Detection of Verb Frames with NooJ

Krešimir Šojat[1], Božo Bekavac[1], Kristina Kocijan[1]

[1] Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

{ksojat, bbekavac, kkocijan}@ffzg.hr

Abstract. This paper deals with semi-automatic extension of CroDeriV with verb valency frames. CroDeriV is a morphological database of Croatian verbs. In its present shape the database comprises 14 500 verbs in infinitive forms. Each verb in CroDeriV is segmented into lexical and derivational morphemes and verbs of the same root are mutually linked. In order to further enrich the CroDeriV with semantic and syntactic information, we have used the NooJ platform to recognize derivationally related verbs, find the verb frames and to speed up the sentence processing.

Keywords: verb frames, NooJ, derivation, morphological grammars, Croatian

1 Introduction

This paper deals with semi-automatic extension of CroDeriV with verb valency frames. CroDeriV is a morphological database of Croatian verbs. In its present shape, the database comprises almost 14 500 verbs in infinitive forms. Each verb in CroDeriV is segmented into lexical and derivational morphemes and verbs of the same root are mutually linked [7]. The database is available for search at http://croderiv.ffzg.hr/Croderiv.

The lexicon structured in this way enables the recognition of derivationally related families of verbs and, at the same time, the detection of derivational spans of particular base forms.[1] In the second phase of its development, CroDeriV is extended with definitions of verbal meanings, i.e. verbal lexemes are analyzed for their meaning structure and divided into lexical units. Each lexical unit is accompanied with one or more sentences illustrating its contextual usage. These sentences also function as a basis for the construction of valency frames, i.e. frames reflecting verbal argument structure.

In this paper we focus on the detection of verbal arguments in sentences using NooJ. We also discuss the structure of existing morphological grammars for Croatian verbs as well as their future development. We experiment with small derivational families of verbs consisting of 4 and 5 derivative forms around a central member of the family – a base form. Such relatively small derivational families enable a careful design of rules or a redesign of already existing ones. The final goal is to detect major constituents in sentences and automatically classify them according to their syntactic function.

[1] CroDeriV resembles databases like CatVar for English and Uni-morph for Russian (http://courses.washington.edu/unimorph and http://clipdemos.umiacs.umd.edu/catvar).

adfa, p. 1, 2011.

The paper is structured as follows: in Section 2 we present the design and structure of CroDeriV in its present form. Section 3 deals with the expansion of CroDeriV with valency frames. In Sections 4 and 5 we describe the detection of derivationally related verbs and verb frames using NooJ. Section 6 briefly describes the corpus used for this purpose and presents some comparative results for two approaches to verbal derivations. The final part of the paper provides an outline for future work.

2 CroDeriV

CroDeriV is a computational lexicon that provides information on the morphological structure of approximately 14 500 Croatian verbs collected from different sources, mainly digital and paper dictionaries [2], [9] and additionally enriched with lemmas from the Croatian National Corpus v3.0 [10] and the Croatian Web Corpus v2 [4].

The Croatian language has a very rich morphology, both in terms of derivation and inflection. Further in this paper we focus on derivational relatedness of verbs, particularly on the derivation of verbs from other verbs. We also discuss how the lexical entries for derivationally related verbs can be extended with data on their argument structure (valency) and morphosyntactic features of arguments. The general purpose of CroDeriV is to obtain a complete morphological analysis of Croatian vocabulary. At its current stage of development, this resource contains only verbal lemmas, whereas the extension with other parts of speech is planned in future development.

Verbs in Croatian can be derived from other verbs via two derivational processes – prefixation and suffixation. Prefixation is far more productive than suffixation. Prefixes are always derivational, whereas suffixes can be derivational as well as inflectional. In the majority of cases, base forms take one prefix. However, one verbal root as a part of a base form can in some rare cases co-occur even with four prefixes. As far as suffixes are concerned, one root usually has two derivational and one inflectional suffix (-ti). This structure can be extended with an additional suffix that denotes a diminutive or pejorative action. Verbs in Croatian can also be formed by compounding, i.e., they can consist of two roots. Although, compounding is not a very productive process as far as Croatian verbs are concerned.[2]

The morphological structure of all verbal lemmas in CroDeriV was determined in several steps. Due to numerous phonological changes, all prefixes were manually analyzed and segmented, whereas the suffixal part was in the first step segmented automatically. However, the results of automatic rule-based segmentation were not satisfactory. The two main problems in the automated processing of Croatian derivation are homography that results in the overlapping of prefixes and suffixes with roots, and numerous phonological changes at the morpheme boundaries resulting in several allomorphs for each morpheme. All results of automatic segmentation were therefore manually checked and all allomorphs were connected to their unique representative morphemes.

