Abstract: In this paper we aim to describe the latest work on the syntactic-semantic classification of verbs in Croatian language. We tried to extract the specific semantics and function of verb prefixes in Croatian and capture these specific features in the Croatian verb valency lexicon CROVALLEX. The syntactic and semantic connection between base verbs and derived prefixed verbs is revealed and modelled as rules which contribute to the enrichment of the CROVALLEX lexicon.

Key Words: verbal prefixes, Croatian Verb Valency Lexicon, verb semantics, syntacto-semantic verb classification, derivational rules, Aktionsart

1 Introduction
The goal of this paper is to present and evaluate the improvement of syntactic-semantic verb classification in Croatian Valency Lexicon (CROVALLEX).

CROVALLEX [5] is the first Croatian verb lexicon that contains valency frames of Croatian verbs crucial for many Natural Language Processing (NLP) tasks, such as parsing, tagging, machine translation, syntactic analysis or word sense disambiguation (WSD) of Croatian language. It contains 1739 verbs associated with 5118 valency frames (which makes an average of 3 valency frames per verb). Those 1739 verbs were selected from the Croatian frequency dictionary [12] according to their number of occurrences. The background theory in CROVALLEX is valency theory, developed by Czech linguists Petr Sgall and his collaborators as the part of the Functional Generative Description [6], used for the description of valency frames in CROVALLEX.

The valency frame in CROVALLEX consists of the set of syntactic arguments that the specific verb demands or grammatically allows. The type of valency relation for each argument is marked up as obligatory “obl” or typical optional “typ”.

Since most of the verbs in Croatian are polysemic, each verb meaning requires unique morphemic form for all its obligatory and optional arguments. That morphemic form is stored in CROVALLEX together with the information about their obligatoriness or optionality. CROVALLEX distinguishes obligatory complements / thematic roles (e.g. Agent-AGT, Patient-PAT, Recipient-REC, Result-RESL, etc.) and typical optional complements (LOC-location, CAUS-cause, HER-heritage, etc).

2 Prefixes and Meanings
The most productive type of verb derivation in Croatian is prefixation, a derivational process that attaches the prefix to a verbal stem resulting in a new lexical item.

The verbal aspect of Croatian verbs is always marked, i.e. the verb infinitive can have either perfective or imperfective form. The prefixation of an imperfective verb can either yield an imperfective verb (davati/to give– do+davati/to add) or a perfective verb (kuhati/to cook – s+kuhati/to finish cooking). On the other hand, the prefixation of a perfective verb can yield only perfective forms (baciti/to throw – iz+baciti/to throw out), since imperfective forms of a perfective verb are built through suffixation (baciti/to throw – bac+a+ti/to throw over and over again).

Although being described in general grammars and dictionaries of Croatian language [2, 3], the current state of knowledge regarding different properties of verb prefixation is inconclusive.

On the other hand, the close relationship between the verbal aspect and Aktionsart (defined as a syntactic-semantic category covering the manner of action, action type, verbal character or aspectual
character) has been revealed in many languages (especially German), as well as the fact that both the aspect and Aktionsart are result of the derivational process.

For the proper interpretation of Aktionsart, the function of prefix in the derived verb is the most important: if the prefix induces the semantic modification of the base verb it is attached to, then it signifies Aktionsart of the derived verb, causing the change of verbal aspect as well [9].

This paper deals with base verbs covered in the current working version of CROVALLEX and their derivatives (prefixed verbs derived from the base verbs) revealing syntactic-semantic relations between them.

Aiming to take only base verbs in CROVALLEX as our starting point for lexicon enlargement, we distinguished 453 base verbs out of 1739 verbs, while the number of base verbs that were not in the first version of CROVALLEX, but had their derivatives in the lexicon was 112. These prefixed verbs were replaced with the base verbs they were derived from and we also added all aspectual counterparts yielding 1453 base verbs in total.

