Extracting Verb Valency Frames with NooJ

Krešimir Šojat
Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
ksojat@ffzg.hr

Kristina Vučković
Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
kvuckovi@ffzg.hr

Marko Tadić
Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

The paper discusses the possibilities of describing verb valency on the basis of local grammars developed in the NooJ format. In this paper we use NooJ to describe the semantic and syntactic valency of app. 120 Croatian verbs belonging to the semantic field of consumption (e.g. *eat, drink, devour, imbibe* etc.). The whole semantic field, consisting of verbal lexical units is viewed as a single semantic frame. The approach relies on the theoretical background of frame semantics used in the development of the FrameNet (Baker et al., 1998; Ruppenhoffer et al., 2006). In such an approach the semantic valency of lexical units is described in terms of core (central) and non-core (peripheral) elements characteristic for the whole frame.

The level of syntax is observed as a level of realization or non-realization of conceptual arguments. As a starting point we use app. 40 sentence types consisting of morphosyntactic combinations possible in Croatian (e.g. *They thought them mathematics* – Nom (nominative) – Acc (accusative) – Acc (accusative)). For each sentence type a local grammar is built, with free word order taken into consideration. At the same time every verb is additionally described in the Croatian NooJ dictionary. Each local grammar is applied to a corpus, and each occurrence or non-occurrence of lexical units in morphosyntactically annotated sentence type is analyzed.

The obtained results show that certain verbs, although in terms of semantic valency can intuitively be described as two argument verbs, are exclusively realized as one argument verbs on the syntactic level. Further results show the importance of non-core frame elements (e.g. means, company) for certain lexical units.

The obtained results are further used for the refinement of verb frames in existing and future verb valency lexica of Croatian verbs.

1 Introduction

Our main agenda is to describe the valency frames of Croatian verbs of consumption as fully as possible. This will allow us to search for non-core (peripheral) elements such as time, place, manner, company, instrument, cause and other in the verb's co-text. In order to do this, we are using core verb valency frames description and then checking the co-text window of 4 phrases[1] that proceed and follow the main verb. The data obtained are used for improving grammars for syntactic and semantic verb co-text recognition.

We start in the Section 2 with the explanation of the theoretical background used in this approach. The Section 3 follows with the description of Croatian verbs of consumption valency main characteristics and the description of data

[1] Chunks have alredy been labeled in the text so the term 'phrase' covers VP, PP and NP chunks.

in our lexicon. Then in Sections 4 and 5 we explain in more detail the syntactic grammars used for detecting verb's co-text. Finally, we conclude with the description of data obtained in the extracted frames and possible future directions.

2 Semantic Frame of Consumption (Ingestion) Verbs

The Berkley FrameNet project is an on-line lexical resource for English based on scenes-and-frames semantics (Fillmore, 1977a; 1977b) and supported by corpus evidence. Ruppenhofer et al. (2006:5) point out that the project's "aim is to document the range of semantic and syntactic combinatory possibilities – valences – of each word in each of its senses […]." The FrameNet lexical database contains app. 10 000 lexical units in nearly 800 hierarchically-related semantic frames.

The lexical unit is defined as a pairing of a word with a meaning, i.e. each sense of a (potentially polysemous) word belongs to a different semantic frame. A semantic frame is conceived as "a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props." (ibid., 2006:5)

Fillmore and Atkins (1994:370) stress that the "frame semantics [...] begins with the effort to discover and describe the conceptual framework underlying the meaning of the word, and ends with an explanation of the relationship between elements of the conceptual frame and their realizations within the linguistic structures that are grammatically built up around the word." Each semantic frame in the FrameNet contains the description of a typical situation or event, lexical units that belong to this frame and typical or expected participants in this event and the circumstances in which the whole event occurs. The commonality or prototypycality of participants is conceived in terms of Fillmore's (1977b) scenes or Schank and Abelson's (1977) scripts. In other words, each frame represents a typical event with typical participants (core or central frame elements) and typical circumstances (non-core or peripheral frame elements).

The sentences from the corpora are annotated on three levels: frame element (semantic role), a grammatical function (e.g. subject or object) and a phrase type (e.g. NP or PP). Ruppenhofer et al. (2006:26) define a core frame element as the "one that instantiates a conceptually necessary component of a frame, while making the frame unique and different from other frames."

On the other hand, "frame elements that do not introduce additional, independent or distinct events from the main reported event are characterized as peripheral. Peripheral FEs [i.e. frame elements] mark such notions as TIME, PLACE, MANNER, MEANS, DEGREE and the like." (ibid., 2006:27). This does not mean that certain frame elements classified as peripheral in one frame cannot be classified as central in other.

