CroDeriV: a new resource for processing Croatian morphology

Krešimir Šojat*, Matea Srebačić[+], Tin Pavelić[+], Marko Tadić*
*Department of Linguistics, Faculty of Humanities and Social Sciences, [+]University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
E-mail: ksojat@ffzg.hr, msrebaci@unizg.hr, tpavelic@ffzg.hr, mtadic@ffzg.hr

Abstract

The paper deals with the processing of Croatian morphology and presents CroDeriV – a newly developed language resource that contains data about morphological structure and derivational relatedness of verbs in Croatian. In its present shape, CroDeriV contains 14 192 Croatian verbs. Verbs in CroDeriV are analyzed for morphemes and segmented into lexical, derivational and inflectional morphemes. The structure of CroDeriV enables the detection of verbal derivational families in Croatian as well as the distribution and frequency of particular affixes and lexical morphemes. Derivational families consist of a verbal base form and all prefixed or suffixed derivatives detected in available machine readable Croatian dictionaries and corpora. Language data structured in this way was further used for the expansion of other language resources for Croatian, such as Croatian WordNet and the Croatian Morphological Lexicon. Matching the data from CroDeriV on one side and Croatian WordNet and the Croatian Morphological Lexicon on the other resulted in significant enrichment of Croatian WordNet and enlargement of the Croatian Morphological Lexicon.

Keywords: computational morphology, CroDeriV, derivational lexicon

1. Introduction

This paper deals with the processing of Croatian morphology and presents CroDeriV – a newly developed language resource. CroDeriV is a database which contains data about the morphological structure and derivational relationships of lemmas in Croatian. As in other Slavic languages, Croatian morphology is rich in both inflectional classes and derivational processes. Whereas inflectional classes are covered extensively by the Croatian Morphological Lexicon (Tadić, 1994, 2005), only a limited number of derivational processes has been taken into account in the computational processing of Croatian, mostly through the development of rule-based stemmers for various information extraction tasks (Ljubešić et al., 2007; Šnajder et al., 2009). A comprehensive approach to the processing of inflectional as well as derivational phenomena is discussed by Ćavar et al. (2008, 2009), who describe *CroMo*, a tool used for both the morphological analysis and the lemmatization of Croatian. As reported, this finite state lexical transducer is based on a database of ca. 250,000 lexical, derivational, and inflectional morphemes. Unfortunately, neither the database nor the tool, are publicly available, nor can they be used elsewhere. During the building of various language resources and tools for Croatian, it has become obvious that a large-scale and publicly available derivational database could significantly speed up their development and improve their quantity as well as quality. At the same time it has also become obvious that such a language resource would provide an excellent foundation for linguistic research on derivational families and morphosemantic relations between words in Croatian.

2. Motivation

Therefore, the main motivation for building a new and extensive database with morphological information about Croatian lexica is the limited scope of derivational problems tackled by existing tools and resources as well as their public unavailability. The original impetus to build CroDeriV comes from the work that has been done on Croatian WordNet (Šojat, 2012), particularly on its verbal part (Šojat & Srebačić, 2014). This work has indicated that the analysis of the morpheme structure and derivational relatedness of a large sample of verbs is a prerequisite for an elaborate and theoretically justified account of their semantic and aspectual relations. In turn, this resulted in verbs being the biggest and best covered lexical category in CroDeriV. Further in this paper we therefore focus on the lexical category of verbs in Croatian, while the full processing of other POS is the next objective in this ongoing project. This paper is structured as follows: in Section 3 we present the design and structure of CroDeriV in its present form. In Sections 4 and 5 we describe the application of CroDeriV to other Croatian language resources and give an evaluation of these experiments. Section 3 deals with the enrichment of Croatian WordNet, and Section 4 with the enlargement of the Croatian Morphological Lexicon. The final part of the paper provides an outline for future work.

3. The design of the CroDeriV database

CroDeriV is a computational lexicon that contains data on the morphological structure of approximately 14 200 Croatian verbs collected from different sources, predominantly machine readable and paper dictionaries (Anić, [4]2004; Šonje, 2000) and further enriched with lemmas from the Croatian National Corpus v3.0[1] (Tadić, 2009) and the Croatian Web Corpus v2[2] (Ljubešić &

