Semantic classification of verbs in CROVALLEX

NIVES MIKELIC PRERADOVIC, assistant professor
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
I. Lucica 3
CROATIA
nmikelic@ffzg.hr

Abstract: - The paper describes the implementation of the Levin's syntactic-semantic classification of English verbs onto verbs in Croatian language. The current Croatian valence verb lexicon (CROVALLEX) containing 1739 verbs associated with 5118 valence frames was enriched with 72 broad semantic classes with two further levels of subdivision (173 classes in total). The research results show that such classification helps in capturing the relation between the syntax and semantics of Croatian verbs in order to reduce the redundancy in the lexicon, but they also show that Levin's classification does not provide a means for full inference of the verb semantics on the basis of its syntactic behavior for Croatian language. The research brought us to conclusion that we need to introduce the more distinctive semantic roles.

Key-Words: - Croatian verb valence lexicon, valence frames, Semantic classes, Verb synsets

1 Introduction

The goal of this paper is to present and evaluate the semantic classification for verbs in Croatian Valence Lexicon (CROVALLEX). The current version of CROVALLEX is available at: http://cal.ffzg.hr/crovallex/index.html).

CROVALLEX is the first Croatian verb lexicon that contains valence frames of Croatian verbs. Although the main goal was to design the valence lexicon of verbs, nouns and adjectives, the current version of a lexicon contains only valence information for verbs. 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 [3]. CROVALLEX contains 1739 verbs associated with 5118 valence frames (which makes an average of 3 valence frames per verb). Those 1739 verbs were selected from the Croatian frequency dictionary [7], according to their number of occurrences.

2 Motivation

Verb valence lexicon is crucial for many Natural Language Processing (NLP) tasks, such as lemmatization, tagging, machine translation, syntactic analysis or word sense disambiguation (WSD).

Regarding the lemmatization, if we look at the Croatian sentences such as (1) and (2), the surface form **prirodu** is either Acc_sg. of the feminine noun *priroda* or Dat_sg. of masculine noun *prirod.*

(1) *Marko voli* ***prirodu[PAT]*** (*Marko loves nature*) and
(2) Marko se raduje ***prirodu*** *maslina[PAT]* (*Marko looks forward to the yield of olives*)

We can lemmatize the surface form using the valence information: the patient [PAT] complement in the valence frame of the verb *voljeti (to love)* in the first sentence cannot have the surface form in Dat_sg, neither can patient [PAT] complement in the valence frame of the verb *radovati se (to look forward)* have the surface form in Acc_sg.

If we look at the tagging process, the NP *bijelim jedrilicama* in the following sentences can be either Dat_pl. or Ins_pl.

(1) Marko se veseli ***bijelim jedrilicama*** *(Marko is delighted with white sailboats)*
(2) Otisnuli su se ***bijelim jedrilicama*** *prema otoku (they casted off to the island with white sailboards*)

The verb *veseliti se (to look forward)* can only have the obligatory complement in the surface form of Dat_pl., while the verb *otisnuti se (to cast off)* cannot take the complement in the surface form of Dat_pl. if the NP is not preceded by the proposition *s.*

Regarding the syntactic analysis, one can see the importance of valence lexicon from the following

example:
*(1) Stavila **ga** je spavati (she put him to sleep)*
*(2)Prestala **ga** je gnjaviti (she stopped bothering him)*

The pronoun *njega* (*him*) in the sentence (1) can only represent the patient functor of the verb that precedes it, since the valence frame of the verb *spavati* (*to sleep*) does not take any obligatory functor apart from agent.
On the other hand, in the sentence (2) the same pronoun represents the complement of the verb that follows it, since the valence frame of the verb *prestati* (*to stop*) can only take the complement in the form of infinitive. If we only take into account the morphosyntactic description, these two sentences are equivalent. Unambiguous sentence structure can be constructed only if we take into account the valence verb frames.
If we look at the WSD process, the following sentences show that the change in sentence meaning is indicated by the verb valence frame change.

