A Method for the Computational Representation of Croatian Morphology

Vanja Štefanec[1], Matea Srebačić[1], and Krešimir Šojat[2]

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

Abstract. In this paper we present the development of the Croatian Derivational Database (CroDeriV), a unique language resource for the Croatian language. We describe the initial stage of its development as well as its redesign according to insight gained from data analyzed thus far. We believe that the data model we have presented will enable us to encode all derivational processes in Croatian. We also believe that our model is sufficiently flexible and abstract that it could be used for other morphologically "rich" languages.

Keywords: CroDeriV, morphological resources, derivation, Croatian.

1 Introduction

In this paper we present the development of the Croatian Derivational Database (CroDeriV). The computational processing of Croatian morphology so far has focused primarily on inflectional phenomena ([10]; [12]; [3]) and the enlargement procedures of the Croatian Morphological Lexicon ([11]). During the process of building Croatian Wordnet ([6]), especially when dealing with derivationally connected members of different synsets ([9]), it became obvious that large-scale data indicating which affixes are used or can be used with particular lexical morphemes do not exist for Croatian. In other words, the data about the derivational spans of particular lexical morphemes have not yet been systematically and extensively presented in the Croatian morphological literature.

A derivational span refers to all attested combinations of derivational affixes and one or more lexical morpheme. Large-scale data on such combinations are necessary not only for the further development of Croatian WordNet, but also for the development of various NLP tools for Croatian, such as stemmers, lemmatizers, Q&A systems, etc.

The CroDeriV database is designed to comprise four major POS, i.e. nouns, verbs, adjectives, and adverbs, and to represent their morphological structure in terms of roots (lexical morphemes) and derivational affixes attached to the roots. The aim of the database is to serve as the basis for future research on Croatian derivational morphology by enabling the recognition of words with the same root of the same or a different POS as well as word-formation processes between words sharing the same root.

M.A. Kłopotek et al. (Eds.): IIS 2013, LNCS 7912, pp. 80–91, 2013.

Lexical entries in CroDeriV contain lemmas analyzed for lexical and derivational morphemes, i.e. each lemma is divided into one or more roots and derivational affixes attached to those roots. In the first phase of the database development, 14,000 verbal lemmas, i.e. verbs in infinitive form, were collected from freely available digital dictionaries of Croatian and semi-automatically analyzed for morphemes.[1] So far, no language resource containing morphologically analyzed lemmas has been developed for Croatian.

In Sects. 2 and 3 we shall briefly discuss related work and word-formation processes in Croatian. In Section 4 we present the present shape of CroDeriV, and in Section 5 we present the reasons for our complete redesign of CroDeriV, as well as the new data model on which it is based. Section 6 comprises future work and conclusions.

2 Related Work

As mentioned, the computational processing of Croatian has been primarily focused on inflection. The Croatian Morphological Lexicon [10] comprises 120,000 lemmas and their inflectional forms. Ćavar et al. [3];[2] describe the development of *CroMo*, a finite state lexical transducer used for morphological analysis and lemmatization. The transducer is based on a database of ca. 250,000 lexical, derivational, and inflectional morphemes. This database is unfortunately not publicly available. Šnajder [8] deals with the procedures of automatic processing and the acquisition of inflectional lexicon for Croatian. Derivational processing is limited to nouns, verbs and adjectives formed by suffixation. The data obtained through lemmatization and stemming are used for further information extraction from raw corpora. Stemmers for Croatian presented in [4] and [5] are developed for the recognition of derivational suffixes and inflectional endings. Although all stemmers are based on linguistic rules, none of them recognizes base forms and derivatives obtained through prefixation. Morphological analyzers for other Slavic languages, e.g. *ajka* for Czech [7] or *Morfeusz* for Polish [13] are also restricted to inflection.

