

Deriving the Graph: Using Affixal Senses for Building Semantic Graphs

Matea Filko

Faculty of Humanities
and Social Sciences
University of Zagreb
matea.filko@ffzg.hr

Krešimir Šojat

Faculty of Humanities
and Social Sciences
University of Zagreb
ksojat@ffzg.hr

Vanja Štefanec

Faculty of Humanities
and Social Sciences
University of Zagreb
vstefane@ffzg.hr

Abstract

In this paper, we will present the semantic graphs as a new way of exploring semantic relationships in CroDeriv. CroDeriv is a morphological database developed for the Croatian language. In previous phases of its development words were segmented into morphemes and derivational links among the base word and the derivative were marked. Currently, we focus on the analysis of affixal meanings. This analysis is the basis for the production of semantic graphs. Semantic graphs are used to capture semantic similarities within various derivational families.

1 Introduction

Language resources dealing with derivational morphology are nowadays being developed for numerous languages (Kyjánek, 2018; Filko, 2020). However, these resources differ significantly in terms of the way they are composed and the type of data they contain. Although the analysis of derivational processes is inextricably linked to semantics, semantic description is usually out of focus in the first stages in the building of such resources due to its complexity and frequent unpredictability. In the existing derivational resources, mainly regular semantic relations between derivationally connected words are marked. For example, Démonette (Hathout and Namer, 2014) provides semantic characteristics of words and indicates whether they denote an action, an agent or a property. The authors of Derivancze (Pala and Šmerk, 2015) mark the semantic type of derivational relations. However, only those relations that have regular and transparent meanings are identified, such as the relation between the action and the agent of the action or an adjective and the properties of the attributes marked by the adjective. Semantic relations have also been added to DeriNet, as the latest phase of its expansion. Ševčíková and Kyjánek (2019) describe a semi-automatic procedure for assigning semantic relationships to units in DeriNet. The semantic relations they have added are in line with the relations listed by Bagasheva (2017), which allows for a later comparison in various languages. At this initial stage, the authors focus on five semantic categories: diminutive and female (for nouns), possessive (for adjectives), and iterative and aspect (for verbs).

In this paper, we present the first steps in the marking of semantic categories of affixes in CroDeriv.¹ The development of CroDeriv in the first phases focused on a complete morphological and derivational analysis of verbs (Šojat et al., 2013), nouns (Šojat et al., 2014; Filko, 2020) and adjectives (Filko and Šojat, 2017). This means that the lexemes were segmented into morphs at the surface layer and all morphs were connected to the corresponding morphemes at the deep layer of presentation (for example, *učiteljica* ‘female teacher’: *uč-i-telj-ic-a* (surface layer) – *uk-i-telj-ic-a* (deep layer)). These procedure is done manually for the approximately 14 000 verbs, 1 500 adjectives and 5 500 nouns, due to low precision rates of the automatic procedures (Šojat et al., 2014).

In the next phase, the starting word in the derivational process was marked, as well as the type of the derivational process that was used for the derivation of particular derivatives (*učitelj* ‘male teacher’ + *-ica* → *učiteljica* ‘female teacher’). Besides, it was indicated whether this is a derivational process that changes the part-of-speech category of derivatives or not (Filko et al., 2020). This procedure is also done

¹CroDeriv is available at croderiv.ffzg.hr.

manually for the subset of verbs (with more than 5 representatives in the derivational family), and for the nouns and the adjectives in the database.

Semantic categories are marked on derivational affixes (see Section 2). Labelling of these categories in graphs within derivational families provides an insight into semantic processes taking part in derivation. Same semantic processes can be realized by various means of derivation and within different derivational families. For example, the derivational semantic pattern **action** → **agent** → **female agent** is realized within the derivational family of the root *uk-* as:

učiti ‘to teach’ → *učitelj* ‘male teacher’ → *učiteljica* ‘female teacher’,

but also within the derivational family of the root *voz-* as:

voziti ‘to drive’ → *vozač* ‘male driver’ → *vozačica* ‘female driver’

or within the derivational family of the root *da-* as:

izdavati ‘to publish’ → *izdavač* ‘male publisher’ → *izdavačica* ‘female publisher’.