[1] http://hnk.ffzg.hr
[2] http://nlp.ffzg.hr/resources/corpora/hrwac

Erjavec, 2011). The primary purpose of this database is to collect in one resource all Croatian verbs detected in the sources mentioned above. The general purpose of such a resource is to obtain a complete morphological analysis of Croatian vocabulary. As mentioned, CroDeriV contains only verbal lemmas at its current stage of development, whereas other POS will be added later. 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.[3] Verbs in Croatian can be derived from other verbs by prefixation and suffixation. Of these two derivational processes, prefixation is far more productive than suffixation.[4] Prefixes are always derivational, whereas suffixes can be derivational and inflectional. In the vast majority of cases, base forms take one prefix. In terms of the morphological structure presented in Table 1 below, one verbal root (**r_1**) as a part of a base form can co-occur with as many as four prefixes (**p_4-p_1**), although this occurs extremely rarely.[5] As far as suffixes are concerned, one root usually has two derivational (**s_2-s_1**) and one inflectional suffix (**-ti**). This structure can be extended with an additional suffix (**s_3**) denoting a diminutive or pejorative action. On top of that, verbs in Croatian can also be formed by compounding, i.e., they can consist of two roots (**r_2-i-r_1**). Thus, the determined maximal morphological structure of Croatian verbs, based on the analysis of a large dataset, is as shown in Table 1.

Prefixes	**Roots**	**Suffixes**
(p_4-p_3-p_2-p_1)	**(r_2-i)-r_1**	**(s_3)-s_2-s_1-*ti***
	pis-	*-a-ti*
'to write$_{ipf}$'		
na-	*-pis-*	*-a-ti*
'to write$_{pf}$'		
pot-	*-pis-*	*-a-ti*
'to sign$_{pf}$'		
pot-	*-pis-*	*-iv-a-ti*
'to sign$_{ipf}$'		
is-pot-	*-pis-*	*-a-ti*
'to sign one by one$_{pf}$'		
is-pot-	*-pis-*	*-iv-a-ti*
'to sign one by one$_{ipf}$'		

Table 1: The morphological structure of Croatian verbs (p = prefix, r = root, i = interfix, s = suffix, () = optional, -ti = infinitive ending).

The maximal morphological structure of compiled verbal lemmas was therefore determined in several steps. Due to numerous and frequent phonological changes all prefixal combinations were manually segmented in the first step. In the second step a rule-based morpheme splitter was applied to the suffixal part. Suffixes in Croatian infinitives can be divided into two broader groups: 1) non-obligatory suffixes, 2) obligatory suffixes. Only one non-obligatory suffix per lemma is possible and it always occupies the first slot on the right side of a stem.[6] In this case obligatory suffixes follow non-obligatory suffixes. Otherwise, they are attached directly to verbal roots. Obligatory suffixes can be divided into 1) suffixes denoting verbal aspect, 2) suffixes denoting conjugational classes, 3) infinitive ending *-ti* or *-ći.* Suffixes denoting verbal aspect and conjugational classes are combined in a predictable manner thus enabling a more accurate design and application of rules. However, the two main problems in the automated processing of Croatian derivation are (1) homography that results in the overlapping of prefixes and suffixes with roots, and (2) numerous phonological changes at the morpheme boundaries resulting in several allomorphs for each morpheme. All results of automatic segmentation were therefore manually checked, and at the same time, all allomorphs, both affixal and lexical, were connected to one representative morpheme. Evaluation of the automatic processing and manual checking is presented in Table 2.

Processed verbs	14 463
Corrected verbs	10 163
Splitter precision	29,73%

Table 2: Precision of automatic processing

This kind of processing enables (1) the recognition of all allomorphs of a particular morpheme and (2) the detection of all affixes that co-occur with particular roots. This procedure further enables the detection of complete derivational families of verbs in Croatian. 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 verb-to-verb derivation. In other words, this procedure enables the detection of full derivational spans of particular verbal roots. Data structured in this way can be used in the enlargement and enrichment of other resources, which is shown in following sections.

4. CroDeriV and Croatian WordNet

As mentioned above, the primary motivation to build CroDeriV database appeared in the process of building Croatian WordNet[7] (Raffaelli et al., 2008, Šojat, 2012). Croatian WordNet (CroWN) is a lexical database that so far has been built by translating and adapting synsets from Princeton WordNet. Since nominal synsets make up almost 75% of the entire CroWN, our goal in the next phase of its development is to make it a more balanced and representative language resource for Croatian,

[3] An extensive statistics of all recorded combinations of affixes and roots is given in Šojat et al. (2012, 2013).

4 Nouns, on the contrary, are mostly derived by suffixation.

[5] Combinations of four prefixes are recorded in only two cases.

[6] Non-obligatory suffixes usually have diminutive or pejorative meaning, e.g. suffix *-uš-* in *pjev-**uš**-ø-i-ti* 'to hum' (< *pjev-**ø**-ø-a-ti* 'to sing').

