CROVALLEX lexicon improvements: Subcategorization and semantic constraints

NIVES MIKELIC PRERADOVIC
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 Croatian valence verb lexicon (CROVALLEX) that contains information on syntactic subcategorization and semantic restrictions of 1739 most frequent Croatian verbs. These 1739 verbs are associated with 5118 valence frames and enriched with 72 broad semantic classes with two further levels of subdivision (173 classes in total). The evaluation shows that syntacto-semantic verb classification helps in capturing the relation between the syntax and semantics of Croatian verbs and therefore reduces the redundancy in the lexicon. Unfortunately, classes in the current version of CROVALLEX do not provide a means for full inference of the verb semantics on the basis of its syntactic behavior. Therefore, in the improved version we plan to introduce the more distinctive semantic roles. In the improved version of CROVALLEX the semantic typing will be based on EuroWordNet Top Ontology. We believe that with such improvements we can solve the problem of sense differentiability and get a finer grained semantic classification of verbs in Croatian language.

Key-Words: - Croatian verb valence lexicon, Valence frames, Syntacto-semantic classes, Verb synsets

1 Introduction

In the natural language processing system lexicon is the most important component, since the size and the quality of the lexicon limits the scope and coverage of the system. Contemporary systems limit the lexicon coverage to specific domains, since the process of building the machine lexicon tends to be much harder than building a paper one. Paper lexicons contain monolingual or bilingual information, while machine lexicons also contain semantic information such as meronymy, holonymy etc. for even more than two languages.

Furthermore, some machine lexicons have syntacto-semantic information, data about restrictions and links to the hierarchical structure that reflects the structure of the human lexicon, but they also serve to support natural language processing on a large number of computer tasks, which means that one should put more work and effort in optimal lexicon design.

Because of the different theories and styles of implementation, existing lexical sources differ in terms of format, coverage, level of detail and description of lexical formalisms [21]. One of the most popular machine lexicons is WordNet [10]. It is designed to display the lexicon of English language in the form of synonym clusters (synsets), with simultaneous display of relations between them.

One component that each machine lexicon should always have, regardless of domain or context, is a module that contains the information required for processing and generation of natural language. Examples of necessary information that each entry in the machine lexicon must have are syntactic subcategorization and semantic restrictions.

Subcategorization specifies the type of syntactic expressions that can be combined with other expressions. For example, the Croatian sentence *Majka je kupila Mariji novu knjigu* [Mother bought Mary a new book] is grammatically correct, because the verb *kupiti* [to buy] takes 3 complements in Croatian; agent-*mother*, direct object or the topic – *book* and the indirect object or the recipient - *Mary*. Thus, the topic must be noun phrase in the accusative case and the recipient must be noun phrase in the dative case.

On the other hand, subcategorization specifies only the syntax restrictions, which means that it allows meaningless sentences such as *Mother bought Mary a new nervousness*, since it has no semantic restrictions. Although this sentence satisfies the syntactic restrictions, it has absolutely no sense.

Semantic constraints classify words into different groups, such as animate/inanimate, abstract/concrete, etc [18]. For example, we know that verb *to wear,* in its meaning "*to have clothing, jewellery, etc. on your body*", takes the noun phrase complement that must be an object that is wearable. We also know that the doer of the verb action must be a living creature. If the doer of the action is not a live entity, he/she can not put something on.

Both the information on syntactic subcategorization and semantic restrictions is contained in the first

Croatian verb valence lexicon – CROVALLEX. It is a lexicon that describes valence - the ability of linguistic units (in this case verbs) to show their combinatorial potential in language statements.

One of the key tasks of natural language processing is to make available information about valence, since the valence of language can not be automatically predicted and it makes it necessary to create a lexicon that will store such information.