The aim of such a procedure is to establish semantic paths, i.e. regular semantic shifts / patterns in the derivation, in addition to the derivational paths. This kind of information is essential for the description of the Croatian morphotactics, which is still an under-investigated area of Croatian linguistics (Filko, 2020).

The paper is structured as follows: in the next section we will explain the principles of assigning semantic categories to affixes. In Section 3, we will describe how affixal meanings are encoded in CroDeriv. In Section 4, the semantic graphs will be introduced. We will present how the semantic graphs can reveal semantic properties of derivational families on the one hand, and particular affixes on the other. We will conclude with final remarks.

2 Affixal meanings

Two basic approaches to affixal meanings differ depending on whether they place emphasis on the process of homonimization or on the process of polysemization of affixes. The process of homonimization splits affixes into two or more separate units, while the polysemization reinterprets homonyms as a single unit (Raffaelli, 2015, 187). As a consequence, homonimization multiplies the number of units, while polysemy results in the meaning networks of particular units, in which it is possible to detect the links between different affixal meanings.² We believe that speakers recognize these links between meanings and that this enables them to use vocabulary economically and effectively. In addition, we believe that such an approach is methodologically more justified because it does not multiply the number of units. Thus, we consider suffixes as polysemous units, i.e. in the analysis we give preference to polysemy rather than homonymy. This approach is well described and substantiated in the reference literature (Rainer, 2014; Aronoff and Fudeman, 2011; Lieber, 2004; Lehrer, 2003; Babić, 2002). As indicated by Filko (2020), this approach is in line with the cognitive-semantic view that polysemy is linguistically and cognitively more economical than homonymy (Raffaelli, 2015).

We determine the meanings on the basis of the synchronic semantic analysis, combining the semasiological and onomasiological approaches at the same time (Bagasheva, 2017). The semasiological approach is manifested in the analysis of the polysemous structures of individual affixes, e.g., nominal suffix *-ba* can have five meanings in its meaning network:

1. action (*ploviti* ‘to sail’ → *plovid-ba* ‘sailing’)
2. result (*skladati* ‘to compose’ → *sklad-ba* ‘composition’)
3. event (*svat* ‘wedding guest’ → *svad-ba* ‘wedding’)
4. location (*nastaniti* ‘to dwell’ → *nastam-ba* ‘dwelling’)
5. non-transparent meaning (*opor* ‘harsh’ → *opor-ba* ‘political opposition’) (Filko, 2020, 172-173).

²In the example of the Croatian suffix *-ba* below, we can notice several cognitive metonymies leading to the diversification of its meaning network, e.g. ACTION FOR RESULT, ACTION FOR PLACE OF ACTION. Numerous types of regular polysemy patterns are recognized in Apresjan (1974).

This example also shows that suffixal meanings can be directly recognized and determined when dealing with semantically transparent motivated words (examples 1-4). On the other hand, semantically non-transparent derivatives and suffixes used in their derivation are treated as a separate group. This holds for derivatives for which it is not possible to clearly determine which part of their meaning is shaped on the basis of the meaning of the suffix (example 5).

The onomasiological approach, however, is used for the determination of the affixes used in the formation of the same semantic categories, e. g. the meaning ‘property, quality’ can be expressed by at least three different nominal suffixes³:

1. -ina (*bijel* ‘white’ → *bjelina* ‘whiteness’)
2. -oća (*slijep* ‘blind’ → *sljepoća* ‘blindness’)
3. -ost (*slab* ‘weak’ → *slabost* ‘weakness’) (Filko, 2020, 191).

