

**Proceedings of the Fourth
International Workshop on
Resources and Tools for
Derivational Morphology**

DeriMo 2023

Dubrovnik, 5-6 October 2023

Edited by:

Matea Filko and Krešimir Šojat

DeriMo 2023

Proceedings of the

Fourth International Workshop

on Resources and Tools

for Derivational Morphology

Editors:

Matea Filko
Krešimir Šojat

5-6 October 2023
Centre for Advanced Academic Studies,
Dubrovnik, Croatia

<https://derimo.ffzg.unizg.hr/>

Copyright © 2023 by the individual authors. All rights reserved.

Published by:

Hrvatsko društvo za jezične tehnologije
(Croatian Language Technology Society)
Ivana Lučića 3
10 000 Zagreb
Croatia

ISBN 978-953-55375-5-7

Preface

This volume contains papers accepted for presentation at DeriMo 2023: The Fourth International Workshop on Resources and Tools for Derivational Morphology, held in Dubrovnik, Croatia, on 5-6 October 2023. The 2023 edition of the workshop continues the discussion on language resources and tools for derivational morphology (and, more generally, on word formation) started in DeriMo 2017 (Milan, Italy) and continued in DeriMo 2019 (Prague, Czechia) and DeriMo 2021 (Nancy, France, held online due to pandemics).

The submission and reviewing processes have been handled by the EasyChair system. The proceedings contain 10 papers selected according to the reviews. In addition, the proceedings include the contribution of the invited speaker, Stela Manova, as an opening paper.

In her contribution, Stela Manova presents a different, form-based view on processing derivational morphology based on a ChatGPT perspective on language, i.e. she tries to demonstrate that the model of derivational morphology that relies on form is less complex than a model that relies on semantics. To prove this point, she conducted a psycholinguistic experiment among 64 native speakers of Polish and 45 native speakers of English. The results demonstrate that the speakers indeed do not need semantics to recognize possible affix combinations in their language, but they rather memorize them in the form of bigrams and trigrams, i.e. the subword units longer than a morpheme. Based on these results, she advocates for the construction of form-focused derivational resources providing information about word structure in terms of bigrams and trigrams of morphemes.

Uliana Petrunina and Hana Filip focus on the automatic detection of the grammatical aspect of Russian verbs based on their morphological form. Verbs that are taken into consideration in their paper are simplex imperfectives and their prefixed and suffixed perfective counterparts. The only suffix that is dealt with is *-nu*, i.e. the one that in all cases yields perfective verbs. The total of 4032 derivational pairs used in the experiment are obtained from various digital databases of Russian verbs. The vector space model pre-trained with a non-contextual method of Distributional Semantics largely succeeded in the identification of the grammatical aspect of verbs in derivational pairs based on their internal, morphological structure.

The contribution of Vojtěch John, Magda Ševčíková, and Zdeněk Žabokrtský presents an interesting approach to automatic root detection and classification, i.e. the detection and classification of one of the basic units obtained by morphological segmentation. They have compared several root identification methods on seven Indo-European languages, namely Czech, German, English, French, Croatian, Italian, and Russian, for which manually segmented and annotated morphological resources are available. The results of their experiments show that simple methods, based on using simple unsupervised heuristics, derivational-tree-based heuristics, and a CRF tagger, enable highly precise automatic root identification in analyzed languages.

Marko Tadić in his work examines the ability of Large Language Models (LLMs) to generate newly derived and compound words. His method is based on the usage of parallel corpora. In this case, the sample of the Croatian part of the Croatian-English Parallel Corpus was translated using the NLTP NMT services at hrvojka.gov.hr. After tokenization with UDPipe, the list of unknown words was extracted from the translated sentences using the Croatian Morphological Lexicon, and the obtained list of 4453 tokens was further manually checked in the existing corpora and dictionaries for Croatian. The total of 321 words unknown to the existing lexica was classified into the following categories: expectable compound, unexpected compound, possessive adjective, alternative derivation, unexpected derivation, or direct alternative calque.

Marta Petrak in her paper deals with nominal prefix *pred-* in Croatian. The analysis is based upon 1006 nouns extracted from the CLASSLA-web.hr corpus and tackles the word-formational and semantic properties of this prefix based on the analyzed nouns. The analysis of word-formation types has shown that nouns with prefix *pred-* in Croatian are formed via prefixation, suffixation, prefix-suffix combination, and back-formation, the last two being rather rare. The semantic analysis yielded three meaning types of the prefix *pred-* in the analyzed derivatives: concrete (spatial) meaning, temporal meaning, and metaphorical / metonymical meaning. The analysis has also found the specific word-formation - semantic pairings, which is in line with the cognitive linguistic tenet of the grammar-lexicon continuum.

Abishek Stephen and Zdeněk Žabokrtský in their contribution investigate the nature of Czech verbs derived from borrowed nominal roots in order to show how synchronic derivational resources can be used in probabilistic analysis of the effects of borrowing in language evolution. They analyze a total of 19854 native and 3972 loanverbs from DeriNet, the word-formational network for Czech, which are attested in corpora. They show that the vast majority of loanverbs in Czech fall into

the conjugational class *-ovat*, and that they build their derivational networks in a similar way as native verbs do, i.e. they have a similar number of derived word forms in their derivational networks, which certainly makes the acceptance of borrowings easier.

Krešimir Šojat and Matea Filko present further development and enrichment of the Croatian Derivation Lexicon - CroDeriv. They discuss morphological segmentation of lexemes at the surface and deep layer and explain the basic principles in this two-layered approach. Further, they present the main derivational processes as well as some that are not described or that are only marginally described in the existing literature. Each derivation process is accompanied by examples. The authors illustrate the structure of derivation families and lexical entries in CroDeriv. The new visualization tool for the presentation of full derivational chains is also presented and briefly described. Finally, problems encountered in the processing of Croatian morphology are indicated and possible solutions in their future work are outlined.

The paper by Nikos Vasilogamvris, Michalis Sfakakis, Giannoula Giannouloupoulou and Maria Koliopoulou analyzes the types of morphological entities in derivational structures of Modern Greek following the MMoOn (Multilingual Morpheme Ontology) model, that enables the modeling of the morphological structures, i.e. the description of morphological entities and their relations. They particularly focus on allomorphy affecting stems and affixes, showing that this phenomenon impacts the derivational processes. In order to enable the generation of new lexical forms, it should be therefore modeled and placed in the derivational environments it affects.

Marco Passarotti and Eleonora Litta present the results of their research on the relation between derivational families and the frequency of their members in Latin corpora. They have analyzed 878 largest derivational families from the Word Formation Latin lexicon with at least 100 occurrences of their members in all the textual resources linked to the LiLa Knowledge Base. They investigated whether root members of the derivational family are also the most frequent one in corpora, and showed that this is the case in 582 families, and in 89 of the remaining 296 families, the most frequent member is also derivationally simple. Derivationally complex members tend to be the most frequent in Index Thomisticus Treebank, which includes a Medieval Latin philosophical treatise by Thomas Aquinas, showing also that different selections of the corpus can influence the results of the research. Finally, they show the importance of the interoperability among the lexical and textual resources, which was made in Lila Knowledge Base for Latin, and enabled the kind of research presented in their paper.

The contribution of Yağmur Öztürk, Izabella Thomas and Snejana Gadjeva describes the creation of two resources for Turkish: Semantürk, an ontology of semantic categories used in the description of morphemes, and DerivBaseTR, a database of Turkish N-to-N derivational suffixes, with their descriptive information, comprising semantic categories from the Semantürk. Authors point out the difficulties in building these resources on top of the existing ones (both textual books and computational resources) and specifically opt for the Open Science perspective in order to improve the research on processing the derivational morphology in Turkish on the one hand, and the NLP in general on the other.

Thomas Samuelsson presents preliminary research on a Corpus-assisted discourse analysis of Russian political news, operationalized as a corpus analysis of derivational prefixes observed as keymorphs. Essentially, the prefixes of nouns, verbs, and adjectives are treated as keywords in this analysis, with their salience in the massive corpus comprising 60 news outlets spanning nine years (2012–2020) estimated with respect to a Russian web reference corpus in the same time frame. The derivational lexicon DeriNet.RU is used to calculate the frequencies of the prefixes in each of the corpora. The analysis reveals some interesting discourse patterns, including socio-political trends, which the authors showcase with examples from the corpus.

We can conclude that the papers included in this volume offer new and valuable perspectives on the building of new derivational resources and the use of existing ones in linguistic research, and we hope that you will enjoy reading them.

At this point, we would like to thank Vanja Štefanec for all the technical support he provided in the preparation and printing of these proceedings. We also thank him from the bottom of our hearts for everything he does for CroDeriv.

Matea Filko
Krešimir Šojat

Program Committee Chairs

Matea Filko	Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
Krešimir Šojat	Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
Fiammetta Namer	UMR 7118 ATILF CNRS & Université de Lorraine, Nancy, France
Stéphanie Lignon	UMR 7118 ATILF CNRS & Université de Lorraine, Nancy, France
Nabil Hathout	UMR 5263 CLLE CNRS & Université Toulouse Jean Jaurès, Toulouse, France

Program Committee Members

Marco Angster	Croatia
Olivier Bonami	France
Nicola Grandi	Italy
Pius ten Hacken	Austria
Eleonora Litta	Italy
Stela Manova	Austria
Claudia Marzi	Italy
Fabio Montermini	France
Sebastian Padó	Germany
Marco Passarotti	Italy
Vito Pirrelli	Italy
Ingo Plag	Germany
Jan Radimský	Czechia
Magda Ševčíková	Czechia
Jan Šnajder	Croatia
Pavel Štichauer	Czechia
Marko Tadić	Croatia
Salvador Valera	Spain
Zdeněk Žabokrtský	Czechia

Local Organizing Committee

Matea Filko	Faculty of Humanities and Social Sciences, University of Zagreb
Krešimir Šojat	Faculty of Humanities and Social Sciences, University of Zagreb
Marko Tadić	Faculty of Humanities and Social Sciences, University of Zagreb
Daša Farkaš	Faculty of Humanities and Social Sciences, University of Zagreb
Jurica Polančec	Faculty of Humanities and Social Sciences, University of Zagreb
Vanja Štefanec	Faculty of Humanities and Social Sciences, University of Zagreb

Organizing Institutions

Croatian Language Technologies Society
Faculty of Humanities and Social Sciences, University of Zagreb
HR-CLARIN - The Croatian National Consortium of CLARIN ERIC Research Infrastructure

Table of Contents

ChatGPT, n-grams and the power of subword units: The future of research in morphology	1
<i>Stela Manova</i>	
Automatic detection of grammatical aspect of Russian verbs based on their morphological properties	13
<i>Uliana Petrunina and Hana Filip</i>	
Identification of root morphs in morphologically segmented data	23
<i>Vojtěch John, Magda Ševčíková and Zdeněk Žabokrtský</i>	
Can Large Language Models Tell Us Something about Derivational Processes?	33
<i>Marko Tadić</i>	
Morphosemantic analysis of Croatian nouns formed with the prefix <i>pred-</i>	39
<i>Marta Petrak</i>	
Understanding Borrowing through Derivational Morphology: A Case Study of Czech Verbs	49
<i>Abishek Stephen and Zdeněk Žabokrtský</i>	
Processing Croatian morphology: roots, segmentation and derivational families	61
<i>Krešimir Šojat and Matea Filko</i>	
Ontological modeling of morphological entities, allomorphy and representation in Modern Greek derivation	71
<i>Nikos Vasilogamvris, Michalis Sfakakis, Giannoula Giannouloupoulou and Maria Koliopoulou</i>	
Of Families and Occurrences. Derivation and Word Usage in Latin	81
<i>Marco Passarotti and Eleonora Litta</i>	
Morphological Resources for the Study of Turkish Derived Nouns	89
<i>Yağmur Öztürk, Izabella Thomas and Snejana Gadjeva</i>	
A keymorph analysis of Russian political news reporting	99
<i>Thomas Samuelsson</i>	

ChatGPT, n-grams and the power of subword units: The future of research in morphology

Stela Manova

University of Vienna

manova.stela@gmail.com

Abstract

Subword units (cf. morphemes in linguistic morphology) are a powerful device for language modeling (cf. Byte Pair Encoding (BPE), a subword-based tokenization algorithm part of the architecture of Large Language Models (LLMs) such as ChatGPT). Based on recent advances in natural language processing, the notion of complexity (the logic of the *Big O* notation in computer science), existing phonology-driven (form-focused) analyses of (derivational) morphology (e.g. Stratal approach) and my own research on affix order in various languages, I maintain that research in morphology should take a form-focused perspective and that novel resources favoring such a change in perspective should be developed. I provide psycholinguistic evidence from a language with poor inflectional morphology (English) and a language with very rich inflection (Polish) that native speakers do not rely on semantic cues for affix ordering in derivation but rather memorize affix combinations as bigrams and trigrams. Speakers seem to treat frequently co-occurring linearly adjacent affixes, be they derivational or inflectional, together, as subword units longer than a morpheme, which is exactly what happens during the subword-based tokenization (BPE) in a LLM. Claims that ChatGPT does not reflect human-like language processing in morphology (and not only) are, most probably, due to the lack of linguistic research that adopts a ChatGPT perspective on language.

1 Introduction

Recently, computer science (CS) has made significant progress and now Generative Pre-trained Transformers (GPT) are used for natural language processing (NLP). A GPT is a type of a large language model (LLM) based on an artificial neural network (transformer architecture) and pre-trained on large data sets of unlabeled text, i.e. a GPT does not use grammar of the type known from linguistic theory. ChatGPT, a LLM chatbot, was launched by OpenAI on November 30, 2022. It has a user-friendly interface and was additionally trained for dialogue with humans. The most surprising feature of ChatGPT from a linguistic point of view (because ChatGPT can accomplish non-linguistic tasks as well) is its ability to generate human-like texts in real-time chat, which has thus raised questions about the correctness of the so-called Chomsky's approach in linguistics that claims for innateness of language. Since this approach has been one of the dominant research paradigms in linguistics for years, the recent advances in NLP are expected to have a significant impact on the future of linguistics as a scientific field. Unfortunately, there has not been any constructive dialogue on these issues: computer scientists are not interested in theorizing but in problem-solving and, as a rule, they do not participate in linguistic discussions; there has been only an exchange (mainly on the *lingbuzz* archive) between psychologists / neuroscientists (Piantadosi 2023) and linguists (Chomsky et al., 2023; Katzir, 2023; Moro et al., 2023; Rawski and Baumont, 2023; Sauerland, 2023). In this exchange, one thing has become clear: linguists do not understand LLMs as an opportunity to see language from a novel perspective. For example:

- If ChatGPT can understand and generate language based only on form (a linear sequence of words in a prompt), form and meaning in language should be in a perfect relationship. As ChatGPT prompts are longer than a word, often even longer than a sentence, the perfect relationship between meaning and form should be visible only if one considers long sequences of words (tokens); tokenization, specifically the Byte Pair Encoding (BPE) algorithm used in LLMs, is introduced in Section 2 below.
- If ChatGPT does not rely on hierarchically organized *trees*, though the latter are a common data structure in CS, this could be an indication that, most probably, there is some

problem with the trees in Chomsky’s approach (linguistic trees have an unnatural direction of growth – from leaves to the root, which is the opposite to how trees grow in CS, see the discussion in Manova, 2022).

- ChatGPT was launched in 2022 and is fluent in an impressive number of languages, Chomsky’s approach celebrated 50 years of linguistics at MIT in 2011 but still cannot generate fluent language. This situation could only mean that, most probably, Chomsky’s theory is unnecessarily complex. As for the millions of parameters in a LLM, just think of the number of neural networks (human brains)¹ Chomsky’s approach has had at its disposal in the years. Complexity is discussed in Section 3 below, see also Manova (2022) in which a ridiculously simple model based on linear structures such as bigrams and trigrams appears more efficient than a syntactic model with hierarchical trees.

Since Chomsky’s approach, among other things, made possible the introduction of the syntax-based Distributed Morphology (Halle and Marantz, 1993; Harley and Noyer, 1999; Embick and Noyer, 2007; Bobaljik, 2017), all the above issues are highly relevant to a morphological event such as *DeriMo 2023: Resources and tools for derivational morphology*, for the proceedings of which this text is meant. Thus in what follows, my focus is on derivational morphology. In Section 4, I demonstrate a form-based analysis of word-formation in two typologically distinct languages, English (with very poor inflectional morphology) and Polish (with very rich inflection) and report the results of a psycholinguistic experiment with native speakers of the two languages. In Section 5 conclusions are drawn and missing resources for research on derivational morphology identified.

2 Subword units

ChatGPT uses `tiktoken` (<https://github.com/openai/tiktoken>), a BPE tokenizer. BPE is a subword-based tokenization algorithm and as such can discover “common subwords”, e.g. pieces such as “ing” in English. A demonstration of tokenization is available at: <https://platform.openai.com/tokenizer>. A LLM, as a rule, operates with a modified BPE and its vocabulary comprises the following types of tokens: single (unique) characters, subword units, whole words, single digits and other special characters. Roughly, similar to what has been established in psycholinguistic research, highly frequent and highly rare pieces of form (tokens) are listed in the LLM vocabulary. ChatGPT has a fixed-size vocabulary of tokens, `cl100k_base`. It has to be noted that the model actually works with numbers, click on token IDs in the tokenization demonstration, the URL just given: a prompt is encoded into a sequence of numbers, when the task is solved, the output is decoded and numbers are again turned into language. Subword units and whole words that are part of the vocabulary are established in terms of the most frequent sequence of adjacent characters in a *n*-gram manner: unique characters are unigrams and as such are listed; highly frequent combinations of two characters (tokens) are bigrams, of three characters – trigrams, etc. Thus, the tokenization is entirely form-based and does not pay any attention whatsoever to semantics (cf. Manova et al., 2020, on from-form-to-meaning versus from-meaning-to-form analyses in morphology, e.g. Distributed Morphology (references in Section 1) and Paradigm Function Morphology (Stump, 2001, 2016; Stump and Finkel, 2013; Bonami and Stump, 2017) are both from-meaning-to-form). LLM tokens (subword units) do not necessarily coincide with morphemes, though the most frequent combinations of adjacent characters can be expected to form either morphemes or words.

Unlike subword tokenization, current studies on and resources for derivational morphology are semantics-based: they operate with word families (Bauer and Nation, 1993, among others), word-formation nests (Burkacka, 2015, and references therein), derivational paradigms (Bonami and Strnadová, 2019; Hathout and Namer, 2019, and references therein), derivational networks

¹ The human brain is a neural network with an unknown number of parameters, see Kozachkov et al. (2023) on how one can “build Transformers using biological computational units”.

(Körtvélyessy et al., 2020), blocking (Aronoff, 1976; Rainer, 2016, and references therein), affix rivalry (Huyghe and Varvara, 2023, and references therein).²

The relevant question is now: Could it be that a model of derivational morphology that relies on form is less complex than a model that relies on semantics? To answer this question, we should first clarify *complexity*.

3 Complexity

In science, a problem often allows for different solutions. The so-called *Big O* notation serves for assessment of the complexity of those solutions in mathematics and CS. The *Big O* notation tells us how an algorithm slows as data grow. That is, complexity is not a property of data (which is the case in linguistics), but of the algorithm (analysis). As an illustration let me evaluate two solutions of a task. Note that the example is meant to help linguists understand the logic of the concept of complexity and is an oversimplification. In CS, the *Big O* notation evaluates the complexity of functions.

Problem: Calculate the sum of the numbers from 1 to 100.

Solution 1: 1+2+3, and so on to 100, i.e. 99 summations are necessary to calculate the sum.

Let us check the behavior of this solution as data grow, e.g. let us increase the amount of the data from 100 to 1000. Following the idea of Solution 1, to calculate the sum of the numbers from 1 to 1000, we have to perform 999 summations. That is, with the growth of the data, more effort is required to come to a solution.

Solution 2: Based on the observation made by the young Gauss that $100+1 = 99+2 = 98+3$, and so on to $51+50$, we can calculate the sum of the numbers from 1 to 100 in two steps: the first step involves addition, the second consists in multiplication: $(1+100)*50=5050$. An increase of the amount of the data from 100 to 1000, does not change the algorithm and we can still calculate the sum of the numbers from 1 to 1000 in two steps: $(1+1000)*500= 500500$.

Both Solution 1 and Solution 2 give the same result, but the first solution is complex and therefore uninteresting, while Gauss's solution is simple and elegant and has been used as a formula for the sum of an arithmetic progression ever since.

How does all this relate to ChatGPT and research in derivational morphology? The ChatGPT approach to language relies on surface forms (for convenience, I will speak of 'phonological information'), see Rule 1; while a linguistics approach usually relies on semantics, see Rule 2.

Rule 1, form-based: If a word A ends in *-a*, attach the suffix B to it.

Rule 2, semantics-based: If X is a particular type of a verb (e.g. an action verb), derive a particular type of a noun Y (e.g. an agent) by the attachment of the productive suffix Z (e.g. *-er*).

Now, the information on which Rule 1 relies is not language-specific and is directly available: for the word A we have to evaluate whether it terminates in *-a* or not. The semantic information on which Rule 2 relies requires additional effort to be discovered and Rule 2 is also language-specific, in the sense that we need some knowledge of the language from which the data come in order to apply this rule. Then, Rule 1 consists of two steps: i) we have to check whether A ends in *-a* and if yes, ii) to attach the suffix B. Rule 2 involves the following steps: a) evaluation whether the word we deal with is a verb; if yes, b) we have to ensure that the verb is of the type we need (an action verb); afterwards c) addition of the productive suffix *-er* to derive an agent noun, if d) the derivation is possible, because e.g. *to edit* is an action verb but does not co-occur with *-er* (moreover, according to linguistic theory *to edit* is a backformation from *editor*, Manova, 2011a). Therefore, we conclude that Rule 2 is more complex than Rule 1.

Before moving to Section 4, in which I demonstrate a form-based analysis of derivational morphology, let us have a look at (1) and (2) which illustrate Rule 1 with real data, from Bulgarian (Slavic). (1) and (2) are not derivational morphology, but a similar rule, though less impressive, for

² Curiously enough, in morphological theory even the definition of *morphome*, a purely morphological form hard to account for in terms of meaning, involves reference to semantics (Aronoff, 1994; Maiden, 2004; Luís and Bermúdez-Otero, 2016; Herce, 2023, among others).

derivation of diminutives is given in Section 4. Bulgarian has a suffixal definite article and indefinite nouns and adjectives in this language may end in *-a*. If semantics is considered, there should be four different *-a* morphemes, cf. the morphosyntactic feature values in (1) and (2), where all *-a* morphemes are bolded and indexed for convenience. The four different *-a* morphemes all select the definite article *-ta* (Manova and Dressler, 2001), though the article has allomorphs, see *selo* ‘village’ in (1d).

- | | | |
|-----|---|--|
| (1) | Nouns: indefinite | → definite |
| | a. sg.fem: <i>bluz-a₁</i> ‘blouse’ | → <i>bluz-a₁-ta</i> ‘the blouse’ |
| | b. sg.masc: <i>bašt-a₂</i> ‘father’ | → <i>bašt-a₂-ta</i> ‘the father’ |
| | c. pl.neut: <i>sel-a₃</i> ‘villages’ | → <i>sel-a₃-ta</i> ‘the villages’ |
| | d. cf. sg.neut: <i>sel-o</i> ‘village’ | → <i>sel-o-to</i> ‘the village’ |
| (2) | Adjectives: indefinite | → definite |
| | sg.fem: <i>krasiv-a₄</i> ‘beautiful’ | → <i>krasiv-a₄-ta</i> ‘the beautiful’ |

4 A form-based analysis of derivational morphology

Undoubtedly, English is the language with the most profoundly studied derivational morphology. (Overviews of research on derivational morphology from a cross-linguistic perspective in Lieber and Štekauer, 2014; Plag and Balling, 2016; and Lieber, 2017). While more recent studies analyze English word-formation based primarily, if not exclusively, on semantics (Lieber, 2004, among many others), previous research known as the *Stratal approach* (Siegel, 1974; Selkirk, 1982; Kiparsky, 1982) is form-focused, see (3): based on phonological information (see the different types of juncture marked by ‘+’ and ‘#’ respectively) forms of affixes are distributed into different strata (classes) so that class II affixes are always outside class I affixes in the word-form.

- (3) English: Stratal approach, from Spencer (1991:79)
- | | |
|----|--|
| a. | Class I suffixes: <i>+ion, +ity, +y, +al, +ic, +ate, +ous, +ive, +able, +ize</i> |
| b. | Class I prefixes: <i>re+, con+, de+, sub+, pre+, in+, en+, be+</i> |
| c. | Class II suffixes: <i>#ness, #less, #hood, #ful, #ly, #y, #like, #ist, #able, #ize</i> |
| d. | Class II prefixes: <i>re#, sub#, un#, non#, de#, semi#, anti#</i> |

Another example of a form-focused analysis is Fabb (1988). This study distributes the English suffixes into four groups as shown in (4):

- (4) English: Suffix-driven selectional restrictions (Fabb 1988)
- | | |
|----|--|
| a. | Group 1: suffixes that do not attach to already suffixed words |
| b. | Group 2: suffixes that attach outside one other suffix |
| c. | Group 3: suffixes that attach freely |
| d. | Group 4: problematic suffixes |

An alternative, form-focused analysis recognizes closing suffixes: a particular suffixal form cannot be followed by other suffixes in a language, Szymanek (2000) for English (and Polish), see also Aronoff & Fuhrhop (2002). Closing suffixes have been established in a number of languages, Manova (2015b) is an overview of research on the topic.

Another highly relevant observation regarding the order of English derivational suffixes is reported in Manova (2011b) and Manova and Knell (2021). Manova (2011b) sees derivational suffix combinations as binary structures of the type SUFF1-SUFF2, where SUFF1 has three valency positions for further suffixation: SUFF2_{Noun}, SUFF2_{Adjective} and SUFF2_{Verb}. The idea of this distribution of outputs according to the lexical-category specification of SUFF2 is based on a mathematical method, Gauss-Jordan elimination. This method serves for solving systems of linear equations numerically, that is, only with the help of elementary operations such as substitution, addition or multiplication. (5) is an example of a system of linear equations.

(5) $2x + y + 2z = 10$

$$\begin{aligned}x + 2y + z &= 8 \\ 3x + y - z &= 2\end{aligned}$$

The goal of Gauss-Jordan is, based only on well-known facts and elementary operations with them, to come to a single option for a variable (the unknown); x , y and z are the variables in (5). If there is only one option for a variable, this option is the solution to the problem.

With respect to affix order in derivation, the well-known information is information about the lexical category specification of an affix, i.e. whether the affix derives nouns (N), adjectives (A) or verbs (V); a single option for a variable means one affix combination of a kind, i.e. a one-to-one mapping of form and meaning. As can be seen from Table 1, this method allows data to be distributed so that in most cases there is one option of a kind, see for N ($-ist_N-dom_N$) and for V ($-ist_N-ize_V$). I label such combinations *fixed*.

SUFF1	Lexical category of SUFF1	SUFF2 classified for lexical category; in brackets, number of types (lemmas) derived with the combination SUFF1-SUFF2	
$-ist$	N	N: $-dom$ (2) A: $-ic$ (631) , $-y$ (5) V: $-ize$ (3)	[<i>fixed combination</i>] [<i>predictable combination</i>] [<i>fixed combination</i>]

Table 1: Combinability of the English suffix $-ist$
(data from Aronoff and Fuhrhop, 2002, based on OED, CD, 1994)

If more than one SUFF2 of the same lexical category is available (see for A in Table 1), one of the SUFF2 suffixes attaches by default, suffix $-ic_A$ in our case: in English, the combination $-ist_N-ic_A$ derives 631 types, while $-ist_N-y_A$ derives only 5 types. I therefore classify $-ist_N-ic_A$ as a *predictable* combination. Regarding default suffixes, having counted suffix combinations in large dictionaries and corpora for different languages, Manova (2011, 2015), Manova and Talamo (2015), and Manova and Knell (2021) maintain that a default suffix derives more than ten types, while SUFF2 suffixes that compete with the default suffix derive ten or fewer types each. Thus, default suffixes are also seen as productive.

Table 2 applies the logic of Gauss-Jordan to a more complex case, the combinability of the Polish suffix $-arz$.³ Polish, unlike English, is an inflecting fusional language and derivational suffixes are often followed by inflection, i.e. in Polish the inflection is obligatory for the well-formedness of a word. All inflectional suffixes in Table 2 are in brackets. Descriptions and analyses of Polish derivational morphology by Polish scholars, as a rule, give derivational suffixes together with the inflection that follows them, either in brackets, as done in Table 2, or unmarked, as a single suffix with the derivational one, $-n(y)$ or $-ny$, respectively; see the first adjectivizing SUFF2 in Table 2. (For a semantics-based analysis of the combinability of Polish derivational suffixes, see Burkacka, 2015; see also the discussion of Polish word-formation in Szymanek, 2010.)

As shown in Table 2, the suffix $-arz$ combines with more than one adjectivizing SUFF2 and a set of nominalizing SUFF2 suffixes. While for the derivation of adjectives, there is only one default suffix, $-sk(i)$ (>10), three different nominalizing suffixes that derive more than ten types can follow the suffix $-arz$: $-czyk$ (>10), $-ni(a)$ (>10) and $-stw(o)$ (>10), all bolded in Table 2 for convenience. The existence of three productive (default) suffixes of the same type (nominalizers) all “competing” for $-arz$ seems to challenge my analysis. Note, however, that the three competing suffixes differ in both form and meaning: $-czyk$ (>10), default for derivation of persons; $-ni(a)$ (>10), default for derivation of places; and $-stw(o)$ (>10), default for abstract/collective nouns. That is, no suffix homophony is involved (homophony is a problem for any form-based analysis). I therefore conclude that all suffix combinations in Table 2 are predictable.

Considering the fact that derivational suffixes in English and Polish seem to form only fixed and predictable combinations, I hypothesized that native speakers should have memorized them and, consequently, should be able to produce them without reference to meaning, that is, based exclusively

³ I thank Bartosz Brzoza for his help with the Polish data.

on form. To test this hypothesis, I designed a psycholinguistic experiment the results of which are reported below. Due to the limited length of this paper, here I present only the results of the native speakers of English and Polish, but the experiment was also conducted with native speakers of German, Italian, Spanish and Slovene, and with advanced non-native speakers of English and German. Overall, the results of all iterations converge. (For curious readers, the scores of the non-native speakers of English are reported in Manova and Knell, 2021; the scores of the native and non-native speakers of German can be found in Brosche and Manova, 2022).

SUFF1	Lexical category of SUFF1	Lexical category of SUFF2	SUFF1-SUFF2 exemplified	Notes
-arz	N	i. ADJ: <i>-n(y)</i> (2)	<i>moc-ar-n(y)</i> ‘strong’	[derives only 2 adjectives]
		ii. ADJ: <i>-ow(y)</i> (1)	<i>gęśl-arz-ow(y)</i> ‘of fiddler’	[derives a single adjective]
		iii. ADJ: <i>-sk(i)</i> (>10)	<i>pis-ar-sk(i)</i> ‘of writer’	[default for derivation of adjectives]
		a. N: <i>-czyk</i> (>10)	<i>piek-ar-czyk</i> ‘baker’s apprentice’	[default for derivation of persons , cf. f]
		b. N: <i>-k(a)</i> (2)	<i>mur-ar-k(a)</i> ‘bricklaying’	[derives only 2 abstract nouns, cf. e]
		c. N: <i>-ni(a)</i> (>10)	<i>kreśl-ar-ni(a)</i> ‘drafting studio’	[derives nouns for places]
		d. N: <i>-nik</i> (1)	<i>piek-ar-nik</i> ‘oven’	[derives a single object]
		e. N: <i>-stw(o)</i> (>10)	<i>księg-ar-stw(o)</i> ‘all booksellers’	[default abstract/collective nouns , cf. b]
		f. N: <i>-yn(a)</i> (5)	<i>mur-arz-yn(a)</i> ‘bad bricklayer’	[derives only 5 nouns for persons, cf. a]

Table 2: Combinability of the Polish suffix *-arz*

Method

64 native Polish speakers and 45 native English speakers were tested, they all participated on a voluntary basis. The questionnaire presented to them consisted of three parts:

- A series of general demographic questions regarding age, gender, nationality, native language(s), other languages spoken, level of education, and experience in a linguistic or other language-related field.
- A small practice to ensure that the participants understood the task properly. The training examples were not part of the test stimuli.
- The main task: 60 suffix combinations (e.g. *-istic* in English, *-arny* in Polish) were presented in a randomized order, and participants were asked to decide intuitively, as quickly as possible, which of the combinations exist and which do not exist as word terminations in the respective language. Of the 60 combinations, 30 exist in the respective language and 30 do not. Of the existing combinations, 15 were productive and 15 unproductive. Of the non-existing combinations, 15 were created from a permutation of an existing combination (reversing the order of the two suffixes such that the combination was not possible in English), and 15 were

created through a spelling manipulation of an existing combination (changing one letter from an existing combination such that the new form does not exist in the respective language). No non-existing combinations included any phonological and/or orthographical impossibilities in the respective language. Participants were given a 10-minute time limit to complete the main task. (On average, the subjects used approximately one third of the time.)

Data Analysis

We used independent t-tests to consider possible significance of overall scores, as well as for stimulus type: existing vs. non-existing and productive vs. unproductive combinations. Figure 1 presents the results of the native speakers of English and Polish.

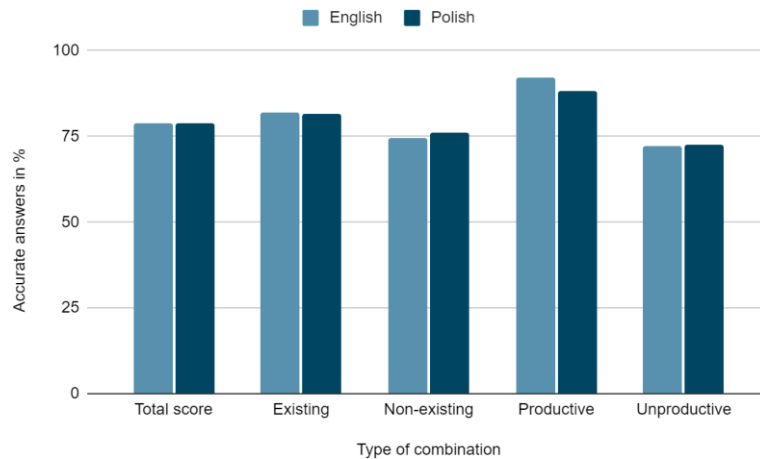


Figure 1: Native speakers' accuracy of recognition of the 60 suffix combinations tested in the experiment (only statistically significant results). Total score = correct answers for all types of suffix combinations tested: existing combinations of two types (productive and unproductive) and non-existing of two types (permutations and manipulations, see Method).

The participants in the experiment did not need semantic cues to process suffix combinability, i.e. they could differentiate between existing and non-existing suffix combinations presented to them without lexical bases such as roots/stems/words. Statistically significant were the differences between existing and non-existing combinations, and between productive and unproductive combinations. As already mentioned, English and Polish differ typologically, in the sense that English has very poor inflectional morphology, while Polish is characterized by a very rich inflectional system. Nevertheless, the results obtained for the two languages are virtually the same, the total score of the correct answers for English is 79% and 78.86% for Polish (Figure 1), though combinations of three suffixes (trigrams, the case of Polish where two derivational suffixes are often followed by inflection) should be easier to recognize than combinations of two suffixes (bigrams, the case of English derivational suffix combinations). In other words, inflection did not seem to have an impact on the processing on suffix combinability in derivation. I therefore conclude that native speakers of Polish see inflection as forming a natural subword unit with the derivational material that precedes it.

Since suffix combinability is not taught at school and all linguistic theories assume that a morphological derivation always starts with a root/stem, depending on the theory, the only plausible explanation why native speakers of English and Polish successfully accomplished a task they should not be able to solve is that they had subconsciously extracted and memorized adjacent suffixes in terms of bigrams⁴ and trigrams, during language acquisition (cf. the training of ChatGPT). Further support to the conclusion that adjacent derivational and inflectional suffixes should be treated together provides Polish diminutive morphology. Polish, like the other Slavic languages (Manova 2015a), derives second-grade diminutives the forms of which contain a sequence of two adjacent diminutive suffixes, bolded in the following example: *dom* 'house' → DIM1 *dom-ek* 'small house' → DIM2 *dom-ecz-ek* 'very

⁴ Analyses of affix order in terms of bigrams are proposed in Ryan (2010) and Mansfield et al. (2020).

small house’. Table 3 presents the combinability of the nominal diminutive suffixes in Polish. The selection of the second diminutive suffix entirely depends on the phonological make-up of the first diminutive suffix: a DIM1 suffix in *-C* is always followed by a DIM2 suffix in *-C*, a DIM1 suffix in *-a* is always followed by a DIM2 suffix in *-a*, and a DIM1 suffix in *-o* is always followed by a DIM2 suffix in *-o*, see Table 3. For the sake of completeness, both *-a* and *-o* are inflection. The selection of the DIM1 suffix is also form-driven in all but one case: the unproductive class of the feminine-gender nouns in *-C* selects DIM1 suffix based not on phonology but on gender, see “Nouns in *-C*” in Table 3. (In Polish, the default ending for feminine nouns is *-a*.)

DIM1 suffixes		DIM2 suffixes	
Nouns in		Productive (attach by addition)	Unproductive (attach by substitution of a DIM1 suffix, i.e. do not combine with DIM1 suffixes)
<i>-C</i>	<i>-ek</i>	<i>-ek</i>	<i>-uszek, -aszek</i>
	<i>-ik / -yk</i>		
	<i>-uszek</i> (unproductive)		
	<i>-iszek / -yszek</i> (unproductive)		
	<i>-aszek</i> (unproductive)		
	<i>-ulek</i> (unproductive)		
	<i>-ka</i> (unproductive, selects feminine nouns)		
<i>-a</i>	<i>-ka</i>	<i>-ka</i>	
	<i>-uszka</i> (unproductive)		
	<i>-iczka / -yczka</i> (unproductive)		
<i>pr-o / -e</i>	<i>-ko</i>	<i>-ko</i>	
	<i>-uszko</i> (unproductive)		

Table 3: Combinability of the DIM suffixes in Polish (from Manova & Winzernitz 2011)

5 Conclusion

Based on the BPE algorithm used for tokenization in LLMs, a mathematical method for problem solving, the so-called Gauss-Jordan elimination, and previous research on affix order (by other authors and my own), I put forward the idea of form-based analysis of derivational morphology and illustrated it with data from two typologically distinct languages, English with very poor inflectional morphology, and Polish with very rich inflection. A psycholinguistic experiment with native speakers of Polish and English confirmed the correctness of the proposal. Native speakers do not need semantic cues to process affix ordering in derivation. They seem to have memorized linearly adjacent affixes, be they derivational or inflectional, as bigrams and trigrams, without reference to semantics, which is exactly what happens during the subword-based tokenization in a LLM. Since morphology works with units of a very small length, the form-meaning correspondences in my analysis (and in (derivational) morphology in general) are not perfect, cf. the long sequences of form used in ChatGPT where form and meaning appear to be in a perfect one-to-one relationship. Nevertheless, a flexible approach, such as the one demonstrated in this paper, i.e. an approach operating with defaults and a fixed reasonable number of exceptions (ten or fewer exceptions in my analysis; exceptions are derived items which due to very low type-productivity should be rote-learned) successfully derives new words from already suffixed ones in English and Polish. Future research is needed to see this approach works with unsuffixed bases.⁵ In this endeavor, form-focused (preferably cross-linguistic) resources for (derivational) morphology providing information about word structure in terms of bigrams and trigrams

⁵ “Automatically discovered set of derivation rules” in Ševčíková and Žabokrtský (2014) can be seen as a step in this direction, as well as Manova (2011a) which is a structural, i.e. form-based, analysis of conversion and subtraction, with a focus on the derivational base. See also Unsupervised Learning of Morphology, Hammarström and Borin (2011).

of morphemes (linear sequences of adjacent subword units) will be essential. Such resources currently do not exist. Thus, claims that ChatGPT does not reflect human-like language processing in morphology (and not only) are, most probably, due to the lack of linguistic research that adopts a ChatGPT perspective on language.

References

- Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. MIT Press, Cambridge, Ma.
- Aronoff, Mark. 1994. *Morphology by itself: Stems and Inflectional Classes*. MIT Press, Cambridge, Ma.
- Aronoff, Mark, and Nanna Fuhrhop. 2002. Restricting Suffix Combinations in German and English: Closing Suffixes and the Monosuffix Constraint. *Natural Language and Linguistic Theory* 20: 451–490.
- Bauer, Laurie, and Paul Nation. 1993. Word Families. *International Journal of Lexicography* 6: 253–279.
- Bobaljik, Jonathan D. 2017. Distributed Morphology. *Oxford Research Encyclopedia in Linguistics*, Oxford: Oxford University Press, <https://doi.org/10.1093/acrefore/9780199384655.013.131>.
- Bonami, Olivier, and Gregory Stump. 2017. Paradigm Function Morphology. In Andrew Hippisley and Greg Stump, editors, *The Cambridge handbook of morphology* Cambridge University Press, Cambridge, pages, 449–481.
- Bonami, Olivier, and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology, *Morphology* 29: 167–197.
- Brosche, Kimberly, and Stela Manova. 2022. German Word Formation and the Organization of the Mental Lexicon. *Annual Conference of the Middle European Master Program in Cognitive Science*. Zagreb, June, https://homepage.univie.ac.at/stela.manova/uploads/1/2/2/4/12243901/poster_suffix_combinations_version.pdf
- Burkacka, Iwona. 2015. Suffix sets in Polish de-nominal derivatives. In Stela Manova, editor, *Affix ordering across languages and frameworks*. Oxford University Press, New York, pages 233–258.
- Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The false promise of ChatGPT. *The New York Times*, <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgptai.html>
- Fabb, Nigel. 1988. English Suffixation is Constrained only by Selectional Restrictions. *Natural Language and Linguistic Theory* 6: 527–539.
- Embick, David, and Rolf Noyer. 2007. Distributed morphology and the syntax/morphology interface. In Gillian Ramchand, and Charles Reiss, editors, *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press, New York, pages, 289–324.
- Halle, Morris, and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth Hale, and Samuel Jay Keyser, editors, *The view from building 20*. MIT Press, Cambridge, Ma, pages 111–176.
- Hammarström, Harald, and Lars Borin. 2011. Unsupervised Learning of Morphology. *Computational Linguistics* 37 (2): 309–350. doi: https://doi.org/10.1162/COLI_a_00050
- Harley, Heidi, and Rolf Noyer. 1999. State-of-the-article: Distributed Morphology. *GLOT International* 4: 3–9.
- Hathout, Nabil, and Fiammetta Namer. 2019. Paradigms in word formation: what are we up to? *Morphology* 29, 153–165.
- Herce, Borja. 2023. *The Typological Diversity of Morphemes: A Cross-Linguistic Study of Unnatural Morphology*. Oxford University Press, Oxford.
- Huyghe, Richard, and Rossella Varvara. 2023. *Affix rivalry: Theoretical and methodological challenges*. *Word Structure* 16: 1–23.

- Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023), lingbuzz/007190
- Kiparsky, Paul. 1982. Lexical Morphology and Phonology. In *The Linguistic Society of Korea, Linguistics in the Morning Calm*. Hanshin Publishing Co, Seoul, pages 1–91.
- Körtvélyessy, Livia, Bagasheva, Alexandra and Štekauer, Pavol. 2020. *Derivational Networks Across Languages*, De Gruyter Mouton, Berlin, Boston.
- Kozachkov, Leo, Ksenia V. Kastanenko, and Dmitry Krotov. 2023. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 120 No. 34, pages = e2219150120, <https://doi.org/10.1073/pnas.221915012>
- Lieber, Rochelle. 2004. *Morphology and lexical semantics*. Cambridge: Cambridge University Press.
- Lieber, Rochelle. 2017. Derivational Morphology. *Oxford Research Encyclopedia of Linguistics*. <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-248>. Retrieved 17 Sep. 2023.
- Lieber, Rochelle, and Pavol Štekauer. 2014. *The Oxford handbook of derivational morphology*. Oxford University Press, Oxford.
- Luís, Ana R., and Ricardo Bermúdez-Otero, editors, 2016. *The morpheme debate*. Oxford University Press, Oxford.
- Maiden, Martin. 2004. Morphological autonomy and diachrony. *Yearbook of Morphology* 2004: 137–175.
- Manova, Stela. 2011a. *Understanding morphological rules*. Springer, Dordrecht.
- Manova, Stela. 2011b. A cognitive approach to SUFF1-SUFF2 combinations: A tribute to Carl Friedrich Gauss. *Word Structure* 4: 272–300.
- Manova, Stela. 2015a. Affix order and the structure of the Slavic word. In Stela Manova, editor, *Affix Ordering Across Languages and Frameworks*. Oxford University Press, New York, pages 205–230.
- Manova, Stela. 2015b. Closing suffixes, In P. Müller, I. Ohnheiser, S. Olsen, and F. Rainer, editors, *Word-Formation in the European Languages*. Vol. 2, Handbooks of Linguistics and Communication Science (HSK) 40/2. De Gruyter Mouton, Berlin, pages 956–971.
- Manova, Stela. 2022. The linear order of elements in prominent linguistic sequences: Deriving Tns-Asp-Mood orders and Greenberg's Universal 20 with n-grams, lingbuzz/006082
- Manova, Stela. 2023. Ordering restrictions between affixes. In Peter Ackema, Sabrina Bendjaballah, Eulàlia Bonet, and Antonio Fábregas, editors, *The Wiley Blackwell Companion to Morphology*. John Wiley & Sons, Hoboken, NJ. DOI: [10.1002/9781119693604.morphcom058](https://doi.org/10.1002/9781119693604.morphcom058)
- Manova, Stela, and Wolfgang U. Dressler. 2001. Gender and Declensional Class in Bulgarian. *Wiener Linguistische Gazette* 67-69: 45–81.
- Manova, Stela, and Mark Aronoff. 2010. Modeling affix order. *Morphology* 20: 109–131.
- Manova, S., and Kimberley Winternitz. 2011. Suffix Order in Double and Multiple Diminutives: With Data from Polish and Bulgarian. *Studies in Polish Linguistics* 6: 115-138.
- Manova, Stela, and Luigi Talamo. 2015. On the Significance of the Corpus Size in Affix-Order Research. *SKASE Journal of theoretical linguistics* 12: 369–397.

- Manova, Stela, Harald Hammarström, Itamar Kastner, and Yining Nie. 2020. What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure* 13: 1–21.
- Manova, Stela, and Georgia Knell. 2021. Two-suffix combinations in native and non-native English: Novel evidence for morphomic structures. In Sedigheh Moradi, Marcia Haag, Janie Rees-Miller, and Andrija Petrovic, editors, *All things morphology: Its independence and its interfaces*. Benjamins, Amsterdam, pages 305–323.
- Mansfield, John, Sabine Stoll, and Balthasar Bickel. 2020. Category clustering: A probabilistic bias in the morphology of verbal agreement marking. *Language* 96: 255–293.
- Moro, Andrea, Matteo Greco, and Stefano F. Cappa. 2023. Large languages, impossible languages and human brains. *Cortex* 167: 82–85.
- OpenAI. 2023. GPT-4 Technical Report. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774) [cs.CL], accessed July 1st, 2023.
- Piantadosi, Steven. 2023. Modern language models refute Chomsky’s approach to language, [lingbuzz/007180](https://lingbuzz.net/007180)
- Plag, Ingo, and Laura Balling. 2016. Derivational morphology: An integrative perspective on some fundamental questions. In Pirelli, Vito, Ingo Plag, and Wolfgang U. Dressler, editors, *Word knowledge and word usage: A cross-disciplinary guide to the mental lexicon*. De Gruyter, Berlin, pages 295–335.
- Rainer, Franz. 2016. Blocking. *Oxford Research Encyclopedia of Linguistics*. Retrieved 19 Sep. 2023, from <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-33>.
- Rawski, Jon, and Lucie Baumont. 2023. Modern Language Models Refute Nothing, [lingbuzz/007203](https://lingbuzz.net/007203).
- Ryan, Kevin M. 2010. Variable affix order: Grammar and learning. *Language* 86: 758–91.
- Sauerland, Uli. 2023. The Impact of Large Language Models on Linguistic Theory and Generative Grammar: A Critical Analysis, [lingbuzz/007217](https://lingbuzz.net/007217).
- Selkirk, Elisabeth. 1982. *The Syntax of Words*. MIT Press, Cambridge, Ma.
- Siegel, Dorothy. 1974. *Topics in English Morphology*. MIT Press, Cambridge, Ma.
- Spencer, A. 1991. *Morphological Theory*. Oxford: Blackwell Publishers.
- Stump, Gregory. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge University Press, Cambridge.
- Stump, Gregory. 2016. *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge University Press, Cambridge.
- Stump, Gregory, and Raphael A. Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge.
- Szymanek, Bogdan. 2000. On morphotactics: Closing morphemes in English. In Bożena Rozwadowska, editor, *PASE Papers in Language Studies*, Aksel, Wrocław, pages 311–320.
- Szymanek, Bogdan. 2010. *A panorama of Polish word-formation*. Wydawnictwo KUL, Lublin.
- Ševčíková, Magda, and Zdeněk Žabokrtský. 2014. Word-formation network for Czech. In Nicoletta Calzolari et al., editors, *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC 2014)*. ELRA, Paris, pages 1087–1093, http://www.lrec-conf.org/proceedings/lrec2014/pdf/501_Paper.pdf.

Automatic detection of grammatical aspect of Russian verbs based on their morphological properties

Uliana Petrunina
Heinrich Heine University
Dusseldorf, Germany
uliana.petrulina@hhu.de

Hana Filip
Heinrich Heine University
Dusseldorf, Germany
filip@hhu.de

Abstract

The goal of this study is to explore whether the properties of the morphological form of Russian verbs can be used to automatically predict their grammatical status as perfective or imperfective. We rely on a vector space model pre-trained with a non-contextual method of Distributional Semantics. The model largely succeeded in correctly identifying the grammatical aspect of derivationally related perfective and imperfective forms based on their morphological form. The study demonstrates that the internal structure of verbs captured by the model can identify whether a given Russian verb form is perfective or imperfective. Our results are especially relevant for computational studies using distributional semantic representations for aspect prediction and analysis of morphological patterns in languages with verbs that exhibit a complex morphological structure.

1 The Main Topic and Questions

This study proposes an approach that automatically differentiates between simplex imperfective and morphologically complex perfective forms in Russian derived from them by means of prefixes or the semelfactive suffix *-nu-*, referred henceforth as ‘derivational pairs’.¹ We pose the following questions:

- i Does the morphological form contribute to distinguishing grammatical aspect of Russian?
- ii How well does the distributional semantics approach handle grammatical aspect detection?

Our approach relies on visualizing the distribution of verb vector representations in a vector space. It is built using a distributional semantic model pre-trained using a fastText non-contextual method, which is specifically adjusted for the analysis of morphological patterns. The method generates vectors for verb lemmas based on their internal subword (morphological) information in the form of character n-grams. It is important to note that Slavic verbs in derivational pairs most often differ in lexical semantics, not just in grammatical aspect, therefore members of a derivational pair may not always have distributionally close vector space representations.

2 Linguistic Assumptions

The perfective-imperfective opposition in Russian, as in other Slavic languages, is largely lexicalized, and manifested in relations between forms that are derivationally related by means of affixation (Dahl, 1985; Filip, 1993/1999, 2000; Wiemer and Seržant, 2017). Examples of derivational pairs we are interested in are given in (1).²

- (1) a. $gavkat' \text{ IMPF} \Rightarrow \text{pro-gavkat}' \text{ PFV}$
bark-INF PX-bark-INF
'(be) bark(ing)' 'bark [several times]'
- b. $gavkat' \text{ IMPF} \Rightarrow gavk\text{-}nu\text{-}t' \text{ PFV}$
bark-INF bark-SX-INF
'(be) bark(ing)' 'bark [once]'

¹The word ‘pair’ in this study refers to two derivationally related verbs (an imperfective verb and an affixed perfective counterpart) and does not refer to ‘an aspectual pair’, in support of the hypothesis proposed by Isačenko (1960); Timberlake (2004).

²We use the following glossing abbreviations: INF (an infinitival form), PX (a prefix), and SX (a suffix).

The derivational pair in (1-a) consists of an underived (simplex or primary) imperfective *gavkat'* with no overt morphological marking of aspect and its perfective counterpart *progavkat'* derived with the perdurative prefix *pro-* ‘through’ which refers to temporal duration (e.g., Tolskaya, 2015; Naumov, 2019). (1-b) illustrates a minor derivational pattern where the semelfactive suffix *-nu-* is added to an imperfective simplex (or primary) imperfective verb *gavkat'* denoting a set of singular events or a plurality thereof, and derives a perfective verb *gavknut'* restricting its denotation to a set of singular events. Russian grammatical aspect, the derivational affixes by which perfective and imperfective verb forms are built, and their semantic, syntactic, and morphological properties, have been studied in theoretical linguistics (e.g., Filip, 2003 and references therein), corpus linguistics (e.g., Janda, 2007), computational linguistics (e.g., Drozd et al., 2015) and rule-based translation (Sonnenhauser and Zangenfeind, 2016). There is an emerging agreement that there are no dedicated markers of perfective aspect in Russian and other Slavic languages, which would consistently mark perfectivity of the verb in all their occurrences; while the semelfactive suffix *-nu-* consistently occurs on perfective verbs, it is a derivational morpheme that only delimits a minor derivational pattern (Filip, 2000, 2003, 2005).

Prefixes have a derivational function, occur on both perfective and imperfective verbs to which they often contribute additional lexical meanings and/or change argument structure (ibid.). For instance, the perfective verb *vyigrat'* ‘to win’ prefixed with the completive *vy-* can select the direct object *priz* ‘prize’ as in *vyigrat' priz* ‘to win a prize’, while its imperfective counterpart *igrat'* ‘to play’ cannot, cf. *igrat' *priz* ‘to play a prize’. As is well-known, prefixes often extend the core meaning of the base verb by adding a variety of modifications, such as spatial and temporal dimensions (largely due to the prepositional origin of most of them) and may also add affective connotations. Semantically speaking, prefixes can be uniformly analyzed as modifiers of eventuality types denoted by verb bases they are applied to (Filip, 2005). While most perfective verbs are morphologically complex, either prefixed or suffixed, as in the examples above, there are a few perfective root (or primary) verbs (e.g., *past'* ‘to fall’, *dat'* ‘to give’). All Slavic languages have morphologically complex secondary imperfectives derived from perfective verbs by the imperfectivizing suffix (realized in allomorphs *-yva-*, *-va-*, *-a-*, as in *perečityvat'* ‘to read over’), but they are not part of this study. When it comes to the semantic analysis of perfectivity and imperfectivity, many rely on Klein’s (1994) idea (within the Reichenbachian tradition) that grammatical aspect concerns the relationship in which: (i) event time is included within topic/reference time (perfective aspect), (ii) topic/reference time is within event time, or overlaps with it (imperfective aspect).

3 Experiment

3.1 Overview of methodology

The experiment for exploring whether the morphological form of Russian verbs can be exploited in automatic determination of their aspectual class was conducted as follows. First, we compiled a list of derivational pairs from existing databases. Second, the distribution of pre-trained word embeddings associated with these perfective and imperfective verbs was presented in a vector space (Mikolov et al., 2013), relying on the distributional hypothesis that linguistic items with similar meanings have similar distributions (Firth, 1957). Vector representations of derivational pairs were constructed using a fastText algorithm integrating with the Continuous Bag of Words architecture (CBOW), which predicts the target word according to its context represented by its n-grams. The visualization of word embeddings was done using the Distributed Stochastic Neighbor Embedding (t-SNE), an unsupervised clustering technique (van der Maaten and Hinton, 2008).

The fastText algorithm (Bojanowski et al., 2017) constructs a semantic vector space capturing semantic relations between words based on their formal similarity and their context similarity.³ It computes non-contextual word embeddings that are unique for each word regardless of context and do not change in downstream tasks (Si et al., 2019; Zhou et al., 2022). In a fastText model, the vector for a word is a sum of all vectors of its n-gram characters. This property enables fastText to achieve higher predictive performance for morphologically rich languages and rare words (Onan, 2020). t-SNE is used for visual verification of generated word embeddings by reducing dimension into a two-dimensional plane for

³Formal similarity is numerical representation of sub-word information, while context similarity is distance in vector space.

fastText embeddings (Sanjanasri et al., 2021) and has been applied to exploring Russian and Finnish inflectional paradigms (Chuang et al., 2023; Nikolaev et al., 2023). Nikolaev et al. use t-SNE for assessing the accuracy of clusters of inflected Finnish nouns in a vector space generated by fastText-based models. By using fastText and t-SNE, we expect the model to be able to separate perfective verbs from their imperfective counterparts by their morphology, rather than by context similarity. We visualize a vector space of derivational pairs to observe if it would exhibit distinct patterns in the form of clusters.

3.2 Derivational data

The data used in the experiment were collected and compiled from existing aspectual databases in Russian: the *Exploring Emptiness (EE)* database, the database of Russian Verbal Aspect (OSLIN database; Borik and Janssen, 2012), and the Essex Database of Russian Verbs and their Nominalizations (Essex database; Spencer and Zaretskaya, 1999).⁴ We compiled our database by extracting entries with derivational pairs of Russian verbs and the affixes by which they are related. We extracted 2899 entries from the Essex Database, 1981 entries from the *EE* database, and 529 entries from the OSLIN database. The data were then transliterated, sorted and cleaned from duplicates. It contains 4032 derivational pairs, as is illustrated in Table 1, 3976 verb forms related by prefixes, and 56 verb forms by the semelfactive suffix *-nu-*.

A sample illustrating the structure of the compiled database is given in Table 2. We see that the imperfective verbs *bajukat'* ‘to sing lullabies, to cradle’, *kapat'* ‘to drip’ and their perfective correspondents *ubajukat'* ‘to lull [to sleep]’ and *kapnut'* ‘to drop, to let fall a drop’ are followed by the type of affix by which they are related (prefix and suffix) and the specific affix form, here *u-* and *nu-*. The latter count reflects the well-known empirical fact that in Russian the number of prefixed perfective verbs is larger than that of verbs formed with the semelfactive suffix *-nu-*.

#Der.Prs.	#Prefixes	#Suffix
4032	3976	56

Table 1: Counts of derivational pairs, prefixes, and the suffix *-nu-* in the compiled database.

IMPF verb	PFV verb	Affix type	Affix
<i>bajukat'</i>	<i>ubajukat'</i>	prefix	<i>u</i>
<i>kapat'</i>	<i>kapnut'</i>	suffix	<i>nu</i>

Table 2: Examples of two entries in the compiled database.

Overall, the database contains about 700 imperfective verbs that have more than one corresponding perfective verb in the compiled database. Table 3 below illustrates an excerpt with a few imperfective verbs that have two or more perfective counterparts. For example, from the imperfective simplex verb

IMPF verb	PFV verb	Affix	#
<i>bespokoit'</i> ‘worry’	<i>obespokoit'</i> ‘trouble’, <i>pobespokoit'</i> ‘bother’	<i>o, po</i>	2
<i>bit'</i> ‘hit’	<i>pobit'</i> ‘beat up’, <i>probit'</i> ‘break through’, <i>razbit'</i> ‘break’	<i>po, pro</i>	3
<i>lepit'</i> ‘mould’	<i>zalepit'</i> ‘seal’, <i>vylepit'</i> ‘mould’, <i>nalepit'</i> ‘stick’, <i>slepit'</i> ‘sculpt’	<i>za, vy, na, s</i>	4
<i>dumat'</i> ‘think’	<i>podumat'</i> ‘think about’, <i>nadumat'</i> ‘imagine’, <i>pridumat'</i> ‘invent’, <i>obdumat'</i> ‘think through’, <i>razdumat'</i> ‘change one’s mind’	<i>po, na, pri, ob, raz</i>	5
<i>mazat'</i> ‘smear’	<i>pomazat'</i> ‘anoint’, <i>vymazat'</i> ‘cover’, <i>izmazat'</i> ‘stain’, <i>zamazat'</i> ‘cover up’, <i>namazat'</i> ‘spread’, <i>promazat'</i> ‘miss’	<i>po, vy, iz, za, na, pro</i>	6

Table 3: Examples of imperfective verbs with two to six prefixed perfective derivationally related verbs in the compiled database.

dumat' ‘to think’ we can derive five perfective verbs by means of five different prefixes, each with a different lexical meaning: *podumat'* ‘to think over’, *nadumat'* ‘to think, to imagine’, *obdumat'* ‘to think through’, *pridumat'* ‘to invent, to come up with’, *razdumat'* ‘to change one’s mind’.

⁴The databases are available at: http://emptyprefixes.uit.no/project_eng.htm, <http://ru.oslin.org>, <https://reshare.ukdataservice.ac.uk/852633/>.

3.3 Visualizing the vector space of Russian aspect

Representations of derivational pairs were constructed on the basis of pre-trained fastText embeddings provided by the RusVectores project. The RusVectores model (Kutuzov and Kuzmenko, 2017)⁵ was trained using the fastText algorithm with the Continuous Bag of Words (CBOW) architecture on the Araneum Russicum Maximum 2018 of 10 billion words from Russian texts crawled from Internet domains; its vocabulary contains 195,782 words with vector size of 300. The vector space representations of verb lemmas sorted by their tags (*impf* versus *pfv*) were then visualized with t-SNE. For the visualization, we used 4032 derivational pairs consisting of 3896 perfective verbs and 1766 imperfective verbs amounting to the total of 5662 verbs.⁶ Figure 1 illustrates the distributional space of perfective (*pfv*) and imperfective (*impf*) verbs visualized by t-SNE. The visualized space contains data points grouped by *pfv* and *impf* aspect forming distinctly separated clusters. The first major cluster to the left is formed by perfective verbs while the second major cluster to the right consists of imperfective verbs surrounded by some perfective verbs, although the right-side cluster at the bottom of Figure 1 is a mixture of imperfective and perfective verbs.

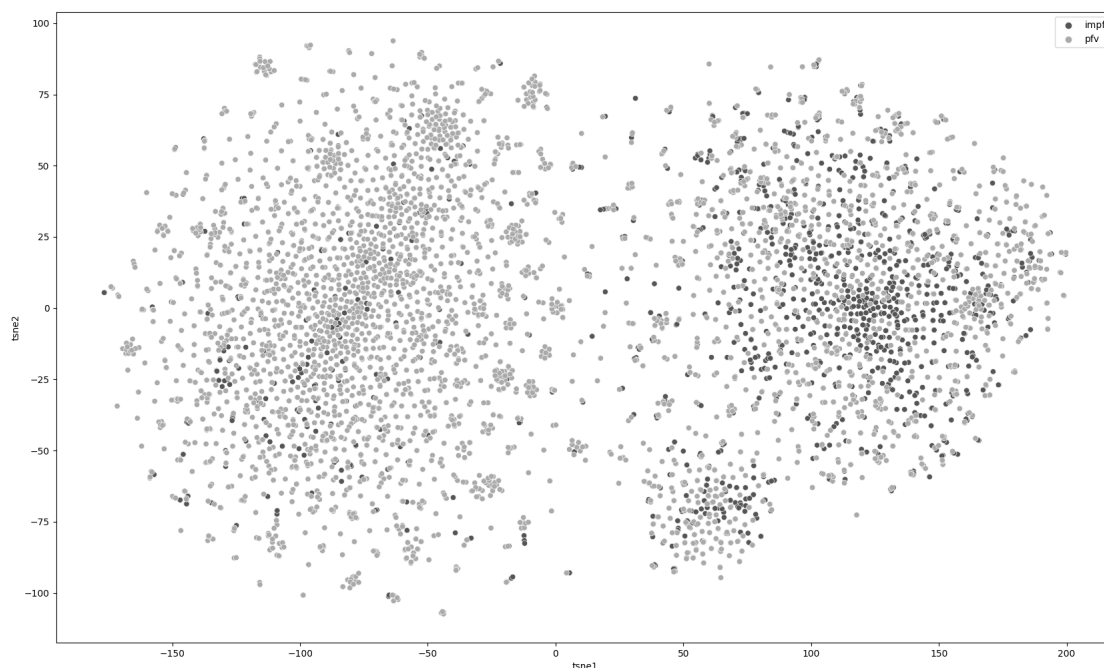


Figure 1: Scattered clusters of perfective (*pfv*; grey dots) and imperfective (*impf*; black dots) verbal lemmas based on high-dimensional vectors of word embeddings from the pre-trained RusVectores fastText model.

We observed quite a few perfective verbs scattered around the main right imperfective cluster, which led us to exploring whether these groupings were motivated by their semantic similarity, i.e., alignment with classes of verbs developed in accordance with their meanings and syntactic behavior by Levin (1993). Following Levin’s (1993) semantically coherent classification of verbs, the clusterings of perfective verbs (grey dots) in the main left and right clusters fall into the classes of verbs listed in Tables 4 and 5 below. As can be seen, the verb classes in both clusters are notionally quite diverse. Change of state verbs such as *obozlit'sja* ‘to become angry’, *obradovat'sja* ‘to become glad’, and *ozveret'* ‘to become engared’ also belong to the class of psychological state. Verbs of these classes occur both in the perfective and imperfective clusters.

The perfective verbs in the minor cluster on the bottom right side are mostly borrowed verbs derived by prefixation from *-ova-* imperfective verbs,⁷ and there are also some non-borrowed verbs. Levin (1993)

⁵araneum_none_fasttextcbow_300_5_2018, available at: <https://rusvectors.org/en/models/>

⁶In this experiment, we used verb lemmas as they were presented in the compiled derivational database.

⁷The borrowed verbs in this cluster are classified as perfective in the Reverse Dictionary of Russian (Ševeleva, 1974, p.

classifies these verbs as verbs of change of state (e.g., *otremontirovat'* ‘to repair’, *proventilirovat'* ‘to ventilate’), and creation and transformation (e.g., build verb *vygravirovat'* ‘to engrave’), among other cases.

Verb class	Example
Manner of speaking	<i>prokvakat'</i> ‘croak’
Measure (price)	<i>vyčislit'</i> ‘estimate’
Putting (fill)	<i>zamaskirovat'</i> ‘cover up’
Cutting	<i>rascarapat'</i> ‘scratch all over’
Psychological state	<i>zainteresovat'</i> ‘interest’
(amuse type)	<i>obesslavit'</i> ‘dishonor’
Social interaction	<i>izbalovat'</i> ‘pamper’
Separating/Disassembling	<i>razdelit'</i> ‘split, separate’
Change of state	<i>obozlit'sja</i> ‘become angry’ <i>prixtvornut'</i> ‘get [a bit] sick’

Table 4: Examples of Levin’s (1993) verb classes for perfective verbs in the *pfv* cluster.

Verb class	Example
Psychological state	<i>pozavidovat'</i> ‘envy’
Desire	<i>vozvzemat'</i> ‘desire’
Social interaction	<i>podrat'sja</i> ‘get into a fight’
Gestures w/ body parts	<i>mignut'</i> ‘wink [once]’
Negative judgment	<i>nakazat'</i> ‘punish’
Change of state	<i>poburet'</i> ‘turn brown’ <i>obradovat'sja</i> ‘become glad’ <i>ozveret'</i> ‘become engared’
Contact by impact	<i>oblobyzzat'</i> ‘kiss’ <i>užalit'</i> ‘sting’ <i>pokusat'</i> ‘bite’

Table 5: Examples of Levin’s (1993) verb classes for perfective verbs in the *impf* cluster.

Change of state verbs are also known as verbs with ‘affected objects’ and creation, contact by impact verbs as ‘effected objects’, both types of objects are traditionally subsumed under the Patient thematic relation. Hence, in so far as these two classes of verbs denote eventualities during the course of which the referents of their direct object arguments undergo some change, they are semantically similar.

3.4 Error analysis

To analyze the errors produced by the RusVectores model, we considered the properties of perfective and imperfective verbs connected to their contextual use. This is why we examined the distributional representations of the verbs based on their corpus frequencies and bi-aspectual uses. First, we labeled perfective verbs with their frequency ranks (high- versus low frequency) based on their corpus frequencies and visualized the distribution via t-SNE. Second, we checked for bi-aspectual perfective/imperfective verbs in the minor right-side cluster as many borrowed verbs were observed in this cluster in Section 3.3.

We used corpus frequencies to analyze the RusVectores’ errors because they affect similarity scores in word embedding. That is, the model would perform better with verbs that have higher frequencies than with verbs with lower frequencies. For example, for the BOW architecture models with increasing frequency counts, similarity score increases generating the same vector for different sentences disregarding context and word order (Asudani et al., 2023). For the list of word list of all perfective verbs, we extracted their raw corpus frequencies from the *Araneum Russicum III Maximum 2019* corpus⁸ using the NoSketch Engine corpus query tool (Rychlý, 2007; Kilgarriff et al., 2014). The frequencies of these verbs were normalized and log-transformed on the scale from one to seven using the Zipf measure proposed by (van Heuven et al., 2014).⁹ This measure converts normalized (item per million) frequencies into more understandable values on the scale from 1 to 7; the values from 1 to 3 are associated with low-frequency words while the ones from 4 to 7, with high-frequency verbs (van Heuven et al., 2014, 1180).

Figure 2 shows that low-frequency verbs are on the left-side cluster while high-frequency verbs are on the right-side cluster (left-side perfective and right-side imperfective clusters on Figure 1, respectively).

611, 607, 604). Their imperfective unprefixed counterparts, *remontirovat'* ‘to repair’, *ventilirovat'* ‘to ventilate’, *gravirovat'* ‘to engrave’, are borrowings and integrated into Russian by means of the *-ova-* suffix. There is no general agreement whether the borrowed unprefixed verbs are imperfective or biaspectual. Their prefixed counterparts are often taken to be perfective (Horiguchi, 2018, 62; Bunčić, 2013, cited in Olsson, 2018), but some treat many of them as biaspectual (Schuler, 1996 and Horiguchi, 2018).

⁸The corpus is based on texts crawled from the Russian Web (more than 19 billion tokens). This is the newer version of the corpus the RusVectores model was pre-trained on. The corpus description is available at: http://aranea.juls.savba.sk/aranea_about/index.html.

⁹For example, for *posmuglet'* ‘to become tanned’ the raw frequency of 4 is normalized, log-transformed (base 10) to -1.056, and scaled to 2, which represents low-frequency rank. For *posmet'* ‘to dare’, the raw frequency of 46899 is normalized, log-transformed to 2.915, and scaled to 6 representing the rank of high-frequency verbs.

High-frequency perfective verbs ranked 5–6 found in the *highFreq* cluster are, for example, *obnaglet'* ‘to become arrogant’, *razbudit'* ‘to wake [someone] up’, *mignut'* ‘to wink [once]’, *užalit'* ‘to sting’, *počtit'* ‘to commemorate’. Some perfective low frequency verbs with ranks 2–3 from the *lowFreq* cluster include *sfantazirovat'* ‘to fantasize’ (rank 3), *srepetirovat'* ‘to rehearse’ (rank 2), *zatorcevat'* ‘to pave with wood blocks’ (rank 2). The vector-space distribution of high- and low-frequency perfective verbs implies that the RusVectores model may show a bias in separating to low- and high-frequency verbs. In general, the model separated perfective and imperfective verbs according to low- versus high frequency (same as in Figure 2), but in this section we only addressed the distributional representations related to the frequency of perfective verbs.

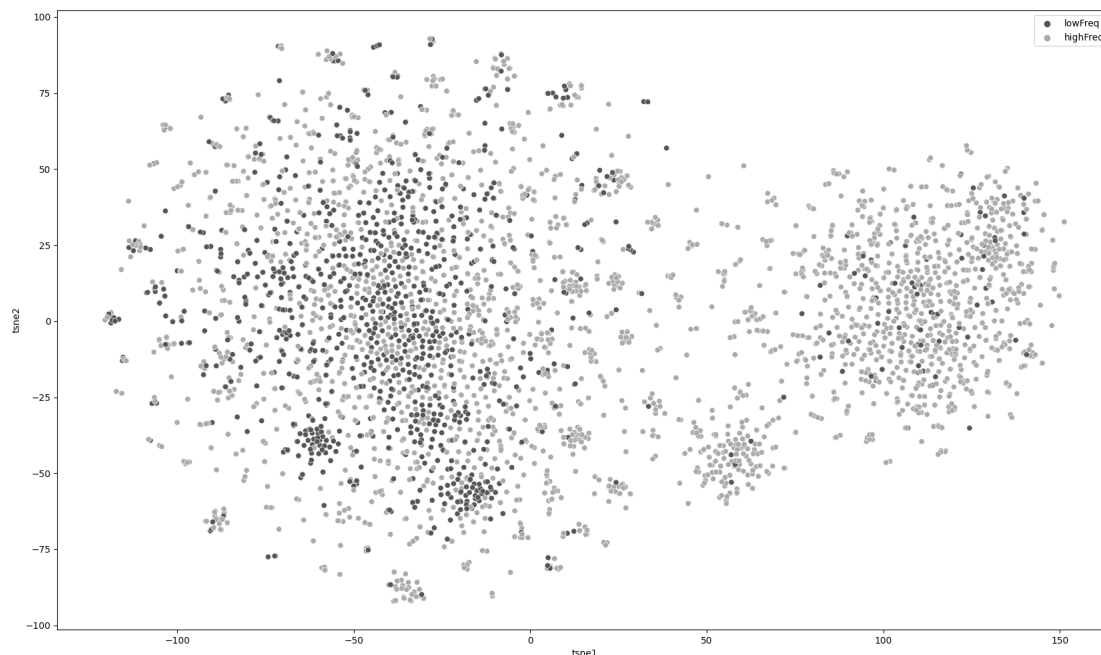


Figure 2: Scattered clusters of high frequency (*highFreq*; grey dots) and low frequency (*lowFreq*; black dots) perfective lemmas based on high-dimensional vectors of word embeddings from the pre-trained RusVectores fastText model. The frequency ranks of these verbs are based on the raw frequency values from the Araneum Russicum corpus log-transformed and scaled by means of the Zipf measure.

As mentioned in Section 3.3, we identified a pattern of borrowed *-ova-* verbs in the minor right-side cluster. We (manually) extracted 221 perfective and imperfective verbs from this cluster and checked if they were bi-aspectual according to the annotation in the verb database based on Zaliznyak’s dictionary (1987).¹⁰ Out of 221 verbs that we analyzed, 49 were marked as bi-aspectual *-ova-* verbs in the verb database, 47 of which were borrowed, and only two non-borrowed (e.g., *zaimstvovat'* ‘to borrow’, *usoveršensstvovat'sja* ‘to improve oneself’). The bi-aspectual borrowed verbs include *kristallizovat'* ‘to crystallize’, *modelirovat'* ‘to model’, *transkribirovat'* ‘to transcribe’, *kooperirovat'* ‘to cooperate’, *orientirovat'* ‘to orientate’, *degustirovat'* ‘to taste’, among others. Many perfective verbs had close distributional properties with their bi-aspectual counterparts and were placed close to each other in the minor cluster. The examples of such derivational pairs (biaspectual–perfective) include *orientirovat'*–*sorientirovat'* ‘to walk someone through’, *degustirovat'*–*prodegustirovat'* ‘to taste’, *zaimstvovat'*–*pozaimstvovat'* ‘to borrow’. It should be noted we observed few *-ova-* verbs in the main right-side cluster compared to the the minor cluster that contained predominantly *-ova-* bi-aspectual verbs. The verbs in the main right-side cluster were mostly perfective and imperfective verbs with stems ending with theme vowels *-e-* (as in *teret'* ‘to rub’), *-a-* (as in *pačkat'* ‘to make dirty’), *-i-* (as in *sverlit'* ‘to drill’).

We may speculate that the clustering that we observe reflects similarities in the distributional properties

¹⁰The database was compiled by Slioussar (2012) based on the grammatical dictionary of Russian (Zaliznyak, 1987) and contains 27409 verbs. Available at: <http://www.slioussar.ru/verbdatabase.html>

of verbs, rather than similarities in their morphological form. It is possible that these are the cases where fastText generated vector representations based more on context similarity than form similarity, which in turn might be due to high frequency of verbal lemmas and the use of biaspectual verbs and their perfective counterparts in similar contexts.

4 Results and discussion

The experiment showed that the distributional model separated perfective and imperfective verbs into two distinct clusters. As the model was built by the non-contextual fastText method, this confirmed our hypothesis that the morphological structure of Russian verbs should be a significant criterion for distinguishing the grammatical aspect of Russian verbs. The semantic examination of clusterings of perfective verbs based on Levin's (1993) classification of verbs revealed diverse semantic classes of these verbs. Although the clusterings of these verbs may have been based on context similarity of these perfective verbs, and therefore have similar lexical meanings, they do not seem to constitute coherent systematic lexical semantic classes.

The error analysis revealed that the RusVectores model had a bias towards corpus frequency of perfective verbs. This resulted in highly frequent verbs being placed in the right-side imperfective cluster, while low-frequency verbs were placed in the perfective left-side cluster. Verbs with higher frequency are likely to have higher similarity scores and cluster with imperfective base verbs. The error analysis also confirmed that the minor cluster consisted mostly of borrowed *-ova-* verbs including bi-aspectual verbs. These observations suggest that perfective verbs tended to cluster with their respective bi-aspectual verbs due to their context similarity as both perfective and bi-aspectual verbs would appear in similar contexts. That is, perfective *-ova-* verbs are semantically closer to their biaspectual counterparts.

For the future work, it would be relevant to assess how semantically close the members of a derivational pair are, and to explain the clusters of perfective verbs observed around the *imprf* cluster. We would use similarity measures (e.g., cosine similarity) to compute how similar perfective and imperfective verbs are to each other based on their distributional representations. We could also carry out a logistic regression analysis to identify which factor(s), morphological/semantic properties or their interaction, predict(s) best the grammatical aspect of the members of the derivational pair. Cosine similarity scores, morphological properties, semantic classes could serve as input to the analysis.

References

- Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. 2023. [Impact of word embedding models on text analytics in deep learning environment: a review](https://doi.org/10.1007/s10462-023-10419-1). *Artificial Intelligence Review* 56(9):10345–10425. <https://doi.org/10.1007/s10462-023-10419-1>.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5:135–146.
- Olga Borik and Maarten Janssen. 2012. A database of Russian verbal aspect. In *Proceedings of the conference Russian Verb, St. Petersburg, Russia*.
- Daniel Bunčić. 2013. Biaspektuelle verben als polyseme: Über homonymie, aspektneutralität und die konative Lesart. *Die Welt der Slaven* 58(1):36–53.
- Yu-Ying Chuang, Dunstan Brown, Harald R. Baayen, and Roger Paul Evans. 2023. Paradigm gaps are associated with weird “distributional semantics” properties: Russian defective nouns and their case and number paradigms. *The Mental Lexicon*.
- Östen Dahl. 1985. *Tense and Aspect Systems*. Blackwell, Oxford.
- Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2015. Discovering Aspectual Classes of Russian Verbs in Untagged Large Corpora. In *2015 IEEE International Conference on Data Science and Data Intensive Systems*. pages 61–68.
- Hana Filip. 1993/1999. *Aspect, situation types and nominal reference*. Ph.D. Thesis, University of California at Berkeley. Garland, New York/London.

