Generation of Verbal Stems in Derivationally Rich Language

Krešimir Šojat*, Nives Mikelić Preradović[+], Marko Tadić*
*Department of Linguistics, [+]Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
E-mail: ksojat@ffzg.hr, nmikelic@ffzg.hr, mtadic@ffzg.hr

Abstract

The paper presents a procedure for generating prefixed verbs in Croatian comprising combinations of one, two or three prefixes. The result of this generation process is a pool of derivationally valid prefixed verbs, although not necessarily occuring in corpora. The statistics of occurences of generated verbs in Croatian National Corpus has been calculated. Further usage of such language resource with generated potential verbs is also suggested, namely, enrichment of Croatian Morphological Lexicon, Croatian Wordnet and CROVALLEX.

Keywords: computational morphology, language generation, morphological lexicon, valency lexicon, Croatian language

1. Introduction

So far the computational processing of Croatian morphology was oriented primarily towards inflectional phenomena (Tadić, 1994; Tadić & Fulgosi, 2003; Tadić, 2005; Ćavar et al. 2009) or various enlargement procedures of Croatian Morphological Lexicon (Oliver & Tadić, 2004). The derivational processing has not been in the focus so far. During the process of building the Croatian Wordnet (Šojat et al., 2005, Raffaelli et al., 2008) and the Croatian Valency Lexicon (Mikelić Preradović, 2008), the necessity to combine these three lines of work became obvious. In this work we primarily focus on the lexical category of verbs, that in Croatian exhibit extremely rich derivational morphology in terms of affixation, and its computational processing.

From the point of linguistic typology and cognitive linguistics, different languages could be regarded as different points in the continuum that spans from predominantly grammaticalized languages on the one side to predominantly lexicalized languages on the other side. The extreme points of this continuum are theoretical constructs – there is neither a completely grammatical, nor a completely lexical language – but different languages exhibit the usage of different means to express the same meaning. Some of them prevail on the lexicalized side of the continuum, and some on the grammaticalized side. Inflection is traditionally regarded as belonging to the grammaticalized side of the spectrum, while languages with the absence of many word-forms for a single lemma belong mostly to the lexicalized side.

Our point of interest here is the derivation of Croatian verbs, more precisely, verbal stems of infinitives and we concentrate only on prefixation. The paper presents the computational method for generation of all possible derivational forms on the basis of the set of productive prefixes in Croatian and the set of verbal roots, thus producing new verbal stems. The output of the procedure is a set of all potential, derivationally related verbal lexemes that are further searched for and attested in Croatian National Corpus. We strongly believe that this methodology could be *mutatis mutandis* applicable to other Slavic or other typologically similar languages. In this respect, this task also could be regarded as a subtask of a larger language generation task.

The general idea stems from the (Halle, 1973) where the theoretical model for a device that would generate all possible morpheme combinations in a given language was presented. Such a device for derivationally rich language as Croatian by far exceeds the scope of this paper. However, parts of the derivational processes can easily be represented by rules that would automatically result in new lexical entries. Since the prefixation has the sufficient level of complexity on its own, we limit the processing of derivational patterns to this phenomenon in this work.
The following section describes the motivation for this task, section 3 describes the method used, followed by section 4 where evaluation is presented. Conclusion and future work is given at the end.

2. Motivation

Generally, the Croatian verbal stems are bearing the fundamental lexical meaning. But, depending on the prefix, the mode and conditions of the denoted action (i.e. phenomena that are usually described as *Aktionsart*), could be specified more precisely. Also, the meaning of the derivational form can differ completely to that of the base form due to the prefix attached.

Furthermore, certain prefixes contribute only to the change of the verbal aspect and have no role in *Aktionsart* of verbs. In this paper these two categories are not distinguished for the sake of the computational processing at the level of derivation only. Its proper treatment would be possible at the semantic level of processing, i.e. in Croatian Wordnet.