3 Word-Formation Processes in Croatian

As in other Slavic languages, word-formation processes in Croatian comprise derivation and compounding. Derivation is a significantly more productive process than compounding, which does not play an important role in Croatian morphology if compared with languages such as German. In some cases it is hard to draw a sharp line between derivation and inflection in Croatian, since, for instance, the formation of gerunds, participles, and comparatives/superlatives are considered to be inflectional processes, whereas the formation of verbal aspectual pairs is treated as derivation.

[1] A similar resource exists for Russian (`http://courses.washington.edu/unimorph/`).

Derivation in Croatian is basically affixation. Affixation can comprise prefixation (*pisati* - ***po****pisati*, ***is****pisati*, ***na****pisati*, ***pre****pisati*), suffixation (*popisati* - *popis****iva****ti*, *pisati* - *pis****ač***, *pis****ar***) and simultaneous prefixation and suffixation (*pisati* - ***s****pisa****telj***).

Compounding is the word-formation process of putting two lexical morphemes together. In Croatian there are three kinds of compounding: bare compounding (stem1 + (interfix) + stem2; stem2 can stand as a separate lexeme; *kuć-e-pazitelj* 'housekeeper'), simultaneous compounding and suffixation (stem1 + interfix + stem2 + suffix, where stems cannot stand as separate lexemes; e.g. *rukopis* 'manuscript' = *ruk-o-pis-Ø*), and finally semi-compounding (two lexemes preserve their meaning, marked with a hyphen between them, e.g. *književno-povijesni* 'litero-historical').

The recognition and description of complex word-formation processes in Croatian is a non-trivial task due to their non-predictability and numerous phonological changes at morpheme boundaries. We tend to overcome these obstacles by recognizing all allomorphs[2] of the same morpheme and linking them to it. Moreover, we do not only want to recognize the specific word-formation rule in question, but also the complete morphological structure of each lemma in CroDeriV. In other words, we want to be able to recognize its root in order to gain insight into complete derivational spans, as has been already pointed out in the preceding sections.

Word-formation processes in Croatian are very similar to word-formation processes in other Slavic languages, particularly to South Slavic languages as e.g. Slovene or Serbian. Therefore, a resource built along the principles as described here could be useful for the development of NLP tools for these languages, as well.

4 The Initial Stage of CroDeriV Development

The main impulse for the building of CroDeriV was the incorporation of derivationally related verbs into Croatian WordNet. This is the reason we decided to start from the verbal part of the lexicon. We have collected approximately 14,000 verbal lemmas (verbs in the infinitive form).

For the purpose of speeding up the morphological analysis, we devised a simple naïve brute-force algorithm based on a small set of linguistic rules. In the first step, we removed 19 productive prefixes. Since prefixation is a recursive process in Croatian (one base can have from zero to four prefixes), this enabled the recognition of the attested prefixal combinations used in verb formation. In the second step, the rules for the removal of suffixes were applied. All Croatian verbs have at least two suffixes denoting aspect and conjugational class before the infinitive ending *-ti* or *-ći*. Optionally, verbs can have another derivational suffix, attached to the root and bearing specific, diminutive or pejorative meanings. Finally, a manual check of the automatic analysis was performed due to the phonological overlapping of affixes and roots, which often resulted in incorrect segmentation.

[2] An allomorph is a variant form of a morpheme.

In this step we also connected all allomorphs to one mutual morph in the underlying representation and added some additional linguistic information about verbal aspect and reflexivity. Moreover, the stems were attached to the roots they are related to. Stems can be either productive (i.e. used in the derivation of at least two verbs) or unproductive (used in a single verb formation).

Therefore, the next step was to devise a data model which would enable us to store this data in a structured way and thus facilitate various types of research. Since the verbal morphological structure is rather rigid, the generalization of this structure for all verbs seemed the most logical. Our decision was to present every lemma as a series of slots which can be either filled or empty. These slots were arranged as follows (P = prefix, St = stem, I = interfix, Su = suffix, E = inflectional ending; square brackets = slot can be empty): [P4] + [P3] + [P2] + [P1] + [St2] + [I] + St1 + [Su3] + Su2 + Su1 + E. In this model, every slot is assigned its own semantics. For instance, Su1 will always contain suffixes which define the verb's inflectional class. However, when we started to explore the possibility of expansion of CroDeriV to the nominal part of the lexicon, a rigid structure with a predefined number and order of slots turned out to be inappropriate. When it comes to nouns, the meanings of either prefixes or suffixes in a particular "slot" are not predetermined (e.g. *šal-ic-|a* 'a cup' vs. *šal-ic-|a* 'a little joke', where *šal-* in the first word is not the same root as *šal-* in the second word, and the suffix *-ic-* in the first one has the meaning of a container, whereas in the second one it has a diminutive meaning), and suffixes of the same form with some shared meaning components can come in different relative distances from the root (*prija-telj-ic-|a* 'a female friend' vs. *lav-ic-|a* 'a female lion'). Since the morphological structure of nouns differs significantly from the morphological structure of verbs, we decided to introduce a completely different data-model which would be able to comprise lemmas of different POS.

5 Redesigning CroDeriV

In the inital phase of CroDeriV development, the morphological structure of the entries was predefined, the derivation was described in the form of a final state, and a derivational process could be computed as a change between two states. In the redesigned database, derivation is represented as a sequence of derivational steps (or phases) which consist of simply adding one combining element (a morpheme or a derivative) to some previous phase. The process starts with a single morpheme, which gets combined (prepended or appended) by only one combining element in each step.

The product of each step is one type of derivative, which inherits the morphological structure from its predecessor and is built upon it. For example, the noun *učiteljica* 'female teacher' is derived form *uč-i-telj-* (*uč-i-telj-|Ø* 'male teacher') by adding the suffix *-ic-* and the inflectional ending *-a*. However, it inherits the complete morphological structure of *uč-i-telj-*, and its underlying representation is *uč-i-telj-ic-|a*. With this design, we automatically solved the problem of

alloting a definite number of slots for morphemes and determining their order (see the *prijateljica - lavica* example). The semantics of a morpheme is not defined by its slot in the morphological structure, but is rather provided by a set of assigned features, which will be further explained later.

It is important to stress that we did not change the way we understand derivation; we simply chose a more flexible model of description. Derivation is comprehended as a sequential process, but in the previous design, it was implicitly coded as the difference between two states. In the redesigned database, the products of each step in the derivational process are stored regardless of whether they:

a) can form words by adding inflectional endings (these will be referred to as *full derivatives*, e.g. *pis-ač-* → *pis-ač-*|*Ø* 'printer'),
b) can productively form other derivatives serving as stems, (e.g., *pis-Ø-iva-* (**pis-Ø-iva-*|*ti*) will serve as a stem for *na-pis-Ø-iva-*|*ti*, *ras-pis-Ø-iva-*|*ti*, etc.),
c) can simply be an intermediate phase in becoming one or the other.

The types of derivatives are described in more detail in the following section.

Since the new model of the database is not based upon a predefined morphological structure, the exact order of the derivational processes had to be established. By establishing the exact order of derivational processes we ensured that sequential building of the morphological structure will always branch in a predictable way and that there will be no derivational phases stored in the database that cannot be considered plausible nor can be defended by any formal morphological theory (e.g. **do-*$\surd$*pis-Ø-* in *na-do-*$\surd$*pis-Ø-a-*|*ti*). In our model we chose the following order: 1) suffixation, 2) prefixation, and 3) compounding. Complex derivational processes described in Croatian grammars, such as prefixal-suffixal derivation, are decomposed into simple phases and executed in the default order.