It is important to emphasize that, when determining the meaning of affixes, we distinguish the lexical meaning and the derivational meaning of words (Rainer, 2005; Babić, 2002). The lexical meaning is regularly much more complex than the derivational meaning. The derivational meaning enables us to determine the type of semantic shifts between words used as stems for adding affixes and derivatives. In these processes, affixes play a very important role. The meaning of affixes is determined within derivational patterns, that is, when determining the meaning of the affix, it is necessary to observe the relationship between the meaning of stems and derivatives (cf. also Bauer and Valera, 2015). For example, Croatian noun *stolar* ‘carpenter’ is derived from the word *stol* ‘table’ and the suffix *-ar*. Thus, the derivational meaning of the noun *stolar* would be ‘the one who makes/produces tables’, and the semantic shift from the stem to the derivative is the one of ‘male agent’. We can conclude that this particular meaning is actually the meaning provided by the suffix *-ar*. This conclusion is further supported by other words in which the meaning ‘male agent’ is mediated by the suffix *-ar*:

kuhar ‘cook’ (← *kuhati* ‘to cook’)
kipar ‘sculptor’ (← *kip* ‘sculpture’)
mesar ‘butcher’ (← *meso* ‘meat’)
ribar ‘fisherman’ (← *riba* ‘fish’)...

Although the lexical and the derivational meaning can be the same, as in *kuhar* ‘the one who cooks’, sometimes they differ in a way that lexical meaning becomes more diversified than derivational meaning. This is the case with the above mentioned noun *stolar*. It has broadened its meaning to denote male agents who produce any kind of wooden furniture, window frames or doors, not only tables. However, the meaning relation between the stem and the derivative is revealed through the derivational meaning. Thus, we have to take the derivational meaning into account when determining derivational patterns and semantic shifts within them, because this particular shift is usually mediated by the affix.

Additionally, the affixal meaning is determined according to only one of possible meanings of polysemous lexemes. For example, the word *upravljač* can denote both an agent ‘controller’ and an object ‘control device’. However, since their morphological and derivational properties are the same, we didn’t want to multiply number of entries in CroDeriv at this point. Thus, we mark only the most prominent meaning (i.e. the meaning that is listed first) as indicated by the extensive online dictionary of the Croatian language available at Hrvatski jezični portal (hjp.znanje.hr), and in the example above, only the meaning of an object will be marked. Although this approach could be criticized, we believe that this decision is methodologically justified when manually annotating the meaning of affixes in several thousands of words.⁴

Related to the principle of determining affixal meanings according to the one meaning of the polysemous lexeme, is the principle of distinguishing the meaning of the suffix from the meaning of derivatives. We believe that suffixal meanings lie between the meaning of stems and derivatives. Although the

³Only 20 most productive nominal suffixes were semantically analyzed at this phase of the research.

⁴It is possible that in the next phases of CroDeriv development lexical units will be analyzed for their polysemous structure. However, this information is already available in Croatian dictionaries, so we currently focus on the data which has not been available so far.

meaning of derivatives is extensively discussed in Croatian grammar books, general semantic categories such as agent, means or location are mediated by suffixes and therefore require more attention (see the examples with the suffix *-ar* above).

We distinguish suffixes that 1) change the part-of-speech category of stems, 2) modify the meaning of stems without changing their part-of-speech category and 3) modify the meaning of stems and change their part-of-speech category. Some suffixes only change the part-of-speech category of stems without any alternation or modification of their meaning. We treat these suffixes as the category of substantivizing, adjectivizing etc. suffixes. The prototypical example of the substantivizing suffix is the suffix *-je*, predominantly used for the derivation of gerunds with the meaning 'action', although its polysemous structure is more complex. Suffixes that modify the meaning of stem words without changing their part-of-speech category are the suffixes that are, for example, used for the derivation of nouns with marked meanings (diminutives, pejoratives, augmentatives) or suffixes used for the derivation of feminine/masculine pairs. The most complex role is played by suffixes that both change the meaning and the part-of-speech category of stems. Such is, for example, a suffix *-ač*, as in *bacač* 'thrower' ← *baciti* 'to throw'. In this derivational process the part-of-speech category was changed from the verb to the noun, and the meaning from 'action' into 'agent'.