Verbs in Croatian are always marked for the verbal aspect, i.e. the infinitive of the verb can be either perfective or imperfective in its form. Prefixation of an imper-

fective verb can either yield an imperfective verb (*osjećati/to feel– su+osjećati/to sympathize*) or a perfective verb (*pisati/to write – pre+pisati/to copy*). Prefixation of a perfective verb can yield only perfective forms (*uzeti/to take – od+uzeti/to take away*), since imperfective forms of a perfective verb are built through suffixation (*uzeti/to take – uz+ima+ti/to take over and over again*).

Imperfective and perfective verbs in Croatian can, in terms of derivation, be roughly divided into four groups. The first group comprises 'pure' verbal stems (e.g. *pisati/to write*). The second group consists of predominantly perfective verbs built through prefixation of verbs from the first group (e.g. ***pre****+pisati/to copy by writing)*. The third group comprises imperfective verbs that exclusively denote the iterativity of the action. Verbs in this group are built through suffixation of the verbs from the second group (e.g. *pre+pis+****iva****+ti/to copy over and over again)*. Finally, the fourth group consists of perfective verbs derived through prefixation of verbs in the third group (e.g. ***is+****pre+pis+iva+ti/to finish copying over and over again)*. Verbs in this group exclusively comprise distributive verbs denoting actions performed by several agents on several objects and/or in successive phases. These verbs almost exclusively have two or even three prefixes, since the prefixation actually takes place on already prefixed verbal stems from the second group.

As far as our approach is concerned, non-prefixed verbs (the first group) is taken as a starting point. In order to determine their actual derivational span, they are combined with each member from the set of prefixes. This set consists of 20 prefixes that are productive in Croatian (*do-, iz-, na-, nad-, o-, ob-, od-, po-, pod-, pre-, pred-, pri-, pro-, raz-, s-, su-, u-, uz-, za-, obez-*). Prefixes ending in a consonant have different forms (allomorphs) depending on the initial letter of the verbal stem. This makes the derived prefixed verb more pronounceable and is known as assimilation. A voiced consonant is replaced by a voiceless one under the influence of the adjacent voiceless consonant. The voiced [*z*] in *iz-, raz-, uz-* is replaced by the voiceless [*s*] in the prefixed verbs *is+plakati, ras+plakati, us+plakati* under the influence of the voiceless [*p*] in *plakati/cry*. A prime example is the prefix *iz-* which also has *iž-, iš-, is-,* and *i-* as different forms. Prefixes ending in a consonant also have alternant forms with the vowel extension –a (*raza-, iza-, uza-, nada-, oba- , oda-, poda-*). These forms with vowel extensions occur before verbal stems beginning with certain consonants or consonant clusters.

Together with their allomorphs, our set of prefixes comprises 52 members in total. Three other prefixes (*mimo-, protiv-/protu- and suprot-)* are not part of this set, since they are no longer productive in Croatian. The translation of the prefixes, their meaning and examples are given in Table1.

As mentioned, the third group of verbs comprises imperfective verbs built from prefixed verbs from the second group through suffixation. The base form of these verbs, i.e. the form of the verbs with detached prefixes, does not exist as independent stem due to suffixation (e.g. *pisivati* in *pre+pis+****iva****+ti*). Nevertheless, these 'dummy' forms can enter into numerous combinations with various prefixes thus forming valid verbal stems. For example, the form *vršavati/to perform iteratively* does not exist on its own, but it can enter into combination with various prefixes and yield verbs as *iz+vršavati/to perform*, *do+vršavati/to bring to an end*, *s+vršavati/to finish*, etc. As far as our work is concerned, we take all verbal stems from the third group and check for their combinatorial possibility with each member from the set of prefixes. As far the verbs from the fourth group are concerned, we seek for two or three prefixes and thereby establish the base form of the verb and all possible combinations of two or even three prefixes as well (e.g. *is-po-na-pijati/to get drunk each by each entirely in a certain period of time*).

