROVALLEX. In order to describe the experiment setup, we shall briefly present the current state of HOBS as well as the rules we implemented to search for valency information through the syntactically annotated sentences.

HOBS is a dependency treebank built along the principles of Functional Generative Description (FGD) [11], 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) [3] 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. The tectogrammatical level refers to the semantic interpretation of disambiguated sentences in the form of tectogrammatical trees with nodes labeled with semantic functors that roughly correspond to the notion of semantic or theta roles. The ongoing construction of HOBS closely follows the guidelines set by PDT, with their simultaneous adaptation to the specifics of the Croatian language. More detailed account of the HOBS project is given in [15].

Currently, HOBS consists of 1855 sentences in the form of dependency trees that were manually annotated with syntactic functions using TrEd [8] as the annotation tool. These sentences, encompassing approximately 50.000 tokens, stem from the magazine Croatia Weekly 100 kw (CW100) that is a part of the newspaper sub-corpus of the Croatian National Corpus (HNK) [14]. The Croatia Weekly sub-corpus was previously XCES-encoded, sentence-delimited, tokenized, lemmatized and MSD-annotated by linguists. Thus, each of the analyzed sentences contains the information on part-of-speech, morphosyntactic category, lemma, dependency and analytical function for each of the word forms.

Table 1. Corpus and treebank stats

	Sentences	Tokens
CW100	4626	118529
Treebank	1855	45994
Experiment	1637	45675



HR	*Unija je već dogovorila neke mjere kako bi pomogla Hrvatskoj.*
EN	*The Union has already arranged some measures in order to help Croatia.*

Figure 1. Sample dependency tree from HOBS

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 [2] [7] to choose from a wide selection of different features in the upcoming experiments with stochastic dependency parsing of Croatian texts. Basic stats for the CW100 corpus and HOBS are given in Table 1.

Sentences in HOBS are annotated according to the PDT annotation manual for the analytical level of annotation [3], with respect to differing properties of the Croatian language. The utilized analytical functions are thus compatible with those of PDT, as illustrated by Figure 1. The list of the most frequently used analytical functions is given in Table 2.

Table 2. Distribution of analytical functions

Function	Count	Percent
Atr	11456	25.08
Adv	4612	10.10
AuxP	4284	9.38
Obj	3029	6.63
Sb	2905	6.36
AuxX	2170	4.75
AuxV	2120	4.64
Coord	1808	3.96
AuxK	1731	3.79
Pred	1396	3.06
AuxG	1131	2.48
Atr_Co	1071	2.34
AuxZ	871	1.91
AuxC	848	1.86
Pred_Co	808	1.77
Pnom	695	1.52
Obj_Co	519	1.14
Sb_Co	475	1.04
Other	3746	8.20

Our verb valency frame extraction strategy was loosely based on the one given in [1] for PDT. According to its well-elaborated sentence filtering scheme, defined for the extraction of valency frames only in sentences considered to encode high-quality information, we decided to filter out sentences with token encoding errors and similar issues that emerged during the treebank building procedure. For further processing we converted the treebank from the feature-structure (FS) format of TrEd into the CoNLL-X treebank format [2]. Filtering was implemented after this step. From the initial set of 1.855 sentences and 45.994 word forms, 1.637 sentences and 45.675 word forms were used in the experiment. 218 sentences were filtered out since they contain 319 so-called bad tokens, mainly decimal numbers, slashes and similar word forms that unfortunately did not pass the translation from the FS to CoNLL-X file format.

The set of rules implemented to extract verb valency frames is small and straightforward. Basically, we search through the tree structure for a given sentence, looking for nodes annotated by analytical functions denoting predicates, or namely by Pred (predicate), Pred_Co (predicate in coordinated sentences) and Pred_Pa (predicate in parenthesis). Upon encountering nodes, i.e. word forms annotated as being predicates, we descend further down the dependency structure, retrieving all the nodes that are both (1) directly dependent of the predicate and (2) the nature of their dependency denotes them as subjects (Sb), objects (Obj), adverbs (Adv) or nominal predicates (Pnom). This basic rule is illustrated by Figure 1. Namely, word form *dogovorila* (en. *arranged*) is (a lexical part of) a verbal predicate and thus chosen by the algorithm, which in turn inspects its direct dependents to retrieve a meaningful valency frame. Thus, the algorithm retrieves *Unija* (en. *Union*) as subject (Sb), *već* (en. *already*) as adverb (Adv) and *mjere* (*en. measures*) as object (Obj). Part of speech and morphosyntactic categories are also retrieved for the selected word forms. The resulting frame instance would appear as given by Table 3.

