Processing Croatian Morphology: Roots, Segmentation and Derivational Families

Krešimir Šojat
Faculty of Humanities
and Social Sciences
University of Zagreb
Croatia
kresimir.sojat@ffzg.unizg.hr

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

Abstract

This paper deals with the development of the Croatian derivational lexicon – CroDeriv. It is a computational database that is designed to store and present morphological data of Croatian words. Each lexical entry in CroDeriv provides information about the morphological structure of words and about derivational links with other words. The database is available for online search according to various parameters. In this paper, we also discuss the linguistic principles we follow in the analysis of words in terms of their morphological structure and grouping words into derivational families. The key element for both procedures, i.e. for the segmentation of words into morphemes and the assignment of words into derivational families, is the accurate recognition of lexical morphemes.

1 Introduction

CroDeriv is a morphological database developed for the Croatian language. Its development took place in several phases. In its first version, CroDeriv contained approximately 15,000 verbs. This version of the lexicon is available for online search at: croderiv.ffzg.hr. In this phase of research and database development, the focus was on the analysis of the morphological structure of verbal lexemes and the structure of the database that would enable queries over various parameters (Šojat et al., 2013). The obtained results proved valuable in many areas, e.g. in the research of verbal aspect, affix ordering, combinations of particular affixes and roots as well as combinations of multiple affixes. The first phase of CroDeriv's development also helped to determine principles for further development of the lexicon. However, the lexicon contained lexemes of only one part of speech (POS), and derivational links among lexemes were not marked. In the second phase of its development, its structure has been expanded with words of other POS, mainly nouns and adjectives, and the representation of derivational links between stems and derivatives as well as explicit marking of word-formation processes has been introduced (Filko et al., 2020).

In this paper, we present further development and enrichment of the existing version of CroDeriV. The paper is structured as follows: in Section 2.1 we discuss morphological segmentation of lexemes in CroDeriv at the surface and deep layer and we explain the basic principles in this two-layered approach. In section 2.2, the main derivational processes are presented as well as some that are not described or that are only marginally described in the existing literature. Each derivation process we describe is accompanied by examples. In section 2.3, we illustrate the structure of derivation families and lexical entries in CroDeriv. In section 3, we discuss some problems we have encountered in our work and outline possible solutions. We finish the paper with the Conclusion and the outline of future work.

2 Morphological analysis

2.1 Segmentation

Each lexical entry in CroDeriv contains information on the morphological structure of lexemes. In other words, each lexeme is segmented into morphemes that it consists of. In the initial phases of CroDeriv's development, this procedure was performed automatically and the results were afterward checked and corrected manually. Due to extensive allomorphy and phonological changes that take part at morpheme boundaries (e.g. assimilation or dropping of phonemes), lexemes are being analyzed and segmented into morphemes manually.

Morpheme is the basic morphological unit. Usually, it is defined as the smallest language sign, i.e. the smallest language unit that can be associated both with the expression on one side and the content on the other (Marković, 2012; Silić and Pranjković, 2005; Barić et al., 1995). In other words, morphemes are the smallest units in the linguistic analysis with their meaning (Haspelmath and Sims, 2010; Booij, 2005). It is important to emphasize that morphemes are abstract units whereas morphs are their physical realization.

Types of morphemes recognized in lexemes are prefixes, lexical morphemes (roots), derivational suffixes, inflectional suffixes, and interfixes for compounds. Each type of morpheme can occur more than once in the morphological structure of lexemes.[1] The following example illustrates multiple prefixation and suffixation in derivation:

s-po-raz-um-je-ti se ’come to an agreement’;

s = prefix; *po* = prefix; *raz* = prefix; *um* = root; *je* = suffix; *ti* = suffix; *se* = reflexive particle

As presented by Filko et al. (2019, 2020), the morphological segmentation of lexemes is based on the two-layered approach: the segmentation at the surface and the deep layer. At the surface layer of analysis, all allomorphs are identified and marked for their type.

For example, the surface layer segmentation of the verb *raščišćavatii* ’to clean up $_{\text{IPF}}$’ can be represented as: *raš-čišć-av-a-ti*; *raš* = prefix; *čišć* = root; *av* = derivational (aspectual) suffix; *a* = derivational (thematic) suffix; *ti* = inflectional (infinitive) suffix.

At the deep layer of presentation, the prefixal allomorph *raš* is connected to its representative morph *raz,* the root allomorph *čišć* to its representative morph *čist,* and the suffixal morph *av* to its representative morph *jav*. The representative morph is the one from which other allomorphs can be established with the least number of morpho-phonological rules. The deep form of the verb *raščišćavati* is thus represented as: *raz-čist-jav-a-ti*.

The same approach – segmentation at the surface and deep layer – is applied to lexemes of other POS. For example, the noun *oglašavanje* ’advertising’, is analyzed at the surface layer as: *o-glaš-av-a-n-j-e*; *o* = prefix; *glaš* = root; *av* = derivational (aspectual) suffix; *a* = derivational (thematic) suffix; *n* = derivational (participle) suffix; *j* = derivational (gerund) suffix; *e* = inflectional suffix. The presentation of the morphological structure at the deep layer is: *o-glas-jav-a-n-j-e*.

