Vanja Štefanec, Krešimir Šojat, and Matea Filko

CRODERIV – SEARCH AND VISUALIZATION INTERFACE

Abstract In this paper, we present the search and visualization interface of the Croatian derivational lexicon – CroDeriv. CroDeriv contains information on the derivational and morphological properties of Croatian lexemes. Each lemma in the lexicon is enriched with its word-formation analysis and morphological segmentation. The search interface enables simple and advanced queries, i.e., by lexemes, by morphological structures, and by word-formation patterns. Moreover, the visualization interface enables the graphical representation of derivational families.

Keywords CroDeriv; Croatian language; morphological analysis; word-formation; morphemes; derivational family

1. Introduction

In this paper, we present the search and visualization interface for CroDeriv[1], an online lexicon that contains data on the morphological structure and derivational relatedness of Croatian lexemes.

Although there are numerous language resources dealing with inflectional morphology for various languages, including Croatian (e.g., Croatian Morphological Lexicon (Tadić & Fulgosi, 2003); hrLex (Ljubešić et al., 2016)), language resources dealing with word-formation morphology are very scarce (see Kyjánek, 2018; Filko, 2020, pp. 109–117). The main difference between derivational resources developed for other languages and CroDeriv is that the former focus almost exclusively on the derivational relatedness of lexemes, disregarding the description of their morphological structure. CroDeriv is therefore a unique language resource on the morphological level since it comprises data on word formation as well as morphological structure. Each lexeme in CroDeriv is segmented into the morphemes it consists of, and each lexical entry also contains information about the derivational relatedness between itself and its derivational base. This kind of representation enables various search (and research) possibilities, which would not be feasible with other derivational resources mentioned above. To make various complex queries across the lexicon possible for a broader audience of users, we developed an intuitive and user-friendly search and visualization interface.

In the following section, we briefly present CroDeriv and the principles of its organization. In Section 3, we describe the search and visualization interface. The paper ends with the concluding remarks.

[1] https://croderiv.ffzg.unizg.hr/

2. CroDeriv

The development of CroDeriv took place in several phases. Its first version contained approximately 15,000 verbal lexemes collected from multiple online dictionaries. In this first phase of lexicon development, the primary goal was the analysis of the morphological structure of lexemes and the development of an appropriate data model that would enable queries over various morphological parameters across the lexicon. This approach proved valuable in many research areas, e.g., in the research of verbal aspect (Šojat et al., 2019), affixal meanings, affix ordering, combinations of particular affixes and roots as well as combinations of multiple affixes (Šojat et al., 2012; Šojat et al., 2013; Filko, 2020). However, the lexicon contained only lexemes of one part-of-speech (POS), and derivational links between them were not explicitly established. The morphological segmentation in this phase was performed semi-automatically using a naïve rule-based method, and the resulting segmentation was manually checked and corrected afterwards due to extensive allomorphy and phonological changes that take part at morpheme boundaries (e.g., assimilation or dropping) leading to a large error rate.

Morphological segmentation was based on the two-layered approach: the segmentation at the surface and the deep layer. Allomorphs were identified and marked for their type, i.e., root, prefix, suffix, or interfix, at the surface layer of analysis. Allomorphs were then associated with their representative morphs at the deep layer of presentation (e.g., *raščistiti* 'to clear off' < *raš-čist-i-ti* (surface layer) ~ *raz-čist-i-ti* (deep layer)). However, the segmentation in this first phase was shallow since lexemes were segmented to morphemes pertaining only to verbal derivational processes (verbal prefixes and suffixes), sometimes resulting in unsegmented complex stems. These unsegmented stems were morphological clusters inherited from non-verb derivational bases. To enable the grouping of lexemes into derivational families, i.e., ones that share the same lexical morpheme, "stems" were also associated with the root morpheme, albeit their content was often non-monomorphemic.

In the present stage, the data model has been completely revised in order to facilitate the expansion of the lexicon to other POS, mainly nouns and adjectives. This data model supports the full morphological decomposition of lexemes, regardless of their POS. At this stage, all data is analysed and segmented into morphemes manually, following the same two-layered approach. Besides, derivational links between lexemes are now explicitly established with identified stems (derivational bases), derivational affixes, and word-formation processes which take part in the derivation of a particular lexeme. Word-formation processes include prefixation, suffixation, simultaneous prefixation and suffixation, ablaut (usually combined with affixation), compounding (usually combined with affixation), etc. (see Filko et al., 2020, pp. 92–93; Šojat & Filko, 2023). Derivational affixes were also labelled for their morphosemantic meaning. Finally, lexemes were assigned their appropriate grammatical categories, i.e., aspect for verbs, gender for nouns, etc.

This design of the data model and the fact that CroDeriv now contains lexemes of all POS facilitates the representation of derivational families into graph-like structures. Derivational families are composed of lexemes that contain the same root, and in some cases, they can comprise more than 200 lexemes. The advantages of such representation authors have already described in Filko et al. (2021). These structures enable the visual representation of derivational families and derivational paths from the lexical morpheme to a certain lexeme and thus make the Croatian derivational morphology more understandable and accessible to the general public.

3. Search and Visualization Interface

CroDeriv data is available to the users through a web application to be used in an internet browser. The architecture of the back-end system has already been described in Filko et al. (2020). Here we shall focus on the front-end interface, built as a Vue.js/Nuxt application. The interface is a part of a CroDeriv project website and it is designed to be intuitive and user-friendly, without the need for any specific usage instructions in the form of user documentation. In this way, we wanted to enable users who may not be technically savvy to also use the lexicon. However, all the features of the interface are described in place through thoughtfully positioned tooltips which contain enough information for users to be able to start using the lexicon right away.

Although we have constructed a special query language for searching the lexicon via API, the search interface eliminates the need for users to learn that language. Instead, the front-end application seamlessly translates users' input into the internal query language.

The search interface enables the user to search the CroDeriv lexicon in several ways: by lexemes, by morphological structures, and by word-formation patterns.

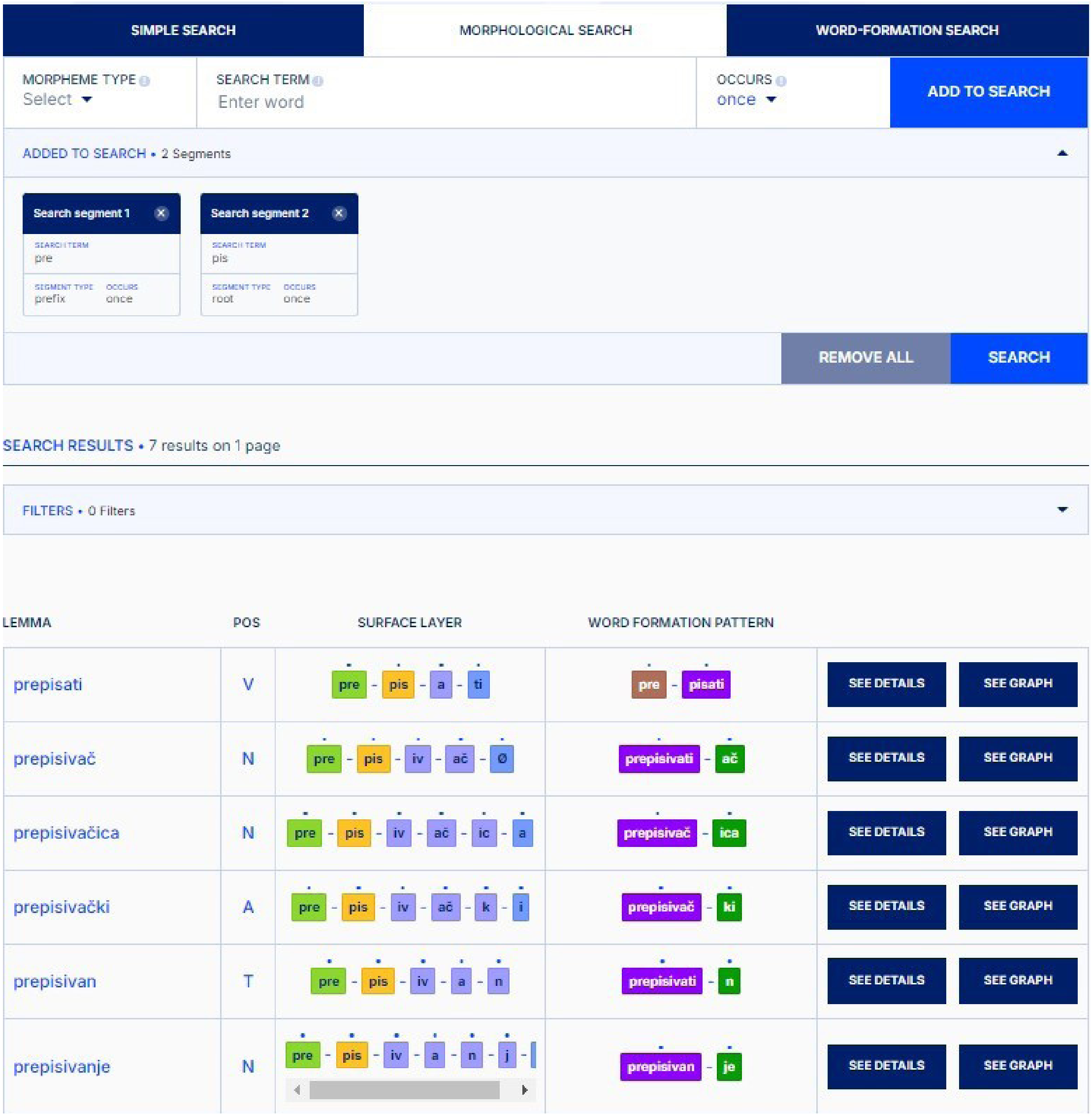
The simplest search is the one by the lexeme, in which the user can search for a particular lexeme or lexemes using the literal text input, or by using POSIX extended regular expressions (e.g., *rad.** will return all lexemes starting in *rad*, regardless of the morphological status of '*rad*' in their morphological structure: *rad* 'work', *raditi* 'to work', *rado* 'gladly', *radost* 'joy'...).

The search by morphological structure is facilitated by the visual search query builder. In this way, the user builds the morphological structure to be searched for against the lexicon database. The query builder allows the user to concatenate the search blocks representing morphemes into a sequence. With the possibility of using regular expressions along with the literal text inputs, the user can create very complex morphological queries, which is a feature that, to our best knowledge, makes this resource unique. For example, the user can search for all single or double prefixed lexemes, within a particular derivational family or across all families, or all lexemes with a particular suffix, or all compound lexemes, or any combination of these conditions. In Figure 1, the results of the search for all lexemes containing the prefix *pre-* followed by the root *√pis* are presented.

The word-formation search functions in a similar way as the morphological search, only with the difference that the building blocks of the search query represent the lexeme's derivational formants, i.e., stems and derivational affixes. This search is especially useful in searching for the same derivational patterns across derivational families or investigating the distribution of polysemous affixes.

All successful searches result in a list of lexemes in a tabular form that satisfy the search conditions (see Figure 1). The result lexeme list obtained by any of the ways of searching, can be additionally filtered by POS and relevant grammatical features. Additional properties associated with every result in this tabular view are lemma, part-of-speech, surface morphological structure, and word-formation pattern. For a more detailed description of every lexeme, the user can click on the link showing a detail page of the lexeme with a complete list of properties, which include deep morphological structure, word-formation process, and senses, i.e., morphosemantic meanings, of the affixes used in a derivation of a particular lexeme.

By clicking on the link for the graph, the user can get the visualization of the derivational path to the lexeme from its lexical root. The visualization is represented in the form of a labelled directed graph in which the nodes represent the lexemes in the derivational path and the edges represent the derivation links. Nodes are labelled with the lemma of the lexeme and the colour of the node denotes its POS. Derivational links are labelled with the name of the word-formation process as well as the sense of the derivational affix.



MORPHEME TYPE: Select

SEARCH TERM: Enter word

OCCURS: once

ADDED TO SEARCH • 2 Segments

Search segment 1
SEARCH TERM: pre
SEGMENT TYPE: prefix
OCCURS: once

Search segment 2
SEARCH TERM: pis
SEGMENT TYPE: root
OCCURS: once

SEARCH RESULTS • 7 results on 1 page

FILTERS • 0 Filters

LEMMA	POS	SURFACE LAYER	WORD FORMATION PATTERN		
prepisati	V	pre - pis - a - ti	pre - pisati	SEE DETAILS	SEE GRAPH
prepisivač	N	pre - pis - iv - ač - Ø	prepisivati - ač	SEE DETAILS	SEE GRAPH
prepisivačica	N	pre - pis - iv - ač - ic - a	prepisivač - ica	SEE DETAILS	SEE GRAPH
prepisivački	A	pre - pis - iv - ač - k - i	prepisivač - ki	SEE DETAILS	SEE GRAPH
prepisivan	T	pre - pis - iv - a - n	prepisivati - n	SEE DETAILS	SEE GRAPH
prepisivanje	N	pre - pis - iv - a - n - j -	prepisivan - je	SEE DETAILS	SEE GRAPH

Fig. 1: Example of the search by morphological structure

4. Concluding Remarks

In this paper, we have presented the search and visualization interface for the Croatian morphological and derivational lexicon CroDeriv. This interface was built to make CroDeriv data available to a broad audience of users, from experts in the field, to students, Croatian language learners, and the general public. Most of the available morphological and derivational resources require a high level of technical expertise since they either have no user interface or utilize complex query languages. Our motivation was to leverage the advantages of a well-thought-out UX design to build an intuitive and easy-to-use, yet powerful interface. We believe that this interface can find its place both as a research platform, as well as a tool for teaching morphology in schools and universities or Croatian as a second and foreign language.

Acknowledgments

This paper and the work presented are partially supported by the Croatian National Consortium of CLARIN ERIC Research Infrastructure (HR-CLARIN).

References

Filko, M. (2020). *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika.* [Doctoral dissertation, Faculty of Humanities and Social Sciences, University of Zagreb]. ODRAZ.

Filko, M., Šojat, K., & Štefanec, V. (2020). The Design of CroDeriv 2.0. *The Prague Bulletin of Mathematical Linguistics*, 115, 83–104. doi: 10.14712/00326585.006

Filko, M., Šojat, K., & Štefanec, V. (2021). Deriving the Graph: Using Affixal Senses for Building Semantic Graphs. In F. Namer et al. (Eds.), *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)* (pp. 120–128). ATILF (CNRS & UNIVERSITÉ DE LORRAINE).

Kyánek, L. (2018). *Morphological resources of derivational word-formation relations.* [Technical report 61; ÚFAL, Charles University, Prague].

Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4264–4270). ELRA.

Šojat, K., Srebačić, M., & Tadić, M. (2012). Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling, 00*(1), 111–142.

Šojat, K., Srebačić, M., & Štefanec, V. (2013). CroDeriV i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika*, *39*(75), 75–96.

Šojat, K., Kocijan, K., & Filko, M. (2019). Processing Croatian Aspectual Derivatives. In I. Mauro Mirto, M. Monteleone, & M. Silberztein (Eds.), *Formalizing Natural Languages with*

NooJ 2018 and Its Natural Language Processing Applications (pp. 50–61). Springer International Publishing.

Šojat, K., & Filko, M. (2023). Processing Croatian Morphology: Roots, Segmentation and Derivational Families. In K. Šojat, & M. Filko (Eds.), *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology (DeriMo2023)* (pp. 61–70). Croatian Language Technologies Society.

Tadić, M., & Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In T. Erjavec, & D. Vitas (Eds.), *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* (pp. 41–46). Association for Computational Linguistics.

Contact information

Vanja Štefanec
Faculty of Humanities and Social Sciences, University of Zagreb
vstefane@ffzg.unizg.hr

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

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