3. The method

Our method consists of several steps. Firstly, a digital version of monolingual dictionary (Anić, 2003) was searched for all verbal headwords (14,106 entries). Verbs were manually inspected and classified as stems without prefixes and stems with prefixes. In this process loanwords from Latin and Greek were put aside because they behave differently and attract Croatised Latin and Greek prefixes. They will be treated in a separate line of work.

This process gave us verbal infinitives without any prefix (e.g. *baciti/throw*), infinitives with one prefix (e.g. *iz+baciti/throw out*); infinitives with two prefixes (e.g. *po+iz+bacati/throw out each by each*); infinitives with three prefixes (e.g. *is+po+pre+bacivati/ throw over each by each iteratively*). The only evidence of Croatian verbal stems with more than three prefixes in Croatian is *pre+po+iz+od+nositi/take away each by each iteratively*), we believe that is has been artificially invented by lexicographer because it can't be detected in any Croatian corpus or web. This is why we limited our approach to only three possible prefixal positions.

All prefixed verbs from the dictionary were stripped off their prefixes and the "pure" verbal lexical morphemes, i.e., roots were obtained. The number of these verbal stems was 6,519 (out of 14,106 verbal headwords). In the second step we generated all possible combinations of 20 productive prefixes, which gave us the list of 348 verb prefixal strings made up of 2 prefixes and the list of 6,500 verb prefixal strings composed of 3 prefixes. These numbers we obtained after 10 rules of morphophonemic adaptations were applied and duplicates removed. The check was also donemanually by searching the Croatian Language Portal (http://hjp.srce.hr, with access to the same lexicographical data as in Anić 2003) and found evidence for 101 prefixal strings made up of 2 prefixes.

In the third step "pure" verbal stems were combined with the prefixal strings of 1, 2 and 3 prefixes. The system used for generation was developed as a set of scripts that take the list of 6,519 stems as input (easily obtainable by deleting of infinitival inflectional ending *–ti* or *–ći*) and the list of prefixal strings (that was divided into 3 lists: list of single prefixes, combination of 2 prefixes and the combination of 3 prefixes).

Prefix	Translation	Meaning	Examples
do-	to, towards, upon	1. to complete an action 2. to make someone / something reach a specific place, size or stage	1. dozreti - ripen, doživjeti-experience 2. dovesti se-drive towards, donijeti - take to, dorasti-grow up, dosegnuti-reach, dograditi-add on / build on, doletjeti-fly to, arrive
iz-/iza-	out, out of, form	1. to leave confined space 2. to separate from 3. to move toward the top 4. to complete an action 5. to perform an action on all consecutive objects	1. izaći -exit, iskočiti - jump out, istrčati - run out 2. izdvojiti - separate from, isključiti - rule out 3. ispeti se - climb toward, izroniti-emerge/rise to the surface 4. izgorjeti - burn out 5. izudarati - beat up, istjerati- expel
mimo-	by, past	to perform an action in the opposite direction, to move past / by someone (something)	mimoići (go by)
na-	onto, on	1. to perform an action from above 2. to perform an action in sufficient or excessive extent 3. to place one object onto another 4. to complete an action 5. to commence an action	1. naskočiti - jump on 2. najesti se - stuff oneself 3. nataknuti - stick on 4. nahraniti - feed / fully satisfy someone's hunger 5. nagristi - bite onto
nad-/ nada-	over	1. to perform an action over something 2. to perform the action of higher intensity than the one expressed by the basic verb	1. nadlijetati - fly over 2. nadvikati-outcry
o-	around	1. to perform an action around something 2. to complete an action	1. ograditi - fence in, okružiti - surround 2. očistiti - sweep, osakatiti - mutilate
ob-	around	1. to perform an action around something 2. to complete an action	1. obgrliti - embrace, obložiti - encase, opkoliti- surround 2. oboriti - knock down
obez-	dis-, de-	to take away the object implicated by the basic verb	obeshrabriti - discourage, obespraviti- deprive one of his rights, obezglaviti - decapitate
od-	apart, away from	1. to complete an action 2. to separate, estrange, remove from something 3. to perform action responding to one expressed by the basic verb	1. otpjevati - sing, odsvirati - play 2. odvojiti - remove, otići - leave, odvesti - take away from 3. odazvati se - respond, odgovoriti - reply
po-	over	1. to commence an action 2. to complete an action 3. to carry out an action onto something 4. to carry out an action to a lesser extent than the action expressed by the basic verb 5. to perform an action on multiple objects or by multiple subjects	1. potrčati - start running 2. poginuti -die, potrošiti-spend, potonuti -sink 3. popločati - pave 4. poraditi - work (for a while) 5. poskakati - jump (one after another), polomiti - break down, pobacati - throw away (one after another)
pod-/ poda-	under, off	1. to complete an action 2. to carry out an action to a lesser extent than the action expressed by the basic verb 3. to carry out an action from below or under that impression	1. podrezati - trim / cut off 2. podcjenjivati - underestimate 3. potpasti - fall under, podmetnuti - palm off, potkupiti - bribe
pre-	over, across	1. to carry out an action to undesirable extent 2. to complete an action 3. to carry out an action over something or to change the place of something 4. to transform from one state to another 5. to repeat the action anew and again 6. to divide something into two or more parts 7. to carry out an action to a lesser extent	1. prejedati se -overeat, prepeći - overdo 2. preboljeti - get over 3. prenijeti - carry over, preletjeti - fly over, preploviti- travel across, pregaziti - run over 4. pretopiti - recast 5. prepričavati - retell 6. pregraditi - partition 7. predahnuti - respite
pred-/ preda-	pre-	to perform an action before something, presupposing something that will happen later	prethoditi – precede, predosjećati – anticipate, predodrediti - predetermine
pri-	to, onto, at	1. to bring an action to a particular result or goal 2. to carry out an action to a lesser extent than the action expressed by the basic verb 3. to bring closer something to someone or something 4. to add something to someone or something 5. to compress and compact	1.primorati - force, prinuditi - coerce, prisiliti - compel, privesti - lead toward 2.primiriti - soothe 3. prinijeti - bring up to see, primamiti - entice, privući - attract, pritegnuti - tighten; prisvojiti - usurp 4. pričvrstiti- attach, prikovati- pin down; prišiti- sew on, privezati - tie down, pridružiti - join, priključiti - hook up, priložiti - enclose 5. prignječiti - pinch, pritisnuti - press
pro-	out, through	1. to commence an action 2. to carry out an action to a lesser extent than the action expressed by the basic verb 3. to carry out an action through something 4. to change something 5. to complete an action 6.rapidly move through something, beside something, by something	1. propjevati - begin to sing, progovoriti - begin to talk, procvasti - flourish 2. protresati - shake out 3. probušiti - pierce, prokopati - dig through, probiti - break through 4. produžiti - lengthen 5. pročitati - read out, prokuhati - boil 6. proći - pass through / by; projuriti - rush by
protiv-/ protu	against, contra	to remove the effect of the action, to respond to someone's action	protusloviti – contradict, proturječiti – disagree
raz-/ raza-	apart, in different directions	1.to perform an action that is contrary to what is expressed by the basic verb 2. to change a state expressed by the basic verb 3. to complete an action 4. to separate, estrange, remove from, free from something 5. to commence an action in an intensive way 6. to disassemble a whole into separate parts	1. razmrsiti- straighten out, razoružati - disarm 2. razblažiti - dilute, razvodniti - water down, razbuditi -awake 3. raznijeti - blow up 4. razmicati - move apart, rastjerati - dispel, razbježati se - scatter, razbaštiniti - disinherit, rastaviti - divorce 5. razljutiti - irritate, razbuktati se -flare 6. razlomiti - break apart, rasuti - disperse
s-/sa-	with, from, off	1. to complete an action 2. to arrange an ensemble into a whole 3. to remove from	1. slomiti- crush 2. skupiti - gather, sastaviti - put together, sljubiti - to bring together, združiti - unite / combine 3. sasuti - pour out, zbaciti - throw off, skinuti - remove from
su-	together, with, along	1. to perform an action together with someone 2. to carry out an action against someone or something	1. surađivati - collaborate, suosnovati - co-found, sufinancirati - co-finance 2. sudariti se - collide, sukobiti se - clash
suprot-	against	to perform an action contrary to one expressed by the basic verb	suprotstavljati se - oppose
u-	in, into, at	1. to penetrate under the surface 2. to get into something 3. develop a property expressed by the basic verb 4. to complete an action 5. to place something into something 6. to perform the action expressed by the basic verb for a longer time	1. urezati - carve, uvući - pull in, uroniti - immerse 2. ugnijezditi se - nest, 3. ubrzati - accelerate 4. ukrasti - steal 5. unijeti - bring in, uvesti - introduce 6. umiriti se - calm down
uz-/uza-	up	1. to commence an action 2. to commence an action in an intensive way 3. to carry out an action from below upwards	1. uzjahati - mount (a horse) 2.ushodati se - bustle 3. uspinjati se - climb up
za-	in, at, down	1. to commence an action 2. to put something inside 3. to attach one object to another 4. to cover something 5. to perform the action expressed by the basic verb for a longer time 6. to carry out an action behind something 7. to complete an action	1. zapjevati - start singing, zazvoniti - start ringing 2. zabosti - stick in 3. zakovati - nail down, zalijepiti - glue down 4. zasuti - fill in /overwhelm 5. zamisliti se - wonder / rethink, zapričati se - chat 6. zabaciti - throw back 7. završiti - complete, zagraditi - fence in

Table 1. translation of the prefixes, their meaning and examples

We manually developed a set of 50 rules for combining the verbal stem and prefix, to take into account characteristic morphophonemic adaptations during the generation of the prefixed verbs (e.g. *iz + pustiti = ispustiti/drop*, *s + baciti = zbaciti/throw off*, *iz + četkati = iščetkati/brush off*, etc.).

Finally, we obtained 125,698 prefixed verbs combining 1 prefix and the verbal stem, 2.183,111 prefixed verbs combining 2 prefixes with the verbal stem and 42.373,500 prefixed verbs combining 3 prefixes and the verbal stem.

All generated combinations of prefixed verbs were then automatically given the inflectional pattern number, following the procedure for generation of word-forms described in (Tadić, 1994 and Tadić, 2005).

The next step consisted of search for the matching word-forms in the 101 Mw in size Croatian National Corpus (Tadić, 2009) against the list of types from that corpus. Out of 9,065,201 verbal tokens of main verbs (excluding auxiliaries) that realised with 143,831 types, only 86,576 were prefixed. The overall statistics of prefixal strings is presented in the Table 2. Also the statistics of detected patterns was collected as well, but it can't be presented entirely due to the limitation in the size the paper. As an illustration, we found verbs, such as *iž+džepariti/ pickpocket till there's nothing left*, that are not mentioned in any of the Croatian monolingual dictionaries, but do exist in the Croatian poetry and prose.

	Number
Verb tokens	9,065,201
Verb types	143,831
Prefixed verbs (types)	86,576
Verbs with 1 prefix (types)	81463
Verbs with 2 prefixes (types)	4894
Verbs with 3 prefixes (types)	219

Table 2. Statistics of occurrences of prefix combinations in types from Croatian National Corpus

The straightforward usage of such a procedure can be recognised easily: the enlargement of existing inflectional Croatian Morphological Lexicon with new lemmas and their word-forms or any Croatian monolingual dictionary with attested verbs. Also, with this approach a valuable pool of possible choices for terminologists is produced.

