CroDeriV 2.0.: Initial Experiments

Krešimir Šojat[1], Matea Srebačić[2], and Tin Pavelić[2]

[1] Faculty of Humanities and Social Sciences, Zagreb
ksojat@ffzg.hr
[2] University of Zagreb
msrebaci@unizg.hr, tin.pavelic@ffzg.hr

Abstract. This paper deals with the processing of derivational morphology in Croatian and focuses on the expansion of CroDeriV – a resource with data on morphological structure and derivational relations. The purpose of CroDeriV is to systematically present the morphological structure and derivational relations of Croatian lexemes, and to use this data for the enrichment and development of existing resources and tools, as well as of new ones. One of the objectives in this ongoing project is to build an analyzer for Croatian capable of analyzing both inflectional and derivational morphemes. In this paper we present the initial experiments towards the enlargement of CroDeriV to include nouns, as well as the development of a morphological analyzer for inflectional and derivational morphemes.

Keywords: CroDeriV, Croatian, derivation, nouns.

1 Introduction

This paper deals with the processing of Croatian morphology and focuses on the expansion of CroDeriV – a resource providing data on the morphological structure and derivational relations of Croatian lexemes. Although detection of the complete morphological structure and all derivational relations could improve the performance of tools used in various natural language processing tasks, the derivational relatedness of Croatian lexemes based on a thorough and detailed analysis of their morphological structure so far have not yet been processed. Derivational relations hold between words that share the same lexical morpheme – i.e., the root – and thus form derivational families. A derivational family consists of lexemes with the same root grouped around a base form. Generally, a lexeme with the simplest morphological structure serves as a base form – i.e., a stem – for various derivational processes. Derivational processes in Croatian primarily refer to affixation and compounding. Affixation mainly consists of prefixation and suffixation. The motivation for building CroDeriV is twofold: (1) to analyze and systematically present the morphological structure and derivational relations of Croatian lexemes resulting from their mutual morphological relatedness, and (2) to use this data for the enrichment of existing resources and tools, as well as for the development of new ones. One of the objectives in this ongoing

A. Przepiórkowski and M. Ogrodniczuk (Eds.): PolTAL 2014, LNAI 8686, pp. 27–33, 2014.

project is to build an analyzer for Croatian capable of full morphological analysis – i.e., the analysis of both inflectional and derivational morphemes grouped around roots.

Inflectional classes are extensively covered by the Croatian Morphological Lexicon (HML). The HML serves as a basis for lemmatization and MSD tagging of Croatian (cf. [12] for the building procedures of the HML and [1] for the evaluation of lemmatization based on the HML). Although it is an extensive inflectional lexicon, it cannot be used for word segmentation, which is necessary for deeper morphological analysis and, consequently, for the detection of derivationally related lexemes and derivational families. The detection of derivational relations is recognized as an important task for the building of resources and tools also for other languages (cf. [4] for English; [6] for Polish; [13] for German). [7] presents an elaborate account of the methods used for the clustering of derivationally related Croatian lemmas from the corpus and also provides an extensive and detailed description of the evaluation metrics. This approach is based on morphological stems and focuses on the suffixal derivation between nouns, verbs, and adjectives. However, it does not take into account the analysis of the morphological structure of lexemes nor the recognition of mutual lexical morphemes within derivational families. An accurate and linguistically justified analysis of Croatian lexemes in terms of morphemes and affixes used in their derivation is one of our main objectives in the building of CroDeriV. In the next section we briefly describe its structure and design.

2 CroDeriV

CroDeriV is a morphological database built by means of combining rule-based processing with manual checking of results. The first phase of CroDeriV's development focused on the processing of Croatian verbs. In its present form, CroDeriV contains 14,326 verbs analyzed for morphemes and grouped into derivational families via mutual lexical morphemes. The total number of lexical morphemes is 3,386. The verbs were analyzed as follows: (1) verbs in infinitive form were collected from various on-line corpora and dictionaries; (2) the collected verbs were automatically segmented into morphemes; (3) the obtained results were manually checked; and (4) verbs sharing the same lexical morpheme were mutually linked. This procedure enabled the recognition of a general morphological structure applicable to all analyzed verbs, as well as the recognition of all lexical and derivational morphemes used in verbal derivational processes. It also enabled the detection of complete verbal derivational families in Croatian, as well as the analysis of possible combinations of affixes and roots, their frequency and productivity.[1]

Most verbs in Croatian are formed from other verbs, thus expanding their morphological structure. For example, the simplex base verb *graditi* 'to build$_{ipf}$' can be prefixed and thus can form the perfective verb *nadograditi* 'to expand by

[1] CroDeriV is freely available for searching at `http://croderiv.ffzg.hr`.

building$_{pf}$'. This verb in turn can be further suffixed to form the derived imperfective: *nadograđivati* 'to expand by building, to annex$_{ipf}$'. A thorough analysis of all verbal derivational processes has yielded the maximal morphological structure of Croatian verbs as follows:

$$(\mathrm{P4}) + (\mathrm{P3}) + (\mathrm{P2}) + (\mathrm{P1}) + (\mathrm{R2}) + (\mathrm{I}) + \mathrm{R1} + (\mathrm{S3}) + \mathrm{S2} + \mathrm{S1} + ti$$

where P = prefix, R = root, I = interfix, S = suffix, ti = infinitive ending, and () = non-obligatory.

Data structured in this way has already proven valuable for various linguistic studies (cf. [10]), as well as for the enrichment of other resources for Croatian – e.g., Croatian WordNet and the HML (cf. [9]; cf.[8]). The next objective in the building of CroDeriV is its extension with other parts of speech. Further on, we present the initial steps in the processing of nouns according to the principles mentioned above and their integration into CroDeriV.

3 The Derivation of Croatian Nouns

Derivational relations in Croatian extend either between words that are the same POS (verb-to-verb, noun-to-noun) or different POS (verb-to-noun, verb-to-adjective, adjective-to-noun, etc.). The most productive word-formation processes are prefixation and suffixation.[2] Like verbs, nouns in Croatian are formed through three basic derivational processes:

(a) **suffixation** (e.g., *pis(ati)* 'write' + *-ac* > *pisac* 'writer'),
(b) **compounding** (e.g., *roman* 'novel' + *-o-* + *pisac* 'writer' > *romanopisac* 'novelist'), and
(c) **prefixation** (e.g., *su-* + *radnik* 'worker' > *suradnik* 'co-worker').

On top of that, nouns are additionally formed through two combined processes:

(a) **compounding** + **suffixation** (e.g., *vatr(a)* 'fire' + *-o-* + *gas(iti)* 'extinguish' + *-ac* > *vatrogasac* 'firefighter'), and
(b) **prefixation** + **suffixation** (e.g., *po-* + *mor(e)* 'sea' + *-ac* > *pomorac* 'sailor').

Apart from these concatenative processes, nouns are also formed via:

(a) **back-formation** (e.g., *dopisati* 'to add by writing' > *dopis* 'letter'), and
(b) **conversion** (e.g., *mlada* 'young, female, adjective' > *mlada* 'bride, noun').[3]

Suffixation is by far the most productive derivational process in the derivation of Croatian nouns.[4]

As mentioned, CroDeriV has been built through the application of rules for morpheme segmentation and the manual checking of results. Generally, the main problems that we have faced in the automated processing of Croatian morphology (cf. Sect. 2) are caused by the following factors: (1) the graphical overlapping of various morphemes, (2) phonological changes at morpheme boundaries,

[2] Compounding is not as prominent and will not be discussed here.
[3] Two extensive accounts of Croatian derivation are [2] and [5].
[4] There are more than 90 productive suffixes used in Croatian noun derivation [2].

(3) phonological changes within lexical morphemes (which frequently result in several allomorphs), and finally, (4) numerous instances of homonymy (for a more detailed account of all instances and examples, cf. [11]). Unfortunately, we face all these problems in the processing of nouns, as well. For example, 36 different suffixal structures end in *-ac* (e.g., *bor-ac* 'fighter', *gled-a-l-ac* 'viewer', *jedanaest-erac* 'penalty kick', etc.). Frequently, suffixal structures of Croatian nouns consists of two or more suffixes (e.g., *grad* 'city' > *grad-i-(ti)* 'to build' > *grad-i-telj* 'builder' > *grad-i-telj-ic-a* 'female builder' > *grad-i-telj-ič-in* 'female builder's'. Since the list of possible suffixal combinations or the rules for the restrictions of these combinations in Croatian do not exist, the design of rules for their recognition and accurate segmentation is challenging and time-consuming work.

4 Experiment

In order to obtain a basic stock of nouns necessary for the expansion of CroDeriV with other POS and to provide a foundation for the development of tools for comprehensive morphological analysis, we have decided to use the nouns from the HML. In the experiment described below, we focused on the subset of the HML's nominal part tagged as common nouns.[5] This choice was motivated by the fact that the HML extensively covers nominal inflectional classes and therefore provides a good source for the detection of the most frequent suffixes and their combinations. The processing was divided into several steps. In the first step, we wanted to detect single nominal suffixes and obtain an initial snapshot of the suffixal productivity. In the second step of the experiment, we focused on nouns derived from verbs through non-concatenative derivational processes, namely back-formation. In the final step of the experiment, we wanted to detect possible suffixal combinations and nominal stems not recognized in the previous steps. The whole experiment is based on data already present in CroDeriV – i.e., roots and stems used primarily in the derivation of verbs from other Croatian verbs. This data was used for matching with nouns from the HML as described below.

4.1 Step 1

In the first step of the experiment, we created a set of rules for the detection and segmentation of single suffixes and applied it to the test sample. The test sample, which consisted of common nouns in the HML, comprised 20,554 nouns. The rules for the recognition and segmentation of single suffixes yielded a total of 4,933 nouns with correctly recognized stems and 22 single suffixes. The five most frequent single derivational suffixes detected in this way are: *-nje* (e.g., *pis-a-nje* 'writing'), *-ač* (e.g., *pis-ač* 'writer'), *-ica* (e.g., *s-pis-a-telj-ica* 'female writer'), *-telj* (e.g., *s-pis-a-telj* 'writer'), *-na* (e.g., *pis-ar-na* 'writing office'). Their respective frequencies are presented in Table 1.

[5] Lemmas in HML are tagged according to MulTextEast specifications v.4.0 ([3]).

Table 1. Five most frequent suffixal combinations

Suffix	*-nje*	*-ač*	*-ica*	*-telj*	*-na*
No. of occurences	2766	224	108	104	100

As the numbers indicate, the suffix *-nje* is the most frequent single suffix of common nouns in the HML. This suffix is used in Croatian almost exclusively for the derivation of gerunds from verbal stems – i.e., from their past participles. The total number of gerunds among the common nouns is 4,586. In other words, almost 25% of all common nouns in the HML are verbal nouns. Although this fact may be surprising as far as the structure of the HML is concerned, it simplified the further processing of this subset of common nouns.

A comparison of verb lists from CroDeriV and the HML revealed that all the verbs from the HML are also present in CroDeriV. Since all the verbs in CroDeriV have been analyzed for morphemes, by slightly adjusting the rules, we were able to automatically determine the full morphological structure of gerunds – i.e., their roots as well as their prefixal structures – inherited from the base verbs. After manual validation, we used these results in the next step of the experiment.

4.2 Step 2

The total remaining number of nouns to be analyzed was thus 15,968. As mentioned in Sect. 3, nouns in Croatian are formed from other POS via affixation, but also through non-concatenative processes, such as back-formation; as in:

dopisati 'to add by writing' → *dopis* 'memo'
opisati 'to describe' → *opis* 'description'
upisati 'to record' → *upis* 'record'

Consequently, all nouns derived from verbs through back-formation inherit the morphological structure of their base forms, apart from their suffixal part. For the detection of nouns derived in this manner, we again matched the data from CroDeriv and the HML. In this way we detected 3,367 common nouns tagged as candidates for the expansion of derivational families in CroDeriV. The manual checking of the results revealed that 1,200 nouns were not correctly segmented, whereas 2,167 nouns were correctly segmented and correctly assigned to the corresponding derivational families from CroDeriV.

Although the recall of 38.61% scored in this part of the processing is low, the high precision of 85.02% enabled their straightforward integration into CroDeriV. The total number of nouns used in further steps was thus reduced to 13,801.

4.3 Step 3

In the final step of the experiment, we randomly chose around 40% of this remaining set of nouns and manually analyzed them for morphemes. The primary objectives of this analysis were to detect possible suffixal combinations as input for rules capable of dealing with multiple suffixes and to detect nominal stems not recognized in previous steps due to the complex morphological structure of their lexemes. The general aim of the whole procedure was to speed up the detection of derivationally related nouns and to check whether they can be linked to derivational families from CroDeriV without further manual segmentation. For this purpose we took the following steps:

(1) we automatically removed manually obtained suffixal combinations and used only nominal stems in further processing – e.g.:
glas-ač-ic-a 'female voter' → *glas* 'voice',
vid-ovit-ost 'clairvoyance' → *vid* 'sight',
(2) we extracted all the stems and roots from CroDeriV, and
(3) we matched them with the list of obtained nominal stems from the HML.

The recall of the whole procedure was 100%. In order to measure precision, we manually evaluated 5,520 randomly selected nouns. Out of this number, 33.88% of the roots from CroDeriV were correctly assigned to nominal stems from the HML. As expected, although the recall was 100%, the precision of the automated root assignment significantly decreased. However, the final results of the conducted experiment can be considered satisfactory. A total of 1,773 roots out of 3,386 roots in CroDeriV (53.4%) has been correctly assigned to at least one noun from HML, thus enabling the automated expansion of derivational families. On the top of that, we obtained 1,753 new nominal roots through manual evaluation, which can be used in the further processing. From the initial set of 20,554 nouns, this simple automated approach assigned the correct root to more than half (12,227) of the nouns.

5 Conclusion

In this paper we have shown the initial steps towards the enlargement of CroDeriV to include another POS, namely nouns. With a combined approach using simple automatic processing and manual checking, we have obtained two noun sets:

(1) a set of nouns which are derivationally related to the verbs in CroDeriV and can be used for the enrichment of already existing derivational families via mutual root, and
(2) a set of nouns that are not derivationally related to the verbs in CroDeriV.

However, many of these nouns are mutually derivationally related and can be used for the inclusion of new derivational families in CroDeriV. Moreover, the list of newly recognized nominal roots, as well as the list of the most frequent nominal suffixes and their combinations, can be used for further improvement of our morphological processing tools.

References

1. Agić, Ž., Tadić, M., Dovedan, Z.: Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica 32(4), 445–451 (2008)
2. Babić, S.: Tvorba riječi u hrvatskome književnom jeziku. HAZU: Nakladni zavod Globus, Zagreb (2002)
3. Erjavec, T.: MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. Language Resources and Evaluation 46(1), 131–142 (2012)
4. Habash, N., Dorr, B.: A Categorial Variation Database for English. In: Proceedings of the Anuual Meeting of the NAACL, pp. 96–102 (2003)
5. Marković, I.: Uvod u jezičnu morfologiju. Disput, Zagreb (2012)
6. Piasecki, M., Ramocki, R., Maziarz, M.: Recognition of Polish Derivational Relations Based on Supervised Learning Scheme. In: LREC 2012 Proceedings, pp. 916–922 (2012)
7. Šnajder, J.: DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian. In: Calzolari, N., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). ELRA, Reykjavik (2014)
8. Šojat, K., Srebačić, M.: Morphosemantic relations between verbs in Croatian WordNet. In: Orav, H., Fellbaum, C., Vossen, P. (eds.) Proceedings of the Seventh Global WordNet Conference, pp. 262–267. GWA, Tartu (2014)
9. Šojat, K., Srebačić, M., Pavelić, T., Tadić, M.: From Morphology to Lexical Hierarchies. In: Vetulani, Z. (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2013 Proceedings), pp. 474–478 (2013)
10. Šojat, K., Srebačić, M., Štefanec, V.: CroDeriV i morfološka raščlamba hrvatskoga glagola. Suvremena lingvistika 39(75), 75–96 (2013)
11. Šojat, K., Srebačić, M., Tadić, M.: Derivational and Semantic Relations of Croatian Verbs. Journal of Language Modelling (1), 111–142 (2012)
12. Tadić, M., Fulgosi, S.: Building the Croatian Morphological Lexicon. In: Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages, pp. 41–46. ACL, Budapest (2003)
13. Zeller, B., Šnajder, J., Padó, S.: DERIVBASE: Inducing and Evaluating a Derivational Morphology Resource for German. In: Proceedings of the 51st Annual Meeting of the ACL (2013)