Table 3. Sample frame instance (*dogovoriti*)

Word form	Lemma	MSD	Function
dogovorila	dogovoriti	Vmps-sfa	Pred
Unija	unija	Ncfsn	Sb
već	već	Rt	Adv
mjere	mjera	Ncfpa	Obj

It should be noted that CROVALLEX does not contain any valency frame information for verb *dogovoriti* at the present moment. This clearly illustrates the usefulness of the extraction method. The resulting frame instance provided by the lookup could now be semi-automatically converted into an actual generic verb valency frame for the verb *dogovoriti* (en. *arrange*) according to FGD by an additional post-processing procedure. This procedure would utilize information on morphosyntactic properties of the word form combined with its analytical function, i.e. the nature of dependency toward the verb. However, such a conversion was put beyond the scope of this preliminary experiment and is discussed further in the text. Currently, we focus on extraction and subsequent manual analysis of retrieved verb frames.

However, a discrepancy should be noted between Figure 1 and Table 3, revealing the need for the implementation of another rule in the presented frame extraction algorithm. Namely,

there is a subordinating conjunction (AuxC) governed by the verbal predicate and the frame instance given in Table 3 does not encode it. In the sentence consisting of elements *Unija (...) dogovorila (...) mjere kako bi pomogla Hrvatskoj* the extracted frame instance does not account for the underlined adverbial phrase, i.e. for the reason of the arrangement (*dogovorila*) of measures (*mjere*) by the Union (*Unija*), yielding the incomplete and incorrect extracted frame in respect to the original treebank sentence. For this reason, we implemented another rule in the algorithm: if nodes denoting subordinating conjunctions (AuxC) and prepositions (AuxP) introducing a subordinated sentence or a prepositional phrase with a syntactic function of subject, object or adverb are encountered and directly dependent of the verbal predicate, the algorithm descents one level further downwards and retrieves nodes that are (1) directly dependent of the subordinating conjunction or preposition and (2) the nature of their dependency is denoted as that of a subject, object, adverb or nominal predicate. As far as the example in Figure 1 is concerned, the algorithm would add another row to Table 3, explicitly denoting the relation between the lexical part of the predicate (*dogovorila*), the subordinating conjunction (*kako*) and the adverbial clause (*kako bi pomogla*).

The following chapter presents the results obtained by the described algorithm on HOBS and discusses their correctness and applicability.

3. Experiment results

Taking 1.637 HOBS dependency trees as the input, the algorithm extracted 1923 verb valency frames for 594 different verb lemmas, which indicates there were sentences from which more than one verbal predicate was encountered, e.g. in sentences with coordinated predicates. The ratio of approximately 3.24 frame instances per verb somewhat diminishes to 2.60 if we exclude the 381 instance of the verb *biti* (en. *to be*), which expectedly dominates the distribution of verbs in Croatian sentences. Other stats for the extracted frame instances are given in Table 4.

The results in the table 4, indicating the predominance of prepositional, sentential and pure adverbs in terms of frame elements, indicate that special attention should be paid to this word type. This refers to further sub-classification of the whole class of adverbial elements into smaller, semantically based groups.

Before further discussion of the obtained results, a note should be taken regarding the evaluation of the presented algorithm. Namely, in this preliminary stage of research in automatic valency frame extraction, we find the evaluation in terms of precision and recall to be somewhat less meaningful. A manual assessment of these measures indicates its high precision (ca 95%), while a more elaborate evaluation will be provided in future research dealing with inducing valency frames from frame instances.

When dealing with adverbs in the form of preposition phrases, these groups can be semi-automatically established on the basis of their morpho-semantic features, whereas other groups require either manual annotation or an expansion of the tagset used for syntactic annotation in terms of adding subclasses like *Adv_manner* or *Adv_time*.

Table 4. Extracted frame instance stats

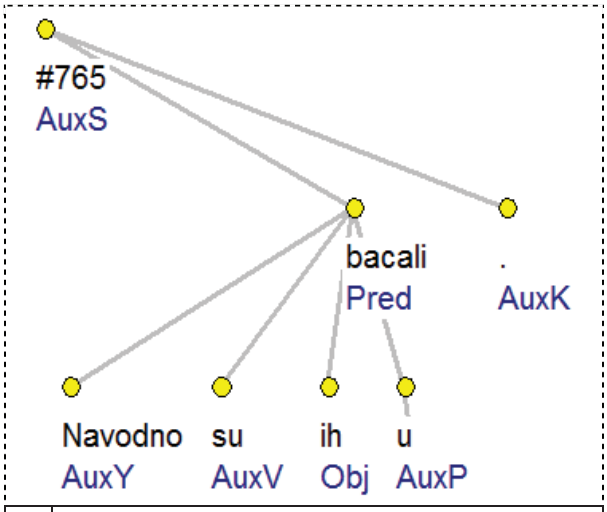
Feature	Count	Percent
Frame instances	1923	100
Pred	1184	61.57
Pred_Co	687	35.73
Pred_Pa	52	2.70
Frame elements	4381	100
Adv	1666	38.03
Obj	982	22.42
Pnom	343	7.82
Sb	1390	31.73
via AuxC or AuxP	1296	100
Adv	1004	77.47
Obj	190	14.67
Pnom	44	3.39
Sb	58	4.47

The second line of the future work should move towards establishing the correlation of a number of frame elements and their semantic tagging used in FGD. Since the arguments (so called *inner participants*) of monovalent and bivalent verbs in FGD are always tagged as Actor and Actor and Patient respectively, the procedure of automatic semantic tagging of frame elements can be relatively straightforward at least in this respect. For the automatic semantic tagging of other three possible inner participants (Addresee, Origin and Effect) rules can be written on the basis of their morpho-syntactic features (morphological cases, types of PPs etc) and on the basis of the number of frame elements annotated as objects.

4. Discussion and future work

Regardless of the relatively small number of processed sentences and thereby determined verb frames in comparison to similar efforts (cf. [1]), the conducted experiment can be used in various aspects. Lexicons such as CROVALLEX can benefit from its results in several ways.

Firstly, their volume can be substantially enlarged in terms of verbs that are currently not contained, but are automatically recognized and lemmatized in sentences from HOBS. Secondly, the description of verb valency is simplified since the method presented above provides not only verb lemmas, but also verb frames determined in annotated corpora. Since the CROVALLEX was compiled manually many non-optional frame elements that are tagged as typical are not present in verb valency frames for particular verb senses.



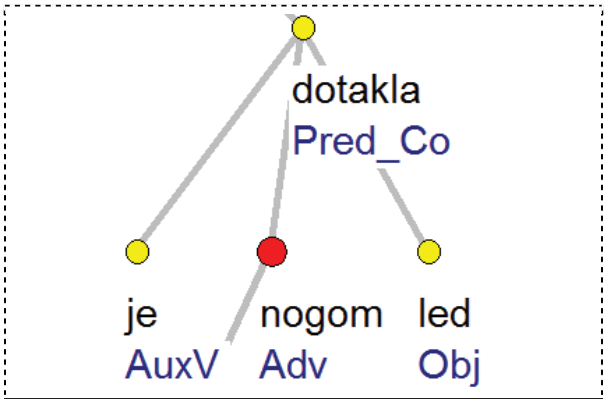
HR	*Navodno su ih bacali u jame …*
EN	*They allegedly threw them into pits …*

3 bacati (bàcati)$_3$ ≈ **zamahom ruke ili ruku udaljivati od sebe; hitati**
-frame: $AGT^{obl}_{0\_or\_1}$ PAT^{obl}_{4} $DIR3^{typ}_{prema+6}$
-example: Dio prosvjednika je komade asfalta bacao prema policajcima
-class: throw/catch

Figure 2. Treebank evidence and CROVALLEX frame for the verb *bacati* (en. *to throw*)

An example is given in Figure 2, where only one type of prepositional phrase (*prema* (en. *towards*) + locative) is given as typical for the given sense of the verb *baciti* (en. *to throw*), but not one given in the figure 2 (*u* (en. *into*) + accusative) and as well as the other others in the automatically extracted verb valency frames.

Further, verb frames determined in the presented experiment can be used for the consistency and accuracy checking of the annotation in frames in already existing lexicons. Figure 3 shows two frames from CROVALLEX for the illustration of valency pattern in corresponding senses of the Croatian verb *dotaknuti* (en. *to touch*). The frame 1 consists of functors AGT + INST defined as *dodirnuti se međusobno* (en. *to touch each other*). The frame 3 consists of functors AGT + PAT defined as *tičući doći u doticaj s čim* (en. *to touch something*). Since we do not feel the semantic difference in these two senses, and since we are convinced that the reflexive *se* in the frame 1 should be interpreted as PAT, lexicographers and researchers can benefit from the closer inspection of annotated trees and extracted verb frames as presented in the figure 3. On the basis of this example as well the others from HOBS we conclude that the frame AGT + PAT + INST would be sufficient and more accurate for the given verb in this sense.



HR	*… ali je nogom dotakla led …*
EN	*… but she touched the ice with her leg …*

1 dotaknuti (dotàknuti)$_1$ ≈ **dodirnuti se međusobno**
-frame: $AGT^{obl}_{0\_or\_1}$ $INST^{typ}_{7}$
-example: Dotaknuli su se rukom
-class: touch

3 dotaknuti (dotàknuti)$_3$ ≈ **tičući doći u doticaj s čim; dodirnuti**
-frame: $AGT^{obl}_{0\_or\_1}$ PAT^{obl}_{4}
-example: Mađari nisu dotaknuli loptu
-class: touch

Figure 3. Treebank evidence and CROVALLEX frames for the verb *dodirnuti* (en. *to touch*)

Detected verb frames can be further used for the elaboration of verb senses and verb usage, especially when dealing with phrasal usage and

idioms, e.g. *ostaviti dojam na nekoga* (en. *to make an impression*), etc.

Finally, useful applications of automatic acquisition of verb valency information go beyond simply creating and enriching lexical resources themselves. Information on verb valency from large coverage verb valency lexicons is shown to be useful in both rule-based and stochastic parsing (cf. [17]). While utilizing verb valency information in rule-based chunking and parsing is, in theory, a rather straightforward procedure, it would be interesting to conduct experiments in using valency frames of Croatian verbs – or valency frames of other languages that share properties with Croatian – as features for training procedures of current state-of-the-art dependency parsers (cf. [2] [7]).

5. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776, 130-1300646-1002 and 130-1300646-0645.

6. References

[1] Bojar O. (2003). Towards Automatic Extraction of Verb Frames. Prague Bulletin of Mathematical Linguistics, 79-80.

[2] 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.

[3] 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.

[4] Hinrichs E, Telljohann H. (2009). Constructing a Valence Lexicon for a Treebank of German. Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories, LOT, Utrecht, Netherlands, pp. 41-52.

[5] McGillivray B, Passarotti M. (2009). The Development of the Index Thomisticus Treebank Valency Lexicon. Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education, pp. 43-50.

[6] 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.

[7] 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.

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

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

[10] Sarkar A, Zeman D. (2000). Automatic Extraction of Subcategorization Frames for Czech. Proceedings of the 18th International Conference on Computational Linguistics, volume 2, pp. 691-697.

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

[12] Šojat K, Vučković K, Tadić M. (2009). Extracting Verb Valency Frames with NooJ. Proceedings of the 2009 International NooJ Conference, in press.

[13] Šojat K. (2008). Sintaktički i semantički opis glagolskih valencija u hrvatskom. PhD thesis, Faculty of Humanities and Social Sciences, University of Zagreb, 2008.

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

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

[16] Tesnière L. (1959). Éléments de syntaxe structurale. Klincksieck, Paris 1959.

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

[18] Žabokrtský Z, Lopatková M. (2007). Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. The Prague Bulletin of Mathematical Linguistics, 87.