2.2 Derivational Processes

Two major word-formation processes in Croatian are derivation and compounding. The main difference between them is that word-formation processes based on derivation involve lexemes with one lexical morpheme, i.e. derivatives share the same lexical morpheme, whereas word-formation processes based on compounding involve lexemes with two or more lexical morphemes. In other words, compounds have usually two or possibly more different lexical morphemes.

Further in this work, we focus exclusively on derivation and discuss relations between lexemes that share the same root. Generally, derivation can be described as a word-formation process that is based on adding one or more affixes to lexical morphemes. Types of affixes recognized in Croatian lexemes are prefixes, suffixes, and interfixes for compounds. That means that the derivation in Croatian is predominantly based on affixation - prefixation, suffixation, or simultaneous prefixation and suffixation. Simultaneous prefixation and suffixation is not interpreted as circumfixation since prefixes and suffixes retain their meaning when used independently in other derivational processes. In other words, we have not come across a single example in which the meaning of a prefix or a suffix when used independently differs from that when used simultaneously. Generally, suffixation is the most productive derivational process. In the development of CroDeriv the following derivational processes were recognized:

[1] There are two exceptions to this rule: 1) multiple prefixation is not possible in compounds, and 2) an inflectional suffix can occur only once in the morphological structure.

1. **suffixation** – addition of single or multiple suffixes or substitution of suffixes
   - *bac(ati)* 'to throw' + *-ač* = *bacač* 'thrower, pitcher'
   - *kazališt(e)* 'theater' + *-ar* + *-ac* = *kazalištarac* 'theater artist'
   - *bac(iti)* PF 'to throw + *-ati* = *bacati* IPF 'to throw'
2. **prefixation** - addition of single or multiple prefixes
   - *nad-* + *moć* 'power' = *nadmoć* 'superiority'
   - *iz-* + *ne-* + *moći* ' be able' = *iznemoći* 'lose power, languish'
   - *pred-* + *s-* + *kazati* ' to tell' = *predskazati* ' to predict'
3. **simultaneous prefixation and suffixation**
   - *ob-* + *nov* 'new' + *-iti* = *obnoviti* 'to renew'
   - *u-* + *sreć(a)* 'happiness + *-iti* = *usrećiti* 'to make happy'
   - *pod-* + *voz(iti)* 'to drive' + *-je* = *podvozje* 'undercarriage'
4. **back-formation + zero suffixation** - subtraction of stems
   - *upis(ati)* 'to enroll' + ø = *upis* 'enrollment'
   - *uvid(jeti)* 'to see, to realize' + ø = *uvid* 'insight'
   - *dokaz(ati)* 'to prove' + ø = *dokaz* 'proof'
5. **SE** - addition of the reflexive particle *se*[2]
   - *dopisivati* 'to add by writing' + *se* = *dopisivati se* 'to correspond'
   - *ograditi* 'to fence off' + *se* = *ograditi se* 'to dissociate'
   - *tužiti* 'to sue' + *se* = *tužiti se* 'to complain'
6. **ablaut** - a systematic variation of vowels in the same root, usually combined with various types of affixation
   - *sagledati* PF 'to perceive' = *saglédati* IPF 'perceive '
   - *pomoći* PF 'to help' = *pomagati* IPF 'to help'
   - *smrdjeti* 'to stink' = *smrad* 'smell, stench'
7. **conversion / zero derivation** - derivation without any change in form of the stem
   - *mlada* 'young (adjective)' = *mlada* 'bride (noun)'
   - *nečist* 'impure (adjective) = *nečist* 'dirt (noun)'
   - *leteći* 'flying (participle, verbal adverb)' = leteći 'flying (adjective)'

These are major processes used in the derivation of Croatian lexemes. However, there are numerous combinations of processes listed above that take place simultaneously, e.g. ablaut + suffixation, prefixation + ablaut, ablaut + back-formation, prefixation + ablaut + suffixation (+ se), and prefixation + se. Since most of these combinations of derivational processes are poorly covered in the existing literature for Croatian, and some of them are not even mentioned at all, we will list a few examples that we came across and that we consider to be relevant:

1. **ablaut + suffixation**
   - *prigovor(iti)*PF + *-ati* 'to complain' = *prigovarati* IPF 'to complain'
   - *bra(ti)* 'to pick' + *-ba* = *berba* 'harvest'

[2] The reflexive particle *se* is not an affix, but it takes part in numerous derivational processes of Croatian verbs and changes the meaning of derivatives. In addition, it is an integral part of the lexeme. In other words, a lexeme does not exist as an independent word without this particle. The particle *se* should be distinguished from the reflexive pronoun *sebe* 'self'. Sometimes they are mixed up because the clitic form of the reflexive pronoun *sebe* is *se*.