[7] Available also in META-SHARE (http://www.meta-share.eu).

primarily by enlarging the number of verbal synsets. The main problem faced during our work on verbs in CroWN is the question how to account for verbal aspectual pairs as well as other derivationally related verbs in Croatian. In order to adequately address these issues, the detection of full derivational spans of base verbs in terms of their possible prefixed and suffixed derivatives has turned out to be necessary. This is particularly important because, in the majority of cases, the relations between derivationally related Croatian verbs cannot be defined as semantic relations as hyponymy and various types of entailments used for linking verbal synsets. [8] Semantic components of verbal derivatives mainly induced by prefixes, such as repetitiveness, distributiveness, beginning or termination of an action, various degrees or quantities, etc. are usually referred to as *Aktionsart* and are typical of Slavic languages. While these semantic components are part of the lexical meaning of verbal lemmas in Croatian, they are usually expressed with particles, phrasal verbs or additional lexical units in English. For example, the verb *zapjevati* 'to start singing' (< *pjevati* 'to sing') is expressed as a single lemma in Croatian, but as a multi-word unit in English. These types of relations, as the one between the base verb *pjevati* 'to sing' and its derivatives, e.g. *zapjevati* 'to start singing', *otpjevati* 'to finish singing', cannot be captured by the semantic relations that are typically used between verbal synsets in CroWN. An additional problem is that the relations between derivationally related verbs cannot be established between verbal synsets, because synsets generally contain verbs that do not share the same lexical morpheme. Due to the expand model (Vossen, 1998) used in the building of CroWN, derivatives denoting temporal or spatial modifications, for example, of base forms are consequently not included. Morphosemantic relations between such verbs are therefore significantly under-represented in CroWN, resulting in a shallow lexical and semantic structure of the verbal part of the lexicon. In order to overcome this deficiency, we explored the possibility of introducing the data from CroDeriV to CroWN. The aim of this procedure is to enrich derivational families of verbs in CroWN according to the derivational relatedness of Croatian verbs. As mentioned earlier, our goal is also to speed up the building of this resource and to enlarge its structure. For this reason we have measured the intersection of verb lists from CroWN and CroDeriV. In the first step we extracted all verbs from verbal synsets in CroWN and removed all metadata, i.e., definitions, usage examples etc., and then matched the obtained list with the list of verbs from CroDeriV. The resulting measures are shown in Table 3.

cov(CroDeriV/CroWN)	98,56%
cov(CroWN/CroDeriV)	32,66%

Table 3: Coverage CroDeriV - CroWN

[8] We use the same semantic relations between verbal synsets as in EuroWordnet and BalkaNet: synonymy, hyponymy/ hypernymy, antonymy, cause, and subevent.

The table indicates that CroDeriV covers almost all verbs treated as morphological types from CroWN. However, more than 2/3 of all verbs in CroDeriV are not included in verbal synsets in CroWN. These numbers show how useful is CroDeriV in the enlargement of CroWN. To further facilitate this procedure, we determined the word families in CroDeriV that contain at least one verb from CroWN. The derivational families were extracted from CroDeriV via mutual lexical morphemes and the verbs from CroWN were labelled accordingly. Other verbs from the detected derivational families were then easily incorporated in CroWN via set of already established morphosemantic relations. [9]

5. CroDeriV and the Croatian Morphological Lexicon

Lexica with morphological information are usually central components of various NLP tools as e.g. lemmatizers and POS/MSD taggers. The Croatian Morphological Lexicon (HML) v5.0., which comprises ca. 120,000 lemmas and all their inflectional forms (over 5 million word forms), can be used in its on-line version[10] both as a lemmatizer and as a generator of inflected forms. The HML is also used as the basis for the CroTag system for the morphosyntactic tagging of texts compliant with the MulTextEast recommendations v4.0. (Agić et al, 2008). In this section we present the automatic merging and expanding of the HML with the data from CroDeriV. In the first step we examined the coverage of lemmas in both resources. The results are presented in Table 4.

	Total numbers
Verbs - HML	8 286
Verbs - CroDeriV	14 192
CroDeriV / HML	6 783
HML / CroDeriV	194
	%
cov(CroDeriV/HML)	**97,66%**
cov(HML/CroDeriV)	**57,01%**
HML enlargement	**81,8%**
CroDeriV enlargement	**1,36%**

Table 4**:** Coverage CroDeriV - HML

The results indicate that a large set of lemmas from CroDeriV is not listed in the HML. In order to include them in the HML we extracted them from CroDeriV and automatically assigned their inflectional patterns. This procedure is based on the same grounds as the one described in the previous section, i.e., the verbs from the HML were matched with those from CroDeriV via mutual lexical morpheme and classified according to derivational families in CroDeriV. This procedure further enabled the

[9] Cf. Šojat et al. (2012, 2013) and Šojat & Srebačić (2014) for an exhaustive list of possible verb-to-verb morphosemantic relations in CroWN.