One of the main goals of this paper is to present a model of this lexicon that is both human and machine readable lexicon which describes the valence of the most frequent verbs in Croatian language. Furthermore, this paper describes the theoretical motivation for the creation of such a lexicon. Finally, the application of such lexicon could be manifold, since it has cultural, educational and scientific component. Also, we believe that this lexicon might be a good starting point for making bilingual valence lexicon as a source for machine translation.

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, 13, 16, and 17]. 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 Accusative singular of the feminine noun *priroda* or Dative singular of the 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 Dative singular, neither can patient [PAT] complement in the valence frame of the verb *radovati se (to look forward)* have the surface form in Accusative singular.

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

(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 Dative plural, while the verb *otisnuti se (to cast off)* cannot take the complement in the surface form of Dative plural 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 [19, 20], 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 [11].
The role of these prepositions can not be determined without the verb valence frame analysis.

3 Structure of the CROVALLEX

Valence theory is an important part of Functional Generative Description (FGD), which takes into account both the syntactic and semantic criteria in a way that the first and second complement of a verb gets estimated in terms of syntactic behavior of the complements, while the other verb supplements are estimated in terms of semantics [22].

Based on the teachings of Tesniere and Fillmore, valence theory was especially important in the 1960s in Czech syntax (developed by P. Sgall and colleagues as a Functional Generative Description, layered approach to natural language processing based on dependency (characteristic of verb to link certain number of syntactic positions to itself) and two-level syntactic markup, where the analytic level provides the surface syntactic representation, while the tectogrammatical level includes a background structure.

Thus, according to the valence theory, the verb in the sentence opens the places that can or must be filled with certain morphological and semantic complements. Verb complements are divided in two groups: inner participants and free modifications. Inner participants can be obligatory or optional, while free modifications may be typical or not.

Omission of the inner participant from the sentence makes the sentence ungrammatical.

Lexical entries of verbs contain at least one, but usually more valence frames, which are defined as a set of syntactic elements (verb complements) that the specific verb requires or grammatically permits. Lexical entry describes the verb in its primary and secondary use (*to absorb:* primary use – to absorb *fluid*, secondary use – to absorb *the price decline*).

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

Syntacto-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.

The first of the three main approaches to the classification of verbs in English that we would like to mention here is WordNet [23].

WordNet is an online available lexical database for English language, developed at Princeton University in the United States, led by George A. Miller [24]. Its development started in early 90s of the 20th century and it is a result of the cooperation of linguists, psychologists and information experts. WordNet is designed as a system that displays the lexicon of English language in the form of synonym clusters (synsets), with simultaneous display of relations between them.

For each noun, verb, adjective, or adverb one may obtain information about their synonyms, antonyms, hyponyms and hypernyms, as well as holonyms, meronyms and coordinated concepts. Additionally,

WordNet provides information on troponyms for verbs and sentence patterns. WordNet currently contains 155287 synonym clusters of nouns, verbs, adjectives and adverbs, where each cluster represents lexicalized concept. Words in a synset are connected with lexical and semantico-conceptual relations. Antonyms connect individual words, while links of the subordinate and superior relations connect all of the synsets. WordNet was created primarily as a semantic network and contains very little syntactic information. But even as the semantic source, it does not contain some

information that is necessary for natural language processing applications, such as explicit predicate-argument structure. Also, the word meanings in WordNet are often too specific; it doesn't have the basic display of semantic components and systematic expansion of basic meanings that create these very specific meanings.

The second of the three main approaches to the classification of verbs in English is VerbNet, developed by Kipper-Schuler in 2006 [25]. It is hierarchically organized computer lexicon of verbs, which is based on Levin's classes with the aim of the systematic organization of entries. VerbNet is domain-independent and has considerable coverage regarding the representation of the verbs in the classification. VerbNet is important because it provides detailed syntacto-semantic descriptions of Levin's classes [26], which are further refined and systemized.

VerbNet shows a way of building lexicons through the narrow and explicit connection between syntax and semantics.