An element is classified as central even in cases when it does not appear in a sentence, but it is conceived as present on a conceptual level. The semantic interpretation of a non-appearing or missing element on the level of syntax can be definite (*definite null instantiation*) or indefinite (*indefinite null instantiation*). [2] The indefinite cases are illustrated by the missing objects of verbs like *eat, drink, sew, bake* etc., when these transitive verbs are used in intransitive (monovalent) constructions.

Verbs like *eat* and *drink* belong to the semantic frame *Ingestion* defined as: "An Ingestor consumes food, drink, or smoke (Ingestibles), which entails putting the Ingestibles in the mouth and taking them further into the body to be absorbed. This may include the use of an Instrument."

The central frame elements are *Ingestor* and *Ingestibles*. The *Ingestor* is defined as the person eating, drinking or smoking, and the *Ingestibles* as the entities that are being consumed by the Ingestor). Periperal frame elements of the semantic frame *Ingestion* in the FrameNet are *Degree, Duration*, *Instrument*, *Manner*, *Means*, *Place*, *Purpose*, *Source*, *Time*.

For the term *Ingestor* we further use the term *Consumer*, and for *Ingestibles* the term *Consumed*.

3 Lexicon

Croatian NooJ lexicon now has 1960 verbs with the lexical information as described in (Vučković et al. 2008). Of that, 102 are verbs of consumption with the following distribution:

- 12 verbs require only consumer (nominative case),

[2] Ruppenhofer i dr. (2006:33): "Sometimes FEs that are conceptually salient do not show up as lexical or phrasal material in the sentence chosen for annotation. [...] The FE that has been identified indicates which semantic role the missing element would fill, if it were present."

- 3 verbs require consumer (nominative case) and what is being consumed (genitive case),
- 34 verbs require consumer (nominative case) and what is being consumed (accusative case),
- 2 verbs require consumer (nominative case) and what is being consumed (instrumental case),
- 14 verbs require either only consumer or consumer and what is being consumed (genitive case),
- 28 verbs require either only consumer or consumer and what is being consumed (accusative case),
- 5 verbs require either only consumer or consumer and what is being consumed (instrumental case),
- 3 verbs require either consumer and what is being consumed (genitive case) or consumer and what is being consumed (accusative case),
- 1 verb requires either consumer and what is being consumed (accusative case) or consumer and what is being consumed (instrumental case).

All the verbs of consumption in our lexicon have been added the '**cons**' label to show they belong to the semantic field of consumption in general. Additional labels for core arguments are added to them in the following manner:

- <+cons1> if the verb needs a consumer (in nominative case)

Ja jedem.
(***I*** *am eating.*)

- <+cons12> if the verb needs a consumer and what is being consumed (in genitive case)

Ona se najela **gljiva**.
(She has stuffed herself with mushrooms.)

- <+cons14> if the verb needs a consumer and what is being consumed (in accusative case)

Ja jedem **ribu**.
(*I am eating* ***fish***.)

- <+cons17> if the verb needs a consumer and what is being consumed (in instrumental case)

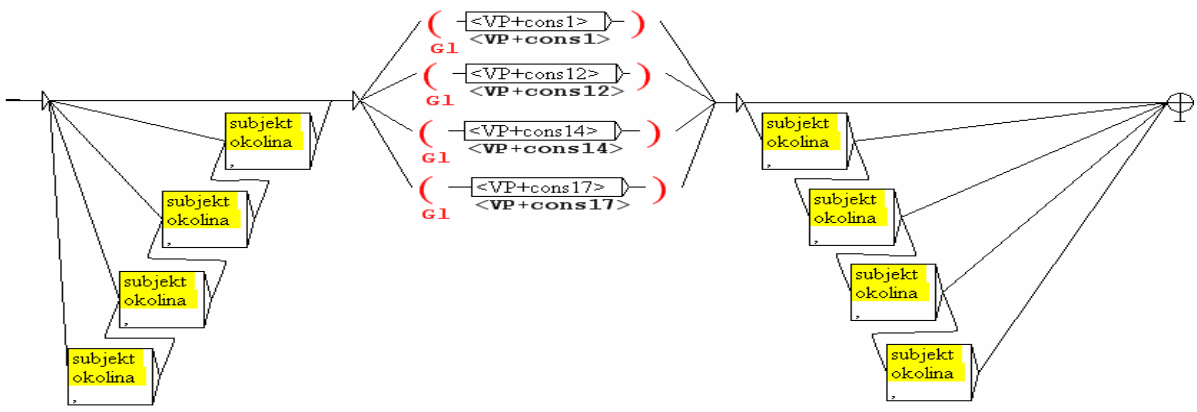
Oni se hrane **kukuruzom**.
(*They are feeding* ***on corn***.)