This will minimize the number of derivational phases necessary for the morphological description of all words in the database. Also, the phases stored in the database do not always reflect the actual derivational stages words undergo. For example, *uč-i-telj-*|*Ø* 'teacher' is derived from *uč-Ø-i-*|*ti* 'to teach', but in this model we cannot produce *uč-i-telj-* from *uč-Ø-i-* since the elision of one suffix, which is not supported as an operation, would be required. So, in the database *uč-i-telj-* will be represented as being derived from *uč-*. Therefore, the derivational phases are simply an economical way of storing morphological data. Although in our derivational model *učitelj* is not represented as directly derived from *učiti*, the direct derivational relation between them is established separately. These "real" derivational relations are established only between lemmata, i.e. full-fledged words.

Besides the process-like description of derivation, the other significant difference between the previous and the present model is that we separated derivation from inflection. These two morphological processes are described in separate tables in the database. The database comprises four main tables:

1. a derivational table,

2. an inflectional table,

3. a morpheme table (including all morphemes, lexical and grammatical),

4. a relations table (modeling the relations between lemmata).

5.1 Derivation

Each entry in a derivational table represents one step in a derivational process in which only one combining element is added. Each entry consists of:

1. a **derivative text** - the surface form of an entry;

2. a **starting derivative** - the derivative from which the formation of the present phase starts. When this field is empty, derivation starts from this point;

3. a **combining element** - a derivational morpheme or morpheme cluster which takes part in a derivational step. This element is appended or prepended to a starting derivative with respect to the derivation type, e.g. the starting derivative *uč-i-telj-* (*uč-i-telj-*|*Ø* 'teacher') and the combining element *ic-* form the full derivative *uč-i-telj-ic-* (*uč-i-telj-ic-*|*a* 'a female teacher') via suffixation;

4. a **derivative type** – thus produced derivatives can be classified as follows[3]:
   (a) *Full derivatives* can produce words by adding inflectional endings but also can continue their derivational process serving as a stem. For example, *uč-i-telj-* can undergo inflection and become *uč-i-telj-*|*Ø* 'teacher', but also can derive *uč-i-telj-ic-* (*uč-i-telj-ic-*|*a* 'female teacher'). Only derivatives classified as *full derivatives* will serve as inflectional stems and therefore are referenced in the inflectional table. If they participate in further derivational processes, they serve as derivational stems (here the word *stem* is used in its strict sense as a linguistic term). In other words, they can serve at the same time as an inflectional and derivational base.

   (b) *Intermediate derivatives* are those that have not finished some non-optional process. They can not form words, nor can they serve as a combining element for derivation - they must continue their derivational process and the choice of combining elements which can be combined with them is very limited. For example, the derivative *-Ø-*, consisted

[3] The purpose of these categories is merely for filtering and they have no deeper linguistic meaning.

only of the verbal aspectual suffix, will be referred to as *intermediate*, because it has to be first appended by the verbal class suffix to participate in any other derivational process;

(c) *Stems* cannot form words but are productive and serve as the basis for further derivation (e.g., *uč-Ø-ava-* → *na-uč-Ø-ava-|ti*, *pod-uč-Ø-ava-|ti*, etc.). Stems normally serve as derivational stems and continue the derivation process which can branch in more than one direction;

(d) *Prefix clusters* and *suffix clusters* are derivatives composed of one or more affixes of the same type. In the process of affixation (prefixation or suffixation) these derivatives will function as a combining element.

5. a **derivational type** - there are seven derivational types recognized in CroDeriV:

   (a) *cloning* (in which a morpheme starts the derivation process),

   (b) *suffixation* (in which a derivative is appended by a suffix cluster),

   (c) *prefixation* (in which a derivative is prepended by a prefix cluster),

   (d) *suffix composition* (in which a suffixal derivative is appended by another suffix),

   (e) *prefix composition* (in which a prefixal derivative is prepended by another prefix),

   (f) *compounding_interfix* (in which a derivative is prepended by an interfix),

   (g) *compounding* (in which a derivative gets prepended by a derivative);

6. a **derivative slug** – a textual representation of the morphological analysis of the derivative's surface form in which morphemes are segmented with hyphens. The purpose of this field is twofold: first is to present the morphological analysis in a human-readable manner, and second is to facilitate faster and more efficient search of database entries according to morphemes they consist of.[4]

7. the corresponding **underlying representation** is a complex structure to which a derivative is linked. It consists of all the morphemes contained within it with their respective order in the structure.