[10] http://hml.ffzg.hr

automatic assignment of inflectional classes to verbs belonging to the same derivational families. However, the assignment of inflectional patterns was possible only in cases when the base verbs were already included in the HML. For example, if the HML contains the verb *hodati* 'to walk', but does not contain its derivative *pre+hodati* 'to walk over', the verb *prehodati* is assigned the inflectional pattern of *hodati* based on the derivational relation between them via the shared root, only adapted for verbal aspect.[11] The word forms of lemmas with the assigned inflectional patterns (Tadić, 1994) can thus be easily generated and incorporated into the HML. In cases when the HML does not contain a particular base verb, the inflectional pattern has to be assigned manually, since there is no base form to serve as a role model for derivatives. Moreover, since CroDeriV enables the recognition of all verbs which share the same stem, inflectional class can be assigned to new verbs if at least one verb with the same stem already exists in HML (cf. Figure 1). The additional output of the experiment is the recognition of lemmas in the HML that are not listed in their full lexical form. This pertains to the particle *se*, which has so far been omitted from the HML. Verbal lemmas are listed without this element, since the tools for the tokenization, lemmatization and MSD-tagging of Croatian currently operate at unigram level.

6. Future work

One of the objectives in our future research is to determine the frequency of lexical and derivational morphemes as well as derivational patterns from CroDeriV using the Croatian National Corpus (HNK). The CroDeriV database will be matched with the HNK v3.1 in order to obtain the data on distribution of lexical and derivational morphemes from this representative corpus. This is important since the information about the frequency and the distribution of morphemes in CroDeriV is based on the fact that each verbal lemma is recorded only once in the database. Although CroDeriV provides the information about the productivity of lexical or derivational morphemes and their combinations, it cannot provide the information about lemmas and derivational patterns in real language, i.e. in corpora. With this kind of information we will be able to calculate the ratio between the recorded productivity of morphemes in CroDeriV and their real occurrence in texts. This kind of information could be eventually inserted into CroDeriV.

7. Conclusion

In this paper we have presented a new resource for processing Croatian morphology – CroDeriV. In its present form it contains only verbs, but even in this phase of its development it is a valuable resource for various linguistic investigations, as well as for the enrichment and improvement of other resources for Croatian. It will be freely available through the META-SHARE[12] platform

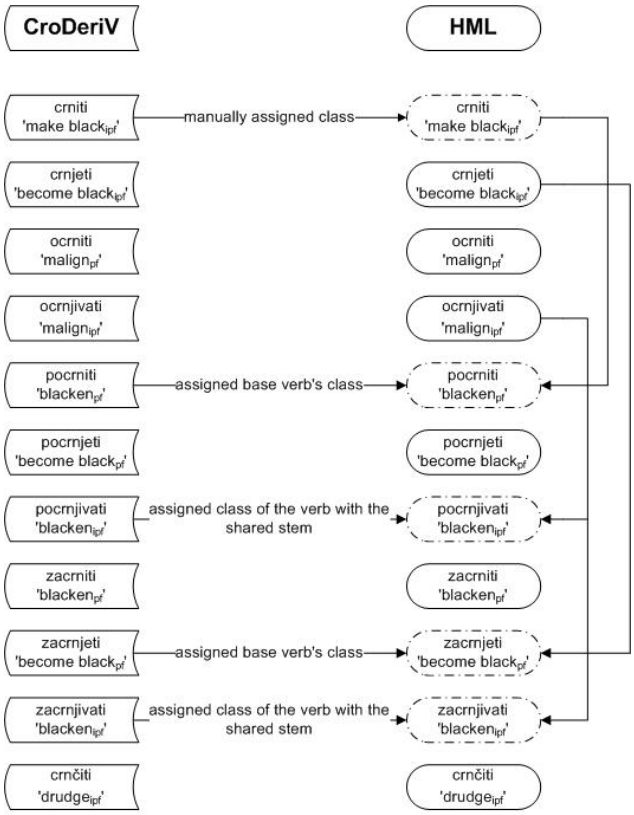

Figure 1: Assignment of inflectional classes

and it is currently available for queries at: http://croderiv.ffzg.hr/. Having in mind the typological closeness of Slavic languages, such a design of a derivational database could be applied to other Slavic languages with relatively simple adaptations. Such databases would enable their cross-lingual comparison in terms of derivational patterns and morphological structure. The information about the morphological structure of words can be used in language generation, question answering, information extraction and machine translation tasks. In this paper we have shown how a new resource can improve existing tools and resources, but also how it can stimulate the development of new ones. The data structured as presented in this paper can also be used for linguistic research of derivational processes, aspectual relations and *Aktionsart* in Croatian. The development of the database is an ongoing project and new lemmas of other parts of speech will be added in the next stage.

8. Acknowledgements

The research that lead to this results was partially supported by the CESAR project (ICT-PSP, Grant 271022) and the XLike project (FP7, Grant 288342), as well as MZOS RH project 130-1300646-0645

9. References

Agić, Željko; Tadić, Marko; Dovedan, Zdravko (2008) Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica. 32, 4;

[11] Verb *hodati* is imperfective and *prehodati* perfective.
[12] http://www.meta-share.eu

pp 445-451.

Anić, Vladimir (2004) Rječnik hrvatskoga jezika, Novi liber, Zagreb, 4th edition.

Ćavar, D., Jazbec, I., Runjaić, S.: Interoperability and Rapid Bootstrapping of Morphological Parsing and Annotation Automata. In: Erjavec, T., _Zganec, G., Jerneja (Eds.) Proceedings of the Sixth Language Technologies Conference, October 16th-17th, 2008 : proceedings of the 11th International Multiconference Information Society - IS 2008, volume C, pp. 80{85. Institut Jožef Štefan, Ljubljana (2008)

Ljubešić, Nikola; Boras, Damir; Kubelka, Ozren (2007) Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. In: Seljan, Sanja; Stančić, Hrvoje (eds.) InFuture 2007: Digital Information and Heritage, Faculty of Humanities and Social Sciences, Zagreb, pp 313-320.

Ljubešić, Nikola; Erjavec, Tomaž (2011) hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In: Proceedings of the 14th International Conference Text, Speech and Dialogue (TSD2011), Plzeň, Czech Republic, 1-5 September 2011, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp 395-402.

Raffaelli, Ida; Tadić, Marko; Bekavac, Božo; Agić, Željko (2008) Building Croatian WordNet. In: Tanács, Attila; Csendes, Dóra; Vincze, Veronika; Fellbaum, Christianne; Vossen, Piek (eds.) Proceedings of the 4th Global WordNet Conference, Global WordNet Association, Szeged, pp 349-359.

Šojat, Krešimir (2012) Struktura glagolskog dijela Hrvatskog WordNeta. Filologija 59; pp 153-172.

Šojat, Krešimir; Srebačić, Matea; Tadić, Marko (2012) Derivational and Semantic Relations of Croatian Verbs. Journal of Language Modelling. 00, 1; pp 111-142.

Šojat, Krešimir; Srebačić, Matea; Štefanec, Vanja (2013) CroDeriV i morfološka raščlamba hrvatskoga glagola. Suvremena lingvistika. 39, 75; pp 75-96.

Šojat, Krešimir; Srebačić, Matea; Pavelić, Tin; Tadić, Marko (2013) From Morphology to Lexical Hierarchies. Vetulani, Zygmunt (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics (LREC 2013 Proceedings). Poznanj: Fundacja Uniwersytetu im. A. Mickiewicza, pp 474-478.

Šojat, Krešimir; Srebačić, Matea (2014) Morphosemantic relations between verbs in Croatian WordNet. In: Orav, Heili, Fellbaum, Christiane, Vossen, Piek (eds.) Proceedings of the Seventh Global WordNet Conference. Tartu : GWA, pp 262-267.

Šonje, Jure (ed.) (2000), Rječnik hrvatskoga jezika, Leksikografski zavod Miroslav Krleža & Školska knjiga, Zagreb.

Šnajder, Jan; Dalbelo Bašić, Bojana; Tadić, Marko (2009) Automatic Acquisition of Inflectional Lexica for Morphological Normalisation. Information Processing & Management. 44, 5; pp. 1720-1731.

Tadić, Marko (1994) Računalna obradba morfologije hrvatskoga književnoga jezika. PhD Thesis. University of Zagreb, Zagreb.

Tadić, Marko (2005) Croatian Lemmatization Server, Southern Journal of Linguistics 29, 1/2; pp 206-217.

Tadić, Marko (2009) New version of the Croatian National Corpus. In: Hlaváčková, Dana; Horák, Aleš; Osolsobě, Klara; Rychlý, Pavel (eds.) After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, pp 199-205.

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