Suffixal meanings are determined in respect to generalized semantic categories, as described in Filko (2020) for nouns, Šojat et al. (2012) for verbs and Filko and Šojat (2017); Bagasheva (2017) for adjectives. Generalized semantic categories are recognized as fundamental by numerous linguists dealing with affixal meanings. The most extensive list of affixal meanings is the list presented in Bagasheva (2017). The semantic categories in her list are determined for various language families, mainly for those in Europe. Generalized semantic categories are additionally specified and divided into subcategories when it is important to determine semantic differences among suffixes (see Šojat et al., 2012).

3 Affixal meanings in CroDeriv

As extensively described in Filko et al. (2020), CroDeriv data model enables lexemes to be explicitly annotated with three layers of description, i.e. morphological, word-formation, and compounding-derivational description. Word-formation description consists of sequence of clusters, which are multi-morphemic units corresponding to notions of stems and derivational affixes. Since they are associated with morphemes they consist of, clusters also serve as a link between morphological and word-formation layer of description.

Similar to morphemes, clusters are stored as independent units in CroDeriv database. They can be of different types so we distinguish stems, and prefixing and suffixing formants, i.e. affixal clusters, which roughly correspond to notions of stems, derivational prefixes and derivational suffixes from morphological theory, respectively. Clusters, but affixal clusters in particular, can be associated with multiple meanings due to polysemic nature of derivational affixes, as already elaborated in Section 2. Instances of clusters, which make up the word-formation description of a particular lexeme, have their associated meanings reduced to one, i.e. the one which is realized in that lexeme.

Compounding segments are objects that form the third layer of description, the compounding-derivational layer. They are composed of sequences of clusters in which there can be only one stem cluster, and one or more affixal clusters. Compounding segments serve as child members in derivational relations, thus allowing compound words to be a part of multiple derivational families, e.g. the lexeme *naredbodavac* 'commander' [lit. 'command-giver'] is a part of two derivational families, *red-* and *da-*. This is reflected in the fact that it is composed of two compounding segments: *naredbo* and *davac*. First compounding segment consists of the stem cluster *naredb*, and the interfixal cluster *o*. Second compounding cluster consists of the stem cluster *dav*, and the suffixal cluster *ac*. Because of their association with clusters on word-formation level, compounding segments also serve as a link between word-formation and compounding-derivational layer of description.

Additionally, we can identify one more, the fourth, lexical, layer on which grammatical information is added to the entire lexeme.

Figure 1 on the example of the word *mrtvačnica* 'mortuary' (← *mrtvac* 'dead'_N) in a simplified manner

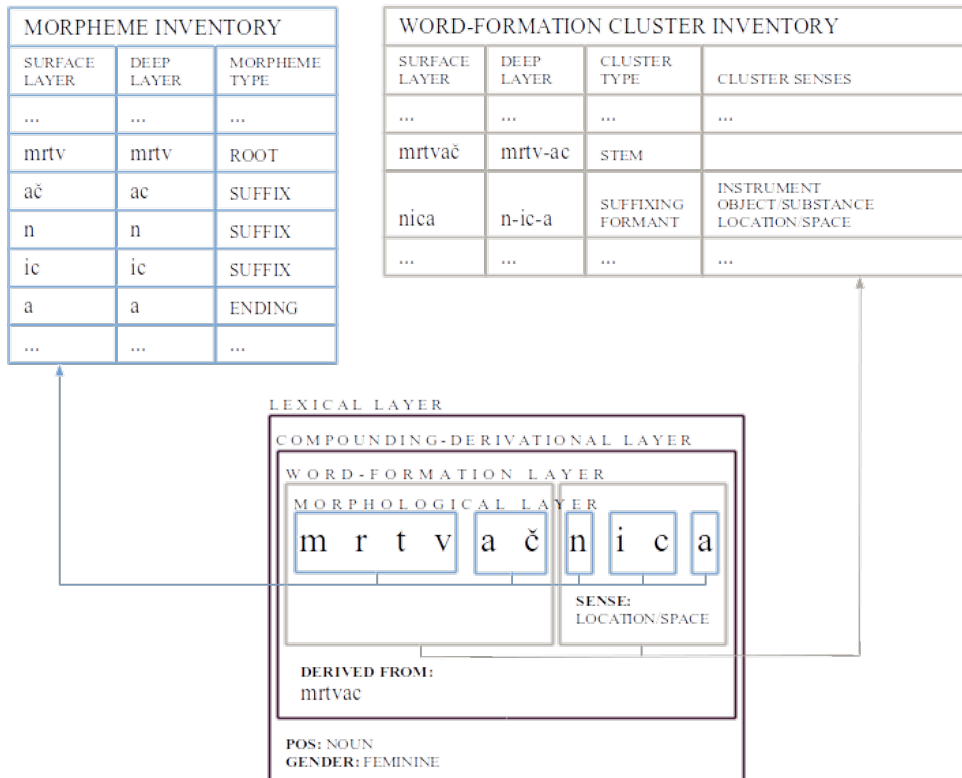


Figure 1: Schematic illustration of how CroDeriv data is organized

schematically illustrates the interconnections between the layers of description in CroDeriv, as well as the type of morphological information included in each of the layers.

The innermost set of borders around groups of letters in the center of the picture represents the morphological layer of description. These borders split the word in allomorphs which are then associated with the allomorphs and morphemes from the morpheme inventory. One layer up is the word-formation layer in which the borders enclose the word-formation units, i.e. “alloclusters”. Similarly, they are associated with the “alloclusters” and clusters from the cluster inventory. At this level, affixal clusters are marked for their sense. The next layer is the compounding-derivational layer which, besides splitting the elements of a potential compound, contains information about the base word in the derivational process. The outermost layer is the lexical layer which adds lexeme’s grammatical information.

As it can be seen from the illustration, in our annotation scheme layers are independent one from another, i.e. annotation on one layer can exist without the annotation on the other, but also higher layers have access to annotation on lower layers which makes the layers interconnected.

Up to this point, ca. 6500 lexemes in our database have been manually analyzed and annotated for word-formation, out of which, sense of the derivational affix was annotated for ca. 3000 lexemes. All publishing-ready CroDeriv data will be made available through regular contributions to the Universal Derivations dataset.

4 Semantic graphs

The interconnection of different layers of description in CroDeriv facilitates exploring other, not explicitly annotated, phenomena that occur at the intersections of these layers. One of them is derivational semantics which can be studied with the help of semantic graphs. The CroDeriv data model allows for structuring the derivational families in two graph representations. One we shall refer to as