[4] This type of "slug" field will be also used in a table in which underlying representations of the entries are stored (see 7.), with a difference that it will contain underlying, instead of a surface form segmented with hyphens. Every "slug" field in the database will be indexed and easily searchable using regular expressions.

A partial representation of the derivational processes including the root *pis-* in a derivational table is shown in Figure 1.

5.2 Inflection

The derivational table represented above is connected to the inflectional table. When a full derivative from the derivational table receives its inflectional ending, the actual word is formed. For example, the full derivative *uč-i-telj-* becomes a masculine noun in the nominative singular only by acquiring the nominative ending *-Ø*. Every entry from this table is a lemma tagged with the following attributes: surface form, reference to the corresponding full derivative in the derivation table, inflectional ending, underlying representation and grammatical categories (POS and MSD). The final set of features or feature values is assigned to a word by particular morphemes.

Feature values are listed in a separate table which is referenced both in the morpheme table and inflectional table. Values are grouped by feature types (morphological, syntactic, semantic, morphosemantic, etc...) and feature names. Values pertaining to the same name are logically mutually exclusive. For example, one of the morphological features is the aspect, which can be perfective, imperfective, or biaspectual. Aspect is encoded by the particular verbal suffix, and the same verb cannot be at the same time perfective and imperfective. This constraint is ignored when features are attached to single morphemes and not to words, because a single morpheme often carries mutually incompatible or exclusive features (e.g., the morpheme *-Ø-* can have either imperfective or perfective meaning, but in a particular word, only one of them is realized). The set of features attached to an inflected form is usually a subset of the union of features attached to the morphemes from which a particular full derivative is combined. Not all of these features are already included in CroDeriV, but they can easily be incorporated at later stages of development, due to its flexible design.

The lemmas, or the entries in the inflectional table, are connected by a set of relations. The new model supports various, but always symmetrical relations. One of the most important relations in the database is, of course, the relation *derives↔is_derived_from*. As we said earlier, the direct derivational relation between two lemmas in some cases cannot be straightforwardly established. In other words, the direct derivational relations have to be explicitly established. By connecting words using derivational relations, we are actually building a network-like structure which illustrates the complete derivational span of a particular lexical morpheme across different POS.

The other types of relations have yet to be included in the CroDeriV.

A graphical representation of the CroDeriV structure is shown in Figure 2.

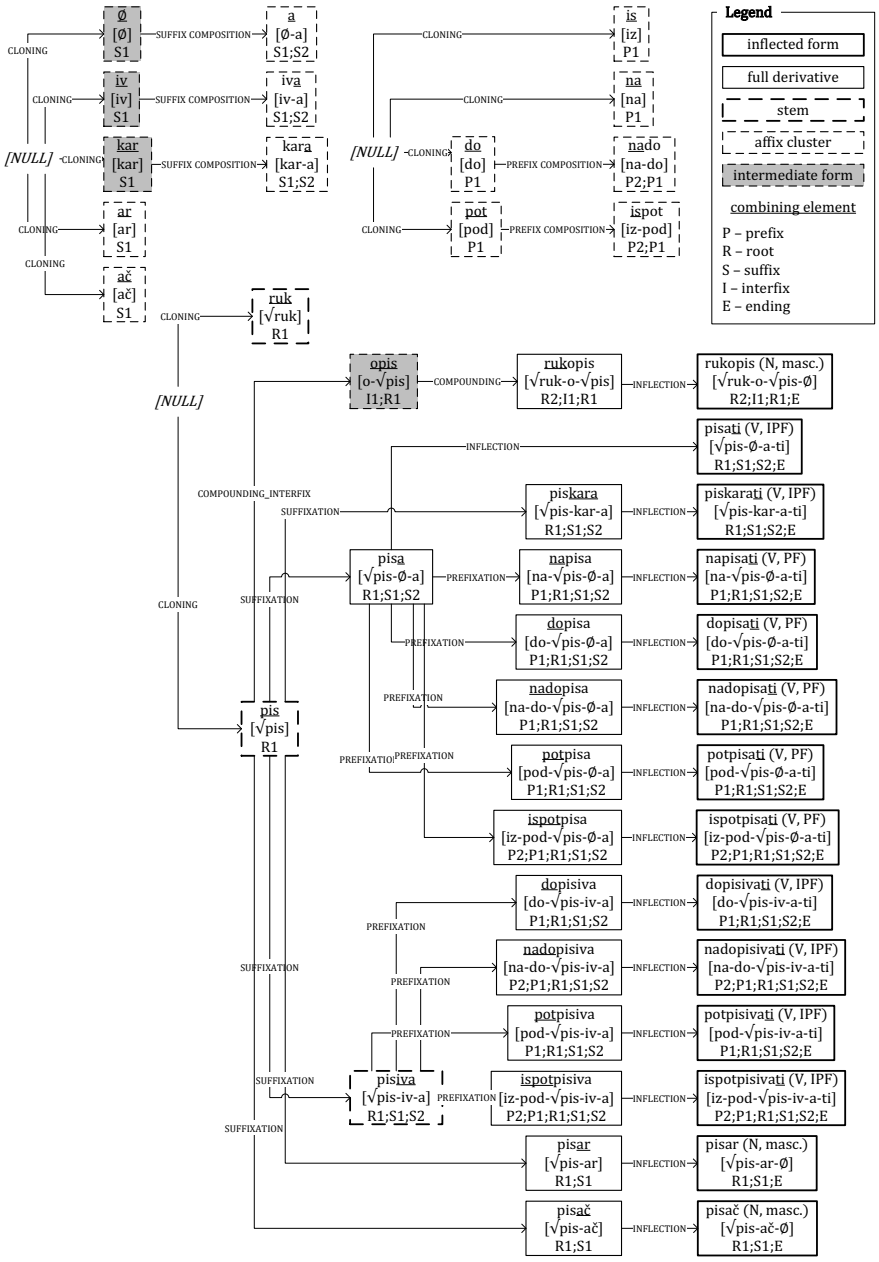

Fig. 1. Schematic representation of the database entries and their mutual relations