*(1) Odgovarali su **na upite** (they answered the inquiries)*
*(2)Odgovarali su **zbog lošeg rada** (they were responsible for bad functioning)*
*(3) Odgovarali su **opisu** (they matched the description)*

Finally, regarding the semantic analysis, we bring the following examples:
*(1) Čistila je **za** ljubimcem (she was cleaning after the pet)*
*(2) Potrčala je **za** ljubimcem (she ran after the pet)*

Preposition *za* preceding the noun in the example (1) indicates location, while the same NP in example (2) indicates the verbal complement (*direction-to*), which is an important difference in the semantically driven approaches.
The role of these prepositions can not be determined without the verb valence frame analysis.

3 Structure of the CROVALLEX

The valence frame in CROVALLEX 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 (verb complements) that the specific verb demands or grammatically allows.
Each frame slot corresponds to one complementation of the given verb. Types of verbal complements (nouns in specific case, adjectives, adverbs, infinitive constructions, prepositional phrases or subordinate clauses) are precisely distinguished.
The type of valence relation for each complement is marked up as obligatory "obl" or typical optional "typ".
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. We distinguish the close list of five obligatory complements (Agent-AGT, Patient-PAT, Recipient-REC, Result-RESL and Origin-ORIG) and 28 typical optional complements.
CROVALLEX also contains additional information about the verbs: definition of the verb meaning, number of meaning for homonymous verbs, aspect (perfective, imperfective, biaspectual), types of verb use (primary, idiomatic), types of reflexivity for reflexive verbs and semantic class.

4 Semantic classes

Syntactic-semantic classes defined by similar morphosyntactic word behaviour and by semantic similarity are very popular in NLP applications. Classes are very useful since they provide insight into the close relationship of verb syntax and semantics, as well as the possibility of generalization over different linguistic features.
CROVALLEX currently contains 72 broad semantic classes with two further levels of subdivision (173 classes in total).
These classes have been originally adopted from VerbNet project [4], which is a large-scale English verb lexicon, based on Levin's verb classes [6] with more fine-grained sets of verbs (82 broad classes, with 395 subclasses).
Levin's classification [6] is the most extensive syntactic-semantic verb classification in English that provides a classification of 3.024 verbs (4.186 senses) into 48 broad/192 fine grained classes. The extended version of Levin's classification constructed by Korhonen [5] incorporates Levin's classes, 26 additional classes by Dorr [2] and 57 new classes for verb types not covered comprehensively by Levin or Dorr.
These verb classes were translated and adopted for the Croatian language.
The motivation for introducing such semantic classification was to capture the relation between the syntax and semantics of Croatian verbs and to capture generalizations over some linguistic properties in order to reduce the redundancy in the lexicon, since Levin provides selectional restrictions attached to the semantic roles.
Another motivation was the proof that it is possible to systematically apply the methodology for analysis of verbs of motion in English onto verbs of motion in Croatian language [9].
The building of semantic classes was done manually with the help of two Croatian monolingual dictionaries [1, 8] and substantiated by the corpus evidence. Without the corpus evidence, it would be hard to observe and

verify the verb's behaviour in context.

Verbs are placed into classes according to their syntactic and semantic features: verbs belonging to the class "*verbs of putting entities in a specific location*" (e.g. verbs *smjestiti-"set", staviti-"place", umetnuti-"insert"*) take a similar sequence of syntactic complements (***Ivan*** *je namjestio/stavio/umetnuo* ***ključ*** *u bravu*) and can be grouped into linguistically coherent class.

The relationship between syntax and semantics is not always perfect as in the example above, nor does this class semantically completely describe its members. But, one can still define the verb classification for the purpose of generalization over the set of their syntactic and semantic features. Classes to a certain extent allow inheritance of word semantics based on its syntactic behaviour, as well as word syntax based on its semantic behaviour.

Lexical classes define the mapping of the verb complements between the surface and the level which shows the structure of the verb and its complements. Classes are desirable components of each system that is based on the predicate-argument structure. Since classes allow generalization over syntactic and/or semantic features, they can be used in natural language processing systems when we lack data to show the behaviour of relevant words. In such a situation we can work with complex structures that contain all the relevant characteristics of the individual words. Classes are also useful when the lexical information must be drawn from a small, specified corpus. These classes can act as a compensation for the lack of necessary information, representing the behaviour of each relevant word.