[2] There are approximately 120 verbal compounds recorded in CroDeriV.

Lexical entries in CroDeriV consist of verbs decomposed into morphemes and linguistic metadata. The metadata in lexical entries indicate verbal aspect and types of reflexivity. The structure provided for all analyzed verbs consists of 11 morpheme slots and covers all combinations of recorded lexical and grammatical morphemes.[3] There are four types of slots for morphemes: (1) derivational prefixes (four slots), (2) roots (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).

This kind of processing enables the recognition of all allomorphs of a particular morpheme and the detection of all affixes co-occurring with particular roots. This procedure also enables the detection of complete derivational families of Croatian verbs. 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. For example, there are four derivatives recorded in CroDeriV of the base form *jedriti* (to sail). These are *do-jedriti* 'to sail to', *od-jedriti* 'to sail away', *pre-jedriti* 'to sail across' and *za-jedriti* 'to start sailing'. This group of verbs constitutes a derivational family consisting of five members. As indicated, all derived forms in this family are produced through prefixation. In Fig. 1 we present the morphological segmentation of this derivational family as structured in CroDeriV.

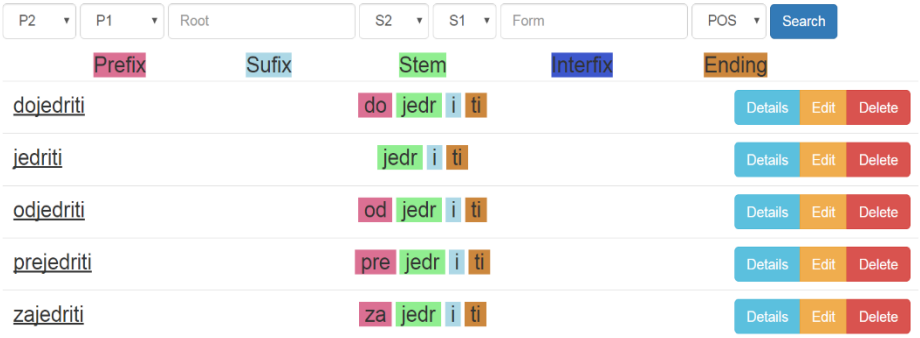

Fig. 1. An example of morphological segmentation and derivational relatedness of verbs from CroDeriV.

3 Verb Frames

As mentioned in Introduction, CroDeriV is being extended with definitions of verbs' meanings. In other words, verbal lexemes are manually analyzed for their meaning and consequently divided into lexical units. This enrichment of lexical entries in CroDeriV is motivated by two reasons. The first one is to enable an accurate and extensive description of derivational processes in Croatian, which cannot be done if the semantic

[3] A more detailed account is given in [6]

component is not taken into account. If we deal only with 'raw' data, as they are presently recorded in CroDeriV, numerous derivational processes cannot be adequately described. For example, the verb *hodati* can have (at least) two senses in Croatian: 'to walk' and 'to date somebody'. The derivative *prohodati* 'to start walking' is semantically related to both of them, but the derivative *prehodati* 'to cover a certain distance by walking' is semantically related only to the first sense. Without definitions of meaning more comprehensible studies in this area are not possible.

Secondly, the introduction of definitions of meaning and the division of verbal lexemes into corresponding lexical units can also enable the development of a large-scale verb valency description to be used in various NLP tasks. As in previous examples, different senses of the same lexeme *hodati* have different valency or argument structures. The lexeme *hodati* is divided into lexical units that correspond to detected senses of verbal lexemes (e.g. *hodati* 1 – to walk; *hodati* 2 – to date somebody). The division of lexemes into lexical units is based on the analysis of sentences from available corpora, mainly the Croatian National Corpus. Each lexical unit is accompanied with one or more sentences illustrating its contextual usage. These sentences also function as a basis for the construction of verb valency frames. Each frame contains information on argument structure characteristic for a particular lexical unit. In the next step of analysis, the arguments are annotated in contextual examples. On this level of processing their syntactic functions and morphological features are annotated in sentences (e.g. SUB/NOM – subject in nominative case; OBJ/PP ACC – prepositional object in accusative).

We believe that the derivational lexicon of verbs enhanced with the data on their valency can provide a valuable information for various linguistic studies and the development of NLP applications. In Fig. 2 we provide an example of an entry for the lexical unit *gledati* 'to perceive by sight'.

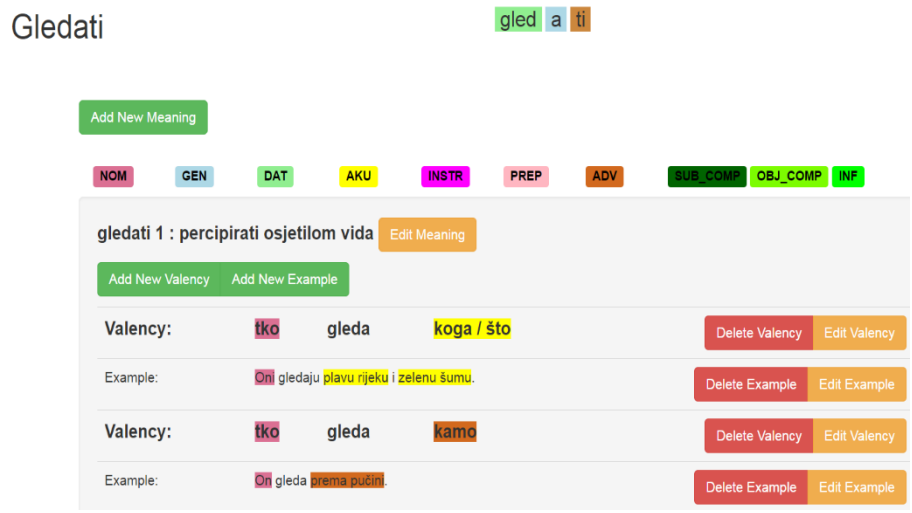

Fig. 2. A verb frame from CroDeriV.

For the visual presentation of sentential elements we use different colors for annotated elements. These elements mainly correspond to constituents. However, the whole procedure is a very time-consuming work if performed completely manually. In order to speed up the building and the development of verb frames we decided to use NooJ. The detection of derivationally related verbs and verb frames with NooJ is described in the following sections.

4 Detection of derivationally related verbs with NooJ

In our research, we decided to use NooJ [5] as an NLP tool since it allows an easy construction of grammars on both morphological and syntactic levels. We needed both in order to recognize derivationally related families of verbs (morphological grammar) before we are able to detect verb valency frames (syntactic grammar). In this chapter, a more detailed description of the morphological grammar will be given and the reasons behind this grammar will be explained. In the next chapter we will do the same for the syntactic grammar while more information on detection of syntactic verbal frames for Croatian using NooJ can be found in [3].

Croatian dictionary of verbs consists of 3 907 verbs (main lemmas). Each verb has a category attribute 'V' and a link to its inflectional paradigm 'FLX=V_PARADIGM'. So far, there are 239 unique inflectional paradigms that combined with the main lemmas produce 342 292 different inflectional forms for verbs (including only simple verb tenses). There are 464 verbs in the dictionary that may appear as reflexive verbs (*ponašati se* 'to behave oneself' or *ponašati se* 'to behave or act'). They are marked with an additional attribute value set '*Prelaz=pov*'. This feature is used in the syntactic grammar for the recognition of compound verb forms (future tenses, past tenses etc.) since an auxiliary and main verb forms may be disconnected with the particle '*se*' denoting the reflexivity of the verb. Recently, we have annotated 91 verbs of perception as the '*prcp*' verbs with additional marking of the perception type (*viz*- vision, *miris* - smell, *sluh* - sound, *okus* - taste and *dodir* - touch). There are also some verbs in our dictionary that have been annotated with their valency frames but this includes only 102 verbs of consummation [8] and 1 739 verbs that were considered the most frequent ones [13].

The existing dictionary is far from being a complete one, and it is clear that it does not hold all of the Croatian verbs. For example, although the verb *bližiti* 'to come near' is in the dictionary, the four verbs that belong to the same family *približiti* 'to bring closer'*, zbližiti* 'to make a bond'*, približavati,* and *zbližavati* are not. Examples like these hold true for numerous derivational families. Luckily, due to the morphological rules used to build derivationally related verbs, it is quite obvious that an implementation of some types of rules should change this situation and automatically 'populate' the list of (missing) verbs. The real question was to choose the right path (for us) that will enable this project. In fact, NooJ offers two solutions for recognizing derivations. One is via direct input of an attribute +DRV next to the dictionary entry and the other via morphological grammar(s). We will try them both and give our pros and cons for both approaches.

The first approach requires that an attribute-value (+DRV=*derivationName*) combination is inserted in-line for each verb in the dictionary and that all *derivationName* values are defined in a separate NOF file (similarly as it is done for +FLX attribute of an each verb). The average number of derivations is 4-5 per verb, but some verbs may take from two to up to even thirty+ derivations. The approximate time for adding a single derivation is 10 minutes (including the time to determine the correct derivational paradigm and define it in the dictionary)[4]. Considering the present number of verbs in our dictionary, this adds up to 39 070 minutes or 81 days (8h/day) of work. Still, we will not be sure if all the possible derivations for each verb are described since some are rarely used and found only in specific genres. On the other side, regardless the (occasional) smaller recall, this approach would provide us with a higher precision.

The second approach gives us all the possible derivations much faster (only 2 days (8h/day)). In spite of the high recall, it gave us a somewhat lower precision than we have hoped for, but enough data that we can learn from in order to improve the grammar and raise the precision. The grammar's detailed description follows, since it is the path we have chosen for this research.

Our morphological grammar has an L1 priority level, which means that it is applied to the text only after all other morphological resources and only to the words not recognized until that point. The grammar's transducer adds the main verb (the one that the new verb is derived from) in the output as the super-lemma, and it uses its syntactic and inflectional information where applicable. The grammar recognizes, not only the main lemma, but all its gender, number and tense dependent endings, i.e. forms.

The new verbs fall into three main categories:

A. verbs built only with a prefix,
B. verbs built only with a suffix,
C. verbs built with a prefix and a suffix.

Thus, our grammar also has three main branches responsible for the recognition of each type (Fig. 3).

[4] It will take even longer if the new paradigm is needed to be defined in the NOF file.

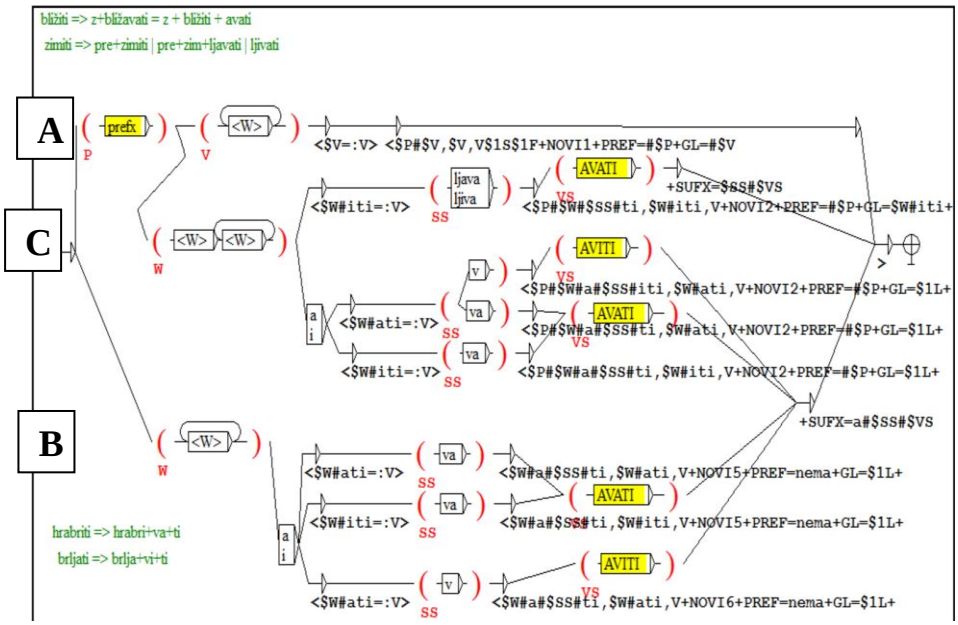

Fig. 3. Morphological grammar for recognizing derivationally related verb families

Firstly, we built the branch for recognizing the type A verbs. The branch finds the predefined prefix and then checks if the remaining of the word is equal to any existing[5] verb form. If such a word is found, the transducer marks the new word as a verb that inherits all the features of the verb it is derived from. Thus, the new word is immediately annotated with the gender, number and tense information. For the purposes of this research, we have added some additional information that include the type of prefix ('+PREF') and the main verb whose family the new verb belongs to ('+GL').