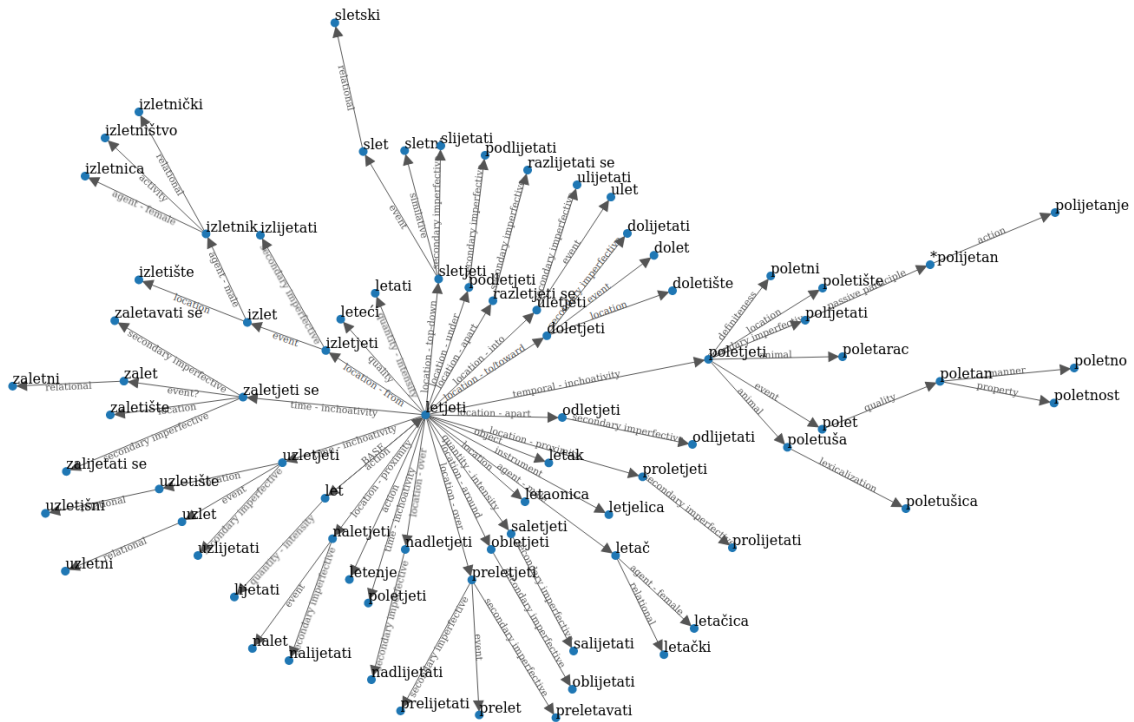


Figure 2: Lexeme-semantic representation of derivational family *let-*.

lexeme-semantic representation (Figure 2), and the other **structure-semantic representation** (Figure 3).

The **lexeme-semantic representation** is a directed acyclic graph with labeled nodes representing lexemes, and labeled edges representing derivational meaning of the lexeme they connect to. This representation is essentially a derivational graph with semantic labels attached to edges and is more-or-less in line with graphs created by Ševčíková and Kyjánek (2019) for Czech. In addition to providing a more informative representation of derivational families, these graphs also show the semantic motivation for the expansion of the derivational tree.

In the Figure 1 example, combining information from the word-formation and compounding-derivational layer enables labeling of the derivational link between the base word and the derivative as

$$mrtvac \xrightarrow{\text{LOCATION/SPACE}} mrtva\check{n}ica.$$

The **structure-semantic representation** is also a directed acyclic graph with semantic labels attached to edges, but here the nodes are labeled with derivational affixes involved in the last derivational step of a particular lexeme.

In the Figure 1 example, combining information from the word-formation layer of both base word and the derivative, and from the compounding-derivational layer of the derivative, reveals the derivational mechanism

$$-ac \xrightarrow{\text{LOCATION/SPACE}} -nica.$$

Structure-semantic graphs, as derivational generalizations of some sort, show derivational mechanisms used for semantic build-up. Having derivational trees represented in such way, by means of approximate graph matching algorithms, this directly allows for 1) quantifying the derivational similarity between derivational trees, 2) exploring the distribution of particular affixes involved in certain semantic change and vice-versa, and 3) calculating causal distribution of derivational mechanisms. The approximate graph

- Mark Aronoff and Kirsten Anne Fudeman. 2011. *What is morphology?*. Fundamentals of linguistics. Wiley-Blackwell, Chichester, West Sussex, U.K. ; Malden, MA, 2nd ed edition. OCLC: ocn608491860.
- Stjepan Babić. 2002. *Tvorba riječi u hrvatskome književnome jeziku*. Number knjiga 2 in Velika hrvatska gramatika. Hrvatska akademija znanosti i umjetnosti : Nakladni zavod Globus, Zagreb, 3., poboljšano izdanje edition. OCLC: ocm53895583.
- Alexandra Bagasheva. 2017. Comparative semantics concepts in affixation. In Juan Santana-Lario and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, Peter Lang, Bern : Berlin : Bruxelles : Frankfurt am Main : New York : Oxford : Wien, Linguistic Insights, pages 33–66.
- Laurie Bauer and Salvador Valera. 2015. Sense Inheritance in English Word-Formation. In Laurie Bauer, Lívia Körtvélyessy, and Pavol Štekauer, editors, *Semantics of Complex Words*, Springer, Heidelberg : New York : Dordrecht : London, number 3 in Studies in Morphology, pages 67–84.
- Matea Filko. 2020. *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika (Intralexical and Interlexical Structures of the Nominal Part of the Croatian Lexicon)*. Phd thesis, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb.
- Matea Filko and Krešimir Šojat. 2017. Expansion of the Derivational Database for Croatian. In Eleonora Litta and Marco Passarotti, editors, *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*. EDUCatt, Milan, pages 27–37.
- Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2020. [The Design of CroDeriv 2.0](https://doi.org/10.14712/00326585.006). *The Prague Bulletin of Mathematical Linguistics* 115:83–104. <https://doi.org/10.14712/00326585.006>.
- Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5):125–168.
- Lukáš Kyjánek. 2018. *Morphological resources of derivational word-formation relations*. Technical Report 61, ÚFAL - Charles University, Prague.
- Adrienne Lehrer. 2003. Polysemy in derivational affixes. In Brigitte Nerlich, Zazie Todd, Vimala Herman, and David C. Clarke, editors, *Polysemy. Flexible Patterns of Meaning in Mind and Language*, De Gruyter Mouton, Berlin : New York, pages 218–232.
- Rochelle Lieber. 2004. *Morphology and lexical semantics*. Cambridge University Press, New York.
- Lihui Liu, Boxin Du, Jiejun xu, and Hanghang Tong. 2019. [G-Finder: Approximate Attributed Subgraph Matching](https://doi.org/10.1109/BigData47090.2019.9006525). In *2019 IEEE International Conference on Big Data (Big Data)*. pages 513–522. <https://doi.org/10.1109/BigData47090.2019.9006525>.
- Karel Pala and Pavel Šmerk. 2015. [Derivancze — Derivational Analyzer of Czech](https://doi.org/10.1007/978-3-319-24033-6_58). In Pavel Král and Václav Matoušek, editors, *Text, Speech, and Dialogue: 18th International Conference, TSD 2015*. Springer, Berlin: Heidelberg, pages 515–523. https://doi.org/10.1007/978-3-319-24033-6_58.
- Ida Raffaelli. 2015. *O značenju: uvod u semantiku*. Biblioteka Theoria / Matica hrvatska Novi niz. Matica hrvatska, Zagreb.
- Franz Rainer. 2005. Semantic change in word formation. *Linguistics* 42(2):415–441.
- Franz Rainer. 2014. Polysemy in Derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford University Press, Oxford, Oxford Handbooks in Linguistics, pages 338–353.
- Yuanyuan Tian, Richard C. McEachin, Carlos Santos, David J. States, and Jignesh M. Patel. 2007. [SAGA: a subgraph matching tool for biological graphs](https://doi.org/10.1093/bioinformatics/btl571). *Bioinformatics* 23(2):232–239. <https://doi.org/10.1093/bioinformatics/btl571>.
- Yuanyuan Tian and Jignesh M. Patel. 2008. [TALE: A Tool for Approximate Large Graph Matching](https://doi.org/10.1109/ICDE.2008.4497505). In *2008 IEEE 24th International Conference on Data Engineering*. pages 963–972. ISSN: 2375-026X. <https://doi.org/10.1109/ICDE.2008.4497505>.
- Magda Ševčíková and Lukáš Kyjánek. 2019. Introducing Semantic Labels into the DeriNet Network. *Journal of Linguistics* 70(2):412–423.
- Krešimir Šojat, Matea Srebačić, and Tin Pavelić. 2014. [CroDeriv 2.0: Initial Experiments](https://doi.org/10.1007/978-3-319-10888-9_3). In Adam Przepiórkowski and Maciej Ogrodniczuk, editors, *Advances in Natural Language Processing*, Springer International Publishing, Cham, volume 8686, pages 27–33. https://doi.org/10.1007/978-3-319-10888-9_3.

Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling* 0(1):111. <https://doi.org/10.15398/jlm.v0i1.34>.

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