The first version of a VerbNet proved to be useful in various fields of natural language processing [27, 28, 29, 30], but that version described only the verb complements in the form of noun phrases and prepositional phrases, and that was its main limitation. VerbNet was in 2004 enriched with 57 new syntacto-semantic classes with detailed descriptions, added by Korhonen and Briscoe [5, 15]. These classes constitute an important addition to the original Levin's classes. Finally, the latest version of VerbNet contains an additional set of 53 classes that were added by Korhonen and Ryant in 2006 [14].

Each verb class in VerbNet lexicon is fully described by the so-called thematic roles (actually deep cases), limits in the verb selection regarding its complements and frames consisting of syntactic description and semantic predicates. Each class in VerbNet also contains a set of syntactic descriptions or syntactic frames that show a possible surface realization of the argument structure.

Semantic restrictions such as *animate/inanimate, person, organization* are introduced to define the types of thematic roles that are allowed as verb complements, while additional restrictions are imposed to indicate the syntactic nature of the constituents that are likely to be associated with a particular thematic role. Syntactic frames can also be restricted defining the adverbs that are allowed in the specific frame. Each frame is assigned with explicit semantic information.

Using the verb classes, VerbNet manages to generalize information on the behaviour of verbs and reduces the effort required to create lexicon as well as the likelihood of introducing errors by adding a new verb to the lexicon.

VerbNet is in its current version freely available online. The comprehensive description and a list of classes based on Levin's classification of verbs in the English language allows the creation of big training corpus that can be used to explore the impact of syntacto-semantic classes on the improvement of results of syntactic parsers and algorithms for solving the word sense disambiguation problem.

The third approach to the classification of verbs in English is represented by FrameNet [31]. Compared with Levin's classification, FrameNet does not group verbs, but groups of words, regarding the conceptual structures (frames).

Combinatorial patterns of these frames are inductively derived from corpus examples. This means that verbs that are grouped together in FrameNet [32] may be similar semantically, but present different syntactic behaviour. Also, verbs that share the same syntactic behaviour can be represented by different semantic frames. In FrameNet [12], verbs such as *to fill* and *to load* share the same frame - *filling*. Additionally, verb *to load* is part of the *placement* frame, while verb *to fill* can be found in the *decorating* frame.

It is necessary to point up that neither Levin nor FrameNet assume that the entire verb syntax reflects the inherent semantics of verbs, although part of the syntax has this power.

VerbNet is dealing with Levin's style verbs, i.e. verbs that take noun phrase complements and complements in the form of prepositional phrases, and therefore offers limited coverage. FrameNet, on the other hand, creates a lexicon of verbs with comparable coverage, but with much more detail concerning the semantics and syntax and verb complements. FrameNet has semantically more consistent categories and richer set of relations among them.

Because of the importance and effectiveness of introduction of syntactic and/or semantic classes in the computer processing of English language, we decided to introduce a detailed syntacto-semantic classification as a part of the valence lexicon for Croatian language.

Therefore, we took over Levin's classification of verbs (underlying one of the largest syntacto-semantic lexicons - VerbNet) and modified it in such a way that the verbs of the Croatian language (more precisely, different meanings of the verb) join a wide range of predefined classes. We hope that the use and analysis of this classification by linguistic experts will prove the usefulness and necessity of syntacto-semantic classes in Croatian computer lexicography and machine translation for Croatian language.

Regarding the structure and the type of information it conveys, CROVALLEX is closest to the second approach. It currently contains 72 broad semantic classes with two further levels of subdivision (173 classes in

total).

All these classes have been originally adopted from VerbNet project [4], 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 syntacto-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

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%

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

The distribution of the 1739 Croatian verbs in the 72 broad semantic classes is presented in Table 1.
Apart from these 72 classes, lexicon also contains 101 subclasses and in the following subsection we offer the detailed list of subclasses with their description.

4.1 Subclasses in CROVALLEX

CROVALLEX offers 72 broad verb semantic classes, which are all listed in Table 1. But, it is also important to mention and describe the 22 subclasses, which are part of these broad classes.
Here is the list of them:

1) Subclasses of the broad class ***Place***: *put* - verbs that place entities at some location; *put_spatial* - verbs of putting entities in the particular spatial configuration; *put_manner* - verbs placing entities in a specific location in a specific manner: *put_direction* - verbs placing entities in a specific location with a defined direction; *pour* - verbs of placing the liquid entities in the container or on the specific surface; *put_around* - verbs of putting other entities around the specific entity; *put_store* - verbs covering surface and placing the entities in the container; *put_result* - verbs of terminated action that is a result of putting the entities in or on the specific place; *put_X_on/in_Y (denominal verbs)* - verbs such as "put X on/in Y" where X is the noun from which the verb is derived

2) Subclasses of the broad class ***Remove****: remove* -

verbs that relate to the removal of an entity from a location; *banish* - verbs relating to the removal of a person from a location; *clear* - verbs of cleaning; *wipe* – verbs that remove things from surfaces or containers have 2 additional subclasses, i.e. wipe_manner (they relate to the method of removing entities from the surface or container) and wipe_instrument (they relate to the instrument of removal of the entities from the surface or container); *steal*- verbs of abduction and property loss; *pit/debone* - verbs of removing entities from something

3) Subclasses of the broad class ***Possesion change***: *give* - verbs of giving; *Contribute/ future_having* - verbs of change of possesion; *fulfil/equip* - verbs of supply and equipment where X gives something to Y that Y deserves, needs or is worthy of; *get/obtain* - verbs of acquisition and obtain that take the benefactive argument; *exchange* - verbs that relate to the exchange of one thing for another

4) Subclasses of the broad class ***Throw/catch***: *throw/catch* – verbs that relate to instantaneous cause of the ballistic motion by imparting a force; *hit* - verbs that relate to the moving of one entity in order to bring it into contact with another entity

5) Subclasses of the broad class ***Contact***: *touch* - verbs that describe surface contact with no implication that the contact is the direct result of the impact; *swat* - verbs that describe surface contact that is the direct result of the impact

6) Subclasses of the broad class ***Creation conversion***: *build/create/prepare* - verbs that describe construction and creation of a product through the transformation of the raw materials; *grow* - verbs describing the transformation of an entity from one form to another; *turn/convert* - verbs of transformation; *perform* - verbs of performance

7) Subclasses of the broad class ***Entity features:*** *appoint* - verbs that denominate the entity; *masquerade* - verbs that relate to the disguise; *characterize/declare* – verbs that relate to observations; *conjecture/ consider* - verbs relating to forming of thoughts, concepts, predictions, judgments and suspicions; *impress* verbs of impression; *suitability/adequacy* - verbs expressing appropriateness and convenience

8) Subclasses of the broad class ***Psychic/emotional:*** *bother* – intransitive verbs denoting psychological action where the entity is concerning oneself; *amuse* - transitive verbs denoting psychological action where thr direct object is living entity; *admire/marvel* – transitive verbs denoting adoration and worship of the specific entity; *mourn/worry* - verbs of intransitive type that express grief or sorrow and reflexive verbs expressing persistent mental uneasiness; *like* – reflexive verbs that denote that some entity is fond of another entity

9) Subclasses of the broad class ***Search_chase***: *chase* - verbs that involve two participants following the same route; *search/stalk/investigate/rummage/ferret* – verbs that relate to searching for entities

10) Subclasses of the broad class ***Social interraction***: *correspond* – verbs that relate to establishment of agreement between entities, *marry* - verbs that relate to an action of joining in marriage of performing a marriage ceremomy; *meet* - verbs relating to the act of encountering living entities; *battle* - verbs that 11) relate to fight and competition

11) Subclasses of the broad class ***Communication***: *transfer_message* - verbs that describe the process of information transfer; *contact* – verbs of shouting; *reveal* – verbs that describe informing process; *speak_manner* - verbs relating to different ways of speaking: *instrument_communication* - verbs that relate to the means of communication; *say/tell/talk* - verbs of speaking; *manner of expression* - verbs describing different ways of expression; *complain* - verbs that express resentment and displeasure; *advise* – verbs of counseling; *confess* - verbs that relate to acknowledgment or admission of faults, misdeeds and crimes; *inquire* – verbs that express demand o request; *reply*- verbs that relate to providing an answer in speech or writing

12) Subclasses of the broad class ***Food_drink:*** *eat* – basic verbs relating to consumation of food and beverages; *chew/gobble* - verbs that denote ways of consuming food and beverages; *dine*- verbs relating to the time of consuming food and beverages; *feed*- verbs relating to provide of food or nourishment; *devour* - verbs that desribe enjoying in the food and drink

13) Subclasses of the broad class ***Body responses:*** *physiological function* - verbs relating to physiological events; *nonverbal_expression* - verbs expressing body communication that does not involve words; *body gestures*- verbs that relate to gesture involving body parts; *snooze*- verbs describing states of consciousness; *body reflex response* - verbs desribing the unintentional body movements; *body_internal_state*- verbs describing the physical condition of the subject, which is a reflection of a particular mental or physical condition; *suffocate* - verbs describing dying from lack of air or oxygen; *verbs of change in physical condition and injuries* have 4 additional subclasses, i.e. *pain*- verbs relating to body pain*; tingle*- verbs describing pain realted to body organs*; hurt*- verbs relating to injuries*; change_bodily_state*- verbs relating to change in physical condition

14) Subclasses of the broad class ***Body Care:*** *verbs of body care and maintenance* have 2 additional subclasses, i.e. *dress*- verbs relating to the maintenance of the whole body and to clothing; *groom* – verbs that describe removing of dirt and parasites from the skin, fur, or feathers of an animal; *body_parts_care*- verbs

expressing care of specific body parts

15) Subclasses of the broad class ***Emission***: *light_emission* - verbs relating to the transmission of light; *sound_emission* - verbs relating to the transmission of sound; *smell_emission* - verbs relating to the transmission of smell; *verbs of substances emission* have 2 additional subclasses, i.e. *substance_emission* verbs that describe the transmitted substance; *substance_emitter*- verbs that describe the activity of the emitter

16) Subclasses of the broad class ***Change of shape and condition***: *break* – verbs describing sudden or violent separation into pieces; *Bend* - verbs related to forcing to assume a different direction or shape; *cook* - verbs related to food preparation; *state_change_external_cause* - verbs describing the changes in entity state due to external causes; *state_change_entity_specific_cause* - verbs that relate to changes in entity state that are specific to an entity; *state_change_calibratable_cause*- - verbs describing the gradient changes in entity state

17) Subclasses of the broad class ***Existence***: *exist* - verbs that relate to general existence; *entity_specific_modes_being* - verbs describing existence specific to a particular entity; *modes_of_being_with_motion* - verbs describing existence that implies motion; *sound_existence*- verbs describing the existence of sound; *swarm/herd/bulge* - verbs describing aggregation of persons or animals; *contiguous_configuration* - verbs describing relations between the two entities

18) Subclasses of the broad class ***Appear/disappear***: *appear*- verbs relating to coming into existence; *disappear* - verbs related to ceasing of existence; *occur* – verbs that describe the process of taking place, happening

19) Subclasses of the broad class ***Motion***: *motion direction across/through* - verbs that describe transcursion, passing over*; motion_direction_inwards*- verbs relating to making en entry, going in*; motion_direction_outwards*- verbs relating to going out, exiting*; upwards motion direction* - verbs that describe upward movement; *motion direction downwards* - verbs that describe downward movement; *motion_direction_circular* - verbs that describe circular movement; *avoid/miss* - verbs that describe moving away from, departin, leaving; *motion direction towards* - verbs that describe approaching, moving towards; *motion_manner* - verbs relating to the manner of movement; *vehicle*- verbs that describe movement by vehicle; *vehicle_with direction*- verbs relating to the movement by vehicle with defined direction; *accompany* - - verbs relating to escort; *motion_swarm*- verbs that describe movement in the crowd