Levin's classes are based on the verb's ability to appear in specific pairs of syntactic frames. Levin describes the syntactic behaviour of a verb with respect to its possible syntactic alternations. Semantic classes are created from the verbs that undergo a certain number of alternations. Alternation means a change in the realization of the verb argument structure, such as: *Ivan je dirnuo pulsirajuće srce -> Ivan me dirnuo u srce (Ivan touched the pulsating heart -> Ivan touched me to the heart).*

Her whole theory is actually concentrated around the idea that grouping words according to the alternation can create semantically coherent classes. She claims that the verbs, both in English and in other languages, can be divided into classes based on the common semantic components. Members of the class share a range of features, starting with the implementation and interpretation of certain complements up to the existence of morphologically related forms.

Levin's verb classification introduces explicit syntactic features of each class. Classes are based on the ability of the verb to appear in pairs of frames that are in some sense semantically preserved. Set of syntactic frames that is attached to each of the classes should reflect the semantic components that limit the permissible complements and verb adjuncts. The basic assumption is that syntactic frames represent a direct reflection of the inherent semantics.

As a result of the implementation of the semantic classification in CROVALLEX, each of the 72 verb semantic classes is described by thematic roles (deep cases) and selection restrictions of its verbs. Each of the classes is also defined by valence frames of its verbs, since they contain a set of syntactic descriptions or

TABLE I
DISTRIBUTION OF THE MOST FREQUENT
CROATIAN VERBS IN THE SEMANTIC CLASSES

CLASS	**% of Verbs**
Communication	24,71%
Motion	22,59%
Possesion change	22,59%
Psychic/emotional action	15,80%
Entity features	15,23%
State change	12,13%
Place	10,63%
Remove	10,46%
Creation_conversion	10,00%
Social interraction	9,89%
Body responses	9,20%
Appear/disappear	9,08%
Begin/continue/stop	8,91%
Existence	6,67%
Emission	5,86%
See/sight/peer/stimulus_subject	5,29%
Measurement/price	5,00%
Change of shape and condition	4,94%
Judgement/praise	4,31%
Contact	4,14%
Learn/understand	3,39%
Combine/join	3,28%
Free/imprisonment	3,22%
Succeed/failure	3,22%
Care/neglect	3,16%
Detract/amend	2,99%
Food_drink	2,93%
Transport	2,87%
Push	2,70%
Murder	2,64%
Linger/rush	2,59%
Throw_catch	2,53%
Allow/admit/adopt	2,47%
Search_chase	2,47%
Consume	2,13%
Organize	2,13%

TABLE I
DISTRIBUTION OF THE MOST FREQUENT CROATIAN VERBS IN THE SEMANTIC CLASSES

CLASS	% of Verbs
Wish	2,13%
Posses/own	1,95%
Force	1,90%
Body care	1,84%
Hold/keep	1,78%
Conceal	1,72%
Separate/split/disassemble	1,72%
Accomplish	1,61%
Defend	1,55%
Destroy	1,49%
Lodge	1,49%
Spatial_configuration	1,49%
Weather	1,49%
Emphasize	1,21%
Acquaint	1,09%
Intentional act	1,09%
Color/illustrate	1,03%
Enforce	1,03%
Modal_verbs	1,03%
Discover	0,92%
Try	0,92%
Exceed	0,86%
Attempt	0,80%
Avoid/miss	0,80%
Control	0,80%
Entity_position	0,63%
Complicate/alleviate	0,57%
Animal_sounds	0,52%
Cut/carve	0,52%
Gore	0,46%
Rest	0,46%
Request	0,34%
Differ	0,29%
Transcribe	0,29%
Accustom	0,23%
Suspect	0,17%

syntactic frames that show a possible realization of complement surface structure.

The distribution of the 1739 Croatian verbs in the 72 semantic classes is presented in Table 1.

It is important to mention that out of 39624 word entries in Croatian frequency dictionary [7], 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 good verb coverage.

The distribution shows that the largest number of these most frequent Croatian verbs falls into classes of communication, motion and change of possession. They are closely followed by verbs of psychic or emotional action and verbs that denote features of the entity.

As to the classification of the verbs in English and Croatian, we can speak of similar languages. Despite the cultural differences in certain parts of the vocabulary of these languages, verbs (specifically the verbs of motion) do not show large differences, since their common background is of empirical nature [9].

However, this does not mean that there is unambiguous equivalence between individual lexemes or between lexemes related to the specific concept.

Examples:

- ***Marko hoda niz ulicu.*** (Marko walks down the street)
- ***Marko hoda ulicama grada.*** (Marko walks the streets of the city)

Semantically speaking, or from the thematic roles point of view, the first example shows the intransitive use of the verb *hodati (*to walk*).* The second example shows the possible pseudo-transitive verb use, since the dative plural complement *ulicama grada* emphasizes crossing the path and relates to the larger distances.

The closer the verb in the group hierarchy gets semantically to the base lexeme (e.g. *šetati se – "to stroll"*compared to *hodati - "to walk"*), the less restrictions it expresses to the concept of distance and the potential typical complements.

The farther the verb in the group hierarchy gets semantically from the base lexeme (e.g. *klipsati – "to trudge"*: *Ranjenik je klipsao hodnicima* [*The wounded man was slogging along the corridors]* but not *Ranjenik je klipsao ulicama grada* [*The wounded man was slogging along the streets of town]*), the more restrictions it expresses on syntagmatic level. In other words, it leads to the restrictions in the typical possibilities of path crossing that are related to the semantic structure of the verbs (typical complements that appear with the verb *klipsati – "to trudge"* express some kind of burden).

As a result of this work, we wanted to generalize about the behaviour of most frequent verbs in Croatian language, using the verb classes. We would also like to reduce the effort required to create lexicons and the likelihood of introducing errors while adding a new verb into the existing lexicon.

Nevertheless, we would like to emphasize that one can not assume that the entire Croatian verb syntax will reflect the inherent semantics of verbs, although part of the syntax has this power.

5 Evaluation

The Levin's classification served as a good starting point, since it brought us to the following conclusion in the evaluation process:

(1) Two different verbs, belonging to the same semantic class, usually have the same valence frame

- Položio je novac u banku AGT[1:obl] PAT[4:obl] DIR3[u+4:opt] (*He put the money in the bank*)
- Uveo je prijatelje u kuću AGT[1:obl] PAT[4:obl] DIR3[u+4:opt] (*He brought his friend into the house*)

(2) The change in verb valence indicates the possible change in verb semantics

- Petar je udario dijete. ACT[1:obl] PAT[4:obl] (*Petar hit the child*)
- Petar je udario u stup. ACT[1:obl] DIR3 [u+4:obl] (*Petar hit at the pole*)

In the current version of CROVALLEX, some complements with the specific surface forms serve as criteria for establishing a reasonably consistent class.
For example, one can prove the appearance of the typical complement DIR1 (direction-from) with its surface forms (*iz+genitive, s+ genitive, od+ genitive*) in the valence frames of the verbs belonging to the *remove and motion_direction_away_from* semantic classes.
The complement DIR1 appears in the valence frames of 213 verbs with particular meanings in the following semantic classes: dissapear (19), remove (59), motion_direction_away_from (48), pit/debone (3), banish (5), push (12), pour (18), throw (16), get/obtain (17), transfer_message (7), free (9).
Although this criterion was reliable for some classes and although there is the obvious relation between the verb semantics and syntactic features of the complements in the valence frames, this criterion is not reliable for all the currently existing classes.
The problem with Levin's classification is that it primarily concentrates on verbs taking NP and PP complements and does not provide a comprehensive set of senses for verb. Many verbs in Croatian are polysemic, and although most of them have a predominating sense in corpus data, there are a significant number of high frequency verbs that cannot be adequately represented with a single sense.

Current obligatory and typical roles, as well as semantic classes, cannot account for all the problems that have arisen from the corpus data.
Example from the corpus 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.)*

Although it looks like the both verbs have the same meaning and same valence frame, if the verb complement for patient (*takvu propagandu*) is not edible, we get the new meaning of the verb.
Furthermore, if the verb complement for agent is not the living entity (*fire*), we also get the new meaning of the verb.
Požar je gutao veliko skladište. (Fire was swallowing the big storehouse.)

Finally, Levin's alternations do not provide consistent classes, which is obvious from the current semantic classification problems in CROVALLEX.