The results of this analysis also will provide the basis for further enlargement of the Croatian Wordnet. This enlargement can rely on verified data that otherwise cannot be determined. After their manual inspection, derived verbs can be treated either as hyponyms of non-prefixed verbs or as members of other lexical hierarchies when the meaning of derived forms differs significantly or totally from the base forms (e.g. *dati/to give – iz+dati/to betray* in one sense or *iz+dati/to release* in other). Since the verbs from the abovementioned second and the third group are members of the same synsets in the Croatian Wordnet, and the only semantic difference among them is iterativity of the denoted action, the described procedure should enable the simple detection of the synset members.

The next possible application of this procedure is enlargement of the Croatian Valency Lexicon (CRO-VALLEX) in terms of automatic population of its list of entries, but also in terms of the influence of derivational processes to possible change of argument structure between base and derived forms of verbs, as described in the next section.

Additional usages of this procedure could be found in different systems for language generation such as dialogue systems (with or without controlled language usage) and machine translation.

4. Evaluation

The automatic generation of possible valid combinations of verbal stems and prefixes in Croatian is currently being evaluated for the purpose of enlarging the Croatian Vallency Lexicon – CROVALLEX.

CROVALLEX contains 1739 verbs associated with 5118 valency frames. Those verbs were selected from 9500 in the Croatian frequency dictionary (Moguš et al., 1999) with their frequency higher than 11. Aiming to take all non-prefixed verbs in CROVALLEX as our starting point for lexicon enlargement, we distinguished 453 non-prefixed verbs out of 1739 verbs, while the number of non-prefixed verbs that are not in the current version of CROVALLEX, but have their prefixed stems in the lexicon is 112.

The valency 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 complement of the given verb. The type of valency relation for each complement is marked up as obligatory "obl" or typical optional "typ". 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 (Direction from-DIR1, Direction to-DIR3, etc).

We hypothesize that verbs sharing the same prefix (more precisely, the same meaning of the specific prefix) will also share the same valency frames and the number of obligatory complements. Furthermore, we also believe that we will be able to add typical optional complements to the valency frame of the verb in a more systematic way, since some of the prefixes usually bound the specific prepositional phrase, such as: *iz+trčati iz, iz+skočiti iz* (*ran from, jump from*) *od+dvojiti od, odlomiti od* (*separate from, break off from*), *pričvrstiti na, prilijepiti za* (*attach to, stick to*), etc.

Example: two different verbs, having the same prefix, share the same type of the prepositional phrase

- Ivan je istrčao iz kuće na ulicu
  AGT[1:obl] DIR1[iz+2:opt] DIR3[na+4:opt]
  (*Ivan ran from the house to the street*)
- Ptica je izletjela iz zaklona pred Ivana
  AGT[1:obl] DIR1[iz+2:opt] DIR3[pred+4:opt]
  (*The bird rushed out of the shelter in front of John*)

Also, one can prove the appearance of typical complements, such as DIR1-Direction from, with their surface forms (e.g. *iz+genitive case, s+ genitive, od+ genitive*) in the valency frames of the verbs with certain prefixes: e.g. ***iz-*** *(izroniti/emerge, iskočiti/jump out, izbiti/erupt, etc.),* ***s-*** *(spustiti/drop, srušiti/knock down, sjuriti se/plunge,* etc*.)* and ***od-*** *(odlutati/wander off odbjeći/elope, otkotrljati/roll off, etc*).

These combinations of prefixes and verbal stems should give 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.

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 act as a compensation for the lack of necessary information, representing the behaviour of each relevant verb.

Our basic idea is to group verbs according to their prefix to create semantically and syntactically coherent synsets. Members of the synset will share a range of features, starting with the implementation and interpretation of certain complements up to the existence of morphologically related forms. In this way, we could reduce the effort required to create lexicons and the likelihood of introducing errors while adding a new verb into the existing lexicon.