20) Subclasses of the broad class ***Linger/rush***: *linger/wait* - verbs that describe the slowness in leaving, especially out of reluctance and remaining or resting in expectation; *rush* - verbs that describe moving or acting swiftly

21) Subclasses of the broad class ***Measurement/price***: *register* - verbs relating to a formal or official recording of items, names, or actions; *cost* - verbs related to amount paid or required in payment for a purchase; *assessment/price* – verbs that describe the classification of something with respect to its worth; *bill* – verbs related to presentation of a statement of costs or charges

22) Subclasses of the broad class ***Begin/continue/stop***: *begin*- verbs relating to taking the first step in performing an action; *complete* - verbs that describe the bringing to a finish or an end*; stop*- verbs that describe discontinuance; *continue* - verbs that describe carring on uninterruptedly and persistance in something

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.

In the improved version of CROVALLEX the semantic typing will be based on EuroWordNet Top Ontology. It is a lattice structure of 63 features that can be combined in feature combinations. The ontology is specifically designed to help the encoding of the lexical semantic relations in a uniform way.

We believe that with such improvements we can solve the problem of sense differentiability and get a finer grained semantic classification of verbs in Croatian language.

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.
[18] Tuzov, V. *Computer Semantics of Russian*. In Proceedings of the 2nd WSEAS Int. Conf. on Simulation, Modeling and Optimization (ICOSMO 2002), Skiathos, GREECE, 2002.
[19] Fragos, K., Maistros, I., Skourlas, C. *Using Conditional Probabilities of Weighted Terms for a Lexicon Based Sense Disambiguation System*. In Proceedings of the 3rd WSEAS Conference in Informatics, Malta, June 2003.
[20] Hung, J.C. *The Semantic Annotated Documents - From HTML to the Semantic Web*. In Proceedings of the 2007 WSEAS International Conference on Computer Engineering and Applications, Gold Coast, Australia, 2007.
[21] Arnold, J. B., Liow, J. S., Schaper, K. A., Stern, J. J., Sled, J. G., Shattuck, D. W., et al. *Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects*. NeuroImage (13). 2001. p. 931–943.
[22] Sgall, P., Hajičová, E., Panevová, J. (1986) The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: Reidel, Prague: Academia.
[23] wordnet.princeton.edu/
[24] Miller, G; Beckwith, R; Fellbaum, C; Gross, D; Miller, K. Introduction to WordNet: An on-line lexical database // International Journal of Lexicography, vol. 3, br. 4. 1990. p.235-244.
[25] VerbNet website. http://verbs.colorado.edu/~kipper/verbnet.html (February 19th, 2010)
[26] Levin, B. English verb classes and alternations: a preliminary investigation. University of Chicago Press. 1993.
[27] Swier, R; Stevenson, S. Unsupervised semantic role labelling. // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004. p. 95–102.
[28] Hensman, S; Dunnion, J. Automatically building conceptual graphs using VerbNet and WordNet. // Proceedings of the 3rd International Symposium on Information and Communication Technologies (ISICT). Las Vegas, NV. 2004. p. 115–120.
[29] Swift, M. Towards automatic verb acquisition from VerbNet for spoken dialog processing. // Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbruecken, Germany. 2005.
[30] Crouch, D; Holloway King, T. Unifying lexical resources. // Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbruecken, Germany. 2005.
[31] Baker, C.F; Fillmore, C.J; Lowe, J.B. The berkeley framenet project. // COLING- ACL'98: Proceedings of the Conference. Montreal. Association for Computational Linguistics. 1998. p. 86–90.
[32] FrameNet website: http://framenet.icsi.berkeley.edu (February 19th, 2010)