Redesign of the Croatian derivational lexicon

Matea Filko
Faculty of Humanities and Social Sciences
University of Zagreb
matea.filko@ffzg.hr

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

Vanja Štefanec
Faculty of Humanities and Social Sciences
University of Zagreb
vstefane@ffzg.hr

Abstract

This paper deals with the redesign of the Croatian derivational lexicon – CroDeriV. In its first online version the lexicon consisted solely of verbs analyzed for morphemes. In further steps of its development, lexemes of other POS (adjectives, nouns) are analyzed, both in terms of their morphological structure and word-formation patterns, and imported into the lexicon. Dealing with new POS as well as the annotation of word-formation patterns among lexemes required the modification of the database structure. In this paper we present a restructured version of the database, adapted to include other POS and to explicitly mark word-formation patterns. These procedures enable precise and refined queries based on various parameters through the online search interface.

1 Introduction

Although the development of language resources dealing with word-formation has begun almost twenty years ago, derivational resources nowadays exist for a relatively limited number of languages (CatVar (Habash and Dorr, 2003) for English; Démonette (Hathout and Namer, 2014) for French; DeriNet (Žabokrtský et al., 2016; Ševčíková and Žabokrtský, 2014) and Derivancze (Pala and Šmerk, 2015) for Czech; Word Formation Latin (Passarotti and Mambrini, 2012; Litta et al., 2016) for Latin; DerIvaTario (Talamo et al., 2016) for Italian; DErivBase (Bajestan et al., 2017; Zeller et al., 2013) for German and DErivBase.HR (Šnajder, 2014) for Croatian). These resources predominantly focus on the annotation of word-formation processes within and across derivational families, i.e. among lexemes with the same root. The majority of them does not take into account the morphemics, in other words, they do not mark the complete morphological structure of lexemes. Procedures applied in their development range from automatic or semi-automatic to completely manual.

Croatian is a Slavic language with rich morphological processes. High-quality language resources dealing with the morphological structure and derivational relations of Croatian lexemes are needed in numerous NLP tasks and they are valuable for various theoretical research. In this paper, we present the development and enrichment of the existing version of the Croatian derivational lexicon – CroDeriV (Šojat et al., 2013).[1] Procedures applied in the building of this lexicon significantly differ from those listed above: 1) the previous version of the lexicon contained only verbs[2]; 2) the focus was on a thorough analysis of the morphological structure of lexemes, whereas word-formation processes among them were not explicitly marked. In the second phase of its development, its structure has been expanded with words of other POS and the representation of derivational relations between stems and derivatives has been introduced. Consequently, the online interface has been adapted to offer a wider range of possible queries.

The paper is structured as follows: in Section 2 we present the current structure of the derivational lexicon and possible queries via online interface; in Section 3 we discuss how the analysis of verbal derivational families used so far can be applied to adjectives and nouns; Section 4 presents the new structure of the database and new query parameters. Section 5 brings concluding remarks and the outline of future work.

[1] The search interface of the current version of the lexicon is available at croderiv.ffzg.hr.

[2] Cf. (Šojat et al., 2012) for the motivation to include only verbs in the first phase of the lexicon development.

Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 71–80,
Prague, Czechia, 19-20 September 2019.

2 Croatian derivational lexicon v. 1.0

In its first version, the derivational lexicon consisted of ca 14.500 verbs collected from two large Croatian corpora (Croatian National Corpus (Tadić, 2009), and Croatian web corpus hrWaC (Ljubešić and Klubička, 2014)) and free online dictionaries. All verbal lemmas, i.e. their infinitive forms, were segmented into morphemes and verbs sharing the same root were grouped into derivational families. As in other Slavic languages, aspect is an inherent category of Croatian verbs (Marković, 2012, 183). Each verb was therefore additionally marked as perfective, imperfective or bi-aspectual.[3]

The morphological segmentation was divided into two steps: 1) automatic segmentation via rules based on lists of various derivational affixes; 2) manual checking of the results which was necessary due to extensive homography and allomorphy of affixes and roots. Thus, all the homographic forms were manually disambiguated and all the allomorphs were linked to single representative morphemes. This line of processing resulted in a two-layer annotation consisting of a surface and a deep layer.

