CROVALLEX: Croatian Verb Valence Lexicon

Nives Mikelic Preradovic, Damir Boras, Sanja Kisicek
Research Assistant, Full Professor, Research Assistant
University of Zagreb, Faculty of Humanities and Social Sciences,, Department of Information Sciences, I. Lucica 3
nmikelic@ffzg.hr, dboras@ffzg.hr, smatic@ffzg.hr

Abstract. *The goal of this paper is to present CROVALLEX – the first Croatian Verb Valence Lexicon. It contains 1739 verbs with 5118 valence frames. It also contains 173 syntactic-semantic classes (72 classes with two further levels of subdivision). Functional Generative Description (FGD) is used as the background theory for the description of valence frames. Syntactic-semantic classes have been derived from VerbNet and modified for Croatian language. Lexicon has the robust and explicitly defined XML data structure which enables browsing through designated levels. CROVALLEX is available at http://cal.ffzg.hr/crovallex/index.html.*

Keywords. Verb valence lexicon, deep cases, syntactic-semantic classes, valence frame.

1. Introduction

The Croatian Valence Lexicon of Verbs, version 2.0008 (CROVALLEX 2.0008) is the first Croatian verb lexicon that contains valence frames of Croatian verbs. Valence is the ability of the language units to show their combinatory potential in language utterance. One of the key goals of the natural language processing is to make valence information available to both human and computer. Since there is no way to automatically predict valence in language, there is a huge need to create machine and human readable lexicons where valence information will be stored. A verb valence lexicon is crucial for many Natural Language Processing (NLP) tasks, such as lemmatization, tagging, syntactic analysis, word sense disambiguation and machine translation. However, before CROVALLEX there was no publicly available high-quality machine-readable lexicon of Croatian verbs. Therefore, the primary goal of CROVALLEX was to build such a lexicon and make it available to other researchers.

Valence theory developed by Czech linguists Petr Sgall and his collaborators as the part of the Functional Generative Description (FGD) is used as the background theory in CROVALLEX for the description of valence frames of selected verbs [2].

Based on the studies of Tesniere [8] and Fillmore [1], valence theory in Czech syntax represents a layered approach to the natural language processing founded on dependency (property of the verb to bind a certain number of syntactic positions to itself) and two-level syntactic markup (where analytical layer includes the surface syntax representation and tectogrammatical layer represents the internal structure). FGD takes into account both syntactic and semantic criteria since it ranks the first and second verb complement with regard to the syntactic behavior of the complements, while the other complements are being ranked with regard to their semantics.

Before CROVALLEX, valence theory in Croatian language was studied only as a part of the project "Contrastive Analysis of English and Serbo-Croatian" led by R. Filipovic. M. Samardzija, participating in that project, established 10 classes of morpho-syntactic verb complements [6] (regarding the case or word class of the complement). He also divided verbs into four valence classes: non-valent verbs (e.g. *grmjeti, kišiti*), single-valent verbs (e.g. *bdjeti, spavati*), double-valent verbs (e.g. *postati, ubrzati*) and triple-valent verbs (e.g. *pitati, staviti*). Shortly after CROVALLEX was published[1], another work on verb valence theory for Croatian (concentrating mainly on verbs of eating and drinking) was published [7].

CROVALLEX – valence lexicon of Croatian language verbs, apart from being based on FGD, formally describes all verb complements that this author suggests but also introduces a few new

[1] Crovallex was published on CD in January 2008. It became online available in May 2008.

complements. These new complements are result of the analysis study of the verb use in the Croatian National Corpus.

2. About the CROVALLEX lexicon

Since CROVALLEX contains 1739 verbs with 5118 valence frames (which makes an average of 3 valence frames per verb). It also contains 173 syntactic-semantic classes (more accurately, 72 classes with two further levels of subdivision). Those classes have been derived from VerbNet (a verb lexicon based on Levin's verb classes which also provides selectional restrictions attached to semantic roles) and specially refined and modified for Croatian language. The motivation for introducing such semantic classification in CROVALLEX 2.008 was the fact that it simplifies systematic checking of consistency and allows for making more general observations about the data.

Those 1739 verbs were selected from the Croatian frequency dictionary [5], according to their number of occurrences. CROVALLEX v 2.008 lexicon is available at the following website:

http://cal.ffzg.hr/crovallex/index.html.

It is important to mention that out of 39624 word entries in Croatian frequency dictionary, 9500 word entries are verbs. Regarding the verb frequency, the number of verbs that have frequency higher than 11 is equal to 1739, while the number of verbs with frequency higher than 1 equals to 6149. Therefore, valence lexicon consisting of 1739 most frequent verbs should provide a good verb coverage.

The idea was to create a lexicon that will contain as much syntactosemantic information as possible in order to use these information in natural language processing. The main goal was to design the valence lexicon of verbs, nouns and adjectives, but the current version of a lexicon contains only valence information of the verbs due to the time constraints.

3. The structure of the valence lexicon

On the topmost level, CROVALLEX 2.008 is divided into word entries. The content of a word entry corresponds to the traditional term of lexeme. Each word entry relates to one or more headword lemmas.

The word entry consists of a sequence of frame entries relevant for the lemma, where each frame entry corresponds to one of the lemma's meanings. Information about the aspect of the lemma is assigned to each word entry as a whole. Verb lemma represents the infinitive form of the verb, which is in case of lexical homonyms and homographs followed by a Roman number in superscript.

Lexical homonyms are groups of two lemmas which have the same spelling and wordform, but considerably differ in their meanings. They also might differ as to their etymology (e.g. *hȉtati*^{I} - *rush vs. hȉtati*^{II} - *throw*), aspect (e.g. *matírati*^{I} *inf. – to make something appear beamless*^{II} *vs. matírati*^{I} *fin.-to defeat*), or conjugated forms *(izvèdem [first person sg.] for ìzvesti*^{I} *– take out vs. izvèzem [first person sg.] for ìzvesti*^{II} *-export)*.

Homographs are groups of two lemmas which have the same wordform, but different accent, and also considerably differ in their meanings. They also might differ as to their etymology (e.g. *ìskapati*^{I} *– leak out drop by drop vs. iskápati*^{II} *- excavate*), aspect (e.g. *ìsplakati*^{I} *fin.-cry one's eyes out vs. isplákati*^{II} *inf.-rinse*), or conjugated forms (*napadnem [first person sg.] for nàpasti*^{I} *- attack vs. napasem [first person sg.] for nàpāsti*^{II} *-graze)*.

Reflexive particle *se* is part of the infinitive only if the verb is derived reflexive (e.g. *vratiti se*) or reflexiva tantum (e.g. *penjati se*).

The primary and the most frequent meanings are listed first, whereas rare and idiomatic meanings are listed last (e.g. the primary meaning for the verb absorb is "to suck up or drink" with example: "a sponge absorbs water", while the idiomatic meaning would be "to take up mentally" with example: "his listeners absorb rubbish". Each frame entry contains a description of the valence frame itself and of the frame attributes. In the beginning, CROVALLEX was designed as a relational database that was later converted to XML format. It has the robust and explicitly defined data structure which enables browsing through designated levels. Apart from XML version, we also automatically built HTML version of the lexicon with Perl code that extracted all XML elements and their attributes and generated the list of interconnected html pages for each verb and its valence frames.

3.1. The surface structures in CROVALLEX

In CROVALLEX 2.008, a valence frame consists of at least one frame slot, although it is more often a sequence of frame slots. It is

defined as a set of syntactic elements (inner participants) that the specific verb demands or grammatically allows.

Each frame slot corresponds to one complementation of the given verb. The following attributes are assigned to each slot: functor, list of possible morphemic forms and type of complementation. Verb complements are often syntactically realized as noun or pronoun in a specific case, adjective or adverb, but they also can be realized as a prepositional phrase, infinitive or subordinating clause.

Single meaning of a verb requires unique morphemic form for all its obligatory and optional complements. That morphemic form is stored in a lexicon together with the information about their compulsoriness/optionality.

Each frame slot in a sentence can be expressed by a limited set of morphemic means, which are called morphemic forms. In CROVALLEX 2.008, the set of possible forms is defined either explicitly, or implicitly. If the form is defined explicitly, then it gets enumerated in a list attached to the given slot. If the form is defined implicitly, no list is specified, because the set of possible forms is implied by the functor of the respective slot.

The list of forms attached to a frame slot may contain values of the following types:

Pure (prepositionless) case. There are seven morphological cases in Croatian. In the CROVALLEX 2.008 notation, they have traditional numbering: 0-hidden nominative, 1 - nominative, 2 - genitive, 3 - dative, 4 - accusative, 5 - vocative, 6 - locative, and 7 - instrumental.

Prepositional case. Lemma of the preposition and the number of the required morphological case are specified (e.g., od+2, na+4, o+6...).

Subordinating conjunction. Lemma of the conjunction is specified. The following subordinating conjunctions occur in CROVALLEX 2.008: što, zašto, kad, kako_bi, kada, jer, kao_da, nego, nego_da, prije, dok, da, čim, kako.

Infinitive construction. The abbreviation 'inf' stands for infinitive verbal complementation and can appear together with a preposition (e.g. 'nego+inf') and with the morphological case (e.g. 'inf+4').

Construction with adjectives. Abbreviation 'adj-number' stands for an adjective complementation in the given case, e.g. adj-7 ('Osjećam se osvježenim' - 'I feel fresh').

Construction with adverbs. Abbreviation 'adv-adverb_word' stands for an adverb complementation in the specific form, e.g. adv-hrabro ('Osjećam se hrabro' - 'I feel brave').

Construction with nominative predicate. Abbreviation 'nom_pred' stands for the complementation that represents nominative predicate, e.g. nom_pred ('Historija je postala legendom' - 'History has become legend').

The lexical unit filling the place of the complement must have some morphosyntactic features (type of word and case), but also has to be semantically compatible with the verb it is attached to.

Any change of the single complement (whether it is change in morphosyntactic features or semantics) results in new valence frame of the verb.

Example from the CROVALLEX for verb “to swallow-gutati”:

- *Jana je gutala kruh. (Jana was swallowing bread.)*
- *Naivna javnost guta takvu propagandu. (The naive public is swallowing such propaganda.)*

If the verb complement for patient (*takvu propagandu*) is not the food, we get the new meaning of the verb.

- *Požar je gutao veliko skladište. (Fire was swallowing the big storehouse.)*

If the verb complement for agent is not the living entity (*fire*), we also get the new meaning of the verb.

There are three main relations regarding the change in the verb meaning and the verb valence in Croatian language (all three are present in CROVALLEX):

- change in the verb meaning does not affect the verb valence (verb “to spook-plašiti”)
  - *Ribolovci mrežom plaše ribe - Fishermen chase fish into the net (“to spook-plašiti” meaning “to chase-tjerati”)*
  - *Surla je plašio djecu paklom i sotonom – Surla spooked kids with Hell and Satan (“to spook-plašiti” meaning “to frighten-strašiti”)*
- change in the verb valence does not affect the verb meaning (verb “to swim-plivati”)
  - *Marko Strahija pliva – Marko Strahija swims (*single-valent verb*)*
  - *Marko Strahija pliva rekord - Marko Strahija swims his record (*double-valent verb*)*
- change in the verb valence and the verb meaning (verb “to drink-piti”)

- *Cijeli Zagreb pije vodu iz podzemlja – The whole Zagreb drinks water from subsoil (*double-valent verb meaning "to swallow liquid"*)*
- *Juraj pije nekontrolirano- Juraj drinks without control (*single-valent verb meaning "to get drunk"*)*

3.2. Functors or deep roles

Functors (deep roles, deep cases) are used for expressing types of relations between verbs and their complementations. Functors are divided into inner participants (*actants*) and free modifications.
Since CROVALLEX has the FGD theoretical background, apart from the surface structure, it also has the deep structure represented by deep cases-functors.
Valence frame of the verb, according to FGD, consists of the obligatory and optional valence complements (arguments/inner participants) and of the obligatory free modifications.
In the design of CROVALLEX we decided to distinguish five inner participants – functors (Agent-AGT, Patient-PAT, Recipient-REC, Result-RESL and Origin-ORIG) as well as the whole sequence of free modifications (modification of place, manner, time, cause, etc).
Functors which occur in CROVALLEX 2.008 are listed in Table 1 and Table 2.

Table 1. Inner participants/functors in CROVALLEX lexicon.

Functor	Example sentence
AGT (agent)	John reads the book.
PAT (patient)	John plays the piano.
REC (recipient)	My mother sent her the money.
RESL (result)	His hard work took him to the victory.
ORIG (origin)	We received the message from the dean.

The inner participants satisfy the criterion which requires that the valence frame of a particular verb does not contain two same inner participants.
Also, inner participants can only occur with some verbs (Ivan [AGT] jede ***ručak*** *[PAT]*- Ivan [AGT] eats ***the lunch*** *[PAT]*).
On the other hand, a free modifications can modify any verb although they tend to occur with some group of verbs more than others, which is semantically motivated.
Furthermore, free modifications can appear within the valence frame of a particular verb more than once (Ivan jede ručak ***na klupi [LOC] pred školom [LOC]*** - Ivan [AGT] eats the lunch [PAT] on ***the bench [LOC***] ***in front of the school [LOC]***).

Table 2. Free modifications/functors in CROVALLEX lexicon.

Functor	Example sentence
ACMP (accompaniment)	My sister visited me with her husband.
AIM (aim)	He left the school to join the army.
BEN (benefactive)	My mother made a cake for me.
CAUS (cause)	My father got angry because I failed the exam.
CNCS (concession)	She still loves him although he lied.
COMPL (complement)	I was sailing the seas as a young researcher.
COND (condition)	I will give you my book if you promise not to lose it.
CONTR (contra)	Tomorrow he plays against the tennis player from Italy.
CPR (comparison)	You will have to study more than you did last time.
DIR1 (direction-from)	My mother just came from the theater.
DIR2 (direction-through)	She drove through the town.
DIR3 (direction-to)	My mother went to the shop.
EXT (extent)	The snow has risen over half a meter.
HER (heritage)	They named the boat after the great sailor.
LOC (locative)	My sister lives in Vienna.
MANN (manner)	She lost the interest in reading very quickly.
INST (instrument)	She sent her the news by email.
DIFF (difference)	The stock prices have risen by about 30%.
OBST(obstacle)	My granny tripped over her toys.
REG (regard)	Regardless of her beauty she still has no boyfriend.
RESTR (restriction)	She will make the lunch for all except John.
SUBS (substitution)	Your boy went to the playground instead of going to school.
TFRWH (temporal-from-when)	I remember her being smart from the high school.
THL (temporal-how-long)	She stayed for her holidays in Italy for the whole month.
THO (temporal-how-often)	She plays the guitar every Saturday.
TOWH (temporal-to	The teacher postponed the

when)	exam to June 11.
TSIN (temporal-since-when)	She didn't study since the last semester.
TWHEN (temporal-when)	She visited us last summer.

Both inner participants and free modifications can be obligatory or optional. In Croatian language, the inner participant AGT (which is the first verb complement) is usually being comprised by the verb itself (*e.g. radim* stands for *I work*, *piše* stands for *He writes*, *jedu* stands for *They eat*, etc), but it does not have to be that way (e.g. *ja radim* can stand for *I work*, *on piše can* stand for *He writes*, *oni jedu* can stand for *They eat*, etc).

CROVALLEX takes all these variations into account. Therefore, the inner participant AGT is labeled as AGT_0/1, suggesting that the agent functor can (but does not have to) be declared in separated syntactic position.

Being a flective language, morphemic surface structures in Croatian (as in many other flective languages) are often tied to specific inner participants.

Agent (AGT) is typically tide to nominative case, Patient (PAT) is tide to accusative case, while Recipient (REC) is tide to dative case. Surface form *od+genitive case* is typical for Origin case (ORIG), while *na+accusative case* or *u+ accusative case are* typical for Result (RESL).

Free modifications also have typical morphemic forms closely connected to semantics. Preposition phrase *prema+locative case* indicates direction towards (DIR3), morphemic form *na+locative case* indicates location (LOC), etc. This makes it much more easy to create the valence lexicon and to fill the valence frames.

The compulsoriness of the inner participants and free modifications is related to the sentence grammaticality.

If the obligatory inner participant/free modification is omitted in sentence, the result is ungrammatical sentence of Croatian language.

The approach to the creation of CROVALLEX lexicon is syntactic and semantic in the same time.

It is syntactic because we adopted the Tesniere's [8] syntactic criterion to determine the first verb complement (always the Agent functor) and the second verb complement (always the Patient functor), regardless of semantics.

It is semantic approach because we determine the rest of the inner participants and free modifications based on the semantic roles, which is Fillmore's approach [1].

To conclude, CROVALLEX distinguishes *inner participants* (that can be obligatory or optional) and *free modifications* (that can be typical or not) with their respective surface forms.

3.3. More on frames and attributes in CROVALLEX

In CROVALLEX 2.008, valence frames consist of inner participants and free modifications that are both obligatory and non-obligatory but typical. Typical inner participants and free modifications are those that are typically related to some verbs (or even to whole classes of them) and not to others. Furthermore, frame attributes are either obligatory or optional. The obligatory attributes have to be filled in every frame. The optional attributes might be empty, usually because they are not applicable.

Obligatory frame attributes are gloss and example. *Gloss* represents a verb or paraphrase roughly synonymous with the given frame/meaning used as a clue for fast orientation within the word entry. *Example* represents the sentence taken from the Croatian National Corpus (http://www.hnk.ffzg.hr/) containing the given verb used with the given valence frame. Optional frame attributes are *class* and *idiom*.

Regarding the aspect of the verb, there are three kinds of verbs in CROVALLEX 2.008: perfective verbs, imperfective verbs and dual aspect verbs. In CROVALLEX 2.008, the value of aspect is attached to each word entry as a whole (i.e., it is the same for all its frames). Dual aspect verbs (i.e. analizirati, bombardirati) are the type of verbs that can be used in different contexts either as perfective or as imperfective.

The focus in CROVALLEX 2.008 is mainly on primary meanings of verbs. But, many frames also correspond to peripheral usages of verbs - these are idiomatic frames with label 'idiom'. An idiomatic frame is characterized either by a substantial shift in meaning (with respect to the primary sense), or by a small and strictly limited set of possible lexical values in one of its complementations.

3.4. Semantic classes in CROVALLEX

Some frames are assigned semantic classes like 'motion', 'transport', 'push', 'meet', 'manner of expression', 'eat', etc. In CROVALLEX 2.008 there are 173 syntactic-semantic classes (more accurately, 72 classes with two further levels of

subdivision). Those classes have been derived from VerbNet - verb lexicon based on Levin's verb classes [3, 4], which provides selectional restrictions attached to semantic roles. Verb classes were specially refined and modified for Croatian language. The motivation for introducing such semantic classification in CROVALLEX 2.008 was the fact that it simplifies systematic checking of consistency and allows for making more general observations about the data. Since it's been proved that it is possible to systematically apply the methodology for analysis of verbs in English onto verbs in Croatian language [10] and since the application of VerbNet semantic classes in English language processing was highly successful, in CROVALLEX we decided to implement as detailed syntactic-semantic classification as possible. Therefore, we took Levin's verb classification and modified it so that the distinct verb meanings in Croatian are assigned to the wide spectrum of the predefined syntactic-semantic classes. We believe that the use and analysis of CROVALLEX by language experts will identify usefulness and necessity of these classes in computational lexicography and machine translation for Croatian language. The complete list and description of all classes and subclasses of Croatian verbs with examples is given in the CROVALLEX manual.

5. Conclusion

During the creation of the CROVALLEX lexicon, the great care was taken to ensure the readability, availability, easy orientation and intelligibility. The main aim of the lexicon was to improve the general linguistic culture and the use of standard Croatian language through intuitive approach to the vast spectrum of the linguistic information. The other aim of the lexicon was to become a module for the automatic processing of the texts in Croatian language, especially for automatic syntactic analysis. Since it became publicly available in January 2008, CROVALLEX was already used by other researchers in the field [9]. Verb valence data, which were added to the Croatian dictionary in NooJ, enhanced recognition of VP-chunks as well as NP-chunks and PP-chunks in a sentence. Those better and improved Croatian chunker results made preparation for building Croatian parser.

In the next version of the lexicon we plan to additionally group verbs into clusters of aspectual counterparts. The cluster will be filled with verbs that are result of (im)perfectivization, if their meaning is close to the meaning and to the valence behavior of the main verb in the cluster. Each cluster will contain the list of individual valence frames and for each valence frame the elements of cluster the frame pertains to will be specified.

The valence of deverbal nouns *(nomina actionis)* and deverbal adjectives will make the interesting add-on in the new version of CROVALLEX, since they keep the valence frame of the source verb, even after the surface transformations.

6. References

[1] Fillmore, Ch. J. FrameNet and the Linking between Semantic and Syntactic Relations. Proceedings of COLING 2002.

[2] Hajičová, E., Panevová, J., Sgall, P. A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank. UFAL/CKL Technical Report. 2002.

[3] Korhonen, A., Briscoe, E. Extended lexical-semantic classification of English verbs. Proceedings of the HLT/NAACL'04 Workshop on Computational Lexical Semantics. Boston, MA. 2004. p. 38-45.

[4] Levin, B. English verb classes and alternations: a preliminary investigation. University of Chicago Press. 1993.

[5] Moguš, M., Bratanić, M., Tadić, M. Hrvatski čestotni rječnik. Zavod za lingvistiku i Školska knjiga, Zagreb. 1999.

[6] Samardžija, M. Valentnost glagola u suvremenom hrvatskom književnom jeziku. Doktorska disertacija. Filozofski fakultet Sveučilišta u Zagrebu, 1986.

[7] Šojat, K. Sintaktički i semantički opis glagolskih valencija u hrvatskom. Doktorska disertacija. Filozofski fakultet Sveučilišta u Zagrebu, 2008.

[8] Tesnière, L. Élements de syntaxe structurale. Paris. 1959.

[9] Vuckovic, K., Mikelic Preradovic, N., Dovedan, Z.Verb Valency Enhanced Croatian Lexicon. Proceedings of NOOJ '08. Cambridge UK: Cambridge Scholars Press/CSP. (In press.)

[10] Žic-Fuchs, M. Semantička analiza glagola kretanja u engleskom i hrvatskom književnom jeziku. Doktorska disertacija. Filozofski fakultet Sveučilišta u Zagrebu, 1989.