- Hana Filip. 2000. The Quantization Puzzle. *Events as grammatical objects, from the combined perspectives of lexical semantics, logical semantics and syntax* 39:39–91.
- Hana Filip. 2003. Prefixes and the Delimitation of Events. *Journal of Slavic linguistics* 11(1):55–101.
- Hana Filip. 2005. On accumulating and having it all: Perfectivity, prefixes and bare arguments. *Perspectives on Aspect. Studies in Theoretical Psycholinguistics* 32:125–148.
- John Rupert Firth. 1957. A Synopsis of Linguistic Theory, 1930–55. *Studies in Linguistic Analysis* Special Volume of the Philological Society:1–31.
- Daiki Horiguchi. 2018. Imperfectivization of borrowed verbs in Russian. *Russian Linguistics* 42:345–356.
- Aleksandr Vasil’evič Isačenko. 1960. *Grammatičeskij stroj russkogo jazyka v sopostavlenii s slovackim – Čast’vtoraja: morfologija* [Grammatical system in Russian as opposed to Slovak. Part 2: Morphology]. Izdatel’stvo akademii nauk, Bratislava.
- Laura A Janda. 2007. Aspectual clusters of Russian verbs. *Studies in Language* 31(3):607–648.
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1):7–36.
- Wolfgang Klein. 1994. *Time in Language*. Routledge, London.
- Andrey Kutuzov and Elizaveta Kuzmenko. 2017. *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*, Springer International Publishing, Cham, pages 155–161. http://dx.doi.org/10.1007/978-3-319-52920-2_15.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.
- Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*.
- Ilya Naumov. 2019. Constraining the distribution of the perdurative in Russian. *Advances in formal Slavic linguistics 2017* 3:205.
- Alexandre Nikolaev, Yu-Ying Chuang, and Harald R Baayen. 2023. A generating model for Finnish nominal inflection using distributional semantics. *The Mental Lexicon*.
- Gustaf Olsson. 2018. The Formation of Aspectual Pairs of Borrowed *ova*-verbs in Russian. *Scando-Slavica* 64(2):228–242.
- Aytuğ Onan. 2020. Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education* 28(1):117–138.
- Pavel Rychlý. 2007. Manatee/Bonito—A Modular Corpus Manager. In *RASLAN*. pages 65–70.
- JP Sanjanasri, Vijay Krishna Menon, Soman KP, Rajendran S, and Agnieszka Wolk. 2021. [Generation of Cross-Lingual Word Vectors for Low-Resourced Languages Using Deep Learning and Topological Metrics in a Data-Efficient way](https://doi.org/10.3390/electronics10121372). *Electronics* 10(12). <https://doi.org/10.3390/electronics10121372>.
- Neikirk Schuler. 1996. *From adaptation to nativization: a synchronic analysis of the category of aspect in borrowed verbs in Russian, Bulgarian and Macedonian*. Ph.D. thesis, The Ohio State University.
- Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association* 26(11):1297–1304.
- Natalia Slioussar. 2012. Nekotorye svedeniya o formoobrazovatel’nyx klassax russkix glagolov [Some information on Russian verb classes]. Ms., Utrecht institute of Linguistics OTS and Saint Petersburg State University.
- Barbara Sonnenhauser and Robert Zangenfeind. 2016. Not by chance. Russian aspect in rule-based machine translation. *Russian Linguistics* pages 199–213.
- Andrew Spencer and Marina Zaretskaya. 1999. *The Essex database of Russian verbs and their nominalizations*. Department of Language and Linguistics, University of Essex.
- Alan Timberlake. 2004. *A Reference Grammar of Russian*. Cambridge University Press.

- Inna K. Tolskaya. 2015. Verbal prefixes in Russian: Conceptual structure versus syntax. *Journal of linguistics* 51(1):213–243.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing Data using t-SNE](http://jmlr.org/papers/v9/vandermaaten08a.html). *Journal of Machine Learning Research* 9(86):2579–2605. <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology* 67(6):1176–1190.
- M. S. Ševeleva. 1974. *Obratnyj slovar'russskogo jazyka [The Reverse Dictionary of Russian]*. Sovetskaja Enciklopedija.
- Björn Wiemer and Ilja A Seržant. 2017. Diachrony and typology of Slavic aspect: What does morphology tell us? *Unity and diversity in grammaticalization scenarios* 16:239.
- Andrey Anatolyevich Zaliznyak. 1987. *Grammatical Dictionary of the Russian Language*. Moscow, Russia .
- Guangyao Zhou, Jingyi Cheng, and Flavius Frasincar. 2022. A hybrid approach for aspect-based sentiment analysis using a double rotatory attention model. *International Journal of Web Engineering and Technology* 17(1):3–28.

Identification of root morphs in morphologically segmented data

Vojtěch John and **Magda Ševčíková** and **Zdeněk Žabokrtský**

Charles University, Faculty of Mathematics and Physics,

Institute of Formal and Applied Linguistics

Malostranské náměstí 25, 118 00 Praha, Czech Republic

Abstract

As a result of the ongoing push for unification, extension and integration of morphological resources, need arises for reliable low-resource morph classification, especially root identification. The paper reports on our experiments with multiple root identification methods with various degrees of supervision, tested on several Indo-European languages, showing, among others, that given morphological segmentation, surprisingly good root identification can be achieved using simple unsupervised statistical methods, the main bottlenecks being compounding and homomorphy resolution.

1 Introduction

The recent push for cross-lingual unification of morphological resources has, among others, brought about the unification of various resources devoted to morphological segmentation (Batsuren et al., 2022b; Žabokrtský et al., 2022), i.e. the task of dividing words into the smallest meaning-bearing units (morphemes or morphs), as well as the closely connected task of morphological classification – dividing the morphemes to classes (of various granularity). Nevertheless, in the available resources, the overall quality and/or completeness of morphological segmentation tends to be higher than that of the morphological classification, or the classification is even missing completely. This is reinforced by the fact that the state-of-the-art morphological segmentation approaches (as witnessed by the 2022 SIGMORPHON shared task; Batsuren et al. 2022a) are based on neural networks and neither include morphological classification nor can be straightforwardly used for obtaining it (even though there are some promising exceptions; e.g. Bolshakova and Sapin 2021, who use neural networks for both morphological segmentation and classification with word-level accuracy of over 90 %). As a result, morphological segmentation of reasonable quality is often easier to obtain than the corresponding morphological classification.

Furthermore, as the tasks of morphological segmentation and classification are closely connected to derivational morphology and as the derivational resources for a given language often contain quite different lexical material than that of the segmentation resources, the degree of their mutual transferability poses an interesting problem. There have been attempts to use derivational trees for obtaining morphological segmentation together with very coarse-grained classification (Bodnár et al., 2020). For an approach using segmented words to build derivational trees, the natural first step seems to be automated morph classification, especially root identification on the pre-segmented data (intuitively, it seems that we could build derivation trees from segmented words using morph classification combined with homonymy and allomorphy resolution); the methods used for root identification would be preferably as little supervised as possible, to minimize requirements on the resources.

The present paper starts with a brief introduction of basic terminology (Section 2) and with an overview of data sources and experiments related to our task (Section 3). Section 4 reports on our experiments with multiple root identification methods with various degrees of supervision, tested on several Indo-European languages. The results analysed in Section 5 document that surprisingly good root identification can be achieved using simple unsupervised statistical methods. Concluding remarks and some ideas for future work are sketched in Section 6.

2 Theoretical background

Although *morpheme* and *morph* are traditional notions belonging to the core linguistic terminology, their definitions vary in the literature. In the present paper, along the lines of [Haspelmath \(2020, p. 117\)](#), *morph* is understood as “a minimal pairing of syntacticosemantic content and a string of phonological segments” and considered as the basic unit of morphological analysis. Morphs are smaller than words (cf. three morphs in *play+er+s*), or identical with them (e.g. *chair* consisting solely of a root morph). Morphs repeat across sets of words, with certain (so-called, *cranberry*) morphs being the exception ([Aronoff, 1976](#)). As morphs are the basic building blocks in inflection and in word-formation processes, they may appear in multiple formal variants in different contexts (allomorphy); cf. the root allomorphs *sheep* and *shep* in the nouns *sheep* and *shep+herd*. *Vice versa*, a particular form can convey different meanings; cf. homonymy of both the root and the inflectional marker in the noun *bear+s* and the verb *bear+s*. In general, words are expected to be fully decomposable into morphs. In the present paper, this task is called *morphological segmentation*, but alternative names are also used (morphemic segmentation, morphemic analysis, etc.).

A *root* morph conveys lexical meaning. Other morphs, if present in the word’s structure, are classified with respect to the root: the root is preceded by one or more *prefixes* (*re-* in *re+play*) and followed by one or more *suffixes* (*-er* in *play+er*); a final suffix that expresses inflectional categories (*-s* in *play+er+s*) can be distinguished by the term *ending*. In words with multiple roots (compounds), *interfixes* are often used to link the roots (*-s-* in the German noun *Arbeit+s+amt* ‘employment office’). In this paper, the task of morph classification is limited to the identification of roots.

The experiments are carried out on seven languages for which morphologically segmented and annotated data are available. Despite the high quality of the data, it should be kept in mind that the segmentation recorded in the data is not always uncontroversial. It depends on the granularity of the analysis, the inclusion of etymological aspects, and other criteria. Similarly, the classification as available in the sources documents that the categories distinguished in theory are sometimes difficult to apply to authentic data. There are always cases in the data that do not fully fit either category and require a decision to be made. One such example is neoclassical formations, which are debated either as multi-root words (compounds), or single-root words where the root is preceded by a prefix(oid) or followed by a suffix(oid). Consistent decision-making is a challenge when annotating individual sources, even more so across sources from different languages. See the classification of morphemes in German verbs and other examples in the error analysis in Section 5.2.

3 Related work

3.1 Data resources

There are several relevant types of data resources, both mono- and multilingual. Instead of enumerating the resources for all the included languages individually, in the following survey we will concentrate on the unified multilingual databases. The corresponding papers usually provide a useful guide to the monolingual resources included in the given project.

First of all, there are morphological segmentation databases. These vary in quality. Some of them, like the multilingual derivational and inflectional database MorphyNet ([Batsuren et al., 2021](#)), are automatically or semi-automatically generated, so they cannot be used as gold data (at least once the accuracy of the classification methods is close to the accuracy of the provided segmentation). Universal Segmentations (UniSegments; [Žabokrtský et al. 2022](#)) is a multilingual collection of language resources containing morphological segmentation. The resources differ in several important respects, including origin (manually or automatically annotated) as well as the presence and granularity of morphological annotation.

Closely connected to (or even overlapping with) these are multilingual morphological lexicons. The largest unification effort to date, the UniMorph project ([Batsuren et al., 2022b](#)), contains in its latest release both morphological segmentation and morphological classification for at least 16 languages. Nevertheless, the segmentation is sometimes dubious or incompatible with our approach to morphological

classification. Thus, for instance, in the Czech data, lemmas of unmotivated words (represented as root nodes in derivational trees) are used instead of root morphs,¹ while in the German data, the words are segmented to morphemes (in the canonical form), not to morphs.

Finally, derivational networks, grouping words that come from the same derivational root, can be used for distinguishing root morphs and derivational affixes. Furthermore, several of these already contain morphological segmentation and classification. Universal Derivations (UDer; Kyjánek et al. 2020) is a multilingual collection of derivational resources, unified to the form of collections of derivational trees. That is, the words are organized in rooted tree structures with the edges representing the derivational relation (*child node* was derived from *parent node*). There is a relevant overlap between resources included in UDer and UniSegments, as some of the derivational resources also contain information relevant to morphological segmentation and classification.

3.2 Morphological segmentation and classification

The methods used for morphological classification vary according to both the quantity and quality of required data; as a rule, the more information is included in the data, the less data is needed. For languages with large and rich resources like Russian, both morphological segmentation and morphological classification can be approached using neural networks (Bolshakova and Sapin, 2021). Even for morphological classification of underresourced languages like Uspaneko (Ginn and Palmer, 2023) or Lezgi (Moeller and Hulden, 2018), neural models have been used with considerable success (around 80 or 90 % accuracy), especially given the very fine-grained tagset. It is to be noted, however, that in the case of Lezgi, where the authors performed both segmentation and classification using both neural network and CRF classifier, the CRF classifier proved to be more successful than the employed neural seq2seq network.

Even though the morph classification as such has not been much concentrated upon, it often appears as a subtask or byproduct of other tasks. Thus Goldsmith (2001) combines minimum description length with several heuristics to get candidate stems and suffixes, while Schone and Jurafsky (2001), or more recently Soricut and Och (2015) induce morphological rules using automatically extracted affixes. Strongly related to morphological classification is interlinear glossing. This task consists in finding morphological glosses (i.e. lexical meaning in the target language and/or morphological categories expressed by the morph), given a morphologically segmented text in a source language and its translation in a target language. Although the current approaches dealing with low-resource languages (Zhao et al., 2020) or CRF (McMillan-Major, 2020) yield interesting results, even there a significant amount of input data with very fine-grained annotation is needed to achieve reasonable accuracy.

4 Experiments

4.1 Data

In our choice of test languages, we were limited primarily by the quality and accessibility of morphological resources for individual languages. The quality of the segmentation resources is very important for the reliability of our results as we will obtain our test data from them. We have therefore selected the languages for which there exist manually segmented and annotated resources included in the UniSegments 1.0 project (Žabokrtský et al., 2022), and we added Czech, for which we have our own manually annotated data. Further, in some of our semi-supervised methods, we use derivational trees. As they are used more or less as a basis for heuristics, there is no need to shun automatically generated data. We have therefore used the derivational resources available in the Universal Derivations project (Kyjánek et al., 2020). The resources are listed in Table 1.

4.2 Methods

As baselines, we have used three simple statistical heuristics. Firstly, we take as roots all the longest morphs of the words (*MaxLen*). In the following methods, if not explicitly described otherwise, if two morphs gain the same score (which should happen very rarely), we pick the first of them. Secondly,

¹Czech lemmas are rarely simplex, monomorphemic words, because even unmotivated words can contain mandatory inflectional affixes.

resource	included in	language	size	lemmas x words	morphs	root tokens	root morphs
CroDeriV	UDer, USeg	Croatian	15 657	Lemmas	65 455	15 819	4 569
MorphoLex	USeg	English	68 624	Words	151 960	77 308	20 153
MorphoLex-FR	USeg	French	15 954	Words	29 087	16 290	11 085
CELEX	USeg	German	51 728	Lemmas	118 920	69 457	16 749
KuznetsEfrimDict	USeg	Russian	73 447	Lemmas	318 647	86 726	6 912
DerIvaTario	UDer, USeg	Italian	10 991	Lemmas	31 246	10 991	5 566
Czech	–	Czech	10 438	Lemmas	40 155	10 438	1 985
Démonette	UDer	French	22 060				
CatVar	UDer	English	82 675				
DeriNetRU	UDer	Russian	337 632				
DeriNet	UDer	Czech	1 027 665				

Table 1: Morphological resources for segmentation and derivation used in our experiments. The Czech data we use are included neither in UDer nor in USeg; information about the structure of the data is included only for the gold segmentation data, not for the derivational tree databases

we label as root morph the morph with the fewest occurrences in the dictionary of segmented lemmas or words (*MinFreq*). This is motivated by the hypothesis that in most of the languages homomorphy between root and non-root morphs is unusual and there is only a limited number of affixes but a large number of root morphs. Thus, in the dictionary, roots will appear in conjunction with the affixes (which are few), and therefore not as often as the affixes, which will appear in conjunction with (many) roots. As our third baseline solution (*MinNeighborEntropy*), following a similar observation, namely that the root morphs predict their neighbouring morphs much better than the affixes, we compute for each morph in the dictionary the entropies of distributions of left and right neighbouring morphs. We then mark as root the morphs with the smallest maximum of the two entropies. It should be noted, however, that the last observation is not self-evident; it would not hold in cases when there is more than one compulsory suffix (or prefix) and when some of the affixes are always surrounded by other affixes. Fourthly (*UnweightedMix*), we combine the first three heuristics in an unweighted way (using the inverse value when required) and use the resulting score. Almost all the above-mentioned methods (except for *MaxLen*) are severely limited by the fact that they can select at most one root morph per word. We have therefore in our last fully unsupervised solution *ProbabMix* used normalized morph scores from the last heuristics *UnweightedMix*, obtaining a probability distribution, subsequently averaging the probabilities (of given morph being root) across the data. Then, we select as root morphs all the morphs achieving at least 5 % probability.²

As the second section of our experiment, we use the information contained in the UDer derivational databases. We have experimented with two approaches: Firstly, we computed the edit distance between each morph and the root of the derivational tree of the current word and all its child nodes, either by itself *DerivRoot* or in combination with the previous three unsupervised heuristics *DerivRoot + UnweightedMix*. Secondly, in the *LongestInDerivTree* method, for some of the languages, we used all the words in the tree to get a rough approximation of the root by finding the longest common part of the words (including a “?” wildcard to partially handle allomorphy).