At the surface, all allomorphs were identified and marked for their type. Possible types of morphemes recognized in Croatian lexemes are prefixes, roots, derivational suffixes, inflectional suffixes and interfixes for compounds. The surface form of the verb *ispuniti* 'to fulfill, to fill out' can be represented as follows:

is-pun-i-ti

is = prefix; *pun* = root; *i* = derivational (thematic) suffix; *ti* = inflectional (infinitive) suffix,

whereas the verb *odobrovoljiti* 'to cheer up' was segmented as follows:

o-dobr-o-volj-i-ti

o = prefix; *dobr* = root2; *o* = interfix; *volj* = root1; *i* = derivational suffix; *ti* = inflectional (infinitive) suffix.

At the deep layer, the prefixal allomorph *is* was connected to its representative morph *iz*. The representative morph is the one from which other allomorphs can be established with the least number of morpho-phonological rules. This kind of analysis enables queries over roots and all derivatives within derivational families, but also over specific affixes or even their combinations (prefixal, suffixal and both) used in various derivational families.[4] The CroDeriV database is available online and it has already been widely used for research and teaching purposes. As indicated, this version of the derivational lexicon is limited in two ways: 1) it is restricted to only one POS; 2) derivational relations between lexemes are not represented. In the following sections, we discuss how the database originally structured for the full analysis of Croatian verbal morphology was modified and expanded.

3 Croatian derivational lexicon v. 2.0

The expansion of the derivational lexicon followed the principles set in previous phases. First, nominal and adjectival lemmas were collected from corpora and online dictionaries of Croatian. In order to obtain a representative sample for further analysis and processing, we chose approx. 6.000 nouns and 1.000 adjectives according to their frequency indicated by the Croatian frequency dictionary (Moguš et al., 1999) and frequency lists generated by corpus management system NoSketchEngine for both representative corpora (Croatian National Corpus and Croatian web corpus hrWaC).[5] Both motivated and unmotivated lexemes were included in our analysis. They were added to the lexicon, in order to capture the word-formational path from the base, unmotivated lexeme, to the final, motivated lexeme. However, named entities were excluded from the lists, since they are not formed via productive word-formation patterns in Croatian (Babić, 2002, 16).

[3] Verbal aspectual pairs are considered separate lemmas in Croatian. Therefore, the so-called thematic suffixes, as *-i-* in *is-pun-i-ti* (see the example in this section), are classified as derivational suffixes (Marković, 2012, 188). Apart from derivation of aspectual pairs, these suffixes are also used to form verbs from other parts of speech, e.g. adjectives or nouns (*pun* 'full' – *pun-i-ti* 'to fill, imperfective' – *is-pun-i-ti* 'to fulfill, perfective'; *rad* 'work' – *rad-i-ti* 'to work, imperfective' – *za-rad-i-ti* 'to earn, perfective'), which is another proof of their derivational status.

[4] The extensive statistics on roots, affixes and their combinations in Croatian is presented in (Šojat et al., 2013).

[5] The procedure of collection and analysis of adjectives is thoroughly described in (Filko and Šojat, 2017). The number of approx. 6.000 nouns was obtained by merging the lists of 5.000 most frequent nouns from the above-mentioned sources.

The next steps consisted of 1) the manual segmentation of lexemes into morphemes, and 2) the analysis of their morphological structure. Our main objective was to establish general rules pertaining to their morphological structure and relevant word-formation processes, both POS-maintaining and POS-changing. The aim of the whole procedure is to enable a rule-based procedure for automatic morphological segmentation to be applied to the rest of the compiled data.[6]

As opposed to verbs, predominantly formed via prefixation or highly-regular suffixation from other verbs (Šojat et al., 2012), nouns and adjectives are mostly formed by means of suffixation. Babić (2002) lists 526 nominal and 160 adjectival suffixes out of the total of 771 suffixes used in Croatian. Although these data are useful in many aspects, the productivity of certain affixes is not provided. Productivity here refers to the number of co-occurrences of an affix and various stems as recorded in data, i.e. the number of different lexemes formed via particular derivational affix. Preliminary research clearly shows that a relatively small subset of suffixes compared to the numbers listed above is used for nominal and adjectival derivation (in our sample, at least). The results will be used for the creation of rules for morphological segmentation in the analysis of the remaining data.

Generally, the morphological segmentation is based on the two-layered approach previously applied to verbs: at the surface layer all possible morphs are identified and marked for their type; at the deep layer allomorphs are connected to the single representative morph, e.g. the noun *učiteljica* 'female teacher' was segmented as follows:

uč-i-telj-ic-a

uč = root; *i*, *telj*, *ic* = derivational suffixes; *a* = inflectional suffix,

whereas the adjective *izlječiv* 'curable' was segmented and processed as follows:

iz-lječ-iv-Ø

iz = prefix; *lječ* = root; *iv* = derivational suffix; *Ø* = inflectional suffix, and the allomorph *lječ* is at the deep layer connected to the representative root morph *lijek*.

The morphological structure of lexemes regardless of their part of speech consists of the following types of morphemes: prefixes, roots, interfixes, and derivational and inflectional suffixes. Each morpheme type can occur more than once in the morphological structure.

In the next step, derivational relations among selected lexemes were annotated. After the lexemes were morphologically segmented and all the allomorphs were linked to representative morphemes, the stem, and the word-formation pattern was determined for each lexeme in the database. Lexemes are POS-tagged and motivated lexemes are derivationally linked to their base lexeme. The derivational connection is established only if there are simultaneous phonological and semantic relations between the base and the derived lexeme (Babić, 2002, 25). In other words, no derivational connection exists 1) between suppletive allomorphs, 2) between two lexemes with diachronically remotely connected meanings. However, if the derivational connection is synchronically transparent, the derivational link is established, inspite of the significant shift in meaning from the base word to the derived lexeme (e.g. *čeznuti* 'to long for' *iščeznuti* 'to vanish'). In that case, the affix sense is marked as idiosyncratic.

The new version of the database thus provides the information on the following word-formational properties:

- word-formation pattern: ***učiteljica*** < *učitelj* + *ica* [suffixation]; ***izlječiv*** < *izliječiti* + *iv* [suffixation]
- allomorph of the stem – stem: *učitelj* – *učitelj*; *izlječ* – *izliječ*
- allomorph of the affix – affix: *ica* – *ica*; *iv* – *iv*
- affix sense: agent, feminine; possibility
- POS of the stem: N; V.[7]

[6] A more straightforward rule-based procedure based on a simple set of rules for the detection and segmentation of single nominal suffixes was applied in (Šojat et al., 2014). However, the main goal of this procedure was to detect words of the same derivational family, not to analyze their morphological structure.

[7] This representation is in line with Babić (2002, 16), probably the most extensive and thorough book on word-formation for a Slavic language, where it is stated that derivational representation should at least show 1) word-formational units (affixes); 2) word-formational stems; 3) types of word-formation processes; 4) meanings of derived words.

The word-formation patterns and affixal senses as presented in our lexicon are explained in more detail in the following subsections.

3.1 Word-formation patterns

We take into account word-formation processes in Croatian that are recorded and described in relevant reference literature:

1. **suffixation:**
   - *pjev(ati)* 'to sing' + *-ač* > *pjevač* 'singer'
   - *glas* 'voice' + *-ati* [8] > *glasati* 'to vote'
   - *učitelj* 'teacher' + *-ev* > *učiteljev* 'teacher's'
2. **prefixation:**
   - *za-* + *pjev(ati)* 'to sing' > *zapjevati* 'to start singing'
   - *do-* + *predsjednik* 'president' > *dopredsjednik* 'vicepresident'
   - *pred-* + *školski* 'school, ADJ' > *predškolski* 'preschool'
3. **simultaneous suffixation and prefixation:**
   - *o-* + *svoj* 'one's own' + *-iti* > *osvojiti* 'to conquer, to win'
   - *bez-* + *sadržaj* 'content' + *-an*> *besadržajan* 'pointless, content-free'
4. **compounding:**
   - *vjer(a)* 'trust' + *-o-* + *dostojan* 'worthy' > *vjerodostojan* 'trustworthy'
   - *zlo* 'evil' + *upotrijebiti* 'to use' > *zloupotrijebiti* 'to misuse, to abuse'
   - *polu* 'half' + *mjesečni* 'monthly'> *polumjesečni* 'semimonthly'
5. **simultaneous compounding and suffixation:**
   - *vod(a)* + *-o-* + *staj(ati)* 'to stand' > *vodostaj* 'water level'
   - *vanjsk(a)* 'external' + *-o-* + *trgovin(a)* 'trade' + *-ski* > *vanjskotrgovinski* 'external trade, ADJ'
6. **simultaneous prefixation and compounding:**
   - *o-* + *zlo* 'evil' + *glasiti* 'to say' > *ozloglasiti* 'to discredit, to bring into disrepute'
7. **back-formation:**
   - *izlaz(iti)* 'to exit' > *izlaz* 'exit'
8. **conversion or zero-derivation:**
   - *mlada* 'young, feminine, ADJ' > *mlada* 'bride, N'
9. **ablaut:**
   - *plesti* = *plet* + *(Ø)* + *(ti)* 'to twine' > *plot* 'fence'.

The word-formation pattern at the same time indicates the type of the word-formation process. In determining the word-formation pattern, we take into account only the last step in the formation of the particular lexeme. For example, the verb *ispunjavati* 'to fulfill, imperfective' is derivationally related to the verb *puniti* 'to fill, imperfective' via indirect derivational connection. However, it is directly formed from the verb *ispuniti* 'to fulfill, perfective'. Therefore, we mark only this last derivational step in the word-formation pattern:

[8] In traditional approaches, thematic suffix and infinitive ending are considered as one word-formational element consisting of two morphemes.

ispun(iti) 'to fulfill, perfective' + *-javati* > *ispunjavati* 'to fulfill, imperfective' [suffixation].
The remote derivational link is available via word-formation pattern of the verb *ispuniti* 'to fulfill, perfective':
is- + *puniti* 'to fill, imperfective' > *ispuniti* 'to fullfil, perfective' [prefixation].
In some cases, it is hard to determine the word-formation pattern due to several plausible possibilities, especially when dealing with suffixation. In these cases, we follow the criteria established in Babić (2002, 38–41):

- if one of the competing solutions increases the overall number of derivational units in Croatian, the other solution should be selected as the more appropriate one;
- if one of the competitive solutions can be applied to the wide range of motivated lexemes, and others cannot, the first solution should be selected as the more appropriate one.

3.2 Affixal senses

Affixes are in our database structured as polysemous units, which is in line with recent approaches to affixal meanings (Babić (2002, 38), Lehrer (2003), Lieber (2004, 11), Lieber (2009, 41), Aronoff and Fudeman (2011, 140–141)). In relation to other constituents of the word-formation pattern, one of the affixal meanings is realized in the final motivated lexeme. For example, verbal prefix *nad-* can have two meanings. It can express:

1. **location** (subtype: *over*), e.g. *letjeti* 'to fly' > *nadletjeti* 'to fly over'
2. **quantity** (subtype: *exceeding*), e.g. *rasti* 'to grow' > *nadrasti* 'to outgrow'.

The detailed typology of possible meanings of verbal prefixes in Croatian is explained in Šojat et al. (2012), whereas possible meanings of the most productive adjectival suffixes are discussed in Filko and Šojat (2017). The inventory of possible affixal meanings for Croatian nouns is designed according to descriptions in Croatian grammar and reference books. Affixes and their meanings are treated differently in Croatian literature. Whereas some grammar books (e.g .(Babić, 2002)) list affixes alphabetically and note their possible meanings, the others (e.g. Silić and Pranjković (2005) and Barić et al. (1995)) list possible meanings of motivated words (e.g. diminutives, locations, instruments, male agents, female agents, animals, etc.) and indicate which affixes can be used for the creation of these meanings. This, in other words, means that suffixes are grouped according to at least one of their meanings. We combined the information from these sources and modified the final polysemous structure of affixes if needed according to the lexemes in our database. The above-mentioned nominal suffix *-ica* can express at least the following meanings[9]:

1. **agent, female**, e.g. *učitelj* 'teacher, male' > *učiteljica* 'teacher, female'
2. **person, both sexes**, e.g. *izbjegao* 'exiled' > *izbjeglica* 'refugee'
3. **animal, female**, e.g. *golub* 'pigeon, male' > *golubica* 'pigeon, female'
4. **diminutive**, e.g. *pjesma* 'song' > *pjesmica* 'ditty, rhyme'
5. **thing**, e.g. *sanjar* 'dreamer, male' > *sanjarica* 'dream book'
6. **drink**, e.g. *med* 'honey' > *medica* 'honey liqueur'
7. **plant**, e.g. *otrovan* 'poisonous' > *otrovnica* 'poisonous plant, mushroom (and venomous snake)'
8. **location**, e.g. *okolo* 'around' > *okolica* 'surrounding'
9. **temporal mark**, e.g. *godišnji* 'yearly' > *godišnjica* 'anniversary'

[9] These are the meanings annotated so far in our material. For a more extensive account, including idiosyncratic meanings, cf. Babić (2002, 183–189)

10. **disease**, e.g. *vruć* 'hot' > *vrućica* 'fever'

11. **literary type**, e.g. *slovo* 'letter' > *poslovica* 'saying'

12. **linguistic term – type of word/sentence**, e.g. izveden 'derived, ADJ' > izvedenica 'derived lexeme'

13. **number of men involved**, e.g. *dvoje* 'two, of different gender' > *dvojica* 'two, of male gender'

14. **anatomical part**, e.g. *jagoda* 'strawberry' > *jagodica* 'cheekbone, fingertip'

In the following section, we present the redesign of the database based on the analysis of the initial set of nouns and adjectives in terms of their morphological structure and word-formation properties.

4 Redesign of the database

In its first publicly available version, the CroDeriV database was structured according to the generalized morphological structure of Croatian verbs, consisting of four slots for prefixes (P), two slots for stems (L), one slot for an interfix between two stems (I), three slots for derivational suffixes (S) and one slot for the inflectional suffix (END). This structure is sufficient to accommodate all Croatian verbs[10]:

(P4) (P3) (P2) (P1) (L2) (I) **L1** (S3) **S2** **S1** **END**.

Online queries are possible across several categories: P2, P1, L1, S2, S1, lemma, and their combinations. Additionally, information about the root is shown by the mouse hover over the stem, and the information about the aspect and reflexivity of verbs is shown by clicking on the Details button. Apart from listing all verbs with the same root, i.e. from the same derivational families, other derivational data among lexemes are not presented. In order to include lexemes of other POS and to show derivational relations among them, the database and online search interface had to be modified. We discuss these modifications in the following subsections.

4.1 Expanding to other POS

The generalized morphological structure of Croatian lexemes differs according to their part of speech. The generalized structure is a theoretical construct that serves to represent the maximum number of slots for morphemes and their combinations across various POS (as presented above for verbs). Generally, the maximum number of prefixes recorded in Croatian lexemes is four, the maximum number of roots is six, and the maximum number of suffixes is seven (Marković, 2013).[11] The first version of the database was structured according to the generalized structure provided for verbs. However, this structure cannot be applied to nouns and adjectives due to the complexity of their suffixal parts. In this stage of work, we have to address issues as: 1) how to present the morphological structure of lexemes in terms of roots, derivational and inflectional morphemes in a consistent manner, regardless of their POS; 2) how to present which lexemes belong to the same derivational families, i.e. have identical lexical morphemes; 3) how to present derivational processes applied between stems and derivatives within families as well as affixes or their combinations thereby used. In order to accommodate the lemmas of different POS, the overall structure of the database was re-organized in a POS-independent manner. Therefore, additional suffixal slots (up to seven) are provided to accommodate the morphologically most complex nouns and adjectives.

Various queries over the database are based on the surface form of the lemma analyzed for morphemes.[12] The type of each segmented morpheme (prefix, root, suffix, interfix) is marked. Additional information pertains to the **part of speech** of the entry and specific grammatical categories. For example, **aspect** and **reflexivity** for verbs, **gender** for nouns and adjectives and **definiteness** for adjectives.

[10] Brackets denote that the segment is optional.

[11] In our material, we recorded up to four prefixes for verbal lemmas, as well as three roots and five suffixes (four derivational + one inflectional) for adjectival lemmas.

[12] Although the surface form of the lemma will be presented, only the queries via morphemes and their combinations, not allomorphs, will be enabled.

4.2 Word-formation relations and word-formation patterns

Apart from the complete morphological structure of lemmas and their grammatical categories, the new version of the Croatian derivational lexicon will include information on their word-formation properties. As indicated in Section 3, we manually marked 1) word-formation patterns, 2) allomorphs and morphs of stems, 3) allomorphs and morphs of affixes, and 4) POS of the base word for each nominal and adjectival lemma in the initial set. This, in turn, enables us to automatically acquire information about: 1) the type of the word-formation process applied (suffixation, prefixation, simultaneous suffixation and prefixation, compounding, simultaneous compounding and suffixation/prefixation, back-formation, conversion, ablaut), 2) the nature of the word-formation process applied (POS-changing or POS-maintaining), 3) the base word / root used in word-formation process.[13]

We plan to include this information for each entry in the lexicon. In the new search interface, the information about grammatical categories (1), morphological structure (2-3), and word-formation properties (4-8) (see the example below) will be available by clicking on the lemma. A link to the base word will be available through the word-formation pattern (4 - poslužiti). The list of all derivatives of the same stem will be available through the link on that stem in the entry (5 - posluži). This will enable users to follow complete derivational paths in both directions: from a root to the final derivative (through the link in 4) and from a particular derived word back to the root (through the link in 5). In the future, we plan to provide links to online dictionaries and inflectional lexica for Croatian and to apply a tool for visualization of derivational relations within families.
The complete structures of entries of different POS are as follows:

1. **lemma:** poslužitelj ‘server’
    - **POS:** N
    - **gender**: masculine
2. **morphological structure – surface layer:** po-služ-i-telj-Ø
   (po = prefix, služ = root, i, telj = derivational suffixes, Ø = inflectional suffix)
3. **morphological structure – deep layer:** po-slug-i-telj-Ø
   (po = prefix, slug = root, i, telj = derivational suffixes, Ø = inflectional suffix)
4. **word-formation pattern:** poslužiti[14] + telj
5. **stem (allomorph of the stem)**: posluži[15] (posluži)
6. **affix (allomorph of the affix)**: telj (telj)
7. **affix sense**: instrument
8. **word-formation process** (POS > POS): suffixation (V > N)
9. **link to the Croatian Language Portal**[16].

---

1. **lemma:** potpisati ‘to sign’
    - **POS:** V
    - **aspect**: perfective
    - **reflexivity**: non-reflexive

[13] If there is the base word, then the lemma is morphologically complex or motivated; if the lemma is formed directly from the root, then it is morphologically simple or unmotivated. However, both motivated and unmotivated words are included in the lexicon, in order to obtain the complete word-formational path of the lexical entries.
[14] The base word is underlined and functions as a link to the entry of that word in the lexicon.
[15] The stem is underlined and functions as a link to all lemmas derived directly from this stem, e.g. *poslužilac*.
[16] Online dictionary of Croatian: www.hjp.znanje.hr.

2. **morphological structure – surface layer:** pot-pis-a-ti
(pot = prefix, pis = root, a = derivational suffix, ti = inflectional suffix)

3. **morphological structure – deep layer:** pod-pis-a-ti
(pod = prefix, pis = root, a = derivational suffix, ti = inflectional suffix)

4. **word-formation pattern:** pod + pisati

5. **stem (allomorph of the stem)**: pisati (pisati)

6. **affix (allomorph of the affix)**: pod (pot)

7. **affix sense**: location: under

8. **word-formation process** (POS > POS): prefixation (V > V)

9. **link to the Croatian Language Portal**.

---

1. **lemma:** beskrajan 'endless'
   - **POS:** A
   - **gender**: masculine
   - **definiteness**: indefinite

2. **morphological structure – surface layer:** bes-kraj-an-Ø
(bes = prefix, kraj = root, an = derivational suffix, Ø = inflectional suffix)

3. **morphological structure – deep layer:** bez-kraj-an-Ø
(bez = prefix, kraj = root, an = derivational suffix, Ø = inflectional suffix)

4. **word-formation pattern:** bez + kraj + an

5. **stem (allomorph of the stem)**: kraj (kraj)

6. **affix1 (allomorph of the affix1)**: bez (bes) **affix2 (allomorph of the affix2)**: an (an)

7. **affix1 sense**: deprivation **affix2 sense**: having the property of [meaning of the base]

8. **word-formation process** (POS > POS): simultaneous prefixation and suffixation (N > A)

9. **link to the Croatian Language Portal**.

5 Concluding remarks and future work

In this paper we presented the redesign of the existing version of the Croatian derivational lexicon and its online search interface, required to include non-verbal lemmas into the lexicon, as well as to represent various derivational properties of Croatian lexemes. The Croatian derivational lexicon v. 2.0 is designed to comprise the information about morphological structures, word-formation patterns and derivational relations among Croatian lexemes. We believe that additional information provided for each lemma, e.g. about grammatical categories or external links to online dictionaries, makes this lexicon even more attractive to users.

As mentioned, we intend to use manually analyzed material to build a rule-based automatic procedure for morphological and word-formation analysis. This will facilitate the analysis of new lemmas and their inclusion in the lexicon.

References

Mark Aronoff and Kristen Fudeman. 2011. *What is Morphology. Second Edition*. Wiley-Blackwell, Chichester.

Stjepan Babić. 2002. *Tvorba riječi u hrvatskome književnome jeziku*. Hrvatska akademija znanosti i umjetnosti : Globus, Zagreb.

Elnaz Shafaei Bajestan, Diego Frassinelli, Gabriella Lapesa, and Sebastian Padó. 2017. DErivCelex: Development and Evaluation of a German Derivational Morphology Lexicon based on CELEX. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*. EDUCatt, Milano, pages 117–127.

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, and Marija Znika. 1995. *Hrvatska gramatika*. Školska knjiga, Zagreb.

Matea Filko and Krešimir Šojat. 2017. Expansion of the Derivational Database for Croatian. In Eleonora Litta and Marco Passarotti, editors, *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*. EDUCatt, Milan, pages 27–37.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for English. In *Proceedings of NAACL-HLT*. AL, Edmonton, pages 17–23.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5):125–168.

Adrienne Lehrer. 2003. Polysemy in derivational affixes. In Todd Z. Herman V. Nerlich, B. and D. D. Clarke, editors, *Polysemy. Flexible Patterns of Meaning in Mind and Language*, De Gruyter Mouton, New York, pages 218–232.

Rochelle Lieber. 2004. *Morphology and lexical semantics*. Cambridge University Press, New York.

Rochelle Lieber. 2009. *Introducing Morphology*. Cambridge University Press, New York.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*. aAccademia University Press, Napoli, pages 185–189.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC - Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*. Association for Computational Linguistics, Gothenburg, pages 29–35.

Ivan Marković. 2012. *Uvod u jezičnu morfologiju*. Number 6 in Biblioteka Thesaurus. Disput, Zagreb. OCLC: 815718585.

Ivan Marković. 2013. O najvećim (i) mogućim hrvatskim riječima. In Stjepan Blažetin, editor, *XI. međunarodni kroatistički znanstveni skup*, Znanstveni zaovd Hrvata u Mađarskoj, Zagreb, pages 43–58.

Milan Moguš, Maja Bratanić, and Marko Tadić. 1999. *Hrvatski čestotni rječnik*. Školska knjiga : Zavod za lingvistiku Filozofskoga fakulteta, Zagreb.

Karel Pala and Pavel Šmerk. 2015. Derivancze — Derivational Analyzer of Czech. In Pavel Král and Václav Matoušek, editors, *Text, Speech, and Dialogue: 18th International Conference, TSD 2015*. Springer, Berlin: Heidelberg, pages 515–523. https://doi.org/10.1007/978-3-319-24033-6$_5$8.

Marco Passarotti and Francesco Mambrini. 2012. First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin. In Nicoletta Calzolari et al., editor, *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. ELRA, Istanbul, pages 852–859.

Josip Silić and Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Školska knj, Zagreb. OCLC: ocm70847560.

Marko Tadić. 2009. New version of the Croatian National Corpus. In Dana Hlaváčková, Aleš Horák, Klara Osolsobě, and Pavel Rychlý, editors, *After Half a Century of Slavonic Natural Language Processing*, Masaryk University, Brno, pages 199–205.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure* 9(1):72–102. https://doi.org/https://doi.org/10.3366/word.2016.0087.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, pages 1201–1211.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. In Nicoletta Calzolari et al., editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. ELRA, Reykjavik, pages 1088–1093.

Jan Šnajder. 2014. DERIVBASE.HR: A High-Coverage Derivational Morphology Resource for Croatian. In Nicoletta Calzolari et al., editor, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. ELRA, Reykjavik, pages 3371–3377.

Krešimir Šojat, Matea Srebačić, and Tin Pavelić. 2014. CroDeriV 2.0.: Initial Experiments. In Adam Przepiórkowski and Maciej Ogrodniczuk, editors, *Advances in Natural Language Processing*, Springer International Publishing, Cham, volume 8686, pages 27–33. https://doi.org/10.1007/978-3-319-10888-9$_3$.

Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling* 0(1):111. https://doi.org/10.15398/jlm.v0i1.34.

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

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limbursk á. 2016. Merging Data Resources for Inflectional and Derivational Morphology in Czech. In Nicoletta Calzolari et al., editor, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. ELRA, Portorož, pages 1307–1314.