As far as our approach is concerned, 1453 base verbs from CROVALLEX were combined with each member from the set of 20 prefixes that are productive in Croatian (do3, iz3, na3, nad3, o3/ob3, obez3, od3, po3, pod3, pre3, pred3, pri3, pro3, raz3, s3, su3, u3, uz3, za).

In Croatian, prefixes ending in a consonant have different forms (allomorphs) depending on the initial letter of the base verb as well as alternant forms with the vowel extension –a (raza3, iza3, uza3, nada3, oda3, poda3).

Therefore, a set of 50 rules for combining the prefix and base verb was manually developed, to take into account characteristic morphophonemic adaptations (e.g. raz+plakati = rasplakati/make someone cry, iz+ćupati = išćupati/pull out, etc.).

Finally, we obtained 28947 prefixed verbs combining prefix and the base verb and searched for the matching word-forms in the several Croatian monolingual and bilingual dictionaries [1, 4, 13, 14, 15].

Parcic (1901) was the only dictionary available in electronic format. The other four dictionaries were manually checked and cross-checked. Regarding the number of attested verbs: 4600 prefixed verbs were found in (Parcic, 1901), 3591 were found in (Sonje, 2000), 4262 verbs in (Anic, 1998), 3327 verbs in (Šarić, Lj., Wittschen W., 2003), and 2541 verbs in (Filipovic, 1996).

When overlapped, the 5 dictionaries yielded 6129 attested prefixed verbs or 21.2% of all derivated prefixed verbs. The interesting fact is that the oldest dictionary (Parcic, 1901) contained 1441 prefixed verbs that could not be found in other 4 dictionaries.

On the other hand, the number of verbs specific only to each of the 4 other dictionaries was as follows: 73 prefixed verbs were found only in (Sonje, 2000), 360 verbs in (Anic, 1998), 56 verbs in (Šarić, Lj., Wittschen W., 2003), and 38 verbs in (Filipovic, 1996). Furthermore, we collected the statistics of derivational characteristics for individual prefixes (see Table 1). Also, we thoroughly analyzed all meanings of 20 prefixes, compared the properties of the base verbs and prefixed verbs and built the model shared by a group of verbs with the same meaning of the prefix. The prefix with the largest set of meanings is za-(in, at, down) with 20 different meanings, while the average number of meanings per prefix is 11.

The process of prefixation establishes connection between different base verbs that start to share the same semantic feature due to the prefix attached. This connection becomes obvious in the valency frames of these verbs, since prefixed verbs tend to share the valency frames, regardless of the valency frame of the base verb.

For instance, some verbs inherit the meaning „to move something or someone from inside out or bottom up“ from the prefix attached, some get the meaning „to join, fix, adhere to“, etc.

We discovered that derived verbs can be treated as hyponyms of base verbs if they do not differ

---

**Table 1. Frequency distribution of each prefix in attested prefixed verbs**

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Translation</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>za</td>
<td>in, at, down</td>
<td>618</td>
</tr>
<tr>
<td>iz</td>
<td>out, out of, from</td>
<td>551</td>
</tr>
<tr>
<td>po</td>
<td>over, throughout</td>
<td>534</td>
</tr>
<tr>
<td>u</td>
<td>in, into, at</td>
<td>491</td>
</tr>
<tr>
<td>na</td>
<td>onto, on</td>
<td>484</td>
</tr>
<tr>
<td>o/ob</td>
<td>around</td>
<td>537</td>
</tr>
<tr>
<td>pre</td>
<td>over, across</td>
<td>373</td>
</tr>
<tr>
<td>pro</td>
<td>out, through</td>
<td>373</td>
</tr>
<tr>
<td>s</td>
<td>with, from, off</td>
<td>371</td>
</tr>
<tr>
<td>pri</td>
<td>to, onto, at</td>
<td>366</td>
</tr>
<tr>
<td>do</td>
<td>to, towards, upon</td>
<td>354</td>
</tr>
<tr>
<td>raz</td>
<td>apart, in different directions</td>
<td>338</td>
</tr>
<tr>
<td>od</td>
<td>apart, away from</td>
<td>301</td>
</tr>
<tr>
<td>pod</td>
<td>under, off</td>
<td>156</td>
</tr>
<tr>
<td>uz</td>
<td>up</td>
<td>128</td>
</tr>
<tr>
<td>nad</td>
<td>over</td>
<td>81</td>
</tr>
<tr>
<td>pred</td>
<td>pre</td>
<td>36</td>
</tr>
<tr>
<td>su</td>
<td>together, with, along</td>
<td>28</td>
</tr>
<tr>
<td>obez</td>
<td>dis-, de-</td>
<td>9</td>
</tr>
</tbody>
</table>