Secondly, we built the branch recognizing the type B verbs. At first, we used the same methodology as in the type A verbs. We split the word in three sections where the second section is one of the possible suffixes. Everything before that, with the addition of an ending '-ti', must match an existing verb in the infinitive form, and everything after the suffix should be a tense, gender and number ending. However, this reasoning recognized too many false positives mostly nouns or adjectives that have the same rooth and the suffix (1st and 2nd sections) but not ending or the 3rd section (for example: *svladavanje* = (svlada+ti) + (va) + (nje) or *čuvarice* = (ču+ti) + (va) + (rice)).

At the end, we built the branch recognizing the type C verbs that uses the same set of predefined prefixes and suffixes and the remaining of the branch is similar to the type B branch. The main difference between them is in the output form of the recognized word.

The grammar proposed here recognizes, in addition to the verbal derivations, some additional words that share similar structures as the ones described for derivations of verbs. We can classify these false positives into four categories:

[5] Where existing means previously built by the inflectional grammar of a dictionary entry.

- nouns: ex. *paravan* [**para**+va+n] the grammar incorrectly recognizes the root as a verb *para+ti* (en. to tear apart) and *para+va+n* as its derivation (using the branch B)
- adjectives: ex. *ekstra* [eks+**tra**] the grammar incorrectly recognizes the root as a verb *tra+ti* (en. to trifle) and *eks+trati* as its derivation (using the branch A)
- verbs: ex. *šarmirati* [š+**armirati**] the grammar incorrectly recognizes the root as a verb *armirati* (en. to reinforce concrete) and *š+armirati* as its derivation (using the branch A)
- foreign words: ex. *delete* [de+**lete**] the grammar incorrectly recognizes the root as a verb *let+(je)ti* (en. to fly) and *de+lete* as its derivation (using the branch A).

If we are to add the missing nouns, adjectives and verbs into our main dictionary, we would be able to solve this problem and subsequently raise the precision of our proposed model. Still, the problem of foreign words would remain, since we do not have these words in the Croatian dictionary. Thus, to completely deal with the problem of precision, it would be safer to add the derivational paradigms directly to the dictionary but to keep the grammar as a very low priority grammar just in the case we have missed some derivations or some new ones have emerged[6].

5 Detection of Verb Frames with NooJ

The second phase in which NooJ was used included the application of an existing chunker for Croatian [12]. After the text is annotated with the noun chunks <NP>, prepositional chunks <PP>, verb chunks <VP>, conjunctions <C> and adverbs <R>, we apply additional syntactic grammar to employ the power of NooJ's transducer and generate an XML-like notation (sentence examples (1), (2), (3), (4)).

The XML-like notation for the sentence: *Jedrili smo sedam dana, imali smo boljih i lošijih jedrenja* - 'We were sailing for seven days, we had better and worse achievements' is given in example (1).

```
<SENTENCE>                                                    (1)
  <VP TYPE="ROOT">Jedrili smo</VP>
    <NP> sedam dana </NP>,
  <VP>imali smo</VP>
    <NP>boljih i lošijih jedrenja </NP>.
</SENTENCE>
```

The XML-like notation for the sentence: *Četvrti dan jedrilo se sa Velikog Iža na Veliki Rat* 'On the fourth day we sailed from Veliki Iž to Veliki Rat' is given in example (2).

```
<SENTENCE>                                                    (2)
    <NP> Četvrti dan </NP>
```

[6] If we observe language as a living thing, it is quite expected that some new derivations will appear in time.

<VP TYPE=”ROOT”>jedrilo se</VP>
<PP> sa Velikog Iža </PP>
<PP>na Veliki Rat</PP>.
</SENTENCE>

The file generated in this manner allows us to produce visual representations of each root verb[7] environment (Fig. 4) as well as the environment of derived verbs (Fig. 3). The chunks are color coded to ensure faster recognition.

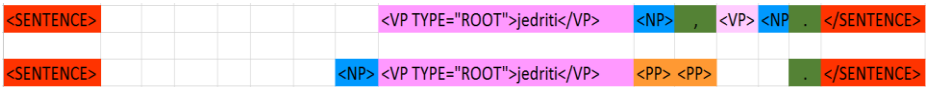

Fig. 4. Example of visual representation of root verb frame

The same XML-like notation is provided for the derived verbs such as for the sentence: *Unatoč lošem ulasku u natjecanje i mnogim ostalim problemima, Fantela I Marinić sun a kraju uspjeli dojedriti do bronce* 'Despite of a bad start and many other problems, Fantela and Marinić won the bronze medal' (example 3).

<SENTENCE> (3)
<PP>Unatoč lošem ulasku
<PP>u natjecanje </PP>
</PP>
<C> i </C>
<NP>mnogim ostalim problemima</NP>,
<NP>Fantela i Marenić </NP>
<VP> su </VP>
<PP> na kraju </PP>
<VP> uspjeli
<VP TYPE=”NOVI1” PREF=”do” GL=”**jedriti**” Verb=”**dojedriti**”> dojedriti
</VP>
</VP>
<PP> do bronce </PP>.
</SENTENCE>

And the sentence: *Još malo i prvi će kišni oblak dojedriti iza obzora i opet ćee padata kiša* - 'Very soon the first cloud will appear on the horizon and it is going to rain again' is presented in example (4).

<SENTENCE> (4)
<R> Još malo </R>
<C> i </C>
<NP> prvi </NP>

[7] In this paper, we will refer to the selection of main verbs used in this research as the root verb and they are marked as TYPE=“ROOT“, while the derivations of these main verbs are marked as TYPE=“NOVIx“. All the other types of verb that may appear in the sentence are only marked as an <VP> chunk with no additional attributes.

<VP> će </VP>
<NP> kišni oblak </NP>
<VP TYPE=”NOVI1” PREF=”do” GL=”**jedriti**” Verb=”**dojedriti**”>dojedriti
</VP>
<PP> iza obzora </PP>
<C> i </C>
<R> opet </R>
<VP>će padati kiša</VP>.
</SENTENCE>

In Fig. 5 we give a visual representation of verb frames for sentences (3) and (4).

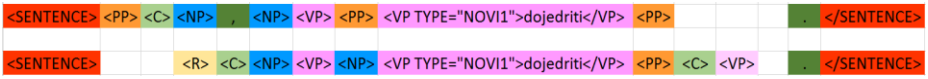

Fig. 5. An example of visual representation of derived verb frame

6 Corpus and the Results of an Experiment

The corpus compiled for Croatian Language Resources for NooJ, as described in [13] is insufficiently large regarding the representation of verbs that belong to our targeted derivational families. An ideal resource for our research would be the Croatian National Corpus (CNC), a representative corpus of the contemporary Croatian standard language. The CNC contains written texts published since 1990. The corpus is automatically lemmatized and MSD tagged using standard heuristic methods as described in [10]. The methodology of coping with unknown words is in more detail explained in [1] together with accuracy statistics which can be relevant for our work.

As CNC comprises over 200 million tokens we decided to extract only the relevant sentences, i.e. the sentences which contain verbs of our potential interest. The extracted subset of CNC in the raw textual format comprises 15 thousand tokens that were imported in NooJ and processed with current version of Croatian lexical module described in [11]. Using such a corpus preparation methodology we deliberately sacrificed some potentially interesting quantitative data about frequencies of certain verbs in the entire CNC. On the other side, we provided focus related corpus optimal for NooJ processing which suits our research needs.

In order to test which approach will score better on the recognition of derived verbs, we have used an even smaller subcorpus of 2 671 tokens. For this experiment, we have used five root verbs and their 4-5 derivations. The Table 1 gives a complete list of the base verbs, derived verbs and the number of their derivations related to gender, number and simple tense that were generated in the inflectional dictionary. This file is automatically produced after joining the dictionary (*.dic) and the morphological derivational description file (*.nof).

Table 1. The number of derivations for the 5 root verbs we have used in the experiment.

Base forms	Derived verbs	Number of derivations
bližiti	*zbližiti, približiti* *zbližavati, približavati*	465
brljati	*zabrljati* *brljaviti, zabrljaviti* *zabrljavati*	243
hrabriti	*ohrabriti, obeshrabriti* *ohrabrivati, obeshrabrivati*	465
jedriti	*dojedriti, odjedriti, prejedriti, zajedriti*	375
kopirati	*fotokopirati, prekopirati, iskopirati,* *nakopirati*	560
zimiti	*prezimiti, uzimiti, zazimiti* *prezmiljavati* *zimovati*	468
TOTAL		2576

We have tested both approaches and the results (Table 2) clearly show in favor of the first test, i.e. adding derivations directly to the dictionary (via +DRV feature). As expected, the precision is somewhat higher, although still not perfect, in the first test which can be explained with the ambiguity of Croatian language. In our example text, all of the verbs that were incorrectly (or ambiguously) marked as derived verbs, where actually either adjectives (*hr. hrabri* – which may be an adjective 'brave' or a verb 'to encourage') or nouns (*hr. zime* – which may be a noun 'winter' or a verb 'to spend a winter'). However, this ambiguity did not come as a surprise to us, since it is very common in Croatian language, but we hope to resolve it on a syntactic level of analysis, especially with the addition of information on verb valency frames.

Table 2. Comparative results for the two approaches.

Measure	**The dictionary approach (+DRV)**	**The morphological grammar approach**
Precision	0,9674	0,9654
Recall	1	0,9398
f-measure	0,9834	0,9524

At this time, we do not offer any statistics on the verb frames since we are still in the process of analyzing and learning from that data.

7 Conclusion

In this paper we have presented a method for the extension and enrichment of a derivational lexicon of Croatian – CroDeriV. In its present form it contains only verbs, whereas other parts of speech will be introduced in future development. CroDeriV is also being extended with definitions of meaning and verb valency information. The project described in this paper has served us as an experiment and a starting point in learning about verb frames with the help of NooJ. As more data will be collected, i.e. more sentences on each verb (root verb and derived verbs), we will be able to define with greater accuracy the exact frame each verb may appear in text. It will take some fine-tuning of the chunker and adding some more identifiers to the chunks in order to make them more informative about the environment the specific families tend to appear in. We believe that this procedure can be valuable for further development of existing language resources like CroDeriV as well as for the development and refinement of existing NooJ resources for the Croatian language.

8 References

1. Agić, Ž., Tadić, M., Dovedan, Z.: Evaluating Full Lemmatization of Croatian Texts. In: Recent Advances in Intelligent Information Systems, Academic Publishing House EXIT, Warsaw, 175—184 (2009)
2. Anić, V.: Rječnik hrvatskoga jezika, Novi liber, Zagreb, 4th edition (2004)
3. Bekavac, B.; Šojat, K.: Syntactic Patterns of Verb Definitions in Croatian WordNet. In Formalising Natural Languages with NooJ : Selected Papers from the NooJ 2011 International Conference (Dubrovnik, Croatia). Kristina Vučković, Božo Bekavac and Max Silberztein (eds.). Cambridge Scholars Publishing, Newcastle., UK: 112-121, 2012.
4. Ljubešić, N.; Erjavec, T.: 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, 395-402 (2011)
5. Silberztein, M.: NooJ Manual, www.nooj4nlp.net. (2003)
6. Šojat, K., Srebačić, M.; Štefanec, V.: CroDeriV i morfološka raščlamba hrvatskoga glagola. Suvremena lingvistika. 39, 75; Zagreb, 75-96 (2013)
7. Šojat, K., Srebačić, M., Pavelić, T., Tadić, M.: CroDeriV: a New Resource for Processing Croatian Morphology. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.). Reykjavik, Iceland : European Language Resources Association, ELRA, 3366-3370 (2014)
8. Šojat, K., Vučković, K., Tadić, M. Extracting verb valency frames with NooJ. In Finite-State Language Engineering with NooJ : Selected Papers from the NooJ 2009 International Conference (Tozeur, Tunisia). Abdelmajid Ben Hamadou, Slim Mesfar and Max Silberztein (eds.). Centre de publication Universitaire : Sfax., Tunisia: 231-242, 2010.
9. Šonje, J. (ed.): Rječnik hrvatskoga jezika, Leksikografski zavod Miroslav Krleža & Školska knjiga, Zagreb (2000)
10. Tadić, M.: New version of the Croatian National Corpus. In: After Half a Century of Slavonic Natural Language Processing. Masaryk University, 199—20. Brno (2009)

11. Vučković, K., Tadić, M., Bekavac, B.: Croatian Language Resources for NooJ. In: CIT. Journal of computing and information technology. 18, 295—301. Zagreb (2010)
12. Vučković, K.: Model parsera za hrvatski jezik. PhD dissertation. Faculty of Humanities and Social Sciences, Zagreb (2009)
13. Vučković, K., Mikelić Preradović, N., Dovedan, Z.: Verb Valency Enhanced Croatian Lexicon. In: Applications of Finite-State Language Processing. Cambridge Scholars Publishing, Newcastle upon Tyne, (2010)