CroComp – Lexicon of Croatian Compounds

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

Introduction

- In this work we focuss on compounds in Croatian and the development of the Lexicon of Croatian compounds - CroComp
- We deal with theoretical aspects of compounding in Croatian and their application in the creation of the Lexicon
  - how to analyze compounds and present their structure
  - how to organize lexical entries in the Lexicon
    - lexical entries in CroComp provide information on the individual elements of each compound, the word-formation pattern used in compounding, and the affixes used in the word-formation process
    - lexical entries also include information about the morphological structure of each compound

Croatian

- South Slavic language → rich inflectional and word-formational morphology
  - there are numerous morpho-phonological processes at morpheme boundaries and vowel alternations within the lexical morphemes (*ablaut*) → frequent allomorphy of affixes and stems
- the main word-formation processes → derivation and compounding
  - along with conversion, blending, acronym formation, and others
- inflection → suffixation
- word-formation → suffixation, prefixation, simultaneous suffixation and prefixation, and ablaut (usually with affixation)
  - compounding → an additional element - a linking vowel or interfix - is commonly used to connect the bases forming the compound

Computational processing of Croatian morphology

- Inflection:
  - several large lexica with paradigms and inflectional patterns (Tadić and Fulgosi, 2003; Ljubešić et al., 2016) used for various NLP tasks
- Two derivational databases:
  1. DerivBase.HR (Šnajder, 2016)
  2. CroDeriv (Šojat and Filko, 2023)
- Compounding → not processed so far
  - not productive as derivation in Croatian (particularly suffixation)
  - not productive as in some other languages (e.g. German)

Croderiv

- The lexical entries provide information about the morphological structure of words and their derivational links with other words
- Each lexeme is segmented into morphemes, the derivational stem used for the derivation is indicated, and the applied derivational process is specified
  - The morphological segmentation of the lexemes is based on a two-layered approach: segmentation at the surface and deep layer. Allomorphs are identified and marked for their type at the surface layer of the analysis → manual analysis
  - Allomorphs are connected to their representative morphs at the deep layer of presentation
    - The representative morph is the one from which other allomorphs can be established with the least number of morpho-phonological rules

Lexical entry in CroDeriv

LEMMA

nadopisivati

PART OF SPEECH

verb

MORPHOLOGICAL STRUCTURE - SURFACE LAYER

na - do - pis - iv - a - ti

MORPHOLOGICAL STRUCTURE - DEEP LAYER

na - do - pis - iv - a - ti

WORD-FORMATION PATTERN

nadopisati - ivati

WORD-FORMATION PROCESS

suffixation (verb > verb)

STEM

nadopis

Processing Compounds

- So far, only derivation has been processed in CroDeriv
  - One of the tasks in its further development is the expansion with compounds
  - For this purpose, a list of approximately 1,300 compound words was collected from monolingual dictionaries and corpora
- Word-formation of compounds
  - We analyze their structure in terms of the lexical stems and affixes used in compounding
  - We also analyze their morphological structure, segmenting the compounds into the morphemes they consist of
  - The results of the analysis are here presented in a computational lexicon designed to store and display morphological data of this kind
    - later to be merged with other derivational data for Croatian

Compounds – theoretical background

- The most comprehensive overview of word-formation in Croatian is given in Babić (2002).
- The author discusses the formation of compound nouns, adjectives, verbs, and adverbs and analyzes them according to two criteria:

1. according to the POS of the elements in the compounds, and
2. according to whether affixes are added in the compounding process

- Based on these criteria, the author determines the main types of compounding in Croatian:

Compounds - types

1. Proper compounds, also referred to as ’pure’ compounds – two words usually joined by an interfix, without any suffixation or prefixation
    - *romanopisac* (roman-o-pisac) ’novelist’
    - *ribolov* (rib-o-lov) ’fishing’
2. Suffixal compounds – combinations of two stems and a suffix
    - *ženomrzac* (žen-o-mrz-ac-) ’misogynist’
    - *častohlepan* (čast-o-hlep-an) ’ambitious, pushy’
3. Prefixal compounds – combinations of two stems and a prefix (or a prefix and a suffix) – only for verbal compounds
    - *omalovažiti* (o-malo-važiti) ’to belittle’
    - *odobrovoljiti* (o-dobr-o-volj-iti) ’to appease, to cheer up’

Fusions or coalescences (in Croatian ’sraslice’)