---

*Recent Advances in Computer Science*
significantly with regard to the meaning. But, the meaning of the prefixed verb can completely differ to that of the base verb due to the prefix attached (see examples 1-6).

Examples:
1. dat/i to give – iz+dati/to betray in one sense or iz+dati/to release in other
2. baciti/to throw; po+baciti - to have an abortion; pred+baciti - to object, pod+baciti - to let down
3. sjeti/to sit – na+sjeti/to sit onto in one sense or na+sjeti/to be deceived in other, pri+sjeti/to get stuck in the troat in one sense or pri+sjeti/to win a seat next to whom in other
4. paliti / to burn – is+paliti/to fire off, s+paliti/to burn down, za+paliti / to burn up
5. gubiti/to lose – po+gubiti/to lose all one after the other in one sense or po+gubiti/to put to death in other, izgubiti/to lose for good
6. pasti/ to fall – na+pasti/to fall onto in one sense or na+pasti/to attack in other

The semantic meaning of prefixed verbs is not a straightforward combination of the base verb meaning and the meaning of the prefix, since both base verbs and prefixed verbs are usually polysemous. Accordingly, the prefixation can be perceived only as a connection between the base verb and the prefixed verb in one of their meanings and not between verbs as a whole. Polysemous base verb might cause polysemy of the prefixed derived verb, but the prefixed verb may also become polysemous as a result of the polysemous meanings of the prefix itself.

3 Prefixes and Verb Classification
As part of our research, we manually examined the syntactic argument structures of the derived prefixed verbs in relation to their base verbs, using the list of 20 productive prefixes, the list of 6129 prefixed verbs derived from CROVALLEX base verbs (and verified for correctness), as well as corresponding 565 base verbs from CROVALLEX.

Our results show that the effect of these prefixes on the argument structure of verbs in Croatian exhibits regularity that can be exploited to drastically reduce the manual work required to produce the high-coverage valency lexicon.

Verbs sharing the same prefix (more precisely, the same meaning of the specific prefix) also share the same valency frames and the number of obligatory and typical optional complements, such as DIR1-direction from or DIR3-direction to, with their surface forms (see examples 7-13 for prefix od- in its meaning “motion direction away from”). Examples 12 and 13 show prefix do- in its meaning “extension of an action over its logical end”, attached to different verbs yielding the same effect (compare do+dati/to piece on; do+pisati/to add to the text; do+liti/to pour more).

Examples:
7. Jedrilica se otisnula iz uvale na pučinu AGT 3obl [DIR1 iz+2] [DIR3 na+4] (The sailboat shoved off the coast to the open sea)
8. Voda je otekla s krova u oluk AGT 3obl [DIR1 ur+2] [DIR3 na+4] (Water ran off the roof into the gutters)
9. Zmija je otajala iz sobe na livadu AGT 3obl [DIR1 ur+2] [DIR3 na+4] (The snake crawled out of the room onto the meadow)
10. Mačka je odskočila s kreveta na pod AGT 3obl [DIR1 na+2] [DIR3 na+4] (The cat jumped off the bed onto the floor)
11. Damir je odšetao iz kazališta u mračnu ulicu AGT 3obl [DIR1 u+2] [DIR3 u+4] (Damir walked away from the theatre into the dark street)
12. Marko je dopisao svoje ime u knjigu žalosti. AGT 3obl PAT 4obl [DIR3 u+4] (Marko added his name to the book of condolences)
13. Ivana je dolila mlijeko u šalicu AGT 4obl [DIR3 u+4] (Ivana poured milk into the cup)

As a consequence, it is possible to add typical optional complements to the valency frame of the verb in a more systematic way.