6 Conclusion

The basic assumption in CROVALLEX is that the verbs belonging to the same semantic class should have the same or similar complements with the same or similar morphosyntactic form in their valence frame.
Unfortunately, Levin's classification does not provide a means for full inference of the verb semantics on the basis of its syntactic behaviour.
Therefore, we would like to introduce the more distinctive semantic roles.
In the improved version of CROVALLEX, we would like to introduce the further division of the verb complements in order to obtain the valence notation with large degree of sense differentiability.

- Marija razbija glavu vazom. [AGT(**animate:1**), PAT(**body_part:1**), INS(**tool:1**]
- Marija razbija glavu glupostima [AGT(**animate:1**), PAT(**body_part:1**), INS(**apstract:1**]
- Kamen razbija prozor [AGT(**inanimate:1**), PAT(**surface:1**)]

The improved version of CROVALLEX will consists of verb synsets, instead of individual verb lemmas.
The verbs in a synset share the same meaning(s) and are part of the aspectual derivational string (aspectual counterparts and prefixed verbs), but do not necessarily contain all aspectual counterparts, since it is not the rare case that aspectual counterparts differ semantically.
Since the average number of verbs in a synset is 3, as a result we expect to get the enriched lexicon with approximately 3500-4000 verbs.
We believe that with such improvements we can solve the problem with two main relations regarding the change in the verb meaning and the verb valence in Croatian language (both derived from corpus data), that the current classification does not account for:

(a) 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")*

(b) 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*)*

Furthermore, in the new version of the lexicon, we would like to give priority to the semantic criteria against the diathesis alternations used by Levin because of the difference between Croatian and English language and because of the difference in data sources (Levin's linguistic literature vs. Croatian corpus data). We would like to obtain verb classes that are semantically more consistent than those we already have.

References:

[1] Anić, V. *Rječnik hrvatskoga jezika (Croatian Language Dictionary)*. Zagreb, 2000.

[2] Dorr, B. *Large-scale dictionary construction for foreign language tutoring and interlingual machine translation.* Machine Translation, 1997. 12(4): pp. 271–325.

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

[4] Kipper, K., Dang, H. T., and Palmer, M. *Class-based construction of a verb lexicon*. In Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000), pp. 691–696.

[5] 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. pp. 38-45.

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

[7] Moguš, M., Bratanić, M., Tadić, M. *Hrvatski čestotni rječnik (Croatian Frequency Dictionary)*. Zavod za lingvistiku i Školska knjiga, Zagreb. 1999.

[8] Šonje, J., Nakić, A. (eds.) *Rječnik hrvatskoga jezika (Croatian Language Dictionary).* Zagreb: Školska knjiga. 2000.

[9] Žic-Fuchs, M. *Semantička analiza glagola kretanja u engleskom i hrvatskom književnom jeziku (Semantic analysis of motion verbs in English and Croatian language)*. Doktorska disertacija. Filozofski fakultet Sveučilišta u Zagrebu, 1989.

[10] Fellbaum, C (ed). *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[11] Hensman, S., Dunnion, J. *Constructing conceptual graphs using linguistic resources.* In Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics, Prague, 2005.

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

[13] Hajič, J., Hladká, B., Pajas, P. *The Prague Dependency Treebank: Annotation Structure and Support.* Proceeding of the IRCS Workshop on Linguistic Databases. Univesity of Pennsylvania, Philadelphia, 201. pp. 105-114.

[14] Kipper, K., Korhonen, A., Ryant, N., Palmer, M. A *Large-Scale Extension of VerbNet with Novel Verb Classes*. Proceedings of EURALEX. Turin, Italy. 2006.

[15] Korhonen, A. *Using semantically motivated estimates to help subcategorization acquisition*. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong. 2000. pp. 216-223.

[16] Pala, K., Ševeček, P. *Valence ceskych sloves*. Sbornik praci FFBU. Brno. 1997. pp. 41-54.

[17] Panevová, J. *Valency Frames and the Meaning of the Sentence*. The Prague School of Structural and Functional Linguistics. / ed. Ph. L. Luelsdorff. Amsterdam-Philadelphia, John Benjamins. 1994. pp. 223-243.