2. **prefixation + ablaut**
    - *pre-* + *zvati se* ’have a name’ = *prezivati se* ’have a surname’
3. **prefixation + ablaut + suffixation**
    - *o-* + *govor(iti)* ’to speak’ + *-ati* = ogovarati ’to slander’
    - *na-* + *vod(i-ti)* ’to lead $_{\text{IPF}}$’ + *-ϕ-ti* = *navesti* ’to lead $_{\text{PF}}$
4. **prefixation + ablaut + suffixation + se**
    - *pre-* + *nov* ’new’ + *-jati se* = *prenavljati se* ’to pretend’
    - *pre-* + *ne-* + *mo(ći)* ’can, be able’ + *-ati se* = *prenemagati se* ’to pretend, to show off’
5. **prefixation + se**
    - *na-* + *jesti* ’to eat’ + *se* = *najesti se* ’to eat one’s fill’
    - *za-* + *trčati* ’to run’ + *se* = *zatrčati se* ’to start running’
6. **prefixation - se** (dropping out of *se*)
    - *u-* + *suglasiti (se)* ’to agree’ = *usuglasiti* ’to agree, to get along’
7. **ablaut + back-formation**
    - *iz(a)bra(ti)* ’to pick’+ ϕ = *izbor* ’choice’
    - *razves(ti se)* ’to divorce’ + ϕ = razvod ’divorce’
    - *opozva(ti)* ’to recall’ + ϕ = *opoziv* ’recall’

This extensive list of derivational processes is made possible by grouping lexemes into derivational families, i.e. the groups of lexemes with the same root. We discuss the structure of derivational families and derivational relations between lexical entries in more detail in the next section.

2.3 Derivational Families

Each derivational family in CroDeriv is structured so that in its center there is a lexeme that represents the central point or origin of the entire family.[3] This central lexeme is unmotivated, i.e. it is not derived from any other stem. These central or core lexemes are derived directly from roots, e.g.: *baciti* ’to throv $_{\text{PF}}$’ from the root *bac*, *ruka* ’hand’ from the root *ruk*, and *nov* ’new’ from the root *nov*. In some cases, roots are identical to actual words in Croatian and in some cases, they are not. We refer to these core lexemes as first-degree derivatives. Derivational families are further modeled in such a way that second-degree derivatives are derived from the core lexeme. Second-degree derivatives are those that, as a rule, differ from the first-degree lexemes only in that they have one or two additional affixes, e.g.:

- *baciti* ’to throw’ - *izbaciti* ’to throw out’, *odbaciti* ’to reject, *ubaciti* ’to throw into’ etc. All second-degree derivatives in this derivational families are derived via prefixation.
- *ruka* ’hand’ - *rukav* ’sleeve’, *rukavica* ’glove’ (suffixation), *rukovati* ’to handle’ (suffixation), *izručiti* ’to extradite’, *uručiti* ’to deliver’ (prefixation), *područje* ’area’, *priručan* ’handy’ (prefixation + suffixation) etc.
- *nov* ’new’ - *novac* ’money’, *novak* ’rookie’, *novost* ’news’ (suffixation), *obnoviti* ’to renew’, *ponoviti* ’to repeat’ (prefixation + suffixation) etc.

Second-degree derivatives provide the basis for further derivational steps in which they serve as the basic lexeme and they are the origin of smaller sub-families or derivational branches. In some cases, second-degree derivatives represent the end of the derivation chain, e.g.: *ruče* ’gymnastic arms’, *rukav* ’sleeve’, *ručerda*, *ručetina* ’hand, (augumentative)’ *ručica*, *ručka* ’handle’, *naručje* ’bosom’, *narukvica*

[3] In rare cases where we cannot base a family on only one lexeme, two lexemes are found at the center of the derivational family.

’bracelet’ are the second-degree derivatives of the stem *ruka* that do not motivate any other lexeme. However, it is much more common for second-degree derivatives to serve as the basis for sub-families that can extend up to seven members in derivational chains. This is the maximum number of derivatives in derivational chains recorded so far. It is possible that this number will increase with the further expansion of CroDeriv. For example:

- 1. *govor* ’speech’ - 2. *govoriti* ’to speak’ - 3. *odgovoriti* - ’to answer, to respond’ - 4. *odgovarati* ’to answer, to match, to account for, to be responsible for’ - 5. *odgovoran* ’responsible’ - 6. *neodgovoran* ’irresponsible’ - 7. *neodgovornost* ’irresponsibility’.
- 1. *glas* ’voice, tone, vote’ - 2. *glasiti* ’to be addressed to, to read’ - 3. *suglasiti se* ’to agree’ - 4. *usuglasiti* ’to agree$_{\text{PF}}$’ - 5. *usuglašavati* ’to agree$_{\text{IPF}}$’ - 6. *usuglašavan* ’agreed upon (participle) - 7. *usuglašavanje* ’harmonization’.