One can also notice the semantic harmony between prefixes, prepositions and typical optional complements. The meaning of the prefix is amplified by the meaning of the preposition. For instance, prefix pre- and preposition preko (over) share the meaning of „moving over something or somebody“. Furthermore, preposition such as preko (over) is always part of the optional complement (such as DIR2_preko) that verbs with prefix pre- typically bind. This harmony is most obvious in the verbs of motion, where some prefixes maintain a distinct meaning of direction or location. The constant use of same preposition (such as u-, na-, uz-) with the same prefix emphasizes the specific (in this case - directional) meaning of the given prefix (see examples 14-19).

Examples:
14. u+accusative, in the valency frames of the verbs with prefixes u- (uroniti/submerge, uskočiti/jump into, uputiti/initiate into)
15. na+accusative in the valency frames of the verbs with prefixes na- (naskočiti/jump on, naljetiti/lay on, naslagati/lay up)
(16) za+accusative in the valency frames of the verbs with prefixes za- (zamaknuti /slink off, zabaciti /throw aside, zaći /go down)
(17) preko+genitive in the valency frames of the verbs with prefixes pre- (prebaciti /get over, prenijeti /pass over, premažati /lay on)
(18) kroz+accusative in the valency frames of the verbs with prefixes pro- (provući /thread, progruri /push through, prokopati /dig through)
(19) uz+ accusative in the valency frames of the verbs with prefixes uz- (uzdići /lift up, uspraviti /bear up, užviti /tower up).

We can conclude that the verb valency properties are closely related to their semantics. The process of prefixation changes the aspect of the verbs, but also their meaning and their valency properties.

4 Verb Classification in Crovallex

CROVALLEX currently contains 72 broad classes with two further levels of subdivision (173 syntactic-semantic classes in total). Those classes have been originally adopted from VerbNet project [8] which is a verb lexicon based on Levin’s verb classes [10, 11] with more fine-grained sets of verbs (82 broad classes with 395 subclasses). These verb classes were adapted for the Croatian language, since most of the alternations used by Levin do not apply for Croatian. The motivation for introducing such semantic classification was to interpret 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.

The Levin’s classification served as a good starting point for syntactic-semantic classification, since it revealed that different base verbs, belonging to the same semantic class, also share identical valency frames (both the number and type of obligatory and optional arguments and their surface morphosyntactic form).

It also helped us to discover that the change in verb valency frame indicates the possible change in the semantics of the verb.

The problem with this classification is that it does not provide consistent classes, which is obvious from the semantic problems in CROVALLEX. A very large number of verbs in Croatian is polysemous, and although most verbs have a predominating sense in corpus data, there is a significant number of high frequency verbs that cannot be adequately represented with a single sense. Table 2 shows that prefixed derivative verbs, compared to their single-valent base verbs, belong to different syntactic-semantic classes, take more specific typical optional complements (direction) or different number of obligatory complements in their valency frames thus becoming double-valent verbs (e.g. with prefix ob- or na-).

The change in prefix (see examples 20-22) indicates the change in both obligatory verb complements and verb semantics (compare misliti/to think; s+misljati/to plot; za+misljati/to imagine)

Examples:
(20) Marija je mislila AGT_1:obl (Marija was thinking)
(21) Marija je smišljala priču AGT_1:obl PAT_4:obl (Marija plotted the story)
(22) Marija je zamišljala more AGT_1:obl PAT_4:obl (Marija imagined the sea)

Additionally, examples in Table 2 show that most of the derivatives sharing the same prefix belong to a single class. Therefore, we wanted to interpret the specific semantics of verb prefixes in Croatian and represent these specific features in CROVALLEX.

