Morphosemantic relations between verbs in Croatian WordNet

Krešimir Šojat
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
ksojat@ffzg.hr

Matea Srebačić
University of Zagreb, Croatia
msrebaci@unizg.hr

Abstract

This paper deals with morphosemantic relations between Croatian verbs and discusses their inclusion in Croatian WordNet. Morphosemantic relations refer to semantic relations between morphologically related verbs, i.e., between verbs from the same derivational family. A derivational family consists of verbs with the same lexical morpheme grouped around a base form. Generally, a verb with the simplest morphological structure serves as a base form for derivational processes. In Croatian, verbs are derived from base forms through prefixation and suffixation. Both derivational processes trigger aspectual and semantic changes. The focus is on semantic relations that regularly appear in various derivational families and consequently in various semantic fields. It is argued that these morphosemantic relations are crucial for the further development of Croatian WordNet.

1 Introduction

Croatian WordNet is a lexical database built through the so-called expand model (Vossen, 1998), i.e., by translating and adapting synsets from Princeton WordNet (further *PWN*) into Croatian. The building of Croatian WordNet (*CroWN*) can roughly be divided into two major phases. The first phase consisted of the translation and adaptation of the so-called basic concept sets from the multilingual projects EuroWordNet (*EWN*) and BalkaNet (*BN*) (cf. (Raffaelli et al., 2008). At present, CroWN contains 10,000 synsets. 8500 of these are from the basic concept sets of EWN and BN. Each synset was manually translated and provided with meaning definitions and usage examples. Synsets contain lexical units of the same part of speech. Since CroWN is a relatively small resource, the second phase of the project is primarily oriented toward its enlargement. Approximately 1500 noun synsets were added using the same procedure. This freely available version of CroWN contains 7391 noun synsets, 2318 verb synsets, and 310 adjective synsets.[1] As the numbers indicate, nouns make up almost 75% of the whole lexicon.[2] Such a strong predominance of this part of speech was a motivation to make CroWN a more balanced and representative resource for Croatian. The second phase of the project is primarily focused on enlarging the number of verbal synsets. However, this task required a re-examination of the building strategy applied so far.

2 Motivation

We decided to re-examine our building strategy for two reasons. The first pertains to differences between English and Croatian verbs that are more significant than was assumed. The second reason is an attempt to speed up the building of CroWN by using other available language resources for Croatian. As far as the first reason is concerned, the lexical hierarchies and word senses from PWN in numerous cases differ significantly from the lexical meaning, number of senses, and sense relations in their Croatian counterparts. For example, the verb *dati* appears in 28 synsets in CroWN (i.e. it is marked for 28 senses), but such a particularization of meaning is a consequence of the adopted expand model, and does not reflect its true semantic structure. Alt-

[1] CroWN can be downloaded from the following site: http://meta-share.ffzg.hr/repository/browse/croatian-wordnet/

[2] Similar situation is frequent in other wordnets. Maziarz et al. (2012) provide detailed statistics of POS distribution across major wordnets.

hough *dati* is a highly polysemous verb in Croatian, we found only 12 different senses of this verb listed in various monolingual dictionaries.[3] Apart from issues concerning conceptual systems and semantic representation, rich derivational processes between Croatian verbs bring about relations that cannot be captured by presently used semantic relations.[4] Verbs in Croatian are derived from other verbs by prefixation and suffixation. Both processes can trigger a change in aspect and the addition of a new semantic component to the base form. In accordance with Binnick (1991), we treat aspect as a grammatical feature, but predictable shifts in meaning, frequently referred to as *Aktionsarten*, as lexical features, pertaining to classes of verbs. This distinction is reflected in the structure of verbal synsets in CroWN. True aspectual pairs, i.e., imperfectives and perfectives denoting completion of an action, are members of the same synset. The lexical meaning of these perfectives differs from imperfectives only in this temporal distinction. Apart from aspectual change, semantic components brought by affixes can produce combinations that, in terms of meaning, can vary from compositional to completely idiosyncratic. E.g., the verb *crtati* 'to draw$_{ipf}$' has a true aspectual pair *nacrtati* 'to draw$_{pf}$', but there are six other prefixed perfectives as well: 1. *pre+crtati* 'to copy (by drawing)$_{pf}$', 2. *pod+crtati* 'to underline$_{pf}$', 3. *o+crtati* 'to outline$_{pf}$', 4. *is+crtati* 'to draw completely$_{pf}$', 5. *u+crtati* 'to draw into$_{pf}$', and 6. *za+crtati* 'to make a plan$_{pf}$'. The same prefixes can be used with other base forms, e.g. *pre+pisati* 'to copy (by writing)$_{pf}$', *pre+slikati* 'to copy (by painting)$_{pf}$. The base form *crtati* 'to draw$_{ipf}$' can also be suffixed, e.g. 1. *crt-k-ati* 'to draw$_{ipf}$, diminutive', 2. *crt-kar-ati* 'to draw$_{ipf}$, pejorative'.[5] Suffixes with diminutive and pejorative meanings can also combine with other base forms. We pose two basic questions: Which semantic regularities can be spotted in combinations of particular affixes and various base forms? and How can thereby established morphosemantic relations be used in our further work? In order to address these issues, as well as to speed up the building of CroWN, we have decided to use data from CroDeriV, a large morphological database of Croatian verbs. In the following section we shall briefly describe this resource (for a full description, see Šojat et al., 2012).

3 CroDeriV

CroDeriV is a computational lexicon containing data on the morphological structure of approximately 14 300 Croatian verbs collected from freely available dictionaries and corpora. The compiled verbal lemmas were analyzed for morphemes with a rule-based approach and the results were checked manually. Each lexical entry in CroDeriV consists of verbs decomposed into morphemes and linguistic metadata. The structure for all analyzed verbs consists of 11 morpheme slots and covers all combinations of recorded lexical and grammatical morphemes. There are four types of slots for morphemes: (1) derivational prefixes (four slots), (2) the lexical part (three slots – in the majority of cases only one is filled, the three slots are provided for verbal compounds of two roots and an interfix), (3) derivational and conjugational suffixes (three slots), and (4) infinitive ending (one slot). The metadata in lexical entries indicate verbal aspect and types of reflexivity.[6] The database enables queries across the full derivational span of a particular base form and provides extensive data about the distribution and frequency of affixes in the derivation of Croatian verbs.[7] In the following section, the underlying analysis of affixal meanings is described.

4 Affixal Meanings

The majority of verbal prefixes in Croatian developed from prepositions, and the original locative component pervades in their meaning. However, they are highly polysemous units and in various combinations they can differently modify the meaning of base forms. For example, the

[3] More examples can be found in Šojat et al. (2012: 111 – 112).

[4] We use the same semantic relations between verbal synsets as in EWN and BN. These relations are synonymy, hyponymy/hypernymy, antonymy, cause, and subevent.

[5] Suffixes *-k-* and *-kar-* occupy the first position on the right side of the verbal root *crt-*. The full morphological analysis of the verbs *crtati, crtkati* and *crtkarati* is thus: *crt-ø-ø-a-ti*, *crt-k-ø-a-ti* and *crt-kar-ø-a-ti* (cf. Section 3).

[6] CroDeriV resembles databases like CatVar for English (`http://clipdemos.umiacs.umd.edu/catvar`) and Unimorph for Russian (`http://courses.washington.edu/unimorph/index.html`).

[7] For data on the productivity and frequency of affixes in Croatian, see Šojat et al. (2013).

verbal prefix *na-* 'on' can have at least eight different meanings (divided further into several subgroups) in combinations with various base forms:
1) pure aspectual meaning: *pisati* 'to write$_{ipf}$' – *napisati* 'to write$_{pf}$'
2) locative meanings:
    a. top-down: *baciti* 'to throw$_{pf}$' – *nabaciti* 'to throw onto$_{pf}$'
    b. proximity: *letjeti* 'to fly$_{ipf}$' – *naletjeti* 'to bump into$_{pf}$'
    c. putting something on something: *slagati* 'to pile$_{ipf}$' – *naslagati* 'to pile one on another$_{pf}$'
3) inchoativity: *trunuti* 'to rot$_{ipf}$' – *natrunuti* 'to begin to rot$_{pf}$'
4) distributivity: *bacati* 'to throw$_{ipf}$' – *nabacati* 'to throw one by one$_{pf}$'
5) sufficiency: *jesti* 'to eat$_{ipf}$' – *najesti se* 'to stuff oneself$_{pf}$'
6) excessiveness: *piti* 'to drink$_{ipf}$' – *napiti se* 'to get drunk$_{pf}$'
7) addition: *gomilati* 'to accumulate$_{ipf}$' – *nagomilati* 'to accumulate a lot of X$_{pf}$'
8) intensitiy:
    a. low intensity: *gristi* 'to bite$_{ipf}$' – *nagristi* 'to bite a bit$_{pf}$'
    b. high intensity: *pisati* 'to write$_{ipf}$' – *napisati se* 'to tire oneself with writing$_{pf}$'

All 19 verbal prefixes recorded in CroDeriV were analyzed in the same manner.[8] This analysis enabled the recognition of the same or similar semantic components shared by different prefixes as well as the division of prefixal meanings into four major semantic groups. The four major groups of prefixal meanings are location, time, quantity, and manner. Each group has several subgroups. An analysis of suffixal meanings yielded an additional semantic group of diminutive and pejorative verbs.[9] Before further discussion, we shall briefly present the morphosemantic relations between verbs in other Slavic wordnets and compare them with the relations used in CroWN.

5 Related Work

Rich derivational morphology in Slavic languages and problems faced in the building of Czech, Bulgarian, and Serbian wordnets are discussed in Pala and Hlaváčková (2007), Koeva (2008), and Koeva et al. (2008). The discussion refers mainly to derivational relations across different parts of speech. Pala and Hlaváčková (2007) list 14 derivational processes in Czech introduced into Czech WordNet as relations between derived and base forms. This results in a "two-level network", where the higher level includes semantic relations between synsets, and the lower level includes derivational relations between single synset members. Although the verb-verb pairs are linked through prefixation, this relation is not used in further analysis. Koeva (2008) points out the relation between verbal aspectual pairs as the most productive derivational relation in Bulgarian and argues that perfective and imperfective verbs in Bulgarian WordNet should be split into separate synsets. While the hypernymy would be based on imperfective verbs only, synsets would be linked with the morphosemantic relation aspect. Relations between prefixed derivatives and base forms, apart from aspectual, are not discussed. The work presented in Koeva et al. (2008) concerning Serbian WordNet refers mainly to derivational relations across different parts of speech. Aspectual pairs in Serbian WordNet are members of the same synset. The most elaborate account of relations between verbs in Slavic is given in Maziarz et al. (2011) and Maziarz et al. (2012). In Polish WordNet 2.0 aspectual pairs are kept apart and lexical hierarchies consist of either perfective or imperfective verbs. Relations between verbs are divided into purely semantic relations (e.g., synonymy, hyponymy, meronymy holonymy, antonymy, processuality, causality, inchoativity, presupposition, and preceding) and derivationally-motivated relations (e.g., pure aspectuality, secondary aspectuality, iterativity, and derivationality). Some of the relations hold between lexical units (word-sense pairs, e.g., antonymy or pure aspectuality), while others hold between synsets (e.g., hyponymy and processuality). In CroWN, pure aspectual pairs are members of the same synset. Pure aspectual pairs are determined primarily by the test of secondary imperfectivization (cf. Jelaska, 2005; Maziarz et al., 2011), but also by additional criteria pertaining to semantics of affixes. The relation of pure aspectuality exists between a base form and a derivative with an affix which does not contain any other semantic components except perfectiveness, e.g. *pisati* 'to write$_{ipf}$' – *napisati* 'to write$_{pf}$' are members of the same synset. The same holds for iterative verbs

[8] These prefixes are: *do-*, *iz-*, *na-*, *nad-*, *o-/ob-*, *obez-*, *od-*, *po-*, *pod-*, *pre-*, *pred-*, *pri-*, *pro-*, *raz-*, *s-*, *su-*, *u-*, *uz-*, and *za-*.

[9] An analysis of prefixal meanings is given in Šojat et al. (2012); suffixal meanings are discussed in Šojat et al. (2013).

and perfective base forms. Although iterative verbs have the additional semantic component of repetitiveness, they differ from their perfectives only in this temporal component. E.g., *pisati* 'to write$_{ipf}$' – *prepisati* 'to copy by writing$_{pf}$' are not members of the same synset. However, *prepisati* 'to copy by writing$_{pf}$' – *prepisivati* 'to copy by writing$_{ipf\text{-}iter}$' are members of the same synset. Each synset member is tagged with one of the following aspect labels: IPF, PF, BI, or ITER, representing imperfective, perfective, bi-aspectual and iterative forms. This distinction is also reflected in different aspectual forms used in definitions, although they are structurally and semantically the same. Finally, all morphosemantic relations in CroWN discussed below hold between single members of synsets, i.e. lexical units, and not between whole synsets.

6 Morphosemantic Relations in CroWN

Morphosemantic relations are based on overlapping components of affixal meanings in combinations with various base forms. The analysis described in Section 4 enabled the classification of affixal meanings into four broad semantic groups for prefixes and one for suffixes. Four major groups for prefixes – location, time, quantity, and manner – are further divided into subgroups (28 in total). Morphosemantic relations and a variety of sub-relations based upon this classification are listed below:

1. PREFIXES:
   a) **location group**: bottom-up, top-down, proximity, through, apart, to/towards, over, into, around, under, re-location, behind, across, from
   b) **time group**: inchoativity, finitiveness, distributivity, preceding
   c) **quantity group**: sufficiency (+/-), excessiveness, intensity (+/-), exceeding, deprivation, addition
   d) **manner group**: inter-connection, change of property.

SUFFIXES:
   a) **diminutive group**: diminutive, pejorative

As far as the semantic impact of prefixes is concerned, relations in the *location* group predominantly hold between verbs of movement, but also between numerous other base verbs and derivatives with spatial relations pervading their lexical meaning (e.g., *udahnuti* 'to inhale' or *uvući* 'to drag into'). Derivatives in *time* group refer to different phases of actions denoted by base verbs (e.g., beginning or termination). The subtype *distributivity* is on the border between the *time* and *quantity* groups since the derivatives denote repetitive actions performed by one or more agents on one or more objects. Since distributive actions are performed iteratively, this relation is listed in the *time* group. Relations from the *quantity* group hold when derivatives denote various degrees of an action (e.g. *naraditi se* 'to tire oneself out (with work)', *najesti se* 'to eat one's fill'). The smallest group – *manner* – contains only two relations denoting changes of properties (e.g., *uprljati se* 'to become dirty') and actions performed in a specific manner (e.g. *sufinancirati* 'to co-finance'). The semantic impact of suffixes is limited to diminutive and pejorative meaning expressed by derivatives (e.g., *jeduckati* 'to nibble'). The aim of this classification is to establish the set of morphosemantic relations and use them within derivational families of verbs in CroWN. To determine which verbs are derivationally related and therefore are candidates for further analysis, we compared the list of verbs from CroWN and CroDeriV. All verbs from the 2318 verbal synsets in CroWN are recorded in CroDeriV.

The full list of verbs from CroWN was filtered into those sharing the same root. The list of verbs recognized as derivatives comprises 2530 base forms and prefixed derivatives. This list was further filtered for verbs marked as aspectual pairs in CroWN. In the next step, prefixed forms were segmented into prefixes and base forms. Thus we obtained 572 base forms and 1476 derivatives as candidates for the assignment of morphosemantic relations. In the final step, the relations were manually assigned to derivationally related verbs from CroWN. When no morphosemantic relation was appropriate due to the idiosyncratic nature of the combinations, we tagged this relation as DERIV (144 verbs).

The result of the whole procedure is a list of 572 base forms and 1186 prefixed verbs marked for morphosemantic relations. There are also 19 lexical units marked as diminutives in CroWN. In CroDeriV, derivational suffixes for diminutives always occupy the first slot to the right of the root (cf. Section 3). Table 1 contains the overall frequency of relations in four major groups as well as the frequency of the three most prominent subrelations for prefixed derivatives (*manner* contains only two subrelations). The last row indicates the frequency of suffixed derivatives.

Group	Freq	Subgroup	Freq
Loc	598	loc_apart	141
		loc_around	87
		loc_from	70
Time	276	time_fin	132
		time_inch	109
		time_distr	28
Quan	190	quan_int	126
		quan_exc	25
		quan_suff	20
Mann	122	mann_prop	88
		mann_conn	34
Dim	19	pejorative	8

Table 1: Frequency of MS relations in CroWN

7 Discussion and Future Work

None of the discussed morphosemantic relations between members of different synsets can be completely subsumed by any of semantic relations between synsets in terms of semantic content. Base verbs and their derivatives are frequently not members of same lexical hierarchies, such as synsets containing derivationally related verbs like *ići* 'to go', *ući* 'to go into', *izaći* 'to go out' and *otići* 'to go away'. Although the verbs *ući* and *izaći* are marked as antonyms, the relatedness of the whole group is not indicated. The relation cause partially overlaps with our morphosemantic relation change of property, but cause cannot encompass reflexive non-agentive counterpart pairs of transitive verbs in Slavic (e.g. *topiti se – otopiti se* 'to melt$_{ipf\text{-}pf}$' *– to become soft or liquid* vs. *topiti – otopiti* 'to melt$_{ipf\text{-}pf}$' *– to cause to become soft or liquid*. The relation of subevent refers to two simultaneous actions or to an action which is a part of the action denoted by another synset, but it does not reflect particular parts of events, such as its beginning or terminating point, as morphosemantic relations of inchoativity or finitiveness do. In order to capture the semantic relatedness between verbs usually scattered across different hierarchies, we have introduced a set of 28 morphosemantic relations. This "two-level network," as defined by Pala and Hlaváčková (2007), along with extensive data from CroDeriV, provides an excellent basis for further work. Although CroDeriV does not contain data about lexical meaning and semantic relations between verbs, information about the morphological structure of verbs proved valuable for the detection of derivationally related verbs and the assignment of morphosemanic relations. Information about complete derivational families is also valuable for the further expansion of CroWN, which is one of our primary goals. It can be used both to complete already present derivational families and to introduce new ones. Finally, the importance of morphosemantic relations in the description of the verbal system in Croatian can be demonstrated with the Croatian verb *gristi* 'to bite$_{ipf}$'. This verb appears in CroWN only in this form, whereas CroDeriV contains ten other derivatives from this base form. Only the derivative marked with the relation DERIVED in Figure 1 below can be straightforwardly connected to other synsets in CroWN via semantic relations. All other derivatives, i.e., 90% of this derivational family, should be connected to this base form primarily by morphosemantic relations as described here. Although the full set of morphosemantic relations as discussed here provides a more densed and fine-grained structure of the Croatian lexicon, we are aware that in numerous cases it is hard to maintain the consistency and clear-cut distinctions among 28 presented relations. However, we are convinced that even a set of morphosemantic relations limited to four major groups of prefixed derivatives (location, time, quantity and manner) and one group of suffixed derivatives (diminutive/pejorative) can substantially enrich wordnets for Slavic languages.

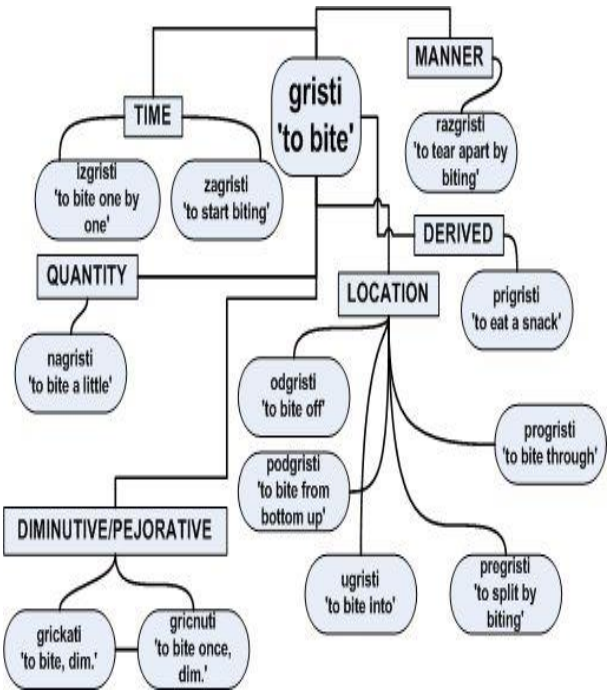

Figure 1: MS relations across a derivational family

Acknowledgements

The research was supported by MZOS RH project 130-1300646-1002 and partially by XLike project (FP7, Grant 288342).

References

Robert I. Binnick. 1991. *Time and the Verb: A Guide to Tense & Aspect*. Oxford University Press, Oxford, UK.

Zrinka Jelaska. 2005. *Hrvatski kao drugi i strani jezik.* Hrvatska sveučilišna naklada: Zagreb.

Svetla Koeva. 2008. Derivational and Morphosemantic Relations in Bulgarian WordNet. *Proceedings of the Intelligent Information Systems 2008*, 359–368.

Svetla Koeva, Cvetana Krstev and Duško Vitas. 2008. Morpho-semantic Relations in WordNet – a Case Study for two Slavic Languages. *Proceedings of the 4th Global WordNet Conference*, 239–254.

Marek Maziarz, Maciej Piasecki, Stanislaw Szpakowicz, Joanna Rabiega-Wisniewska, Bozena Hojka 2011. Semantic Relations between Verbs in Polish WordNet 2.0. *Cognitive studies*, 11:183–200.

Marek Maziarz, Maciej Piasecki and Stanisław Szpakowicz. 2012. An Implementation of a System of Verb Relations in plWordNet 2.0. *Proceedings of the 6th Global WordNet Conference*, 181–188.

Karel Pala and Dana Hlaváčková. 2007. Derivational Relations in Czech WordNet. *Proceedings of the Workshop on Balto-Slavonic Languages*, 75–81.

Ida Raffaelli, Marko Tadić, Božo Bekavac, Željko Agić 2008. Building Croatian WordNet. *Proceedings of the 4th Global WordNet Conference*, 349-360.

Krešimir Šojat, Matea Srebačić and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, 0 (1): 111-142.

Krešimir Šojat, Matea Srebačić and Vanja Štefanec. 2013. CroDeriV i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika*, 39 (75): 75 – 96.

Piek Vossen. (Ed.) 1998. *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dodrecht, Boston, London.