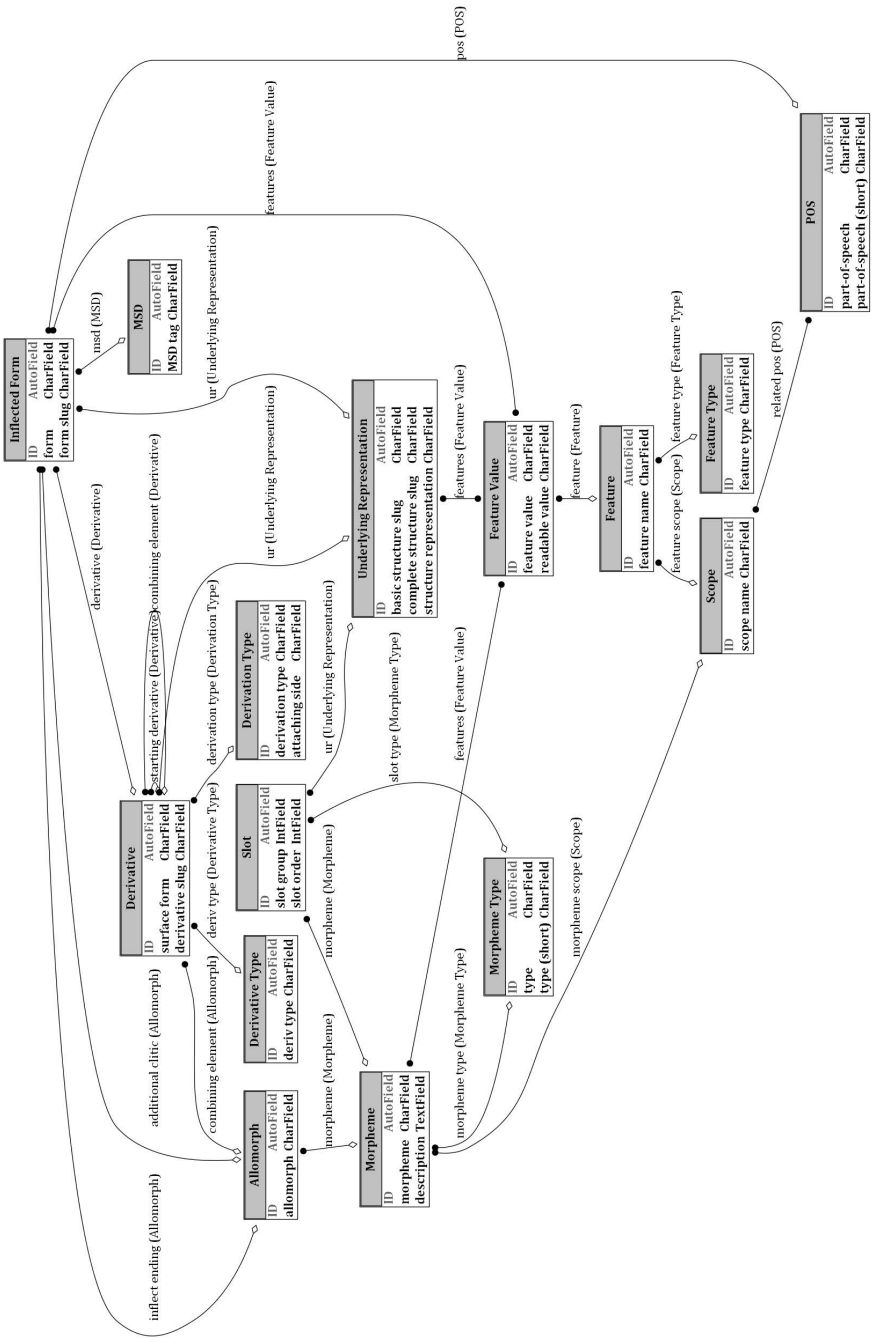

Fig. 2. CroDeriV Data Model

6 Future Work and Conclusion

As we have already stated, CroDeriV is a unique language resource for the Croatian language. Even in the earliest stage of its development, it has been shown to be an extremely valuable source of data for different kinds of linguistic research. For now, we have carried out only a preliminary analysis on the frequency of particular derivational affixes and their attested combinations. These results have already provided new insight into derivational processes in Croatian, since this kind of information cannot be found in Croatian grammars, nor have studies of this type, to the best of our knowledge, ever been done on a representative language data sample. Our plan is to go further: we would like to explore how certain affixes (or their combinations) are good predictors of grammatical (morphological, syntactic, semantic, etc...) features and how they are correlated. One of the features that especially interests us is the alteration of verbal aspect which can occur under different derivational processes. Another interest of ours is to see how syntactic features, such as verbal syntactic frames, are influenced by prefixation. Furthermore, we would like to explore the possibilities of (semi-)automatic expansion of Croatian WordNet by making use of morphosemantic features of derivational affixes. Also, our plans go towards building language tools on the basis of this data, such as a morphological analyzer which would be able to perform both deep and shallow analysis. A good morphological (derivational) analyzer would be useful not only as a part of other language tools, but also for the further expansion of the database.

In this paper we have presented the redesign of CroDeriV, the Croatian Derivational Database. Despite the fact that it is merely a technical solution, we tried to build it in accordance with morphological theories. We believe that the data model we have presented will enable us to encode all derivational processes in the Croatian language. We also believe that our model is flexible and abstract enough to be used for other morphologically "rich" languages.

Derivational morphology is an important part of lexicalization processes in Slavic languages. In predominantly analitical languages, as e.g. English, derivational morphology does not have such a prominent role when compared to Slavic. We believe that derivational databases as the one described in this paper could facilitate the development of NLP tools for Slavic languages or improve the quality of existing ones. We are convinced that the theoretical model that CroDeriV is based upon can be applied to other Slavic languages with only minor modifications. This especially pertains to South Slavic languages as e.g. Slovene and Serbian. [5]

Acknowledgments. This work has been supported by XLike (Cross-lingual Knowledge Extraction), a project in the area of Language Technologies (ICT-2011.4.2) funded by the European Community's Seventh Framework Programme FP7/2007-2013.

[5] This resource will be freely available as soon as possible, probably under CC BY-NC-SA 3.0 licence.

References

1. Bernhard, D., Cartoni, B., Tribout, D.: Evaluating Morphological Resources: a Task-Based Study for French Question Answering. In: Proceedings of the International Workshop on Lexical Resources at ESSLII, Slovenia (2011)
2. Ćavar, D., Jazbec, I., Runjaić, S.: Interoperability and Rapid Bootstrapping of Morphological Parsing and Annotation Automata. In: Erjavec, T., Žganec, G., Jerneja (eds.) Proceedings of the Sixth Language Technologies Conference, Proceedings of the 11th International Multiconference Information Society, IS 2008, October 16-17, vol. C, pp. 80–85 (2008)
3. Ćavar, D., Jazbec, I., Stojanov, T.: CroMo Morphological Analysis for Standard Croatian and its Synchronic and Diachronic Dialects and Variants. In: Finite-State Methods and Natural Language Processing - Post-proceedings of the 7th International Workshop FSMNLP, pp. 183–190. IOS Press, Italy (2009)
4. Ljubešić, N., Boras, D., Kubelka, O.: Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. In: Seljan, S., Stančić, H. (eds.) INFuture2007: Digital Information and Heritage, pp. 313–320. Odsjek za informacijske znanosti Filozofskoga fakulteta, Zagreb (2007)
5. Pandžić, I.: Oblikovanje korjenovatelja za hrvatski jezik u svrhu pretrazivanja informacija. MA thesis. University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics
6. Raffaelli, I., Tadić, M., Bekavac, B., Agić, Z.: Building Croatian WordNet. In: Proceedings of the 4th Global WordNet Conference, pp. 349–359. Global WordNet Association, Szeged (2008)
7. Sedláček, R., Smrž, P.: Automatic Processing of Czech Inflectional and Derivative Morphology. In: FI MU Report Series. Masaryk Univesity: Faculty of Informatics (2001)
8. Šnajder, J.: Morfološka normalizacija tekstova na hrvatskome jeziku za dubinsku analizu i pretraživanje informacija. PhD thesis. University of Zagreb, Faculty of Electrical Engineering and Computing (2008)
9. Šojat, K., Srebačić, M., Tadić, M.: Derivational and Semantic Relations of Croatian Verbs. Journal of Language Modelling. O.1, 111–142 (2012)
10. Tadić, M., Fulgosi, S.: Building the Croatian Morphology Lexicon. In: Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages, pp. 41–46. ACL, Budapest (2003)
11. Tadić, M., Oliver, A.: Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In: LREC 2004 Proceedings, pp. 1259–1262. ELRA, Paris-Lisabon (2004)
12. Tadić, M., Bekavac, B.: Inflectionally Sensitive Web Search in Croatian using Croatian Lemmatization Server. In: Proceedings of ITI 2006 Conference, SRCE, Zagreb (2004)
13. Woliński, M.: Morfeusz a practical tool for the morphological analysis of Polish. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining 2006 Conference, pp. 511–520. Wisła, Poland (2006)