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Valency</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base verbs: juriti (to rush), trčati (to run), letjeti (to fly), ploviti (to sail)</td>
<td>AGT_1:obl</td>
<td>Motion manner</td>
</tr>
<tr>
<td>Prefixed verbs: pojuriti (to start rushing), potrčati (to start running), poletjeti (to start flying)</td>
<td>AGT_1:obl [DIR3_prema +3]</td>
<td>Rush</td>
</tr>
<tr>
<td>Prefixed verbs: odjuriti (to rush off), otrčati (to run off), odljeteti (to fly away)</td>
<td>AGT_1:obl [DIR1_od+2] [DIR3_prema +3]</td>
<td>avoid/ miss</td>
</tr>
<tr>
<td>Prefixed verbs: izjuriti (to rush out), istrčati (to run out), isploviti (to sail out)</td>
<td>AGT_1:obl [DIR1_iz+2] [DIR3_prema +3]</td>
<td>motion direction outwards</td>
</tr>
<tr>
<td>Prefixed verbs: utrčati (to run into), uploviti (to sail into), uletjeti (to fly into)</td>
<td>AGT_1:obl [DIR3_u+4]</td>
<td>motion direction inwards</td>
</tr>
<tr>
<td>Prefixed verbs: oprtrčati (to run around), obletjeti (to fly around), oploviti (to sail around)</td>
<td>AGT_1:obl PAT_4:obl</td>
<td>contiguous location</td>
</tr>
<tr>
<td>Prefixed verbs: nadjuriti (to throw out)</td>
<td>AGT_1:obl PAT_4:obl</td>
<td>Banish</td>
</tr>
</tbody>
</table>

Table 2. Changes in valency and semantics caused by prefixation in motion verbs
prefix (more precisely, the same meaning of a prefix) also have the same or similar complements with the same or similar morphosyntactic form in their valence frame. Furthermore, some complements with the specific surface forms also served as criteria for establishing a reasonably consistent syntactic-semantic class, where several prefixes (such as iz-, s-, and od-) share the same meaning.

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 valency frames of the verbs belonging to the following syntactic-semantic classes: disappear, remove, motion direction away from, pit/debone, banish, push, pour, throw, get/obtain, transfer message, free. Also, prefixes iz-, s- and od- share the following meanings: separation, movement away from, completion of an action where the patient is depleted, completion of an action to the end with no possibility for further action, creation, gaining or obtaining and the change of state.

These combinations of prefixes and base verbs gave us insight into the close relationship between the syntax and the semantics of the verbs sharing the same prefix, as well as the possibility of generalization over different linguistic features and the possibility of reducing the redundancy in CROVALLEX.

In the working version of CROVALLEX we introduced explicit syntactic and semantic features of each verb synset. Synsets are based on the ability of the prefixed 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 synsets reflects the semantic components that limit the permissible complements and typical optional complements. As a result, every verb synset is described by thematic roles (deep cases) and selection restrictions of its prefixed verbs. Every synset is also defined by valency frames of its verbs, since they contain a set of syntactic descriptions.

Since such verb synsets contain all the relevant characteristics of the individual verb, they allow generalization over syntactic and/or semantic features of these verbs. They can also act as a compensation for the lack of necessary information, representing the behaviour of each relevant verb. If grouped according to their prefix, verbs create semantically and syntactically coherent synsets where members of the synset share a range of features, starting with the implementation and interpretation of certain complements up to the existence of morphologically related forms.

As a result, it is possible to reduce the effort required to enlarge our lexicon and the likelihood of introducing errors while adding a new verb into the existing lexicon.

5 Conclusion and Future Work

This paper presents the results of our attempt to capture and model the connections between base and prefixed verbs in the valency lexicon of Croatian verbs - CROVALLEX.

We can conclude that the basic function of prefixation in Croatian is not pure perfectivisation, but the derivation of a new verb with the lexical meaning more or less different from that of the base verb.