Consumer in all these cases may or may not be mentioned i.e. it is optional in all the descriptions since it can be understood from the verb form. Thus, the full description of our verbs of consumption in the lexicon looks like this:

jesti,V+FLX=JESTI+Prelaz=pov
+cons+cons1+cons14

meaning that the verb 'jesti' (to eat) is a reflexive verb (+Prelaz=pov) of consumption (+cons) with the two possible co-texts. One is with only a consumer (that may or may not be mentioned in the sentence) and the other one is with a consumer in nominative case and something being consumed in accusative case.

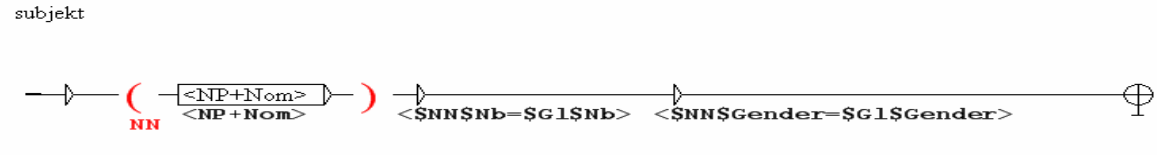
Picture 1*: the main graph for detecting verb's co-text*



4 Syntactic Grammar for Detecting Verb's Co-Text

Special grammar for detecting verb's co-text was build. The main graph (see ***Picture 1***) uses 2 subgraphs to detect if there is a comma <,>, a subject <*subjekt graph*> or something else <*okolina graph*> that proceeds and/or follows the main verb.

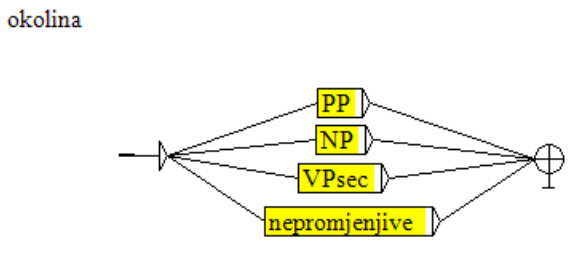
***Picture 2**: the subgraph for subject*



The subgraph for detecting subject checks if the subject and the main verb agree in gender and number (see ***Picture 2***).

All remaining options are described in the next subgraph named <okolina> (see ***Picture 3***).

***Picture 3**: the subgraph describing all remaining options*



This subgraph has four subsubgraphs where <NP> subsubgraph checks for all types of noun phrases, <PP> subsubgraph checks for all types of preposition phrases, <VPsec> subsubgraph checks for all types of verb phrases and <nepromjenjive> checks for all other nonflective word classes like adverbs and conjunctions.

5 Extracting Frames

After applying our grammar to the text, we export the data into an xml file and observe it as if in a table with 4 places preceding the main verb and 4 places following it. The data for our sample sentence is given in Table 1.

Kao **i većina drugih, ta obitelj nikad ne jede u Branimirovoj već hranu nosi** kući.

(Like many others, that family never eats in Branimirova street but carries their food home.)

-4	-3	-2	-1
i	većina drugih	ta obitelj	nikad
<C>	<NP+Nom>	<NP+Nom>	<R>

0
ne jede
<VP+cons1>

+1	+2	+3	+4
u Branimirovoj	već	hranu	nosi
<PP+L>	<C>	<NP+Acc>	<VP>

Table 1

The verb is always in the position 0 while the words that proceed it have the **-** prefix and the words that follow it have the **+** prefix.

The following two sentences give us the examples of possible problems on deciding what role in the sentence to give to <PP+G> construction.

A: Ona se tako hrani poradi svoga siromaštva što ga ne smije otkriti kćeri. (Table 2)

(She feeds herself in such a manner due to her powerty that she must not disclose to her daughter.)

-4	-3	-2	-1
	ona	se	tako
	<NP+Nom>	<VP>	<R>

0
hrani
<VP+cons1>

+1 poradi svoga siromaštva	+2 što	+3 ga	+4 ne smije otkriti
<PP+G>	<PRO>	<NP+Acc>	<VP>

Table 2

B: Prije početka susreta jeli su kroasane i voće i pili voćne sokove. (Table 3)

(Before the beginning of the meeting they ate croassans and fruit and drank fruit juices.)

-4	-3	-2	**-1 Prije početka susreta**
			<PP+G>

0 jeli su
<VP+cons14>

+1 kroasane i voće	+2 i	+3 pili	+4 voćne sokove
<NP+Acc>	<C>	<VP+cons14>	<NP+Acc>

Table 3

Both sentences have a prepositional phrase in genitive <PP+G> as a complement (the fields are shaded in gray). However, these two phrases, although in the same gender, only appear to serve the same role in the sentence. The first <PP+G> is followed with a pronoun of question and then with some other word forms (in this case <NP+Acc> and <VP>). Such combination is marked as an adverbial of cause <ADV+cause> in a sentence built of a prepositional phrase in genitive and additional attribute of that prepositional phrase.

On the other hand, the second <PP+G> is actually an adverbial of time in a sentence since its preposition is marked as a preposition of time in the lexicon.

Thus, our Table 2 can be remarked as Table 4 and Table 3 as Table 5.

-3 Ona	**-2 se**	-1 tako	**0 hrani**	**+1 poradi svoga siromaštva što ga ne smije otkriti kćeri.**
<NP+ CONSMER>	**<VP>**	<ADV +manner	**<VP +cons1>**	**<ADV +cause>**

Table 4

-1 prije početka susreta	**0 jeli su**	**+1 kroasane i voće**	+2 i	+3 pili	+4 voćne sokove
<ADV +time>	**<VP +cons14>**	**<NP+ CONSUMED>**	<C>	<VP +cons14>	<NP+Acc>

Table 5

6 Conclusion and Future Work

The peripheral frame elements play an important role in the sentence semantics when dealing with the verbs of consumption. As shown in the examples above, verbs acquire additional senses in sentences where the core elements are omitted or are not emphasized through semantic links with peripheral frame elements. This pertains particularly to peripheral elements such as PLACE, INSTRUMENT or COMPANY, e.g.:

Every day he lunches at the best restaurant in the city [PLACE].

He eats only with his hands [INSTRUMENT] and never uses a fork and a knife.

On Monday and Friday he lunches with the chairman of the board [COMPANY].

Our plan for future work can be divided into three separate stages. The first stage will include building local grammars for recognizing syntactic verb valency frames including the full morphosyntactic description of all phrases and not only PP chunks. The second stage includes grammars for recognizing semantic verb valency frames that will include both core and peripheral frame elements. In the third stage we will check if described syntactic and semantic frames can be copied into other semantic fields. If they prove to be reusable, this will enable us to describe verbs of other semantic fields much faster and this will lead us to improved development of a parser for Croatian sentences.

Acknowledgments

This work was done within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants 130-1300646-1776 and 130-1300646-1002.

References

Steven Abney. 1996. *Partial Parsing via Finite-State Cascades*, Journal of Natural Language Engineering 2(4):337-344.

C. F. Baker, Ch. J. Fillmore, J. B. Lowe. 1998. *The Berkley FrameNet Project*, online: http://framenet.icsi.berkley.edu

Ch. J. Fillmore, N. T. S. Atkins. 1994. *Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography,* in Computational Approaches to the Lexicon (eds. B.T.S. Atkins, A. Zampolli), Oxford UP, Oxford, 449-377.

Ch. J. Fillmore. 1977a. *The Case for Case Reopened*, in Syntax and Semantics 8. Grammatical Relations, (eds. P. Cole, J.M. Saddock) Academic Press, New York, 59-81.

Ch. J. Fillmore. 1977b. *Scenes-and-Frames Semantics,* in Linguistic Structures Processing, (ed. A. Zampolli) North-Holland Publishing Company, Amsterdam / New York / Oxford, 55-81.

Maurice Gross. 1993. *Local grammars and their representation by finite automata*, Data Description - discourse (ed. M. Hoey), Harper-Collins, London, 26-38.

J. Ruppenhoffer, M. Ellsworth, M. Petruck, Ch. Johnson, J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*, online: http://framenet.icsi.berkley.edu

R. C. Schank, R. P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*, Lawrence Erlbaum Associates, New Jersey.

Max Silberztein. 2003. *NooJ Manual,* online: http://www.nooj4nlp.net (200 pages).

Krešimir Šojat. 2008. *Sintaktički i semantički opis valentnosti hrvatskih glagola (Syntactic and Semantic Description of Valencies of Croatian Verbs)*, PhD dessertation, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Kristina Vučković. 2009. *Model parsera za hrvatski jezik (Parser Model for Croatian Language)*, PhD dissertation, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.

Kristina Vučković, Nives Mikelić Preradović, Zdravko Dovedan. 2008. *Verb Valency Enhanced Croatian Lexicon*, NooJ 2008 Conference, Budapest, (in print).

Kristina Vučković, Marko Tadić, Zdravko Dovedan. 2008. *Rule Based Chunker for Croatian*, in Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC'08, (eds. N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, D. Tapias) Marrakech, ELRA, pp. 2544-2549.