Processing Croatian Morphology: Roots, Segmentation and Derivational Families

Krešimir Šojat & Matea Filko

Faculty of Humanities and Social Sciences

University of Zagreb, Croatia

Introduction

- This work deals with the development of the Croatian derivational lexicon – CroDeriv
  - 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
  - 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 → accurate recognition of lexical morphemes

CroDeriv

- 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 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

CroDeriv

- 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

Outline of the talk

- In this work, we present further development and enrichment of the existing version of CroDeriV.
- 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.
- The main derivational processes are presented as well as some that are not mentioned or that are only marginally described in the existing literature.
- We illustrate the structure of derivational families and lexical entries in CroDeriv.
- We discuss some problems we have encountered in our work and outline possible solutions.

CroDeriV – segmentation

- Each lexical entry in CroDeriv contains information on the morphological structure of lexemes
- Each lexeme is segmented into morphemes that it consists of.
  - In 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.

Types of morphemes

- Morpheme is the basic morphological unit.
  - usually, it is defined as the smallest language unit that can be associated both with form and meaning
  - morphemes are abstract units whereas morphs are their physical realization
- Types of morphemes recognized in Croatian are:
  - prefixes
  - lexical morphemes (roots)
  - derivational suffixes
  - inflectional suffixes
  - interfixes (for compounds)

Morphemes and segmentation

- Each type of morpheme can occur more than once in the morphological structure of lexemes.
- 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

The following example illustrates multiple prefixation and suffixation in derivation:

- *s-po-raz-um-je-ti se* 'come to an agreement'
- *ne-s-po-raz-um* 'disagreemnet'

Segmentation – two layers

- 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. At the deep layer, all allomorphs are linked to their representative morph.
- SURFACE LAYER: raščišćavati – raš-čišć-av-a-ti
- DEEP LAYER: raš → raz; čišć → čist; av → 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

Segmentation – two layers

- The same approach – segmentation at the surface and deep layer – is applied to lexemes of other POS.
- For example, the noun *oglašavanje* 'advertising',
- surface layer: 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.
- deep layer: o-glas-jav-a-n-j-e

Word-formation processes

- Two major word-formation processes in Croatian are derivation and compounding.
  - derivation → involve lexemes with one lexical morpheme, i.e. derivatives share the same lexical morpheme
  - word-formation → involve lexemes with two or more lexical morphemes.
    - In other words, compounds have usually two or possibly more different lexical morphemes.
- Further, we focus on derivation and discuss relations between lexemes that share the same root.

Derivation

- Derivation can be described as a word-formation process that is based on adding one or more affixes to lexical morphemes.
  - affixes → prefixes, suffixes, and interfixes for compounds.
- 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.
    - 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.
- Suffixation is the most productive derivational process.

Derivational processes

- In the development of CroDeriv the following derivational processes were recognized:

- suffixation – addition of single or multiple suffixes or substitution of suffixes
  - *bac(ati)* 'to throw' + *ač* = *bacač* 'thrower, pitcher'
  - *kazališt(e)* 'theater' + *ar* + *ac* = *kazališ**t**arac* 'theater artist'
  - *bac(iti)* 'to throw' + *ati* = *bacati* 'to throw'

Derivational processes

- 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’
- simultaneous prefixation and suffixation
  - *ob* + *nov* + -iti = *obnoviti* 'to renew’
  - *u* + *sreć(a)* 'happiness + *iti* = *usrećiti* 'to make happy’
  - *pod* + *voz(iti)* 'to drive' + *je* = *podvozje* 'undercarriage'

Derivational processes

- 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'
- addition of the reflexive particle se
    - 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*
  - *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'

Derivational processes

- ablaut - a systematic variation of vowels in the same root, usually combined with various types of affixation
  - *sagledati* 'to perceive' = *saglédati* 'perceive'
  - *pomoći* 'to help' = *pomagati* 'to help'
  - *smrdjeti* 'to stink' = *smrad* 'smell, stench'
- conversion / zero derivation - derivation without any change in form of the stem
  - *mlada* 'young (adjective)' = *mlada* 'bride (noun)'
  - *nečist* 'impure (adjective) = *nečist* 'dirt (noun)'
  - *letećі* 'flying (participle, verbal adverb)' = *letećі* 'flying (adjective)'

Derivational processes

- These are major processes used in the derivation of Croatian lexemes.
  - 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 many 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:

Derivational processes

- ablaut + suffixation
  - *prigovor(iti)* + *ati* 'to complain' = *prigovarati* 'to complain’
  - *bra(ti)* 'to pick' + *ba* = *berba* 'harvest’
- prefixation + ablaut
  - *pre* + *zvati se* 'have a name' = *prezivati se* 'have a surname’
- prefixation + ablaut + suffixation
  - *o* + *govor(iti)* 'to speak' + *ati* = *ogovarati* 'to slander’
  - *na* + *vod(i-ti)* 'to lead' + *ø-ti* = *navesti* 'to lead’
- 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'

Derivational processes

- prefixation + se
  - *na* + *jesti* 'to eat' + *se* = *najesti se* 'to eat one's fill’
  - *za* + *trčati* 'to run' + *se* = *zatrčati se* 'to start running’
- prefixation - se (dropping out of *se*)
  - *u* + *suglasiti (se)* 'to agree' = *usuglasiti* 'to agree, to get along’
- ablaut + back-formation
  - *iz(a)bra(ti)* 'to pick'+ *ø* = *izbor* 'choice’
  - *razves(ti se)* 'to divorce' + *ø* = *razvod* 'divorce’
  - *opozva(ti)* 'to recall' + *ø* = *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.

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.
  - In rare cases where we cannot base a family on only one lexeme, two lexemes are found at the center of the derivational family.
- 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 throw’ from the root ***bac***,
  - *ruka* 'hand' from the root ***ruk***,
  - *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

- 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
  - *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.

Root SĚK-

I	II	III	IV	V	VI	VII	PP	DP
KORIJEN SĚK-								
sjeći, V < sjek + ti (S)							sje-ći	sěk-ø-ti
	sjekao, GPR < sje(ći) + l (S)						sjek-a-o-ø	sěk-ø-l-ø
	sječen, GPT < sje(ći) + en)S)						sječ-e-n-ø	sěk-e-n-ø
		sječenje, N < sječen + je (S)					sječ-e-n-j-e	sěk-e-n-j-e
	sjecište, N < sje(ći) + ište (S)						sjec-išt-e	sěk-išt-e
		sjecišni, A < sjeciš(te) + ni (S)					sjec-iš-n-i	sěk-išt-n-i
	sječa, N < sje(ći) + ja (S)						sječ-a	sěk-j-a
	siječanj, N < sje(ći) + anj (S)						siječ-anj-ø	sěk-nj-ø
		siječanjski, A < siječanj + ski (S)					siječ-anj-sk-i	sěk-nj-sk-i
	sječivo, N < sje(ći) + ivo (S)						sječ-iv-o	sěk-iv-o
	sjekač, N < sje(ći) + ač (S)						sjek-ač-ø	sěk-ač-ø
	sjekira, N < sje(ći) + ira (S)						sjek-ir-a	sěk-ir-a
		sjekirica, N < sjekir(a) + ica (S)					sjek-ir-ic-a	sěk-ir-ic-a
	sjekotina, N < sje(ći) + otina (S)						sjek-ot-in-a	sěk-ot-in-a
	sjekutić, N < sje(ći) + utić (S)						sjek-ut-ić-ø	sěk-ut-ić-ø
	sječimice, ADV < sje(ći) + imice (S)						sječ-imice	sěk-imice
	sjeckati, V < sje(ći) + kati (S)						sjec-k-a-ti	**sěc-k-a-ti**
		sjeckao, GPR < sjecka(ti) + l (S)					sjec-k-a-o-ø	**sěc-k-a-l-ø**
			sjeckalica, N < sjeckal + ica (S)				sjec-k-a-l-ic-a	**sěc-k-a-l-ic-a**
		sjecnuti, V < sje(ći) + nuti (S)					sjec-n-u-ti	**sěc-n-u-ti**

Root VID-

A	B	C	D	E	F	G	H	I
I	II	III	IV	V	VI	VII	PP	DP
KORIJEN VID-								
vid, N < vid + ø (S)							vid-ø	vid-ø
	vidik, N < vid + ik (S)						vid-ik-ø	vid-ik-ø
		vidikovac, N < vidik + ovac					vid-ik-ov-ac-ø	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-ø	vid-ov-it-ø
		vidovnjak, N < vidov(it) + njak (S)					vid-ov-njak-ø	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-ø	vid-ě-l-ø
			vidjelac, N < vidjel + ac (S)				vid-je-l-ac-ø	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-ø	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-ø	vid-n-ø
			vidnost, N < vid(a)n + ost (S)				vid-n-ost-ø	vid-n-ost-ø
		vidljiv, A < vid(jeti) + ljiv (S)					vid-ljiv-ø	vid-ljiv-ø
			vidljivost, N < vidljiv + ost (S)				vid-ljiv-ost-ø	vid-ljiv-ost-ø

Derivation – lexical entries

- 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*.
- 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.
  - 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 (maximum number of derivatives in derivational chains recorded so far. It is possible that this number will increase with the further expansion of CroDeriv.)

CroDeriv – lexical entry

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

Segmentation & derivation

- 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.
  - 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 implemented in stems.

Deep layer – multiple roots

- Problems:
  - *sjećі* 'to cut’ - sje-ći / sěk-ø-ti
  - *sjeknuti* 'to cut (deminutive)' - sjek-n-u-ti / sěk-n-u-ti
  - *sjeckati* 'to cut (deminutive)' - sjec-k-a-ti / **sěc-**k-a-ti
  - *sjecnut*i 'to cut (deminutive)' - sjec-n-u-ti / **sěc**-n-u-ti
- Another derivational family:
  - *pucati* 'to crack, to fire' - puc-a-ti / puk-a-ti
  - *puckati* 'to crack (deminutive)' - puc-k-a-ti / **puc**-k-a-ti
  - *puknuti* 'to crack, to fire' - puk-n-u-ti / puk-n-u-ti
  - *pucnuti* 'to crack (deminutive)' - puc-n-u-ti / **puc**-n-u-ti
    - We here list different root morphs in the deep structure since there there is no morpho-phonological rule that could explain the change of the root *puk / puc*, or *sěk* / *sěc* before suffix –k or –n in contemporary Croatian

Deep layer – solution

- 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.
- A similar problem with lexemes derived by ablaut:
    - *brati* 'to pick' - *berba* 'harvest' - *birati* 'to choose - *izbor* 'choice
    - *teći* 'to flow' - *protjecati* 'to flow' - *protok* 'flow'
- A similar solution as in the examples above:
    - *brati* (**bra**-*ø*-ti, root: bra) / *berba* (**ber**-b-a, root: bra) / *birati* (**bir**-a-ti, root: bra) / *izbor* (iz-**bor**-*ø*; root: bra)

Homographic roots

- The next problem we encountered relates to the homographic roots.
- *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':
  - 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 is necessary…

No roots at surface layer

- The problem we have not tackled yet concerns one of the biggest families - the one which contains verbs like:
  - *ići* 'to go' and *otići* 'to leave', as well as *doći* 'to come' and *dolaziti* 'to come'.
- 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:
  - *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 and the suffix.
  - *doći* 'to come' - do (prefix) – XXX – ći (suffix)

Suppletive stems

- 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' and *dolaziti* 'to come', *naći* 'to find' and *nalaziti* 'to find', *ući* 'to enter' and *ulaziti* 'to enter' 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.

THANK YOU
FOR YOUR ATTENTION!