In CroDeriv’s lexical entries, we do not record the full derivational chain. We mark only the last derivational step, that is, only the stem from which a particular lexeme is derived is indicated. For example, in the lexical entry for the noun *neodgovornost* we only indicate that it is derived from the adjective *neodgovoran*. The full structure of lexical entries in CroDeriv is presented in Filko et al. (2019, 2021).

In Table 1 below, we show how the lexical material is processed and prepared for input into CroDeriv. The examples are from the derivational family structured around the root SĚK. Its meaning is associated with cutting and dismembering. The first-degree derivative is the verb sjeći ’to cut’.[4] We use the symbol *ě* for the reflexes of Proto-Slavic *jat* in the contemporary Croatian language. In this way, we solve the problem of numerous surface allomorphy and connect all reflexes to the representative *ě* at the deep layer. Note that there are four allomorphs at the surface layer of the same root in only nine examples in Table 1 below (SL column). At the deep layer there is only one representative morph - *sěk*, except in the last example. We will discuss this and similar cases in the next section.

I	II	SL	DL
sjeći, V - sjek + ti (S)		sje-ći	sěk-ø-ti
	sjecište, N - sjek(ti) + ište (S)	sjec-išt-e	sěk-išt-e
	sjekotina, N - sjek(ti) + otina (S)	sjek-ot-in-a	sěk-ot-in-a
	sječa, N - sjek(ti) + ja (S)	sječ-a	sěk-j-a
	siječanj, N - sjek(ti) + anj (S)	siječ-anj-ø	sěk-nj-ø
	sječivo, N - sjek(ti) + ivo (S)	sječ-iv-o	sěk-iv-o
	sjekira, N - sjek(ti) + ira (S)	sjek-ir-a	sěk-ir-a
	sjekutić, N - sjek(ti) + utić (S)	sjek-ut-ić-ø	sěk-ut-ić-ø
	sjeckati, V - sjek(ti) + kati (S)	sjec-k-a-ti	sěc-k-a-ti

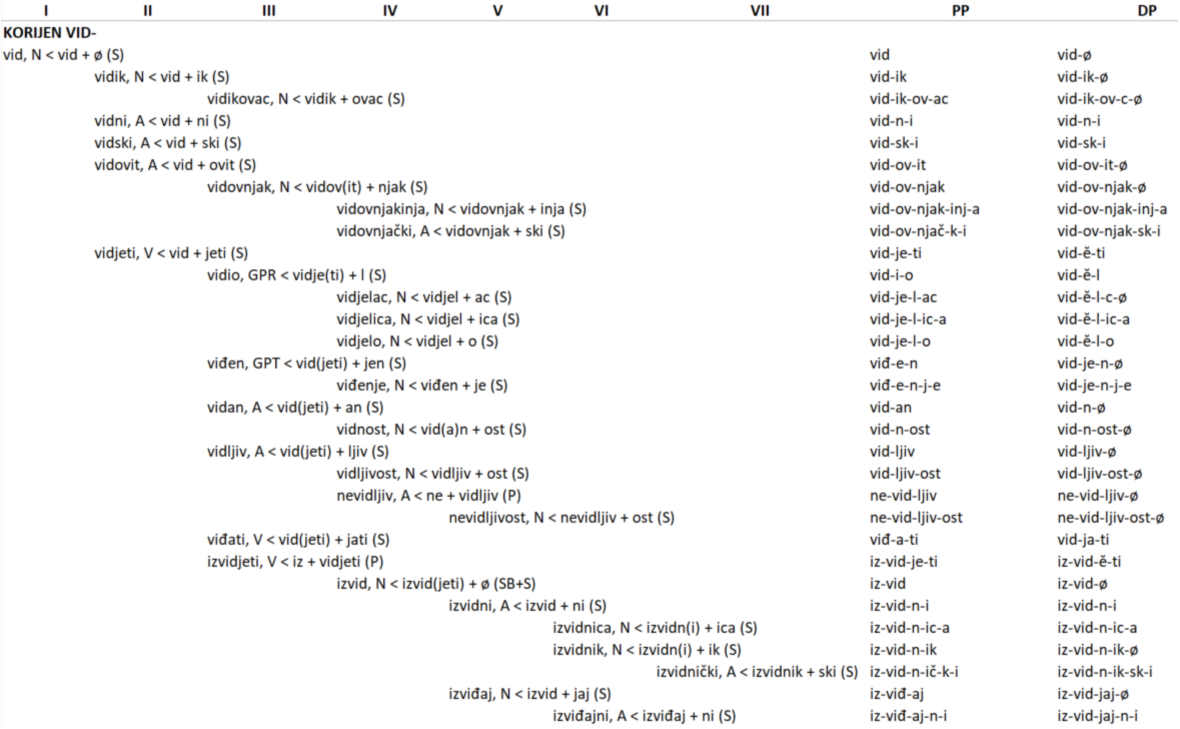
Table 1: An example from the derivational family of the root *SĚK* ’to cut’[5]

3 Discussion

In the previous sections, we indicated that each lexeme in CroDeriv is morphologically segmented and that the segmentation is performed at two layers - surface and deep. We also mentioned that in CroDeriv we combine two types of morphological data, i.e. in addition to the morphological segmentation for each lexeme, we record the word-formation relations with other lexemes as well as word-formation processes by which the lexemes were created. In Figure 1, we present a part of the derivational family of the root *VID* ’sight, to see’. For each lexeme, we provide information on the word class (*N = noun, V = verb, GPR = active past participle, GPT = passive past participle* etc.), stem, affixes that participate in

[4] I= first-degree derivatives, II = second-degree derivatives, SL = segmentation at the surface layer, DL = segmentation at the deep layer (cf. Section 2.1), V = verb, N = noun, (S) = suffixation.

the derivational process, and the type of the derivational process (*S = suffixation, SB+S = subtraction + zero suffixation, P = prefixation* etc.). Columns I, II, III, etc. indicate whether a lexeme is a first-, second- or third-degree derivative. The final two columns refer to morphological segmentation at surface (PP) and deep (DP) layer. We have indicated that in CroDeriv's lexical entries, we do not provide information about full derivational chains. Instead, we provide information about the stem that served for the derivation of that lexeme. Information about the full derivation chain can be found using the visualization tool available to users of the lexicon.



I	II	III	IV	V	VI	VII	PP	DP
KORIJEN VID-								
vid, N < vid + ø (S)							vid	vid-ø
	vidik, N < vid + ik (S)						vid-ik	vid-ik-ø
		vidikovac, N < vidik + ovac (S)					vid-ik-ov-ac	vid-ik-ov-c-ø
	vidni, A < vid + ni (S)						vid-n-i	vid-n-i
	vidski, A < vid + ski (S)						vid-sk-i	vid-sk-i
	vidovit, A < vid + ovit (S)						vid-ov-it	vid-ov-it-ø
		vidovnjak, N < vidov(it) + njak (S)					vid-ov-njak	vid-ov-njak-ø
			vidovnjakinja, N < vidovnjak + inja (S)				vid-ov-njak-inj-a	vid-ov-njak-inj-a
			vidovnjački, A < vidovnjak + ski (S)				vid-ov-njač-k-i	vid-ov-njak-sk-i
	vidjeti, V < vid + jeti (S)						vid-je-ti	vid-ě-ti
		vidio, GPR < vidje(ti) + l (S)					vid-i-o	vid-ě-l
			vidjelac, N < vidjel + ac (S)				vid-je-l-ac	vid-ě-l-c-ø
			vidjelica, N < vidjel + ica (S)				vid-je-l-ic-a	vid-ě-l-ic-a
			vidjelo, N < vidjel + o (S)				vid-je-l-o	vid-ě-l-o
		viđen, GPT < vid(jeti) + jen (S)					viđ-e-n	vid-je-n-ø
			viđenje, N < viđen + je (S)				viđ-e-n-j-e	vid-je-n-j-e
		vidan, A < vid(jeti) + an (S)					vid-an	vid-n-ø
			vidnost, N < vid(a)n + ost (S)				vid-n-ost	vid-n-ost-ø
		vidljiv, A < vid(jeti) + ljiv (S)					vid-ljiv	vid-ljiv-ø
			vidljivost, N < vidljiv + ost (S)				vid-ljiv-ost	vid-ljiv-ost-ø
			nevidljiv, A < ne + vidljiv (P)				ne-vid-ljiv	ne-vid-ljiv-ø
				nevidljivost, N < nevidljiv + ost (S)			ne-vid-ljiv-ost	ne-vid-ljiv-ost-ø
		viđati, V < vid(jeti) + jati (S)					viđ-a-ti	vid-ja-ti
		izvidjeti, V < iz + vidjeti (P)					iz-vid-je-ti	iz-vid-ě-ti
			izvid, N < izvid(jeti) + ø (SB+S)				iz-vid	iz-vid-ø
				izvidni, A < izvid + ni (S)			iz-vid-n-i	iz-vid-n-i
					izvidnica, N < izvidn(i) + ica (S)		iz-vid-n-ic-a	iz-vid-n-ic-a
					izvidnik, N < izvidn(i) + ik (S)		iz-vid-n-ik	iz-vid-n-ik-ø
						izvidnički, A < izvidnik + ski (S)	iz-vid-n-ič-k-i	iz-vid-n-ik-sk-i
				izviđaj, N < izvid + jaj (S)			iz-viđ-aj	iz-vid-jaj-ø
					izviđajni, A < izviđaj + ni (S)		iz-viđ-aj-n-i	iz-vid-jaj-n-i

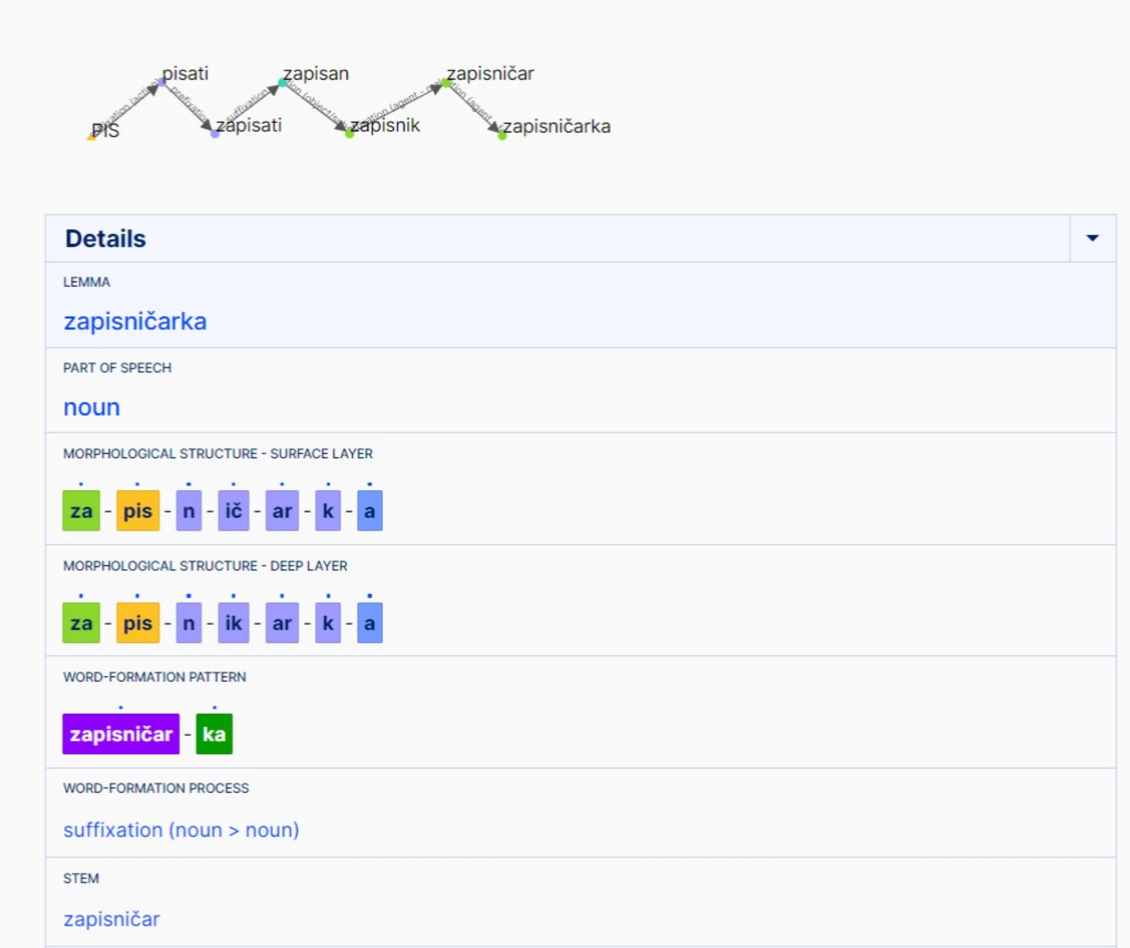
Figure 1: The excerpt of the derivational family for the root *VID-*

In Figure 2, we give an example of how the entry in CroDeriv is structured. We also show a visualization tool used in the new CroDeriv's online search interface that shows the full derivation chain for the lexeme *zapisničarka* 'scorer, clerk (female)'. The full derivational chain is as follows:

- *pisati* 'to write' - *zapisati* 'to write down'- *zapisan* 'written down (participle)'- *zapisnik* 'record, minutes' - *zapisničar* 'scorer, clerk (male)' - *zapisničarka* 'scorer, clerk (female)'.

Such two-sided processing of Croatian morphology has many advantages: 1. it provides an insight into the morphological structure of lexemes; 2. the segmentation at the deep layer enables easier and more precise recognition of all root allomorphs and their linking to representative morphs; 3. the segmentation at the deep layer also enables easier and more precise recognition of all affixal allomorphs; 4. the segmentation provides an excellent insight into morpho-phonological processes and changes occurring in the Croatian language.

The approach that combines segmentation and marking of word-formation relations between lexemes is based on the assumption that the elements participating in each word-formation process cause morpho-phonological changes precisely in that process. Such an approach to word formation in Croatian is new since it does not assume the existence of stems in which certain morpho-phonological processes have already been carried out before a certain word-formation process began. The basic assumption from which we start is that if there are one or more morpho-phonological changes, e.g. triggered by the addition of affixes, any such change occurs in that process. In other words, they are not inherited or already



Details

LEMMA

zapisničarka

PART OF SPEECH

noun

MORPHOLOGICAL STRUCTURE - SURFACE LAYER

za - pis - n - ič - ar - k - a

MORPHOLOGICAL STRUCTURE - DEEP LAYER

za - pis - n - ik - ar - k - a

WORD-FORMATION PATTERN

zapisničar - ka

WORD-FORMATION PROCESS

suffixation (noun > noun)

STEM

zapisničar

Figure 2: The lexical entry *zapisničarka* in the new CroDeriv search interface

implemented in stems.[6] Unlike the approach to Croatian morphology in CroDeriv, such an approach is represented in many works on word formation in Croatian and related Slavic languages (Babić, 2002; Klajn, 2002, 2003).

Although there are many advantages to the approach we advocate, there are certain cases that raise questions. We have stated that in Table 1 above, in the last example, the root at the deep layer is not connected to the morph that is representative of other root allomorphs. This also applies to examples 3. and 4. below;

1. *sjeći* ’to cut’ - sje-ći / sěk-ø-ti
2. *sjeknuti* ’to cut (deminutive) - sjek-n-u-ti / sěk-n-u-ti
3. *sjeckati* ’to cut (deminutive)’ - sjec-k-a-ti / sěc-k-a-ti
4. *sjecnuti* ’to cut (deminutive) - sjec-n-u-ti / sěc-n-u-ti

We will give a few more examples from another derivational family:

1. *pucati* ’to crack, to fire’ - puc-a-ti / puk-a-ti
2. *puckati* ’to crack (deminutive)’ - puc-k-a-ti / puc-k-a-ti
3. *pucnuti* ’to crack (deminutive)’ - puc-n-u-ti / puc-n-u-ti

[6] In terms of morpho-phonological rules, we largely follow Marković (2013), a modern, precise, and extensive account of Croatian morpho-phonology.

The reason why we here list different root morphs in the deep structure is that there is no morpho-phonological rule that could explain the change of the root *puk* to *puc* before the diminutive suffix *-k* in the contemporary Croatian language. The same holds for the root *sěk* in the examples above. In addition, in example 2 for the root *sěk*, the deep-layer segmentation is *sěk-n-u-ti*. In example 4, the deep-layer segmentation is *sěc-n-u-ti*. In other words, we have two different root allomorphs in the same phonological environment. The same holds for example 3 for the root *puk*. Here again, the deep-layer segmentation is *puc-n-u-ti*, although there is a lexeme *puknuti* ’to crack, to fire’ which is at the deep layer segmented as *puk-n-u-ti*.

Marković (2013, p. 140, 146) considers such examples to be ”pre-sibilarized” or ”pre-iotized”. He states that in many similar examples ”we have a possible sibilarization, however, it is probably more elegant to connect them with a pre-sibilarized verb root” (Marković, 2013, p. 140). The author does not provide an additional explanation, and we interpret this to mean that pre-sibilarization or pre-iotation were carried out in the earlier stages of language development and cannot be explained by the rules that apply in the contemporary language (cf. Mihaljević, 1991). In Croderiv we use a solution in which both root allomorphs are listed at the deep layer, but the second one in parentheses. We consider such and similar lexemes as members of the same derivational family.

We encountered a similar problem with lexemes derived by ablaut. For example:

1. *brati* ’to pick’ - *berba* ’harvest’ - *birati* ’to choose - *izbor* ’choice,
2. *teći* ’to flow’ - *protjecati* ’to flow’ - *protok* ’flow’,

Here, the question also arises as to which of the root allomorphs to take as the representative one, since morpho-phonological rules cannot justify the selection of only one. In CroDeriv we use a similar solution as in the examples above. The segmentation at the deep layer is as stated in the above examples for the roots *sěk* and *puk*, but one of the root allomorphs is taken to be representative and listed in parentheses. In this way, we can present such lexemes as members of the same derivational family.

The next problem we encountered relates to the homographic roots. For example, the verbs *leći* ’to lie down’, *ležati* ’to lay down’, and *leći* ’to lay (eggs), to brood.’ have the homographic root *leg*. Both lexemes *leći* have the same deep-layer presentation: *leg-ø-ti*. The deep-layer presentation for *ležati* is *leg-a-ti*. Similar homography occurs with numerous other roots, for example, *kupiti* ’to buy’ and *kupiti* ’to gather’. Both first-degree lexemes and many derivatives in their derivational families are semantically very similar. We solve the problem with homographic roots by marking them with a different number: kup_1 for lexemes semantically associated with buying, kup_2 for lexemes associated with gathering, kup_3 for lexemes associated with bathing, kup_4 for lexemes associated with docking etc. As for their semantic distinction, checking in etymological dictionaries (Skok, 1971, 1972; Matasović et al., 2016, 2021; Snoj, 2003) is the only way to solve such problems.

The last issue we will discuss here refers to the structuring of derivational families composed of suppletive stems. The problem we have not tackled yet concerns one of the biggest families in terms of the number of its members - the one which contains verbs like *ići* ’to go’ and *otići* ’to leave’, as well as *doći* ’to come $_{\text{PF}}$’ and *dolaziti* ’to come $_{\text{IPF}}$’.

We can assume that the lexical morpheme in the verb *ići* ’to go’ is *id*, and the segmentation at the surface and deep layer could be shown as follows:

- *ići* ’to go’ - i-ći / id-ø-ti

As a rule, a lexical morpheme is defined as one that is obligatory in every word. If we look at the verb *doći* ’to come’, the lexical morpheme does not exist at the surface layer. Instead, the surface structure of this verb consists of the prefix *do-* and the suffix *-ći*, which is an allomorph of the infinitive ending *-ti*:

- *doći* ’to come’ - do (prefix)-XXX-ći (suffix).

Apart from the problematic surface layer, it also remains unclear how to represent the deep structure of this lexeme; perhaps as: *do-id-ø-ti*. The same issue appears with numerous lexemes with the same root: *naći* ’to find’, *ući* ’to enter’, *proći* ’to pass’... It also remains unclear which rule can be used to explain such a structure. Furthermore, the aspectual pairs *doći* 'to come $_{PF}$' and *dolaziti* 'to come $_{IPF}$', *naći* 'to find $_{PF}$' and *nalaziti* 'to find $_{IPF}$', *ući* 'to enter $_{PF}$' and *ulaziti* 'to enter $_{IPF}$' are derived from suppletive stems. Again, there is no morpho-phonological rule that holds for the contemporary Croatian which could be used for the explanation of suppletive stems. As a possible solution, the procedure described in the examples above can be used: 1. to keep separate deep-layer roots in segmentation, and 2. to provide a root taken to be representative in parenthesis in order to enable the assignment of these lexemes into the same derivational family. In this way, the search for morphologically and semantically related lexemes in CroDeriv would be enabled.

4 Conclusion

In this paper, we have briefly presented the structure of the CroDeriv, the derivational lexicon for Croatian which provides information about the morphological structure of words and about derivational links with other words, thus forming the derivational families. Since the structure of the lexical entries in CroDeriv has been explained in more detail in previous work (e.g. (Filko et al., 2020), here, we have focused on the derivational processes in Croatian that have not yet been recognized in the existing literature. These processes emerged when the Croatian words were analyzed in the format used in CroDeriv. Moreover, such a formal analysis has forced us to find both computationally applicable and theoretically plausible solutions for unsolved (and even theoretically untackled) problems in Croatian morphology and word formation in order to include very frequent, but irregular lexemes in CroDeriv. Only a handful of the most interesting ones were presented here due to the limitations of this paper, but we can foresee that even more problems of this kind will emerge with further analysis of the data. We hope that the procedure and general rules applied in the examples presented here could be, with or without the modifications, applied to future issues, as well.

References

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

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

Geert Booij. 2005. *The Grammar of Words. An introduction to Linguistic Morphology*. Oxford Textbooks in Linguistics. Oxford University Press, New York.

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

Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2020. The Design of Croderiv 2.0. *The Prague Bulletin of Mathematical Linguistics* 115:83–104. https://doi.org/10.14712/00326585.006.

Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2021. Deriving the graph: Using affixal senses for building semantic graphs. In Fiammetta Namer, Nabil Hathout, Stéphanie Lignon, Magda Ševčíková, and Zdeněk Žabokrtský, editors, *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology*. Karlovo sveučilište, Prag, pages 120–128.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*. Understanding Language. Hodder Edducation, London, 2nd edition.

Ivan Klajn. 2002. *Tvorba reči u savremenom srpskom jeziku: prvi deo: slaganje i prefiksacija*. Zavod za udžbenike i nastavna sredstva : Institut za srpski jezik SANU ; Matica srpska, Beograd: Novi Sad.

Ivan Klajn. 2003. *Tvorba reči u savremenom srpskom jeziku: drugi deo: sufiksacija i konverzija*. Zavod za udžbenike i nastavna sredstva : Institut za srpski jezik SANU ; Matica srpska, Beograd: Novi Sad.

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

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

Ranko Matasović, Dubravka Ivšić Majić, and Tijmen Pronk. 2021. *Etimološki rječnik hrvatskoga jezika*, volume Sv. 2, O-Ž. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Ranko Matasović, Tijmen Pronk, Dubravka Ivšić, and Dunja Brozović-Rončević. 2016. *Etimološki rječnik hrvatskoga jezika*, volume Sv. 1, A-Nj. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Milan Mihaljević. 1991. *Generativna i leksička fonologija*. Školska knjiga, Zagreb.

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

Petar Skok. 1971. *Etimologijski rječnik hrvatskoga ili srpskoga jezika*, volume Knjiga 1. Jugoslavenska akademija znanosti i umjetnosti, Zagreb.

Petar Skok. 1972. *Etimologijski rječnik hrvatskoga ili srpskoga jezika*, volume Knjiga 2. Jugoslavenska akademija znanosti i umjetnosti, Zagreb.

Marko Snoj. 2003. *Slovenski etimološki slovar*. Modrijan, Ljubljana.

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