Analyzing all productive derivational prefixes for verbs in Croatian, we proved that the prefix entails not only the change in the verbal aspect, but, more importantly, the change in the lexical meaning of the verb, the change in valency frame (both the number of arguments and the surface forms) as well as the syntactic-semantic class.

When determining the typology of semantic features of each productive prefix in Croatian we realized that the features implicated by prefixation usually mean temporality (inchoative, punctuative, durative, iterative), locativeness, direction (adlative, ablative, periphrastic), dimension (intensive, diminutive), state change (resultative, mutative), transformativity, distribution, comitativity, etc.

Unfortunately, such classification does not provide a means for full inference of the verb semantics. Therefore, we would like to introduce the more distinctive semantic roles in our future work.

Our current goal is to improve the syntactic-semantic classification of verbs in the lexicon and decrease redundancy in CROVALLEX. The final goal of our model is to allow the automatic prediction of valency frames, as well as syntactic-semantic properties of prefixed verbs that are still to be annotated.

Although we revealed the obvious relation between syntactic-semantic classes based on Levin’s classification and meanings of the prefixes in Croatian, we are still not able to describe all verb meanings or to account for semantic problems that arise from the polysemous nature of the verb.

Example for the verb “nedostajati-to miss”:

(23) Nives nedostaje more [AGT$_3$;obl PAT$_1$;obl] (Nives misses the sea)
(24) Nives nedostaje moru [PAT$_1$;obl AGT$_3$;obl] (The sea misses Nives)
(25) Moru nedostaje Nives [AGT$_3$;obl PAT$_1$;obl] (The sea misses Nives)
More nedostaje Nives [PAT₁:obl AGT₃:obl]

All sentence constructions (23-26) are syntactically and morphologically valid, but sentences (24) and (25) are not semantically possible, since the agent is not the living entity (only in the poetic context (24) and (25) might be semantically acceptable). We aim to resolve this type of problem introducing the verb subcategorization features (see examples 27 and 28).

(27) Nives nedostaje more [AGT₃:obl (person:1) ENT₁:obl(body of water:1)]

(28) More nedostaje Nives [ENT₁:obl (body of water:1) AGT₃:obl(person:1)]

In the improved version of CROVALLEX, we plan to introduce the semantic typing based on the VERBALLEX lexicon in order to get a finer grained semantic classification, while building a bilingual valency lexicon.

Verballex is a valency lexicon of Czech verbs developed at the University of Brno [7] that includes approx. 10 500 Czech verbs with semantic classes inspired by the Levin’s classification (similar to CROVALLEX).

Although CROVALLEX and VERBALLEX share approx. the same number of the thematic roles, VERBALLEX also distinguishes approx. 1000 subcategorization features, each of them being related to Princeton WordNet and Czech Wordnet.

Confronting Czech-Croatian pairs of verb valency frames, we aim to capture the semantic part of the Czech valency frames and apply the subcategorization features to Croatian frames with minor modifications.

We believe that with such improvements we can solve the problems that the current version of lexicon does not account for. For example, the change in the verb meaning does not always affect the verb valency in Croatian, but the current version of CROVALLEX does not capture these distinctions.

Example for the verbs “nasjesti”:

(29) Nives je sjela na klupe
AGT₁:obl(person:1)[LOC₄:obl(destination:1)]
(Nives sat on a bench)

(30) Kontinentalna ploča je nasjela na oceansku.
ENT₁:obl(surface:1), LOC₄:obl(surface:1) (The continental plate sat on the oceanic plate)

(31) Nives je nasjela na laž
AGT₁:obl(person:1), ABS₄:obl(abstraction:1)
(Nives fell for a lie)

Subcategorization features such as person:1, destination:1 and abstraction:1 in sentences (29-31) distinguish different verb meanings very well.

Enriching obligatory arguments with subcategorization features will ensure the valency notation with higher degree of sense differentiability.

References: