PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 115 OCTOBER 2020 83–104

The Design of Croderiv 2.0

Matea Filko, Krešimir Šojat, Vanja Štefanec

Faculty of Humanities and Social Sciences, University of Zagreb

Abstract

This paper deals with methods applied in the expansion and design of CroDeriv – the Croatian derivational lexicon. The first version of the lexicon contained only verbs that were segmented and analyzed for morphemes. The database is available online. In a further development, other parts-of-speech (adjectives, nouns) are imported into the lexicon. All imported lexemes are analyzed in terms of their morphological structure and word-formation patterns. Due to new parts-of-speech, and a new type of information, the modification of the database structure was necessary. Here, we present a restructured version of the database, adapted to include other POS, and to explicitly mark word-formation patterns among derivationally related lexemes. We focus on underlying principles for precise and refined queries based on various parameters through the online search interface.

1. Introduction

Croatian is a South Slavic language with very rich inflectional and derivational morphology. Whereas inflection is based almost exclusively on suffixation, various combinations of derivational affixes take part in word-formation. All morphological processes are characterized by frequent affixal as well as root allomorphy. Croatian inflectional morphology is extensively covered by several large lexica with paradigms and inflectional patterns used mainly in natural language processing (NLP) tasks such as lemmatization, morphosyntactic description (MSD) and part of speech (POS) tagging etc. The quantity of language resources dealing with word-formation is significantly smaller. This holds not only for Croatian but also for other languages worldwide. Moreover, derivational resources exist for a relatively limited number of languages, although the development of such resources has begun almost twenty years

Corresponding author: matea.filko@ffzg.hr

Cite as: Matea Filko, Krešimir Šojat, Vanja Štefanec. The Design of Croderiv 2.0. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 83–104. doi: 10.14712/00326585.006.

ago (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 derivational resources generally focus on the annotation of word-formation processes within and across derivational families, i.e. among lexemes that share the same root. Generally, they do not provide the account of the morphological structure of words, i.e. they do not present their morphemic make-up. Procedures applied in their development range from automatic or semi-automatic to completely manual.

As mentioned, Croatian is a Slavic language with rich morphological processes both in terms of inflection and derivation. High-quality language resources dealing with the morphological structure and derivational relations of Croatian lexemes are useful for numerous NLP tasks, but they are also valuable in various theoretical work. In this paper[1], we present the expansion and redesign of the current version of the Croatian derivational lexicon – CroDeriv (Šojat et al., 2013).[2] Procedures applied in the building of its first version differ from those listed above: 1) this version of CroDeriv contained only verbs, i.e. other POS were not included[3]; 2) the focus was on a thorough analysis of the morphological structure of lexemes, whereas word-formation relations among them were not marked. In the second phase, CroDeriv has been expanded with words of other POS and the representation of derivational relations between base words and derivatives has been introduced. Consequently, 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 first version of CroDeriv 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 other POS, i.e. to adjectives and nouns, and extended in new directions. Section 4 presents the new structure of the database and new query parameters. In Section 5 concluding remarks and the outline of future work are given.

[1] This is an extended and significantly modified version of the paper ”Redesign of the Croatian derivational lexicon” presented at the DeriMo 2019 Conference in Prague and published in the Proceedings as Filko et al. (2019).

[2] The search interface of the lexicon is available at `http://croderiv.ffzg.hr/`.

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

2. Croatian derivational lexicon v1.0

The first version of CroDeriv contained ca 14,500 verbs[4] 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 automatically segmented into morphemes via a rule-based approach. The results were afterwards manually checked and edited. This procedure enabled the recognition of lexical morphemes / roots shared by various verbs as well as affixes used in their derivation. The recognition of mutual lexical morphemes enabled the creation of verbal derivational families, i.e. verbs with the same roots were grouped into derivational families accordingly. Morphological analysis of verbs also enabled the analysis of affix frequency and various combinations of derivational and lexical morphemes. Queries over such combinations are available online. Each lexical entry, i.e. verbal infinitive, is accompanied by additional information regarding its aspect. As in other Slavic languages, aspect is an inherent verbal category (Marković, 2012, 183); therefore, each verb was marked as perfective, imperfective, or bi-aspectual.[5] In cases of homography, lexical entries were disambiguated on the basis of aspectual properties and separated (one marked as imperfective, the other as perfective).

One of CroDeriv's distinctive features is the fact that lemmas are segmented into morphs, and morphs are linked to representative morphemes. The morphological segmentation of lemmas in CroDeriv consisted of two steps: 1) automatic segmentation via rules based on the list of various derivational affixes; 2) manual checking of the results necessary due to extensive homography and allomorphy of affixes and roots. In this process, we recognized and manually disambiguated all the homographic forms of various morphemes. Parallelly, we linked various allomorphs to single representative morphemes. The underlying principle for this line of processing is a two-layer approach consisting of a surface and a deep layer.

At the surface, the first step is the segmentation into morphs. The procedure enables that all allomorphs of a certain morpheme are identified and marked for

[4] This version is therefore referred to as *CroDeriV*.