4. Conjoined words bearing all the inflectional markers from corresponding syntactic phrases, without affixation

- Such compounds are created by the fusion of phrases or firm collocations, usually when the meaning of both elements is unified and thus becomes independent
- The elements that appear in a compound word can appear in the same or reverse order in phrases
- Fusions are characterized by the absence of interfixes and suffixes
  - *hvale* ’praise’ + *vrijedan* ’worthy’ = *hvalevrijedan* ’praiseworthy’
    - from phrases *hvale vrijedan* or *vrijedan hvale*
  - *dan* ’day’ + *gubiti* ’waste, lose’ = *dangubiti* ’waste time’
    - from phrases *dan gubiti* or *gubiti dan*.

Semi-compounds

5. Semi-compounds – elements joined by a hyphen in writing, without any inflectional marker on the first element

- *bob-staza* ’bobsleigh track’
- *boks-meč* ’boxing match’
- *čarter-let* ’charter flight’
- *remek-djelo* ’masterpiece’
- *vagon-restoran* ’dining car’

and many others, nowadays mostly of English origin.

- Some authors consider the hyphen to be a type of interfix
- Regardless of the spelling norm, semi-compounds are frequently written without a hyphen
- For now, we do not include this type of compounds in CroComp. For now, we are not dealing with such compounds. Compounding types 1. – 4. are included in the current version of CroComp.

CroComp

- One of the tasks in the future development of CroDeriv is the expansion with compound words.
- For this purpose, we compiled an initial list of 1,300 compounds that belong to the main POS – nouns, adjectives, verbs, and adverbs.
- In the first step of the analysis we aimed to determine which words served as the bases for the formation of the compound.
- In the second step we aimed to determine whether affixes were used in the process.

CroComp

- Further, we aimed to identify which word-formation patterns are applied in the formation of compounds in Croatian.
- The word-formation pattern is determined on the basis of which POS are in the first and second positions in the compounds and which affixes are used in the compounding.
- We have identified the following POS that can appear as the first or second element in the compounds:
  - **1. element**: 1. noun, 2. adjective, 3. pronoun, 4. num, 5. verb, 6. participle, 7. adverb
  - **2. element**: 1. noun, 2. pronoun, 3. verb, 4. participle, 5. preposition, 6. numeral
    - The label *numeral* encompasses cardinal and ordinal numbers, as well as the so-called numeral nouns

CroComp

- So far, we have analyzed

  802 nouns

  484 adjectives

  22 verbs

  4 adverbs

  - which together make up a total of 1,312 compounds
- These data show that compounding is productive for nouns and adjectives and that there are patterns that allow the creation of new compound words
- Compound verbs and adverbs are very rare in Croatian, and new compounds in these POS are seldom formed in the contemporary language

Nouns

- Based on the combinations of members from the groups of the first and second elements, along with the addition of various interfixes and suffixes, we identified 28 different word-formation patterns for nouns.
- The noun formation patterns were grouped into six main types based on the first element in the compounds, that is, according to the part of speech to which the first element belongs.
  - ’Interfix’ or ’suffix’ in parentheses in main types means that in some patterns it is not realized.

Nouns

- **1. *noun*** *+ interfix + verb / noun / participle (+ suffix)* (524):
- noun + interfix + verb + suffix (350):
  - *pismonoša* (pism-o-noša) ’postman’; *nogomet* (nog-o-met) ’football’
- noun + interfix + noun (146):
  - *brodovlasnik* (brod-o-vlasnik) ’shipowner’; *romanopisac* (roman-o-pisac) ’novelist’
- noun + interfix + noun + suffix (23):
  - *hodočašće* (hod-o-čašće) ’pilgrimage’; *vukodlak* (vuk-o-dlak) ’werewolf’
- noun + interfix + participle + suffix (5):
  - *krvoproliće* (krv-o-proliće) ’bloodshed'

Nouns

- **2. *adjective*** *+ interfix + noun / verb (+ suffix)* (108):
- adjective + interfix + noun + suffix (61):
  - *osnovnoškolac* (osnovn-o-školac) ’elementary school student’
  - *dugoprugaš* (dug-o-prugaš) ’long-distance runner’
- adjective + interfix + noun (36):
  - *suhozid* (suh-o-zid) ’drywall’
  - *zločin* (zl-o-čin) ’crime’
- adjective + interfix + verb + suffix (11):
  - *mladoženja* (mlad-o-ženja) ’groom’
  - *svetogrđe* (svet-o-grđe) ’sacrilege'

Adjectives

- Using the same method as with nouns, we identified 27 different word-formation patterns for adjectives, which we grouped into six main types based on the word class found in the first part of the compound.

- ***adjective*** *+ interfix + noun / adjective / participle (+ suffix)* (219):
- adjective + interfix + noun + suffix (194):
  - *čistokrvan* (čist-o-krvan) ’purebred’; *tamnook* (tamn-o-ok) ’dark eyed’
- adjective + interfix + adjective (13):
  - *crnobijel* (crn-o-bijel) ’black and white’; *gluhonijem* (gluh-o-nijem) ’deaf-mute’
- adjective + interfix + noun (9):
  - *bjeloput* (bjel-o-put) ’pale-skinned’; *dugovrat* (dug-o-vrat) ’long-necked’
- adjective + interfix + participle + suffix (3):
  - *jedinorođeni* (jedin-o-rođeni) ’only-begotten’; *živorođeni* (živ-o-rođeni) ’live-born

Adjectives

- ***adverb*** *+ verb / noun / participle / adjective (+ suffix)* (61):
- adverb + adjective (30):
  - *visokoobrazovan* (visoko-obrazovan) ’highly educated’; *tamnocrven* (tamno-crven) ’dark red’
- • adverb + verb + suffix (16):
  - *dalekosežan* (daleko-sežan) ’far-reaching’; *brzoplet* (brzo-plet) ’rash, impulsive’
- adverb + noun + suffix (9):
  - *maloljetan* (malo-ljetan) ’underage’; *višeglasan* (više-glasan) ’polyphonic’
- adverb + participle (4):
  - *brzorastući* (brzo-rastući) ’fast-growing’; *dobrostojeći* (dobro-stojeći) ’well-off’
- adverb + noun (1):
  - *višečlan* (više-član) ’multi-member’
- adverb + participle + suffix (1):
  - *novorođeni* (novo-rođeni) ’newborn

Verbs

- Using the similar method as with nouns and adjectives, we identified 8 different word-formation patterns for verbs and grouped them into 4 main types.
- The criterion for establishing the main types is based on whether affixes are used in the compounding process, and if so, which type of affixes are applied.
  - The affixes used in compound verbs are prefixes, interfixes, and suffixes. The prefixes are *o-* and *u-*. The interfix is *-o-*. The suffix is *-iti (se)*.

Verbs

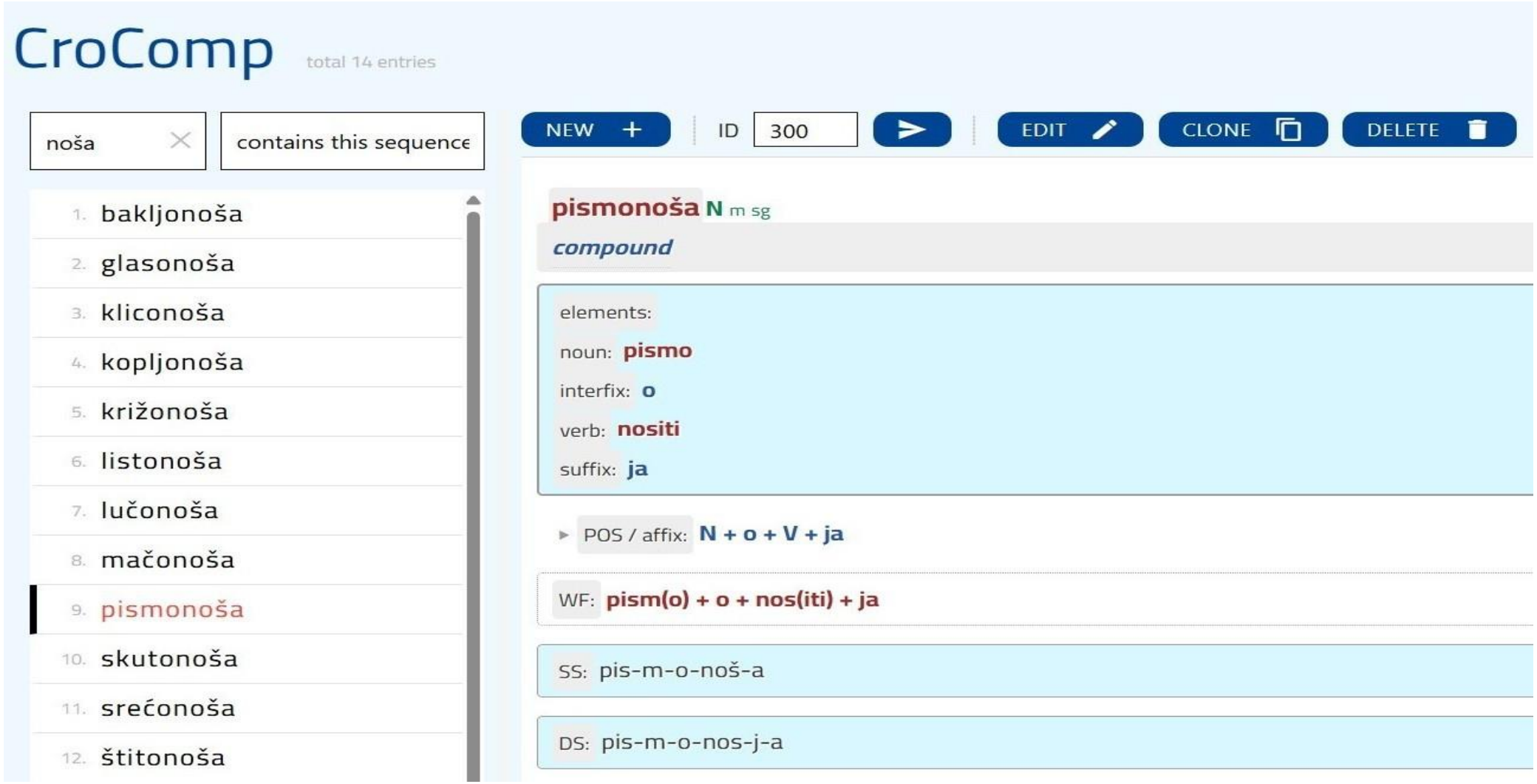
- 1. ***pure compounding***:
- adverb + verb
  - *krivotvoriti* (krivo-tvoriti) ’to forge’; *zlouporabiti* (zlo-uporabiti) ’to abuse’
- 2. ***prefixation + compounding***:
- prefix + adverb + verb
  - *odugovlačiti* (o-dugo-vlačiti) ’to procrastinate’; *omalovažiti* (o-malo-važiti) ’to belittle’
- 3. ***prefixation + compounding + suffixation***:
- prefix + adjective + interfix + noun + suffix
  - *odobrovoljiti* (o-dobr-o-voljiti) ’to appease’
- 4. ***compounding + suffixation***:
- num-card + interfix + noun + suffix + se
  - *dvoumiti se* (dv-o-umiti se) ’to be in doubt’

Adverbs

- So far, we have only 4 compound adverbs on our list. We have noted three word-formation patterns. The interfix is *-o-*, and the suffix is *-ke*.
- numeral + interfix + noun + suffix (2):
  - *četveronoške* (četver-o-noške) ’on all fours’; *dvonoške* (dvo-noške) ’on two legs’
- adverb + preposition (1):
  - *maloprije* (malo-prije) ’a short while ago’
- adverb + noun (1):
  - *sutradan* (sutra-dan) ’the next day’

- From the main types of word-formation patterns shown above, it is evident that in compounds where the adverb is the first element, we do not record an interfix.

Lexical entries in CroComp



Conclusion

- This paper presents the development and structure of CroComp, a computational lexicon of Croatian compound words.
- Building on previous work in Croatian derivational morphology, CroComp focuses specifically on compounding, as this type of word-formation has not been tackled within the scope of computational processing of Croatian morphology so far.
- The lexicon currently includes 1,312 compounds, categorized by part of speech and analyzed according to their morphological structure, including the identification of stems, affixes, and word-formation patterns.
- The classification of patterns was based on the parts of speech of the compound elements and the presence and type of affixes used in the compounding process.

Conclusion

- As expected, our analysis confirmed that compounding is most productive in the formation of nouns and adjectives, with a wide variety of patterns and affix combinations
  - verbal and adverbial compounds are significantly less frequent
  - coordinative compounds are significantly less frequent than subordinative
- The lexicon is compiled in the online dictionary writing system Lexonomy
- CroComp is publicly available and can be viewed at:

  https://lexonomy.zzl.ffzg.unizg.hr/twgi63yh

THANK YOU FOR YOUR ATTENTION!