The version of CROVALLEX that is under construction will introduce explicit syntactic and semantic features of each verb synset. Synsets will be 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 should reflect the semantic components that limit the permissible complements and typical optional complements. Example:

(1) Two different verbs, sharing the same semantic prefix (or the same meaning of the same prefix, if prefix has more than one meaning), usually have the same valency frame. For example, prefix *iz-* in its meaning *"to separate from",* attached to different verbs will yield the same effect (compare *is+ključiti/remove, iz+dvojiti/separate)*:

- Trener ga je isključio iz igre
  AGT[1:obl] PAT[4:obl] DIR1[iz+2:opt]
  (*Coach removed him from the game*)
- Izdvojila je kukolj od pšenice
  AGT[1:obl] PAT[4:obl] DIR1[od+2:opt]
  (*She separated the tares from the wheat*)

(2) The change in prefix indicates the possible change in both obligatory verb complements and verb semantics (compare plakati/cry; o+plakati/mourn, is+plakati/cry)

- Marija je plakala
  ACT[1:obl]
  (*Marija cried*)
- Marija je **o**plakala Petra.
  ACT[1:obl] PAT[4:obl]
  (*Marija mourned Peter*)
- Marija je **is**plakala more suza.
  ACT[1:obl] PAT[4:obl]
  (*Marija cried sea of tears*)

As a result, every verb synset will be described by thematic roles (deep cases) and selection restrictions of its prefixed verbs. Every synset will also be defined by valency frames of its verbs, since they contain a set of syntactic descriptions.

Furthermore, we want to generalize about the behaviour of the most frequent prefixes and prefixed verbs in Croatian language, using the verb synsets. We plan to discover the prefixes that increase the number of verb arguments (valency-increasing prefixes), as well as prefixes that reduce the number of verb arguments (valency-reducing prefixes). Valency-increasing prefixes can further be divided into prefixes that add (usually only one) obligatory complement to a verb, and prefixes that add optional complement to a verb. If the added complement is obligatory, it changes the intransitive verb into transitive, while the optional complement represents an indirect object with the meaning of location, direction, recipient, etc. We want to reveal the connection between the valency-increasing and decreasing prefixes and the change in verb semantics, which is the case in the following examples.

Example:

(1) Valency-increasing prefix adds one obligatory verb complement to the stem verb and changes the verb semantics (compare: *pasti/fall – na+pasti/attack; sjesti/sit – prisjesti/get stuck; živjeti/live – proživjeti/experience*)

- Marija je pala
  ACT[1:obl]
  (*Marija fell*)
- Marija je napala Petra.
  ACT[1:obl] PAT[4:obl]
  (*Marija attacked Peter*)

On the other hand, valency-reducing prefixes reduce the number of obligatory verb complements by one, changing the transitive verb into intransitive.

(2) Valency-decreasing prefix removes the second obligatory verb complement of the stem verb and changes the verb semantics (compare: *brisati/wipel - s+brisati/run away; maknuti/move- umaknuti/scamper*)

- Marija je brisala pod.
  ACT[1:obl] PAT[4:obl]
  (*Marija wiped the floor*)
- Marija je zbrisala.
  ACT[1:obl]
  (*Marija ran away*)

We also want to determine a set of valency-neutral prefixes that only change the verb aspect, but do not change the valency of a verb. Valency-neutral prefixes do not add nor subtract the obligatory complements to the valency frame of a verb (e.g. prefix *is-* in *isprati/wash off*: both the stem verb *prati - to wash* and the prefixed verb *isprati- to wash off* have 2 obligatory complements). We hypothesize that valency-neutral prefixes make the largest set, followed by valency-increasing prefixes, but the

number of prefixes belonging to each set and the frequency are yet to be determined.

5. Future work

Work planned for the future includes testing of the all possible occurrences of generated verbal stems and their word-forms in a large Croatian Web Corpus - hrWaC (Ljubešić & Erjavec, 2011) that would give us the statistics and frequency of the existing occurrences in the largest existing Croatian corpus. Also, the simple series of scripts could be rewritten in a different environment such as NooJ.

Furthermore, the improved version of CROVALLEX will consists of verb synsets based on prefixation, instead of individual verb lemmas. The basic assumption in CROVALLEX is that prefixed verbs belonging to the same synset should have the same or similar complements with the same or similar morpho-syntactic form in their valency frame. The verbs in a synset will share the same meaning(s), but will not necessarily contain all aspectual counterparts, since it is not the rare case that aspectual counterparts differ semantically.

We also plan to discover the average number of verbs sharing the same prefix and belong to the same synset, in order to calculate the final number of verbs in the enriched CROVALLEX lexicon.

6. Conclusion

We have presented a procedure for generating prefixed verbs in Croatian comprising combinations of one, two or three prefixes. The result of this generation process is a pool of derivationally valid prefixed verbs although not necessarily occuring in corpora. The statistics of occurences in Croatian National Corpus has been calculated. Further usage of such language resource with generated potential verbs is also suggested, namely, enrichment of Croatian Morphological Lexicon, Croatian Wordnet and CROVALLEX.

7. Acknowledgements

This work was supported partially within the projects 130-1300646-0645, 130-1300646-1002 and 130-1301799-1999 funded by the Ministry of Science, Education and Sports of the Republic of Croatia and partially by ICT-PSP project CESAR, grant 271022, funded by European Comission.

8. References

Anić, V. (2003). *Veliki rječnik hrvatskoga jezika.* Zagreb: Novi liber.

Ćavar, D., Jazbec, I.-P., Runjaić, S. (2009) Efficient Morphological Parsing with a Weighted Finite State Transducer. *Informatica* 33(1), pp. 107--113

Halle, M. (1973). Prolegomena to a Theory of Word Formation. *Linguistic Inquiry*, 4, pp. 3--16.

Ljubešić, N., Erjavec, T. (2011) rWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In *Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011)*, Plzeň, Czech Republic : Lecture Notes in Artificial Intelligence 6836, Springer, pp. 395--402.

Mikelić Preradović, N. (2008). Pristupi izradi strojnog tezaurusa za hrvatski jezik. PhD thesis in information sciences. University of Zagreb, Faculty of Humanities and Social Sciences.

Moguš, M., Bratanić, M., Tadić, M. (1999) *Hrvatski čestotni rječnik*. Zagreb: Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu – Školska knjiga.

Oliver, A., Tadić, M. (2004) Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon: ELRA, pp. 1259--1262.

Raffaelli, I., Tadić, M., Bekavac, B. Agić, Ž (2008) Building Croatian WordNet. In: *Proceedings of the 4th Global WordNet Conference*, Szeged: Global WordNet Association, pp. 349—359

Šojat, K., Bekavac, B., Tadić, M. (2005) Zašto nam treba hrvatski Wordnet? In Granić, J. (ed) *Semantika prirodnog jezika i metajezik semantike*. Zagreb-Split: HDPL, pp. 733-743.

Tadić, M. (1994) Računalna obradba morfologije hrvatskoga književnoga jezika. PhD thesis in linguistics. University of Zagreb, Faculty of Humanities and Social Sciences.

Tadić, M. (2005) The Croatian Lemmatization Server. *Southern Journal of Linguistics*. 29 (1/2); pp. 206—217.

Tadić M. (2009). New version of the Croatian National Corpus. In Hlaváčková, D., Horák, A., Osolsobě, K., Rychlý, P. (eds.) *After Half a Century of Slavonic Natural Language Processing*. Brno, Masaryk University, pp. 199--205.

Tadić, M., Fulgosi, S. (2003) Building the Croatian Morphological Lexicon In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, Budapest: ACL, pp. 41-46.