[5] Verbal aspectual pairs are considered separate lemmas in Croatian. Moreover, Croatian words are limited to one inflectional suffix per word, and in case of verbal infinitives, this slot is filled with infinitive ending *-ti*. Thematic suffixes are also used in the formation of verbs from other POS, e.g. from adjectives or nouns (*pun* 'full' – *pun-**i**-ti* 'to fill$_{\text{IMPF}}$' – *is-pun-**i**-ti* 'to fulfill$_{\text{PF}}$'; *rad* 'work' – *rad-**i**-ti* 'to work$_{\text{IMPF}}$' – *za-rad-**i**-ti* 'to earn$_{\text{PF}}$'). Therefore, thematic suffixes, as ***-i-*** in *is-pun-i-ti*, are classified as derivational (Marković, 2012; Silić and Pranjković, 2005; Barić et al., 1995). However, some authors point out that the status of thematic suffixes is not clear. Thus, Manova (2015) recognizes following domains in the structure of Slavic word: (PREFIX)-BASE-(DERIVATIONAL SUFF)-(THEMATIC MARKER)-(INFLECTIONAL SUFF). As opposed to our approach, thematic suffixes are here neither derivational nor inflectional. However, we believe that every suffix is (more or less typical) member of the derivational or inflectional domain. Research on Croatian thematic markers has shown that they have more derivational than inflectional properties, thus, we consider them as members of the derivational domain.

their type. Possible types of morphemes recognized in Croatian lexemes are derivational prefixes, roots, derivational suffixes, inflectional suffixes, and interfixes for compounds. For example, the surface form of the verb *ispuniti* 'to fulfill, to fill out' would be presented as:

$$\left[is\right]_{\text{prefix}} \left[pun\right]_{\text{root}} \left[i\right]_{\text{derivational (thematic) suffix}} \left[ti\right]_{\text{inflectional (infinitive) suffix}}$$

whereas the compound verb *odobrovoljiti* 'to cheer up' is analyzed as follows:

$$\left[o\right]_{\text{prefix}} \left[dobr\right]_{\text{root}_2} \left[o\right]_{\text{interfix}} \left[volj\right]_{\text{root}_1} \left[i\right]_{\text{derivational suffix}} \left[ti\right]_{\text{inflectional (infinitive) suffix}}$$

At the deep layer, we link the prefixal allomorph *is* 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 and their combinations (prefixal, suffixal, and both) used in various derivational families.[6] However, this version of CroDeriv is limited in two ways: 1) it is restricted to only one POS, and 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 v2.0

The expansion of CroDeriv is based on nominal and adjectival lemmas collected from corpora and online dictionaries of Croatian. We chose approx. 6,000 nouns and 1,000 adjectives according to their frequency indicated by the Croatian frequency dictionary (Moguš et al., 1999). We also used frequency lists generated by the corpus management system NoSketchEngine for both representative corpora (Croatian National Corpus and Croatian web corpus hrWaC).[7] Named entities were excluded from the list, since they are formed via non-productive word-formation patterns (Babić, 2002, 16). The obtained list of lemmas was used as a representative sample for further analysis and processing.

In order to incorporate lexemes of other POS and simultaneously mark word-

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

[7] 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 methodology is explained in Filko (2020).

formation relations among them, the database needed to be restructured. The structure of the database remained morpheme-based,[8] i.e. we consider morphemes as basic meaningful units. Further, we assume that words have an internal structure. This *intra*-lexical structure is predictable to a certain degree, at least for certain POS. Following the two-layer approach discussed above, lemmas in Croderiv 2.0 are analyzed for morphs and morphemes. However, in this phase of development, we introduce a new type of information, i.e. the links indicating derivational relations between lexemes. As far as the database structure is concerned, this means that connections between base words and derivatives are explicitly marked and annotated. More details about the annotation scheme and underlying principles are given in the following sections. The introduction of new POS resulted in the expansion of derivational families already present in CroDeriv 1.0 and the establishment of new ones. The new ones are based on nominal and adjectival roots, previously not recorded in verbal families. The online interface for CroDeriv 2.0 enables graphical presentation of derivational relations. In other words, the online interface for CroDeriv 2.0 is designed to present graphical visualization of *inter*-lexical relations within derivational families. More details will be given below.

As mentioned, the morphological analysis follows the two-layered approach from Croderiv 1.0, and consists of two steps: 1) morph analysis at the surface layer, and 2) morpheme analysis at the deep layer (see Figure 1, the upper branch). However, the annotation of derivational relations among lexemes required an additional and different kind of analysis, i.e. the analysis of word-formation links and patterns. The distinction between morphological and word-formation analysis is exemplified in Figure 1. The results of word-formation analysis are available through the Croderiv 2.0 online interface. The new interface also provides information on 1) the type of the word-formation processes, and 2) affixal senses for the affixes detected in word-formation patterns of analyzed lemmas.[9] A detailed presentation of lexical entries in CroDeriv 2.0 is given in Section 3.3 below.

In the following subsections we describe these data more closely and focus on basic principles governing the morphological and the word-formation analysis applied in CroDeriv.

3.1. Morphological analysis

The morphological analysis of new lexical material consisted of 1) the manual segmentation of lexemes into morphs and morphemes, i.e. morph and morpheme analysis, and 2) the categorization of obtained results. The morphological structure

[8] As opposed to word-based approaches, cf. Stewart (2016, 5).

[9] The basic unit in our lexicon, following the approach in Croderiv 1.0 is lemma, i.e. infinitive form for verbs, nominative singular for nouns, nominative singular masculine for adjectives.

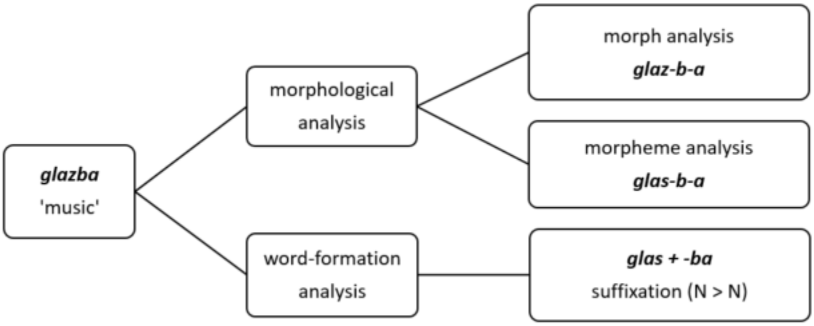

Figure 1. Morphological vs. word-formation analysis

of Croatian lexemes, regardless of their POS, consists of the following types of morphemes: prefixes, roots, interfixes, derivational and inflectional suffixes.[10] Each morpheme type can occur more than once in the morphological structure of lexemes, except inflectional suffixes. The morph and morpheme analysis is the prerequisite for the detection of both unmotivated and motivated lexemes, needed for the annotation of word-formation patterns (see Section 3.2). Generally, motivated lexemes are morphologically more complex than unmotivated, i.e. they have at least one morpheme extra in comparison to unmotivated ones. Besides, one of the aims of the manual segmentation of the representative sample is to develop a procedure for automatic segmentation in future.[11]

As opposed to verbs, usually formed via prefixation or highly-regular suffixation from other verbs (Šojat et al., 2012), nouns and adjectives are predominantly formed via 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 frequency of certain affixes is not provided. Frequency 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 a particular derivational affix. Preliminary research showed that a relatively small subset of suffixes compared to the numbers listed above is actually used for nominal and adjectival derivation in our representative sample.[12] As indicated, we plan to use these results for the development of a morphological parser for Croatian.

[10] Prefixes are always derivational.

[11] A procedure based on a 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.

[12] Filko (2020) shows that only 221 different nominal suffixes (out of 526 listed in Babić (2002)) occur in the morphological structure of 5,536 most frequent nouns in Croatian.

The morphological segmentation of new POS is based on the two-layered approach 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. When applied to nouns and adjectives, the analysis of the noun *učiteljica* 'female teacher' looks like this:

$$\begin{bmatrix} u\check{c} \end{bmatrix}_{\text{root}} \left(\begin{bmatrix} i \end{bmatrix} \begin{bmatrix} telj \end{bmatrix} \begin{bmatrix} ic \end{bmatrix} \right)_{\text{derivational suffixes}} \begin{bmatrix} a \end{bmatrix}_{\text{inflectional suffix}}$$

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

$$\begin{bmatrix} iz \end{bmatrix}_{\text{prefix}} \begin{bmatrix} lje\check{c} \end{bmatrix}_{\text{root}} \begin{bmatrix} iv \end{bmatrix}_{\text{derivational suffix}} \begin{bmatrix} \emptyset \end{bmatrix}_{\text{inflectional suffix}}$$

and the allomorph *lječ* is at the deep layer connected to the representative root morph *lijek*.

The analysis of morphs and morphemes is based on the following principles:[13]

- Morph analysis must be complete (no morpho-phonological residue is allowed). This means that all phonemic material of the analyzed lemma is distributed to at least one morph.
- The detection of morphs is based on commutation. This method enables the recognition of all units that 1) reoccur (e.g. *uči-***telj** 'teacher', *vodi-***telj** 'leader, presenter', *gleda-***telj** 'viewer'), or 2) stand in the opposition with other units (e.g. *uč-***i***-ti* 'to learn', *hod-***a***-ti* 'to walk', *vid-***je***-ti* 'to see').
- As in other Slavic languages, numerous phonological changes occur at morpheme boundaries. Criteria for the analysis of various allomorphs, resulting from various phonological processes, are not precise. This means that in many cases it was difficult to establish straightforward links between certain parts of the phonemic material and morphs. To resolve this problem in a unified manner, the following rule was determined: if there is a fused phonemic material, allocate as much as possible of this material to the stem (see footnote 15). For example, in the word *tajništvo* 'secretariat' ← *tajnik* 'secretary' + *-stvo*, two interpretations at the surface layer are possible: *taj-n-i-štv-o* or *taj-n-iš-tvo*, depending on the allocation of the phoneme *š* to the stem or to the suffix.[14] We have decided to resolve all similar situations in favour of stems.

[13] The detailed elaboration of principles and solutions to specific problems in Croatian is given in Filko (2020).

[14] For the detailed explanation of the phonological change in this example see Marković (2013, 25, 125).

- After the surface layer morphs were detected, we determined their representative morphs at the deep layer, i.e. those to which allomorphs are connected. Hereby we follow the approach from CroDeriv 1.0: the representative morph is the one from which other allomorphs can be established with the least number of morpho-phonological rules. However, if a representative morph cannot be established via phonotactic criteria, the following frequency-based criterion is applied: the representative morph is the morph which most frequently appears in the morphological structure of various derivationally related lexemes.

The results of this analysis are reflected in the overall structure of lexical entries in CroDeriv 2.0 (see Section 3.3). As mentioned, this analysis enables the recognition of motivated lexemes and word-formation patterns. The underlying principles for the word-formation analysis are described below.

3.2. Word-formation analysis

The main goal of our word-formation analysis is to mark derivational relations among lexemes. From the theoretical point of view, for each motivated lexeme in the database we need to determine: 1) corresponding word-formation elements, and 2) a word-formation pattern. By word-formation elements in Croatian, we refer to prefixes, stems, interfixes, and suffixes. By word-formation patterns, we refer to suffixation, prefixation, simultaneous suffixation and prefixation, compounding, simultaneous compounding and suffixation, simultaneous prefixation and compounding, back-formation, and conversion / zero-derivation. Word-formation elements and patterns are presented in more detail below.

3.2.1. Word-formation elements

The first objective in the word-formation analysis is to determine word-formation elements. This step is necessary for the recognition of word-formation patterns (see Section 3.2.2). In Croatian, the following types of elements are recognized:

1. **stem**:[15] ***čaš***-*a* 'glass', where *-a* is the inflectional suffix
2. **prefix**: ***ne***-*čist* 'dirty' ← *ne-* 'non-' + *čist* 'clean'
3. **interfix**: *par-****o****-brod* 'steamboat' ← *par*(*a*) 'steam' + *-o-* + *brod* 'ship, boat'
4. **suffix**: *šljiv-****ik*** 'plum yard' ← *šljiv*(*a*) 'plum' + *-ik*

[15] Following Marković (2012), we define *stem* as a segment consisting of one or more morphs to which derivational affixes are added. Stems can be equal to roots, as in *vid-jeti* 'to see' < *vid* 'sight', or they can consist of a root + one or more morphs, as in: *vidje-lica* 'psychic'< *vidjeti* 'to see'. Thus, we determine roots during the morphological analysis, and stems as a part of word-formation analysis. Derivational stems are sometimes equal to inflectional stems, e.g. *kum* 'godfather' > *kum-a* 'godmother', where *kum-* is both derivational and inflectional stem. However, in the word *čašica* 'small glass' < *čaša* 'glass', the derivational stem is *čaš-*, whereas the inflectional stem is *čašic-*. Inflectional and derivational stem are also called inflectional and derivational base.

The main problem we encountered in this analysis pertains to the status of certain suffixes. First, there are suffixes that at the same time can be interpreted as derivational as well as inflectional. For example, the suffix *-a* in the above example for the stem *čaš-* functions as a derivative suffix for the derivation of nouns, but also as an inflectional suffix for forming the nominative case, singular, femininum. Note that the main difference between this example and the thematic marker in verbs is that thematic marker is followed by an inflectional suffix (see footnote 5). Second, some suffixes that can be distinguished as different morphemes at the morphological level are added simultaneously as one derivational suffix (see the examples below). Therefore, we distinguish between (possibly complex) suffixes as word-formation elements, and (simple) suffixes in the morphological structure. The difference between them is that suffixes as word-formation elements can be morphologically complex, consisting of derivational and inflectional morphemes. Further, suffixal word-formation elements can contain more than one derivational suffix and an inflectional suffix. The motivation for this decision is twofold: 1) to resolve the status of suffixes, such as of *-a* discussed above – on the level of morphological analysis they are marked as inflectional, and 2) to indicate that groups of morphemes are simultaneously used as elements in various word-formation processes. First, we present the structure consisting of one derivational and one inflectional morpheme. The morphological structure of the Croatian noun *čistoća* 'cleanness, purity' consists of two suffixes: one derivational (*-oć-*) and one inflectional (*-a*), but there is only one word-formation suffixal element (*-oća*):

- morphological analysis (MA): *čist-oć-a*
- word-formation analysis (WFA): *čist* + *-oća* → *čistoća*

As mentioned, word-formation suffixes can consist of two (or more) derivational suffixes and one inflectional suffix (in the rightmost position). Below we list examples for adjectives, verbs and nouns, as analyzed in CroDeriv 2.0:

- *vođen* 'led'
  MA: *vođ-e-n-Ø*[16] (surface layer) *vod-je-n-Ø* (deep layer)
  WFA: *voditi* 'lead' + *-jen* → *vođen*
- *prepisivati* 'to copy'$_{\text{IMPF}}$
  MA: *pre-pis-iv-a-ti* (surface layer) *pre-pis-iv-a-ti* (deep layer)
  WFA: *prepisati* 'to copy'$_{\text{PF}}$ + *-ivati* → *prepisivati*
- *administracija* 'administration'
  MA: *administr-ac-ij-a* (surface layer) *administr-at-ij-a* (deep layer)
  WFA: *administrirati* 'to administer' + *-acija* → *administracija*

Generally, complex suffixal elements, as listed in the WFA lines above, are composed of invariant affixal combinations. Being fixed combinations, we treat them as single

[16] Zero suffix is here inflectional. Compare with genitive case: *vođ-e-n-a*.

units at this level of analysis and presentation. We intend to expand this line of research in the future. The MA lines above present the morphological analysis at the deep and the surface layer. We indicated that in many cases it is hard to determine morpheme boundaries and functions due to various phonological processes. In the following example, we demonstrate how such cases are resolved and how links between morphological and word-formation analysis are established: the noun *radništvo* 'working class' ← *radnik* 'worker' + *-stvo* consists of two word-formation elements: stem *radnik* and suffix *-stvo*. At the surface MA layer, due to morpho-phonological changes, the stem *radnik* is realized via its allomorph *radniš*, while the word-formation suffix *-stvo* is realized via its allomorph *-tvo*. At the deep layer, these allomorphs are connected to their representative forms and used for the presentation of word-formation elements. The connection of allomorphs to their representative forms at the deep layer enables the recognition of words formed via same word-formation patterns, i.e. derived via same prefixes or suffixes (e.g. *ribarstvo, radništvo* are both formed via denominal suffixation with suffix *-stvo*), and the recognition of words derived from the same stem (e.g. *radništvo, radnica, radnikov, suradnik* are formed from the stem *radnik*).

3.2.2. Word-formation patterns

Apart from word-formation elements, we also determine the type of word-formation pattern for each motivated entry in our lexicon. Word-formation patterns indicate the links between base words and various derivatives. Lexical entries provide the information on word-formation processes applied in word-formation patterns. We take into account the following word-formation processes in Croatian:

1. **suffixation:**
   - *pjev(ati)* 'to sing' + *-ač* → *pjevač* 'singer'
   - *glas* 'voice' + *-ati*[17] → *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'$_{\text{ADJ}}$ → *predškolski* 'preschool'$_{\text{ADJ}}$
3. **simultaneous suffixation and prefixation:**
   - *o-* + *svoj* 'one's own' + *-iti* → *osvojiti* 'to conquer, to win'
   - *bez-* + *sadržaj* 'content' + *-an* → *besadržajan* 'pointless, contentless'
4. **compounding:**
   - *vjer(a)* 'trust' + *-o-* + *dostojan* 'worthy' → *vjerodostojan* 'trustworthy'
   - *zlo* 'evil' + *upotrijebiti* 'to use' → *zloupotrijebiti* 'to misuse, to abuse'

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

- *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'$_{\text{ADJ}}$
6. **simultaneous prefixation and compounding:**
   - *o-* + *zlo* 'evil' + *glasiti* 'to say' → *ozloglasiti* 'to discredit, to bring into disrepute'
7. **back-formation:**[18]
   - *izlaz*(*iti*) 'to exit' → *izlaz* 'exit'
8. **conversion or zero-derivation:**
   - *mlada* 'young'$_{\text{ADJ+FEM}}$ → *mlada* 'bride'$_{\text{N}}$
9. **ablaut:**
   - *plesti* = *plet* + (∅) + (*ti*) 'to twine' → *plot* 'fence'.

In lexical entries, only the last step in the formation of a particular lexeme is presented. Although the verb *ispunjavati* 'to fulfill'$_{\text{IMPF}}$ is (remotely) derivationally related to the verb *puniti* 'to fill'$_{\text{IMPF}}$, their derivational connection is indirect since it is derived from the verb *ispuniti* 'to fulfill'$_{\text{PF}}$. We mark only the last derivational step in the word-formation pattern. Therefore:
ispun(*iti*) 'to fulfill'$_{\text{PF}}$ + *-javati* → *ispunjavati* 'to fulfill'$_{\text{IMPF}}$ [suffixation].
The remote derivational link is available via word-formation pattern of the verb *ispuniti* 'to fulfill'$_{\text{PF}}$:
is- + *puniti* 'to fill'$_{\text{IMPF}}$ → *ispuniti* 'to fullfil'$_{\text{PF}}$ [prefixation].

Derivational connections between motivated lexemes and their base lexemes are based on the following principle:

1. If there are simultaneous phonological and semantic relations between stems of two lexemes, two lexemes are derivationally connected (Babić, 2002, 25); e.g. *čist* 'clean' → *čist-oća* 'cleanness'.
   This principle holds in the vast majority of cases. However, in some cases stems need to be determined based on other criteria:
2. *lost* stems[19] and affixes: if a stem is synchronically not present in any other lexeme, but its suffix is clearly recognizable in the morphological structure of other

[18] Although some authors consider similar cases as examples of conversion or zero derivation (see next item), we define *conversion* as a process with no segmental or suprasegmental changes (Marković, 2012, 81). Thus, we consider cases with segmental changes as different word-formation processes. Therefore, we treat this case as back-formation as a type of subtraction.

[19] Lost stems are to be found in the so-called base-less derivatives (Gaeta and Ricca, 2003), which should synchronically be considered as simplex, since they cannot be related to any other existing base, but their suffixes are clearly recognizable from the morphological structure of the derivative in their typical senses. Lost stems are similar to the notion of unrecoverable bases (Talamo et al., 2016), and, at the word-formation level, they are similar to cranberry morphemes at the level of morphological analysis.

lexemes, this stem is taken into consideration in further processing. For example, the stem in *vrab*-ac 'sparrow' does not exist as a lexeme in Croatian, but the suffix *-ac* is normally used in the word-formation of nouns denoting male animals (e.g. *žaba* 'frog'$_{\mathrm{FEM}}$ → *žabac* 'frog'$_{\mathrm{MASC}}$). We refer to this type of stems as lost stems.

3. *paradigmatic* stems and affixes: if a stem cannot be synchronically associated with any existing base lexeme, but still, it occurs in at least two derivatives (Talamo et al., 2016, 84), this stem is taken into consideration in further processing. For example, the stem *dub* is recognized in the lexemes *dubok* 'deep' and *dubina* 'depth', regardless of the fact that the word *dub* does not exist. The same derivational relation is recognized in other pairs of lexemes, in which the stem functions as a separate word:
   dubok 'deep' vs. *dubina* 'depth' vs. **dub*
   širok 'wide' vs. *širina* 'width' vs. *šir* 'width $_{\text{expressive}}$'
   visok 'high' vs. *visina* 'height' vs. *vis* 'height $_{\text{expressive}}$'.
   We refer to this type of stems as paradigmatic stems.[20]
4. *possible* stems and affixes: in many derivational families, word-formation patterns cannot be established in a straightforward manner due to *missing links* between members of families. These links can be theoretically postulated as *possible* words, completely compliant to morphological structure and derivational processes in Croatian. Thus, if a base word actually does not exist, but it could be formed via regular and productive word-formation patterns, this stem is taken into consideration in further processing. Such cases are usually related to verbal participles and gerunds. In example 1 below, the past participle is attested and used for further derivation. In example 2, the past participle is not attested, i.e. it actually does not exist. However, its morphological structure is analogous to attested forms, it is marked as such and used in the database structure:[21]
   1) *pjevati* 'to sing' → *pjevan* 'sung' → *pjevanje* 'singing'
   2) *sjećati se* 'to remember' → **sjećan* 'remembered' → *sjećanje* 'remembrance, memory'.

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;

[20] The difference between paradigmatic and lost stems is visible in the graphical representation of derivational families - paradigmatic stems serve as the basis for the word-formation of two or more words, while only one word is derived from the lost stems.

[21] This line of processing is similar to the approach used in DeriNet 2.0 and their *fictitious lexemes*, which are defined as "lexemes that are attested neither in the corpora nor in the dictionaries but, based on structural analogies, fill a paradigm gap in the derivational family" (Vidra et al., 2019, 82).

- if one of the competing solutions can be applied to a wider range of motivated lexemes than the other, this solution should be selected.

3.2.3. Affixal senses

Morphological processing in CroDeriv enables the recognition of various combinations of affixes and roots and therefore provides an excellent basis for research. The research on the semantic impact of affixes in word-formation processes shows that derivational affixes frequently behave in a similar manner in various derivational families. In other words, derivational prefixes and suffixes similarly or even identically affect the meaning of derivatives in different derivational families. This means that the meaning structure of derivational affixes can be decomposed and its meaning components, i.e. affixal senses, can be (more or less) determined. We intend to incorporate this information into lexical entries. In our database, affixes are structured as polysemous units, which is in line with recent approaches to affixal senses (Babić (2002, 38), Lehrer (2003), Lieber (2004, 11), Lieber (2009, 41), Aronoff and Fudeman (2011, 140–141)). Taking into account other elements in word-formation patterns, one of the affixal meanings is realized in motivated lexemes. For example, the verbal prefix *nad-* can have two senses. 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 semantic analysis of Croatian verbal prefixes is given in Šojat et al. (2012), whereas the most frequent adjectival suffixes are discussed in Filko and Šojat (2017). A detailed semantic analysis of highly frequent nominal suffixes is presented in Filko (2020).[22] The inventory of affixal senses is based on data from Croatian grammar and reference books. As expected, affixes and their senses are treated differently in Croatian literature. Whereas some authors (e.g. Babić (2002)) list affixes alphabetically and note their possible senses, 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. In other words, they group affixes according to at least one of their meaning components. We combined the information from these sources and modified polysemous structures of affixes according to recorded lexemes in the database. For the nominal suffix *-ica* the following senses were determined (new ones may appear in future analysis):[23]

1. **agent, female**, e.g. *učitelj* 'teacher'$_{\text{MASC}}$ → *učiteljica* 'teacher'$_{\text{FEM}}$

[22] Bagasheva (2017) presents the comprehensive list of semantic categories which should be applicable for the study of affixal derivation, at least in European languages. Her set of 51 comparative semantic concepts in affixation is used as a starting point in the *Cross-linguistic research into derivational networks* project. First results of this project are presented in Körtvélyessy (2019).

[23] These are the senses recorded so far in our material. For a more extensive account, including idiosyncratic combinations, see Babić (2002, 183–189)

2. **person, both sexes**, e.g. *izbjegao* 'exiled' → *izbjeglica* 'refugee'
3. **animal, female**, e.g. *golub* 'pigeon'$_{\text{MASC}}$ → *golubica* 'pigeon'$_{\text{FEM}}$
4. **diminutive**, e.g. *pjesma* 'song' → *pjesmica* 'ditty, rhyme'
5. **thing**, e.g. *sanjar* 'dreamer'$_{\text{MASC}}$ → *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'
10. **disease**, e.g. *vruć* 'hot' → *vrućica* 'fever'
11. **literary type**, e.g. *slovo* 'letter' → *poslovica* 'proverb'
12. **linguistic term – type of word/sentence**, e.g. izveden 'derived'$_{\text{ADJ}}$ → izvedenica 'derivative'
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'

To sum up, the new version of the database provides the information on the following word-formation 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.[24]

3.3. The structure of lexical entries

The information discussed in Sections 3.1 and 3.2 is encoded for each entry in the lexicon. The new search interface will provide the information about grammatical categories (1), morphological structure (2-3), and word-formation properties (4-8) (see the example for the lemma *poslužitelj* and others below). 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 accessible through another link attached to the stem (5 - posluži). This will enable users to follow complete derivational paths in both directions: from roots to derivatives (through the link in 4) and from various derivatives back to roots (through the link in 5). In future, we plan to provide links to online dictionaries and inflectional lexica for Croatian for additional information.

[24] 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. For the morphological analysis of these entries see Section 3.1.

The complete structure of entries of different POS is as follows:

Nouns

1. **lemma:** poslužitelj ‘server’
   - **POS:** N
   - **gender**: masculine
2. **morphological structure – surface layer:**
   $\left[\text{po}\right]_{\text{prefix}} \left[\text{služ}\right]_{\text{root}} \left(\left[\text{i}\right] \left[\text{telj}\right]\right)_{\text{derivational suffixes}} \left[\ \right]_{\text{inflectional suffix}}$
3. **morphological structure – deep layer:**
   $\left[\text{po}\right]_{\text{prefix}} \left[\text{slug}\right]_{\text{root}} \left(\left[\text{i}\right] \left[\text{telj}\right]\right)_{\text{derivational suffixes}} \left[\varnothing\right]_{\text{inflectional suffix}}$
4. **word-formation pattern:** poslužiti[25] + telj
5. **stem (allomorph of the stem)**: posluži[26] (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**[27].

Verbs

1. **lemma:** potpisati ‘to sign’
   - **POS:** V
   - **aspect**: perfective
   - **reflexivity**: non-reflexive
2. **morphological structure – surface layer:**
   $\left[\text{pot}\right]_{\text{prefix}} \left[\text{pis}\right]_{\text{root}} \left[\text{a}\right]_{\text{derivational suffix}} \left[\text{ti}\right]_{\text{inflectional suffix}}$
3. **morphological structure – deep layer:**
   $\left[\text{pod}\right]_{\text{prefix}} \left[\text{pis}\right]_{\text{root}} \left[\text{a}\right]_{\text{derivational suffix}} \left[\text{ti}\right]_{\text{inflectional suffix}}$
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**.

[25] The base word is underlined and functions as a link to the entry of that word in the lexicon.

[26] The stem is underlined and functions as a link to all lemmas derived directly from this stem, e.g. *poslužilac*.

[27] Online dictionary of Croatian: `http://hjp.znanje.hr/`.

Adjectives

1. **lemma:** beskrajan 'endless'
   - **POS:** A
   - **gender**: masculine
   - **definiteness**: indefinite
2. **morphological structure – surface layer:**
   $\left[\text{bes}\right]_{\text{prefix}} \left[\text{kraj}\right]_{\text{root}} \left[\text{an}\right]_{\text{derivational suffix}} \left[\ \right]_{\text{inflectional suffix}}$
3. **morphological structure – deep layer:**
   $\left[\text{bez}\right]_{\text{prefix}} \left[\text{kraj}\right]_{\text{root}} \left[\text{an}\right]_{\text{derivational suffix}} \left[\text{Ø}\right]_{\text{inflectional suffix}}$
4. **word-formation pattern:** bez + kraj + an
5. **stem (allomorph of the stem)**: kraj (kraj)
6. **affix$_1$ (allomorph of the affix$_1$)**: bez- (bes-)
   affix$_2$ (allomorph of the affix$_2$): -an (-an)
7. **affix$_1$ sense**: deprivation
   affix$_2$ 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**.

In the following section, we focus on 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 CroDeriv database

Unlike many existing derivational lexicons and databases, which mostly focus on presenting derivation as connections between lexemes and thus building derivational trees or graphs (Kyjánek et al., 2019), CroDeriv is primarily devised as a morphological resource. It means that derivational relationships are seen as a result of a specific change in the morphological structure between two lexemes, and as such recorded and presented in the database structure.

The integration of new data required a redesign of the database. The first version of CroDeriv contained only verbs and the data model was therefore built upon the generalized morphological structure of Croatian verbs. Croatian verbs, in various affixal combinations, can take up to four prefixes, three derivational suffixes, and one inflectional suffix. The lexical part contains one or two lexical stems and an optional interfix. The first data model thus provided 9 slots for affixal allomorphs, connected to their respective morphemes, and two slots for lexical stems connected to their respective forms at the deep layer. Apart from the fact that this data model could not accommodate lexemes of other POS, it suffered from other shortcomings, as well.

In this design, derivational relationships were not explicitly marked. Further, the search engine used for CroDeriv 1.0, due to simplified presentations of the generalized morphological structure, showed only stems, and not their morphological structures. These structures can be complex, especially for verbs derived from nouns and adjectives. For example, the relationship between *služiti* and *službovati* could not be established, because the derivational path looks like this:

služiti 'to serve'$_V$ → *služba* 'service'$_N$ → *službovati* 'being in civil service'$_V$

Although *služiti* and *službovati* share the same root *slug* and belong to the same derivational tree, the two verbs are derived from different stems: *služ-* (comprised only of the root *slug*) and *služb-* (comprised of the root *slug* and the nominal suffix *-b-*). This and other problems in terms of the limitations of the data model were tackled in Štefanec et al. (2013).

The new data model is a combination of principles taken from the previous models and new ones gained from detailed analysis of data as described in this paper. The description of word-formation properties is stored separately from the morphological structure of lexemes whereas derivational connections between them are explicitly created. We believe that this model has enough descriptive power to accommodate and describe the entire Croatian lexicon.

4.1. The new CroDeriv data model

In the new model, following the theoretical approach to the morphological analysis presented in Section 3.1, the lexemes are analyzed for morphemes. Technically, the morphemes are presented as sequences of characters (empty sequences corresponding to zero-morphs are also possible). These sequences at the surface layer are identified as allomorphs and connected to their respective morphemes at the deep layer.

The word-formation description in the data model is presented in the form of building blocks called clusters. Clusters are multi-morphemic units that reflect word-formation processes and roughly correspond to stems/affixes, as presented in Section 3.2.1. The only difference between them is that suffixal clusters do not contain inflectional suffix, which is stored separately. Further, there are no discontinuous clusters. This means that simultaneous prefixation and suffixation is based on simultaneous adding of two types of clusters.

The new design of the database is capable of dealing with compound lexemes by the introduction of the notion of compounding segments. Compound lexemes are split into two or more compounding segments, where the compounding segment on the left side consists of the stem and the interfix, while compounding segment on the right contains the other stem and suffixes. E.g., the compound lexeme *knjigovežnica* 'bindary' is split into two compounding segments: *knjigo* + *vežnica*. Compounds consisting of more than two compounding segments are split as follows:

starocrkvenoslavenski 'Old Church Slavonic' = *staro* + *crkveno* + *slavenski*.

Compounding segments are objects that can be connected to other lexemes by derivational links. If two or more compounding segments can be identified in a lexeme, this lexeme has more than one parent in the database structure. However, due to complexity problem of querying graphs, only one connection will be marked as primary to keep the derivational network as a tree-like structure.

4.2. Technical solutions

The new CroDeriv system is a database-driven server application, developed in Django, a high-level Python web-framework[28] with Django REST framework toolkit[29]. The application supports data querying and retrieval via REST over HTTP. Default data retrieval format is JSON. UDer format[30] is also supported, where applicable, and it can be requested by the client using content negotiation principles.

Data is stored in a PostgreSQL relational database in a normalized form. Since graph-like structures are extremely expensive to query, PostgreSQL Materialized Views were used to increase time efficiency. Materialized Views, as normal Views, use the database rule system, but their result persists in a table-like form until refreshed. This means that highly complex data structures can be transformed in a way which is more redundant but facilitates easy querying, and that this time-expensive operation of transforming will be done sufficiently rarely, probably only after some content is added or changed. In the views, the lexemes' morphological and word-formation structures were pre-computed into easily searchable representations, and paths to every node (i.e. lexeme) in the graph were linearized and stored in a flattened form. On top of that, indexes were added to all searchable fields in the view, which resulted in significant improvement in search latency.

4.3. Querying the CroDeriv database

Beside the possibility to search the database using simple queries, which is the option interesting mostly to the general public, a simple query language, similar to corpus query language (CQL), was constructed which will enable more complex and refine queries.

CroDeriv system supports two general types of queries: lexeme-structural and tree-predecessor. The first type searches for lexemes with particular morphological or word-formation structure. For example, query

```
[prefix="pre"]
```

would return all lexemes starting with a prefix *pre-*. Also, query

[28] `https://www.djangoproject.com/`

[29] `https://www.django-rest-framework.org/`

[30] See Vidra et al. (2019) for detailed description of the format.

```
[morpheme=".+"]*[root="pis"]
```

would return all lexemes that contain the root *pis*. Similarly, when searching for lexemes with particular word-formation pattern, query

```
[prefix="pre"]
```

would return all lexemes that were derived from another lexeme by means of prefixing with *pre-*. Also,

```
[cluster=".+"]*[suffix="inj"]
```

would return all lexemes that were derived with a suffixing word-formation element *-inj-*. It is important to notice that lexeme-structural search queries match results always from the beginning.

The second type of search searches for lexemes in a particular derivational path. For example, query

```
{pos="A"}{pos="N"}{pos="V"}
```

would return all verbs derived from nouns, which were derived from adjectives. Also, query

```
{aspect="biaspectual"}{reflexivity="reflexive"}
```

would return all reflexive verbs derived from biaspectual verbs. This type of queries matches results from the end, i.e. it is possible to search only up the derivational tree, and not down.

Finally, the two types of queries can also be combined. For example, query

```
{aspect="biaspectual"}{aspect="perfective",
    morpho=[morpheme=".+"]*[root="ču"]}
```

would return all perfective verbs that contain the root *ču* and are derived from biaspectual verbs.

5. Concluding remarks and future work

In this paper, we presented the design of CroDeriv 2.0 and its online search interface, required to include non-verbal lemmas as well as to present various derivational properties and relations of Croatian lexemes. CroDeriv 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, will make this lexicon even more attractive to users.

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

Acknowledgements

This research is supported by the short-term support of the University of Zagreb.

Bibliography

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

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

Bagasheva, Alexandra. Comparative semantic concepts in affixation. In Santana-Lario, Juan and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, Linguistic Insights, pages 33–66. Peter Lang, Bern : Berlin : Bruxelles : Frankfurt am Main : New York : Oxford : Wien, 2017.

Bajestan, Elnaz Shafaei, Diego Frassinelli, Gabriella Lapesa, and Sebastian Padó. 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*), pages 117–127, Milano, 2017. EDUCatt.

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

Filko, Matea. *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika* (*Intralexical and Interlexical Structures of the Nominal Part of the Croatian Lexicon*). Phd thesis, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, 2020.

Filko, Matea, Krešimir Šojat, and Vanja Štefanec. Redesign of the Croatian derivational lexicon. In Žabokrtský, Zdeněk, Magda Ševčíková, Eleonora Litta, and Marco Passarotti, editors, *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 71–80, Prague, 2019. Charles University.

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

Gaeta, Livio and Davide Ricca. Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data. *Italian Journal of Linguistics / Rivista di Linguistica*, 15(1):63–98, 2003.

Habash, Nizar and Bonnie Dorr. A categorial variation database for English. In *Proceedings of NAACL-HLT*, pages 17–23, Edmonton, 2003. AL. doi: 10.3115/1073445.1073458.

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

Kyjánek, Lukáš, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages. In Žabokrtský, Zdeněk, Magda Ševčíková, Eleonora Litta, and Marco Passarotti, editors, *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 101–110, Prague, 2019. Charles University.

Körtvélyessy, Lívia. Cross-linguistic research into derivational networks. In Žabokrtský, Zdeněk, Magda Ševčíková, Eleonora Litta, and Marco Passarotti, editors, *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 1–4, Prague, 2019. Charles University.

Lehrer, Adrienne. Polysemy in derivational affixes. In Nerlich, Brigitte, Zazie Todd, Vimala Herman, and David D. Clarke, editors, *Polysemy. Flexible Patterns of Meaning in Mind and Language*, pages 218–232. De Gruyter Mouton, New York, 2003. doi: 10.1515/9783110895698.217.

Lieber, Rochelle. *Morphology and lexical semantics*. Cambridge University Press, New York, 2004. doi: 10.1017/CBO9780511486296.

Lieber, Rochelle. *Introducing Morphology*. Cambridge University Press, New York, 2009. doi: 10.1017/CBO9781316156254.

Litta, Eleonora, Marco Passarotti, and Chris Culy. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 185–189, Napoli, 2016. Accademia University Press. doi: 10.4000/books.aaccademia.1799.

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

Manova, Stela. Affix Order and the Structure of the Slavic Word. In Manova, Stela, editor, *Affix Ordering Across Languages and Frameworks*, pages 205–230. Oxford University Press, January 2015. doi: 10.1093/acprof:oso/9780190210434.003.0009.

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

Marković, Ivan. *Hrvatska morfonologija*. Number 7 in Biblioteka Thesaurus. Disput, Zagreb, 2013.

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

Pala, Karel and Pavel Šmerk. Derivancze — Derivational Analyzer of Czech. In Král, Pavel and Václav Matoušek, editors, *Text, Speech, and Dialogue: 18th International Conference, TSD 2015*, pages 515–523, Berlin: Heidelberg, 2015. Springer. doi: 10.1007/978-3-319-24033-6_58.

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

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

Stewart, Thomas W. *Contemporary Morphological Theories. A User's Guide*. Edinburgh University Press, Edinburgh, 2016.

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

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

Vidra, Jonáš, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. DeriNet 2.0: Towards and All-in-One Word-Formation Resource. In Žabokrtský, Zdeněk, Magda Ševčíková, Eleonora Litta, and Marco Passarotti, editors, *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89, Prague, 2019. Charles University.

Zeller, Britta, Jan Šnajder, and Sebastian Padó. 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)*, pages 1201–1211, Sofia, 2013. Association for Computational Linguistics.

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

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

Šojat, Krešimir, Matea Srebačić, and Marko Tadić. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, 0(1):111, 2012. doi: 10.15398/jlm.v0i1.34. URL `http://jlm.ipipan.waw.pl/index.php/JLM/article/view/34`.

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

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

Štefanec, Vanja, Krešimir Šojat, and Matea Srebačić. A Method for the Computational Representation of Croatian Morphology. In Kłopotek, M. A. et al., editor, *Language Processing and Intelligent Information Systems*, pages 80–91. Springer, 2013. doi: 10.1007/978-3-642-38634-3_10.

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

Address for correspondence:
Matea Filko
`matea.filko@ffzg.hr`
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia