Filozofski
fakultet
Sveučilišta u
Zagrebu

Redesign of the Croatian derivational lexicon

Matea Filko, Krešimir Šojat, Vanja Štefanec

Faculty of Humanities and Social Sciences, University of Zagreb

{matea.filko, ksojat, vstefane}@ffzg.hr

Intro

- derivational resources – limited number of languages (22 – Kyjánek 2018)
  - English: CatVar
  - French: Démonette
  - Czech: DeriNet, Derivancze
  - Latin: Word Formation Latin
  - Italian: DerIvaTario
  - Spanish: DeriNet.ES
  - Persian: DeriNet. Fa
  - Polish: The Polish Word-Formation Network
  - German: DErivBase
  - Croatian: DerivBase.HR, CroDeriv...
- what makes CroDeriv different from these resources?

CroDeriv

- first version:
  - only verbs ☺
  - not exactly a derivational resource – focus on a thorough analysis of the morphological structure of lexemes
  - word-formation processes were not explicitly marked

- current version:
  - lexemes of all major POS: verbs, adjectives, nouns, adverbs
  - **complete morphological structure + word-formation patterns + derivational relations**
  - new online interface

CroDeriV 1.0 – recap

- croderiv.ffzg.hr
- 14.500 verbs in infinitive form
  - collected from online corpora and dictionaries
  - information about aspect and reflexivity is also encoded for each verb
- complete **morphological structure**
  - all verbs analyzed for morphemes
  - verbs with the same root mutually connected
    - 3 286 roots
  - recognition of **derivational families**
    - recognition of affixes used in derivational processes with particular roots
      - **their combinations / distribution / frequency**

CroDeriv 1.0 – recap

1. surface layer – morphological analysis

- ***pis**-a-ti – pre-**pis**-a-ti – pre-**pis**-iv-a-ti – is-pre-**pis**-a-ti – is-pre-**pis**-iv-a-ti – po-is-pre-**pis**-a-ti*
- ***let**-je-ti – iz-**let**-je-ti – iz-**lijet**-a-ti*

2. deep layer – allomorph detection

  - _is- = iz-_ _let* = lijet*_
- all allomorphs are linked to the single representative morpheme
  - _is-, iš-, i-, iz- =_ **_iz-_** _let*, lijet* = let*_
- all verbs of the same root are mutually connected – derivational families
  - homographic roots are recognized and marked as e.g. _rib1_, _rib2_…
  - _rib*-ar-i-ti_ ‘to fish’ vs. _rib*-a-ti_ ‘to scrub’

3. stem detection

- enables the recognition of the derivational path of the particular word from the root to the final lexeme
- encoded in the database, but not visible via search interface

CroDeriv 1.0 – recap

- overall structure provided for all verbs – 11 slots:
  - prefixal part: 4 slots
  - lexical part: 3 slots: 2 lexical morphemes + interfix (compounded verbs)
  - suffixal part: 3 slots + infinitive ending (*ti)*

(P4) + (P3) + (P2) + (P1) + (L_2) + (I) + L_1 + (S3) + S2 + S1 + *ti*

pis + Ø + Ø + a + *ti* *pisati* 'to write'

pis + uck + Ø + a + *ti* *pisuckati* 'to write, dim.'

po + is + pre + pis + Ø + iv + a + *ti* *poisprepisati* 'to copy all over by writing, distr.'

P = prefix; **L** = lexical morpheme / stem; **I** = interfix; **S** = suffix; **()** = non-obligatory

- this kind of (closed and regular) structure cannot be applied to other POS
  - each slot in verbal morphological structure has its function
  - this is not the case with nouns and adjectives

CroDeriv 2.0

- complete redesign of the database structure:

1. morphological structure has to be represented as more flexible

  - no strictly defined slots
  - predominant word-formation processes:
    - verbs = prefixation
    - nouns, adjectives = suffixation

this results in completely different morphological structures

2. complete word-formation analysis has to be included in CroDeriv 2.0

  - word-formation rules, patterns, processes and paths were only implicitly marked in CD 1.0
    - often impossible to derive them from morphological analysis

3. full derivational families have to be recognized and visualized

CroDeriv 2.0

- adjectival and nominal lemmas were collected from corpora and online dictionaries of Croatian
    - ca. 1.000 adjectives and 6.000 nouns as a representative sample according to their frequency
        - *Croatian frequency dictionary* (Moguš et al., 1999)
        - frequency lists generated by corpus management system NoSketchEngine for both representative corpora (Croatian National Corpus and Croatian web corpus hrWaC)
    - both motivated and unmotivated lexemes
    - adverbs are included in the most diversified derivational families (for the time being)
    - NE are excluded

CroDeriv 2.0 – morphological analysis

- **manual segmentation** – two layered approach as applied to verbs
  - surface layer: all possible morphs are identified and marked for their type

    uč-i-telj-ic-a ‘female teacher’

    uč = root; *i, telj, ic* = derivational suffixes; *a* = inflectional suffix
    iz-lječ-iv-Ø ‘curable’
    iz = prefix; *lječ* = root; *iv* = derivational suffix; *Ø* = inflectional suffix
  - deep layer: allomorphs are connected to the single representative morpheme

    uk-i-telj-ic-a ***iz-lijek-iv***
- morphological structure regardless of POS: prefixes, roots, interfixes, (derivational and inflectional) suffixes
  - each morpheme type can occur more than once

CroDeriv 2.0 – derivational analysis

- word-formation pattern/process:
  - ***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

CroDeriv 2.0 – word-formation processes

- **suffixation**
  - *pjev(ati)* 'to sing' + *-ač* > *pjevač* 'singer'
  - *glas* 'voice' + *-ati* > *glasati* 'to vote'
  - *učitelj* 'teacher' + *-ev* > *učiteljev* 'teacher's'
- **prefixation**
  - *za-* + *pjev(ati)* 'to sing' > *zapjevati* 'to start singing'
  - *do-* + *predsjednik* 'president' > *dopredsjednik* 'vicepresident'
  - *pred-* + *školski* 'school, ADJ' > *predškolski* 'preschool'
- **simultaneous suffixation and prefixation**
  - *o-* + *svoj* 'one's own' + *-iti* > *osvojiti* 'to conquer, to win'
  - *bez-* + *sadržaj* 'content' + *-an*> *besadržajan* 'pointless, content-free'

CroDeriv 2.0 – word-formation processes

- **compounding**
  - *vjer(a)* ‘trust’ + *-o-* + *dostojan* ‘worthy’ > *vjerodostojan* ‘trustworthy’
  - *zlo* ‘evil’ + *upotrijebiti* ‘to use’ > *zloupotrijebiti* ‘to misuse, to abuse’
  - *polu* ‘half’ + *mjesečni* ‘monthly’> *polumjesečni* ‘semimonthly’
- **simultaneous compounding and suffixation**
  - *vod(a)* + *-o-* + *staj(ati)* ‘to stand’ > *vodostaj* ‘water level’
  - *vanjsk(a)* ‘external’ + *-o-* + *trgovin(a)* ‘trade’ + *-ski* > *vanjskotrgovinski* ‘external trade, ADJ’
- **simultaneous prefixation and compounding**
  - *o-* + *zlo* ‘evil’ + *glasiti* ‘to say’ > *ozloglasiti* ‘to discredit, to bring into disrepute’

CroDeriv 2.0 – word-formation processes

- **back-formation**
  - *izlaz(iti)* ‘to exit’ > *izlaz* ‘exit’
- **conversion / zero-derivation**
  - *mlada* ‘young, feminine, ADJ’ > *mlada* ‘bride, N’
- **ablaut**
  - *plesti* = *plet* + *(Ø)* + *(ti)* ‘to twine’ > *plot* ‘fence’

CroDeriv 2.0 – affixal senses

- affixes = **polysemous units**
  (Babić (2002), Lehrer (2003), Lieber (2004, 11), Lieber (2009, 41), Aronoff and Fudeman (2011))
  - one of the affixal meanings is realized in the final motivated lexeme
  - e.g. verbal prefix *nad-* can express two meanings:
    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'
- typology of possible meanings:
  - verbal affixes: Šojat et al. 2012
  - the most productive adjectival suffixes: Filko and Šojat 2017
  - the most productive nominal suffixes: in preparation (Filko, PhD thesis)
    - according to descriptions in Croatian grammar and reference books and modified according to the lexemes in our database

CroDeriv 2.0 – affixal senses – suffix -*ica*

1. **agent, female**, e.g. *učitelj* ‘teacher, male’ > *učiteljica* ‘teacher, female’
2. **person, both sexes**, e.g. *izbjegao* ‘exiled’ > *izbjeglica* ‘refugee’
3. **animal, female**, e.g. *golub* ‘pigeon, male’ > *golubica* ‘pigeon, female’
4. **diminutive**, e.g. *pjesma* ‘song’ > *pjesmica* ‘ditty, rhyme’
5. **thing**, e.g. *sanjar* ‘dreamer, male’ > *sanjarica* ‘dream book’
6. **drink**, e.g. *med* ‘honey’ > *medica* ‘honey liqueur’
7. **plant**, e.g. *otrovan* ‘poisonous’ > *otrovnica* ‘poisonous plant, mushroom (and venomous snake)’

CroDeriv 2.0 – affixal senses – suffix -*ica*

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* ‘saying’

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

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

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

CroDeriV 2.0 – structure of the entry

1. lemma
   - POS
   - gender/aspect/reflexivity/definiteness
2. morphological structure – surface layer
3. morphological structure – deep layer
4. word-formation pattern: base word(s) + affixes
5. stem (allomorph of the stem)
6. affix (allomorph of the affix)
7. affix sense
8. word-formation process (POS > POS)
9. link to the Croatian Language Portal

grammatical categories

morphological structure

derivational properties

link to the entry of the base word in CroDeriv

link to the list of all lemmas derived from this stem

CroDeriV 2.0 – structure of the entry – N

link to the visualization
of the derivational family

1. **lemma:** poslužitelj ‘server’
    - **POS:** N
    - **gender**: masculine
2. **morphological structure – surface layer:** po-služ-i-telj-Ø
(*po* = prefix, *služ* = root, *i, telj* = derivational suffixes, *Ø* = inflectional suffix)
3. **morphological structure – deep layer:** po-slug-i-telj-Ø
(*po* = prefix, *slug* = root, *i, telj* = derivational suffixes, *Ø* = inflectional suffix)
4. **word-formation pattern:** poslužiti + telj
5. **stem (allomorph of the stem)**: posluži (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**

CroDeriV 2.0 – structure of the entry – V

1. **lemma:** potpisati ‘to sign’
    - **POS:** V
    - **aspect**: perfective
    - **reflexivity**: non-reflexive
2. **morphological structure – surface layer:** pot-pis-a-ti

(*pot* = prefix, *pis* = root, *a* = derivational suffix, *ti* = inflectional suffix)

3. **morphological structure – deep layer:** pod-pis-a-ti

(*pod* = prefix, *pis* = root, *a* = derivational suffix, *ti* = inflectional suffix)

4. **word-formation pattern:** pod + pisati
5. **stem (allomorph of the stem)**: pisati (pisati)
6. **affix (allomorph of the affix)**: pod (pot)
7. **affix sense**: location: under
8. **word-formation process** (POS > POS): prefixation (V > V)
9. **link to the Croatian Language Portal**

CroDeriV 2.0 – structure of the entry – A

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

2. **morphological structure – surface layer:** bes-kraj-an-Ø

(*bes* = prefix, *kraj* = root, *an* = derivational suffix, *Ø* = inflectional suffix)

3. **morphological structure – deep layer:** bez-kraj-an-Ø

(*bez* = prefix, *kraj* = root, *an* = derivational suffix, *Ø* = inflectional suffix)

4. **word-formation pattern:** bez + kraj + an
5. **stem (allomorph of the stem)**: kraj (kraj)
6. **affix1 (allomorph of the affix1)**: bez (bes) **affix2 (allomorph of the affix2)**: an (an)
7. **affix1 sense**: deprivation **affix2 sense**: having the property of [meaning of the base]
8. **word-formation process** (POS > POS): simultaneous prefixation and suffixation (N > A)
9. **link to the Croatian Language Portal**

CroDeriV 2.0 – structure of the entry – C

1. **lemma:** brodograditelj
    - **POS:** N
    - **gender**: masculine

2. **morphological structure – surface layer:** brod-o-grad-i-telj-Ø

(*brod*, *grad* = root, *o* = interfix, *i, telj* = derivational suffixes, *Ø* = inflectional suffix)

3. **morphological structure – deep layer:** brod-o-grad-i-telj-Ø

(*brod*, *grad* = root, *o* = interfix, *i, telj* = derivational suffixes, *Ø* = inflectional suffix)

4. **word-formation pattern:** brod + o + graditi + telj

5. **stem (allomorph of the stem)**: brod (brod)|gradi (gradi)

6. **affix1 (allomorph of the affix1)**: i (i) **affix2 (allomorph of the affix2)**: telj (telj)

7. **affix1 sense**: verbal action **affix2 sense**: agent, masculine

8. **word-formation process** (POS > POS): simultaneous compounding and suffixation (N, V > N)

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

Demo

- http://193.198.214.203/root/let/

Concluding remarks

- CroDeriv 2.0
  - redesigned database
    - words of all major POS
      - compounds included!
    - morphological structure
    - word-formation patterns
    - derivational relations among Croatian lexemes
  - new visual design and online search interface – more attractive to users

Thank you!