Finally, for comparison with the supervised methods, we have trained a CRF tagger as implemented in the nltk package (Bird, 2006), on training data from UniSegments; that is, we treated the segmented words as sentences and the morphs as tagged words (with only roots and non-roots being distinguished as the tagset categories).

5 Evaluation

5.1 Evaluation methods

For our experiments, we have used data in Croatian (Table 2), German (Table 3), English (Table 4), Italian (Table 5), Russian (Table 6), French (Table 7), and Czech (Table 8). We have run our experiments on 5 000 randomly selected segmented words from each of the languages; for the only supervised method,

²The hyperparameters were selected arbitrarily and could probably be improved, given large-enough development data; that would, nevertheless, change the setting from unsupervised to (semi-)supervised

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	98.6 %	100 %	99.3 %	99.5 %
MaxLen	86.0 %	91.9 %	97.2 %	93.4 %
MinFreq	97.3 %	98.7 %	98.0 %	98.3 %
MinNeighborEntropy	97.1 %	98.5 %	97.8 %	98.0 %
UnweightedMix	96.7 %	98.1 %	97.4 %	97.7 %
ProbabMix	91.9 %	95.8 %	99.0 %	96.8 %
DerivTree	95.8 %	97.2 %	96.5 %	96.7 %
DerivTree + UnweightedMix	97.1 %	98.5 %	97.8 %	98.1 %
LongestInDerivTree	97.3 %	98.7 %	98.0 %	98.2 %
CRF tagger	98.3 %	98.7 %	99.1 %	98.8 %

Table 2: Croatian

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	57.5 %	100 %	78.2 %	85.3 %
MaxLen	59.6 %	94.5 %	80.3 %	84.1 %
MinFreq	55.7 %	97.6 %	76.1 %	83.2 %
MaxNeighborEntropy	55.7 %	97.6 %	76.1 %	83.1 %
UnweightedMix	55.8 %	97.7 %	76.3 %	83.3 %
ProbabMix	83.4 %	97.0 %	92.7 %	93.7 %
DerivTree	55.7 %	97.7 %	76.2 %	83.2 %
DerivTree + UnweightedMix	55.9 %	97.8 %	73.4 %	83.4 %
CRF tagger	92.2 %	97.3 %	98.0 %	97.1 %

Table 3: German

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	87.8 %	100 %	93.9 %	95.9 %
MaxLen	84.2 %	95.2 %	93.5 %	93.2 %
MinFreq	85.3 %	97.3 %	91.3 %	93.3 %
MinNeighborEntropy	85.4 %	97.4 %	91.4 %	93.4 %
UnweightedMix	85.6 %	97.7 %	91.6 %	93.6 %
ProbabMix	91.0 %	97.3 %	96.5 %	96.3 %
DerivTree	85.2 %	97.2 %	91.2 %	93.2 %
DerivTree + UnweightedMix	85.5 %	97.5 %	91.5 %	93.5 %
CRF tagger	94.0 %	97.7 %	97.6 %	97.2 %

Table 4: English

the CRF tagger, we have additionally selected another set of 5 000 words as training data. The sizes of the train and test sets were selected so that all the methods can be tested on the same data and (for the supervised method) the size of training data is the same for all the methods (as for the unsupervised methods, the test set is the train set). Since many of our methods only select the best candidate for the root (all apart from the *CRF Tagger*, *MaxLen* and *ProbabMix*), we have also run an oracle experiment (*OracleOneRoot*), selecting at most one root morph for each word.

We use four evaluation metrics, one on the word-level (accuracy) and three on the morph level (resp. root-level): precision, recall, and F-measure, averaged over the words (so that every word has the same weight). For the morph-level metrics, we formulate the task rather as root identification than morph classification to gain a rough error analysis. Thus, for most of the languages, for instance, precision significantly higher than recall would mean that most of the errors were false negatives; i.e. a root was identified incorrectly as a non-root.

5.2 Error analysis

In the evaluation, we take the test data at the face value. Nevertheless, it should be noted that some of the measured errors might be actually due to errors in the data. Firstly, the provided segmentation might be incorrect. For example, the German data contain the word *übersichtlich* segmented erroneously as *über+sich+tllich*; this caused wrong classification of the morph *-tllich* as root by the *MinFreq* baseline, as the erroneous morph appears very infrequently in the data. Second, the errors might be caused by (seemingly) arbitrary decisions in the morph classification in the data. For example, the German data,

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	100 %	100 %	100 %	100 %
MaxLen	67.7 %	75.8 %	84.2 %	78.5 %
MinFreq	97.5 %	97.5 %	97.5 %	97.5 %
MinNeighborEntropy	96.8 %	96.8 %	96.8 %	96.8 %
UnweightedMix	96.6 %	96.7 %	96.7 %	96.7 %
ProbabMix	90.8 %	94.4 %	98.0 %	95.6 %
DerivTree	87.7 %	87.7 %	87.7 %	87.7 %
DerivTree + UnweightedMix	96.1 %	96.1 %	96.1 %	96.1 %
CRF tagger	96.2 %	97.1 %	97.9 %	97.3 %

Table 5: Italian

Russian	Word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	82.5 %	100 %	91.1 %	94.1 %
MaxLen	60.6 %	81.2 %	85.6 %	80.4 %
MinFreq	76.4 %	93.2 %	84.7 %	87.5 %
MinNeighborEntropy	74.9 %	91.7 %	83.2 %	86.0 %
UnweightedMix	76.9 %	93.8 %	85.3 %	88.1 %
ProbabMix	80.1 %	92.0 %	94.8 %	92.0 %
DerivTree	72.3 %	88.8 %	80.5 %	83.2 %
DerivTree + UnweightedMix	78.1 %	94.9 %	86.4 %	89.2 %
CRF tagger	90.2 %	96.0 %	95.2 %	95.0 %

Table 6: Russian

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	97.8 %	100 %	98.9 %	99.2 %
MaxLen	87.2 %	92.0 %	94.0 %	92.4 %
MinFreq	94.6 %	96.8 %	95.7 %	96.1 %
MinNeighborEntropy	94.7 %	96.9 %	95.8 %	96.2 %
UnweightedMix	94.7 %	96.9 %	95.8 %	96.1 %
ProbabMix	92.9 %	96.5 %	97.8 %	96.7 %
DerivTree	94.6 %	96.8 %	95.7 %	96.0 %
DerivTree + UnweightedMix	94.8 %	97.0 %	95.9 %	96.2 %
LongestInDerivTree	94.8 %	97.0 %	95.9 %	96.3 %
CRF tagger	94.4 %	97.0 %	96.8 %	96.7 %

Table 7: French

method	word-level accuracy	average precision	average recall	average F-measure
OracleOneRoot	100 %	100 %	100 %	100 %
MaxLen	76.1 %	86.1 %	97.4 %	89.7 %
MinFreq	96.1 %	96.1 %	96.1 %	96.1 %
MinNeighborEntropy	96.1 %	96.1 %	96.1 %	96.1 %
UnweightedMix	97.2 %	97.2 %	97.2 %	97.2 %
ProbabMix	95.4 %	97.6 %	99.9 %	98.4 %
DerivTree	97.7 %	97.7 %	97.7 %	97.7 %
DerivTree + UnweightedMix	98.6 %	98.6 %	98.6 %	98.6 %
CRF tagger	97.6 %	98.5 %	99.5 %	98.8 %

Table 8: Czech

containing annotations like *aus+(führ)+en*, but also *(unter)+(führ)+en*,³ do not seem to draw any clear borderline between prefixes and roots. Some of the undesirable features of the data might however also favor the systems. One of these is undersegmentation; in some cases, the words are not segmented at all, making root identification trivial. Thus, for instance, the English data contain clearly undersegmented words like (*bishopric*), (*salsify*) or (*wringing*).

One of the main limitations of most of the baseline solutions is the inability of the heuristics to recognize multiple root morphs in the same word; this, while not an issue for languages and word categories where compounds are scarce (like Croatian verbs) did significantly decrease the accuracy of the algorithm in languages where compounds are common (e.g. German; compare the average precision and recall; compare also with results of the oracle experiment). For example, on Czech, the best performance was

³The morphs labeled as roots are in brackets.

language	avg morphs per word	compounds	root-affix homomorphy	avg word-level precision	best word-level precision
Czech	3.84	0.0 %	0.1 %	94.3 %	98.6 %
German	2.48	42.5 %	1.6 %	64.3 %	92.2 %
English	2.23	12.2 %	0.6 %	87.0 %	94.0 %
French	1.83	2.2 %	0.4 %	93.5 %	94.8 %
Croatian	4.18	1.4 %	0.2 %	95.0 %	98.3 %
Italian	2.82	0.0 %	0.5 %	91.2 %	97.5 %
Russian	4.33	17.4 %	1.5 %	76.2 %	90.2 %

Table 9: Morphological complexity

method	Czech	German	English	French	Croatian	Italian	Russian
MaxLen	76.1 %	88.5 %	92.0 %	88.8 %	86.8 %	67.7 %	68.9 %
MinFreq	96.1 %	96.8 %	97.1 %	96.8 %	98.7 %	97.5 %	92.4 %
MinNeighborEntropy	96.1 %	96.7 %	97.2 %	96.8 %	98.5 %	96.8 %	90.6 %
UnweightedMix	97.2 %	96.9 %	97.4 %	96.8 %	98.1 %	96.6 %	93.1 %
ProbabMix	95.4 %	93.7 %	95.4 %	94.4 %	92.3 %	90.8 %	85.2 %
DerivTree	97.7 %	96.9 %	96.9 %	96.7 %	97.2 %	87.7 %	87.5 %
DerivTree + UnweightedMix	98.6 %	97.1 %	97.3 %	96.9 %	98.5 %	96.1 %	94.5 %
CRF tagger	97.6 %	93.0 %	96.3 %	96.4 %	98.4 %	96.2 %	94.3 %

Table 10: Word-level accuracy on data without compounding

achieved by *DerivTree + UnweightedMix*, which selects only one root, while for German, even the simplest baseline (*MaxLen*) able to select more than one root performed better than the oracle. Approximately the same effect, although on a much smaller scale, can be observed for English and Russian. Furthermore, for languages rich in compounding, the *ProbabMix* method performed significantly better than all the remaining non-CRF heuristics. However, if we remove the compounds from the test data, the word-level accuracy changes significantly (see Table 10). In such a setting, both the CRF tagger and *ProbabMix* are outperformed by other methods for all the languages; the best methods are then either the simple statistics (*MinFreq* or *UnweightedMix*) or *DerivTree + UnweightedMix*.

Although in most of the metrics and most of the languages, the CRF tagger yields the best results, in all but two of the languages (Czech and French) some unsupervised method is more accurate than those using derivational trees. Furthermore, the difference in performance between the heuristics and the CRF classifier is often almost negligible. Interestingly, the results do not seem to be affected by the morphological complexity of the languages, as documented by Table 9.

Another interesting question is the influence of homomorphy and allomorphy resolution. Homomorphy might affect the performance either indirectly (in the computation of the heuristics) or directly, as is the case for the *ProbabMix* method, which presupposes no homomorphy between roots and affixes. Allomorphy might cause errors especially for the methods using derivational trees, where the edit distance between the morphs and the root word is used. It should be noted, however, that allomorphy might be irrelevant or even work in favour of some of the methods (e.g. *MinFreq*). Both homomorphy and allomorphy are very hard to detect in a completely unsupervised setting, although some approaches could possibly be adapted from the comparable task of word sense disambiguation.

While we do not possess any reliable method to detect allomorphy-related errors even with the gold data, homomorphy of root and non-root morphs is easily detectable in the gold data. As listed in Table 9, the languages vary in the percentage of instances of root-affix homomorphy in the test data. A comparison of these with the percentage of homomorph misclassification (in Table 11) shows that even the indirect influence of homomorphy is considerable in the statistics – both the *UnweightedMix* and the *DerivTree + UnweightedMix* err disproportionately often in homomorph classification, although the disproportion is not so marked as for the *ProbabMix* and the CRF tagger. It is also noteworthy that the CRF tagger is in some cases more prone to homomorphy-related errors than the simple *ProbabMix* method (German, Russian).

method	Czech	German	English	French	Croatian	Italian	Russian
UnweightedMix	6 %	8 %	8 %	6 %	14 %	17 %	22 %
ProbabMix	16 %	23 %	18 %	15 %	49 %	24 %	39 %
DerivTree + UnweightedMix	4 %	8 %	8 %	6 %	12 %	16 %	22 %
CRF tagger	12 %	40 %	19 %	11 %	32 %	23 %	45 %

Table 11: Homomorphy-related errors

6 Conclusion

We have compared several root identification methods on seven Indo-European languages, using simple unsupervised heuristics, derivational-tree-based heuristics, and a CRF tagger. The experiments show that simple unsupervised statistical methods are sufficient for cross-linguistically highly precise root identification. While the results can be slightly improved using derivational trees, the CRF taggers, trained on a small dataset, usually achieved further improvement. The main bottlenecks of the current methods seem to be compounding, homomorphy resolution (for the CRF tagger), and potentially allomorphy resolution (for the derivational trees).

In the future, as the unsupervised heuristics proved to provide unexpectedly good results, we would like to further probe their possible combination with other methods, possibly as sources for generating data, on which a neural classifier could be pre-trained. We would also like to concentrate on (preferably, low-resource) homomorphy and allomorphy resolution, drawing inspiration from the approach by Harsha et al. (2022).

Secondly, we would like to concentrate on increasing the granularity of the classification. Given morphological lexicons for the respective languages, derivational databases could then be used similarly as in John and Žabokrtský (2023). We would also like to mine other available high-quality multilingual resources containing morphological information, notably the Universal Dependencies (Nivre et al., 2020), which contain rich morphological annotation in the form of so-called Universal Features.

7 Acknowledgements

This work has been supported by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101). It has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure.

References

- Mark Aronoff. 1976. *Word Formation in Generative Grammar*. The MIT Press, Cambridge.
- Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022a. The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Seattle, Washington, pages 103–116.
- Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. [MorphoNet: a large multilingual database of derivational and inflectional morphology](#). In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Online, pages 39–48. <https://doi.org/10.18653/v1/2021.sigmorphon-1.5>.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George

- Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoeck, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022b. **UniMorph 4.0: Universal Morphology**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 840–855. <https://aclanthology.org/2022.lrec-1.89>.
- Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Association for Computational Linguistics, Sydney, Australia, pages 69–72.
- Jan Bodnár, Zdeněk Žabokrtský, and Magda Ševčíková. 2020. Semi-supervised induction of morpheme boundaries in Czech using a word-formation network. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*. Springer, pages 189–196.
- Elena I. Bolshakova and Alexander S. Sapin. 2021. Building a Combined Morphological Model for Russian Word Forms. In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*. Springer, Berlin, Heidelberg, page 45–55.
- Michael Ginn and Alexis Palmer. 2023. Taxonomic loss for morphological glossing of low-resource languages. *arXiv preprint arXiv:2308.15055*.
- John Goldsmith. 2001. **Unsupervised Learning of the Morphology of a Natural Language**. *Computational Linguistics* 27(2):153–198. <https://doi.org/10.1162/089120101750300490>.
- N. Sree Harsha, Ch. Nageswar Kumar, Vijaya Krishna Sonthi, and K. Amarendra. 2022. **Lexical ambiguity in natural language processing applications**. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. pages 1550–1555. <https://doi.org/10.1109/ICEARS53579.2022.9752297>.
- Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology* 30(2):117–134.
- Vojtěch John and Zdeněk Žabokrtský. 2023. The unbearable lightness of morph classification. In Kamil Ekštejn, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*. Springer Nature Switzerland, Cham, pages 105–115.
- Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. 2020. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics* 115:5–30.
- Angelina McMillan-Major. 2020. **Automating gloss generation in interlinear glossed text**. In *Proceedings of the Society for Computation in Linguistics 2020*. Association for Computational Linguistics, New York, New York, pages 355–366. <https://aclanthology.org/2020.scil-1.42>.
- Sarah Moeller and Mans Hulden. 2018. **Automatic glossing in a low-resource setting for language documentation**. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pages 84–93. <https://aclanthology.org/W18-4809>.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. **Universal Dependencies v2: An evergrowing multilingual treebank collection**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 4034–4043. <https://aclanthology.org/2020.lrec-1.497>.
- Patrick Schone and Daniel Jurafsky. 2001. **Knowledge-Free Induction of Inflectional Morphologies**. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. <https://aclanthology.org/N01-1024>.
- Radu Soricut and Franz Och. 2015. **Unsupervised Morphology Induction Using Word Embeddings**. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1627–1637. <https://doi.org/10.3115/v1/N15-1186>.

- Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. [Automatic inter-linear glossing for under-resourced languages leveraging translations](#). In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pages 5397–5408. <https://doi.org/10.18653/v1/2020.coling-main.471>.
- Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards Universal Segmentations: UniSegments 1.0. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. ELRA, Marseille, France, pages 1137–1149.

Can Large Language Models Tell Us Something about Derivation Processes?

Marko Tadić

University of Zagreb,
Faculty of Humanities and Social Sciences,
Zagreb, Croatia
marko.tadic@ffzg.unizg.hr

Abstract

The paper presents the preliminary research on usage of Large Language Models (LLMs), primarily using translation models in Neural Machine Translation (NMT) process, to generate newly derived and compound words. The method for detecting and classifying newly generated words by usage of NMT translation models, is being presented.

1 Introduction

Recently there has been a clear shift from knowledge-based and human-engineered methods towards data-driven architectures, which has led to the progress in the field of Language Technology (LT). One recent aspect associated with the paradigm shift in language processing is the use of pretrained Large Language Models (LLMs). Large-scale monolingual and/or multilingual textual data is used to train LLMs. Pre-trained LLMs, like BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), and XLM-RoBERTa (Conneau et al., 2020), have offered a framework for using the knowledge acquired during the training process to be later applied to newer tasks. One such task could be the usage of LLMs in detection of derivational morphology phenomena and, if possible, their classification and description. In this respect this paper presents results of a preliminary research that tries to determine whether a LLM can be used to detect derivationally and compositionally newly generated words in a language. The detection process in essence boils down to the usage of a Neural Machine Translation (NMT) model pretrained on parallel data (Croatian-English parallel corpus) to translate one side of already humanly translated Croatian-English parallel corpus: English into Croatian again. The resulting translation has been matched with the existing Croatian lexica in order to detect newly coined words, i.e. words that are unknown to the existing lexical resources. These words are distributed in several categories, their overall frequency is presented and the results are being discussed.

The paper is structured as follows. The section 2 presents the scarce related work where LLMs were used in derivational morphology, while in the section 3 the used language resources are described. In the section 4 the methodology is detailed and in section 5 results are presented accompanied by discussion. The conclusions and possible future directions of research are provided in section 6.

2 Related Work

So far the usage of LLMs in processing derivational morphology has been quite scarce. Cotterell et al. (2017) and Deutsch et al. (2018) proposed neural architectures that represent derivational meanings as tags. In Edmiston (2020) experiments, which probe the hidden representations of several BERT-style models for morphological content, are being presented and discussed. The most prominent work in this task so far is provided by Hofmann et al. (2020a, 2020b, 2021) where the authors use the auto-encoder to check the morphological well-formedness (MWF), finetune the BERT into DagoBERT that is capable of generating new derivations, and use finetuning to improve

BERT's interpretation of complex words.

All these papers based their research on monolingual LLMs, while to the best of our knowledge, the approach proposed in this paper is the first one that uses the LLMs in multilingual context, i.e. using the translation model and LLM for the target language in order to investigate and extrinsically evaluate the generation of derivatives and compounds in a target language. Starting from the parallel corpus enables us to keep the content variable under control since in a parallel corpus the translation equivalents at the level of sentences, i.e. translation units (TUs), are explicitly marked by unique sentence IDs and are considered to convey the same overall sentence meaning.

3 Language Resources

In this research we used the following language resources:

The parallel corpus that was used for experiments is the Croatian-English Parallel Corpus (CW) described in (Tadić, 2000). It is a unidirectional corpus of newspaper articles published in *Croatia Weekly* between 1998 and 2000, translated by professional translators and language proofed by three different English native speakers.

For the machine translation of English sentences into Croatian, the NMT models developed within the CEF-project National Language Technology Platform (NLTP)¹ were used through the online interface² of its Croatian installation (Vasiļevskis et al. 2023). The baseline NMT model were trained on DGT parallel English-Croatian data and then finetuned with additional 0.76 million Translation Units (TUs), including the whole CW parallel data set. The NMT models are typical Transformer based models that were produced by Tilde³ are described in (Krišlauks and Pinnis 2020). However, unlike in the described en→pl translation model training, that also used backtranslation due to the noisy input data, for training en→hr and hr→en translation models only the Transformer base model configuration was used since the data were composed of only clean human translations. The first instance of these translation models for en→hr and hr→en pairs was used in EU Council Presidency Translator⁴ in 2020 and it received BLEU scores of 36.93 and 41.30 respectively. For the NLTP Croatian installation, these translation models were enriched with additional training data of approximately 1 million tokens and has been used in this experiment.

For tokenization of translated sentences the UDPipe pipeline⁵ has been used with the Croatian set UD2.10 selected.

For detection of the unknown words the Croatian Morphological Lexicon (HML)⁶, an inflectional lexicon with 110,000+ lemmas and 6M+ wordforms, accessible as an online service, was used. Its features were described in detail in (Tadić, 2005).

During the checking of unknown words a number of existing Croatian language resources were used starting with corpora: Croatian National Corpus (HNK)⁷ and Croatian Web Corpus (hrWaC)⁸. The online lexica used for checking were: Hrvatski jezični portal⁹, Croatian Special Field Terminology¹⁰, Croatian Terminology Portal¹¹, Jezikoslovac¹², Croatian Glosbe¹³ online dictionary, CroDict¹⁴ online dictionary, Croatian Encyclopedia¹⁵, co-textual search engine Kontekst¹⁶ set to Croatian queries, and common search engines Google and DuckDuckGo. Also, as the final means

¹ <https://nltp-project.info>

² <https://hrvojka.gov.hr>

³ <https://www.tilde.com>

⁴ <https://hr.presidencymt.eu>

⁵ <https://lindat.mff.cuni.cz/services/udpipe/>

⁶ <https://hml.ffzg.hr>

⁷ <https://filip.ffzg.hr>

⁸ <http://nlp.ffzg.hr/resources/corpora/hrwac/>

⁹ <https://hjp.srce.hr>

¹⁰ <https://struna.ihjj.hr>

¹¹ <https://nazivlje.hr>

¹² <https://jezikoslovac.com>

¹³ <https://hr.glosbe.com>

¹⁴ <https://crodict.hr>

¹⁵ <https://enciklopedija.hr>

¹⁶ <https://kontekst.io>

used for finding a lexical evidence the paper version of the *Veliki rječnik hrvatskoga standardnoga jezika* (Jojić et al., 2015) was used.

4 Methodology

In this section the methodology used in the research is described in detail.

4.1 Translation of English Sentences

The CW was obtained from META-SHARE¹⁷ in TMX format and the sample of 10,000 TUs was selected and English sentences from aligned pairs were extracted. The unique sentence IDs were preserved in order to be able to refer back to the original Croatian sentences (*hr* mark in examples) when needed.

The English sentences were translated using the Croatian installation of the NLTP NMT services at hrvojka.gov.hr. The source 10,000 English TUs (*en* mark in examples) had 234,278 tokens, while the translated Croatian TUs (*hr-t* mark in examples) had 193,020 tokens.

4.2 Tokenisation with UDPipe and Matching with the Croatian Morphological Lexicon

The *hr-t* set of sentences was tokenized using the UDPipe online services and the results were downloaded in CoNLL-U format. Only the first column of that format was used in this research. However, the annotation information from the remaining nine columns could be used for future investigations on e.g. quality of lemmatization, particularly when it comes to the unknown and for UDPipe system unseen words. This might be one of directions for the future research, but it certainly surpasses the limits of the current paper.

The token list from the first column was uploaded to the HML requesting the lemmatization of each token. In the case of unknown token, the HML returns #NIL#, so it was easy to extract words unknown to HML.

4.3 Inspection and Classification of Unknown Words

The list of #NIL# tagged tokens, 4453 in total, was then manually inspected for evidence. Every token not being evidenced in any of aforementioned corpora, lexica or search engines was marked and classified in accordance with the preliminary classification scheme. The scheme and basic statistics is presented in Table 1.

Before the manual inspection it was decided that certain types of unknown words will not be taken into account: 1) named entities; 2) translation errors (e.g. direct transfer of the original English word); 3) deverbative nouns ending in *-nje* since they are highly productive in Croatian¹⁸; 4) highly productive negated adjectives and nouns (e.g. *nekoristan*, *nekompetencija*); 5) highly productive compounds written usually with dash (e.g. *makedonsko-hrvatski*, *ne-Hrvat*). On the other hand, we put a strong emphasis on detecting compounds written without dash since they usually express stronger bond between compounding parts.

5 Results and Discussion

Here each of the categories of the classification scheme is described and exemplified. :

- expectable compound: compounds that could be expected having in mind possible combination of compounding parts, e.g. en: *self-denying* / hr-t: *samoopovrgavajući*, en: *late antique* / hr-t: *kasnoantika*
- unexpected compound: compounds that are partial errors in translation but convey the general meaning, e.g. en: *five-movement* / hr-t: *petokretni* instead of hr: *petostavačni*, en: *Euro game* / hr-t: *euroigre* instead of hr: *europske igre*;

¹⁷ <https://meta-share.org>

¹⁸ This decision could be questioned since investigating this highly regular and productive derivational pattern in Croatian (and many other Slavic languages as well) could reveal some of the underlying mechanisms that LLMs are dealing with when trained at the subword level. However, this topic might deserve the paper on its own while here we wanted to tackle the widest possible number of different phenomena at this preliminary pilot level.

- possessive adjective of names: highly productive derivation, but sometimes with unexpected derivations, e.g. en: *Boka Croats* / hr-t: *bočki Hrvati* instead of hr: *Hrvati iz Boke*, en: *Klein's* / hr-t: *Kleinski* instead of hr: *Kleinov*;
- alternative derivation: derivation that uses different, but possible, derivation affix, e.g. en: *lace-makers* / hr-t: *čipkaši*, en: *broker* / hr-t: *burzer*;
- unexpectedable derivation: derivations that are partial errors in translation, but convey the general or alternative meaning, e.g. en: *swallow* (bird) / hr-t: *gutljica*, en: *voucher holders* / hr-t: *imatelji vaučera*;
- direct alternative calque: derivations or compounds that directly conveys the English word and tries to translate its parts and/or adapt it phonetically and morphologically in Croatian, e.g. en: *underworld organisations* / hr-t: *podsvjetske organizacije* instead of hr: *mafijaške organizacije*, en: *Knights Hospitallers* / hr-t: *Hospitalari* instead of hr: *ivanovci*.

Category	Tag	Frequency
expectable compound	so	17
unexpectedable compound	sn	11
possessive adjective (-ov/-ski/-čki)	pp	164
alternative derivation	dz	76
unexpectedable derivation	dn	15
direct alternative calque	pz	38
total		321

Table 1: Words unknown to the existing lexica and their classification scheme with basic statistics.

The initial 4453 words marked with #NIL# as the result of matching with HML, were scaled down after the manual inspection and lookup for evidence in different language resources to the total of 321 cases. Most of the tokens unknown to HML were named entities and clear translation errors.

The 321 occurrences of newly generated words represent 7,21% of all unknown words. This might look like a small number, but this should be regarded as a percentage of total number of lexical entries used in the sampled 10,000 TUs, i.e. 193,020 hr-t tokens. These 7,21% cases are the spots in the English text that for some reason invoked the translation LM to come up with derivation or composition in order to convey the basic meaning. Was it invoked because of the lacunae in Croatian lexicon where in the English such lexical items exist? Certainly not since the manual inspection confirmed that in many cases in the original Croatian source such lexical items exist.

Does the LLMs have intrinsic preference to generate derivations or compounds because of limited lexicon used in the training process? What is really being conveyed with this language means and their selection in the process of machine translation using LLMs? Is it the similar content running in two parallel texts, or approximation of its similarity represented through LLM-based MT, that affects also such lower language levels as derivational morphology?

The individual examples for expectable categories might look quite surprising to a native speaker of Croatian, but after careful inspection of the English source, the expected and alternative derivatives and compounds generated in translations are morphologically well-formed (see examples above).

6 Conclusions and Future Directions

We presented the preliminary investigation that tried to detect the amount and types of possible newly derived and compound words produced by a LLM. The LLMs (particularly translation

models where we have the experimental variable of the same content in two languages under control), that are being trained to take into account the subword segments, in their performance are now being able to signal the spots where the transferred content could be represented by derivational or compositional means available in the target language. In this respect the LLM generates the derivations or compositions not yet registered in any lexica of the target language by following the derivational and compositional rules of that language and thus producing MWF words. At this spots are LLMs signaling us something? It seem like they are pinpointing the nodes in the total combinatorial capacity of a language at the derivational/compositional level, the nodes in the derivational/compositional network of morpheme combinations, that exist *in potentia*, but are not (yet) filled with an accepted combination of morphemes. These nodes were certainly not filled with lexical entries in the training material, but still the LLM has envisaged their existence. Can LLMs help us in recognizing the topology of this network or it is just another way of representation of the derivational/compositional complexity in language?

This production of neologisms is particularly characteristic for translation pair en→hr since these two languages differ typologically, namely English is more analytical and tends towards phrasal solutions, while Croatian is more synthetical and tends towards derivational solutions. It would be interesting in the future to investigate the reverse direction of translation, i.e. hr→en and then check the ability of the same translation LM to generate derivatives/compounds in English and to provide their classification and statistics.

If humans would generate such new words, representing in fact new lexical entries, we would tend to consider this a creative use of language. Can we treat such words the same way when they are being generated by LLMs?

Although the research presented in this paper didn't produce fully automatic method of detecting newly generated derivations and compositions, this could represent one of directions for future research. We have a parallel corpus at our disposal and the difference between the humanly produced original text in Croatian and NMT produced translated counterpart from English into Croatian could be automatically compared for differences.

Moreover, following (Hofmann et al. 2020b), we need further intrinsical evaluation to find out how input segmentation impacts the derivational knowledge available to a LLM. This might suggest that the performance of LLMs could be improved if a morphologically informed vocabulary of units (e.g. derivationally segmented) were used in the training phase. At this stage of training of LLMs, we don't really know how the subword segmentation is being produced and to what extent the division into segments really corresponds to the real morphological boundaries.

It would certainly be most useful if we could make use of existing LLMs in the computational processing of derivational/compositional morphology and even more so if we could perhaps be able to train a new LLM tailored to be sensitive on derivational/compositional information.

Acknowledgments

The work reported here was supported by the European Commission through the CEF Telecom Programme (Action No: 2020-EU-IA-0082 National Language Technology Platform, NLTP, Grant Agreement No: INEA/CEF/ICT/A2020/2278398) and by the Ministry of Science and Education of the Republic of Croatia through the support for the Croatian CLARIN Research Infrastructure Consortium.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901. <https://doi.org/10.48550/arXiv.2005.14165>

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

- Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.acl-main.747.pdf]
- Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 714–720, Copenhagen, Denmark, Association for Computational Linguistics. [https://aclanthology.org/D17-1074.pdf]
- Daniel Deutsch, John Hewitt, and Dan Roth. 2018. A distributional and orthographic aggregation model for English derivational morphology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1938–1947, Melbourne, Australia. Association for Computational Linguistics. [https://aclanthology.org/P18-1180.pdf]
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, Association for Computational Linguistics. [https://aclanthology.org/N19-1423.pdf]
- Daniel Edmiston. 2020. A Systematic Analysis of Morphological Content in BERT Models for Multiple Languages. <https://doi.org/10.48550/arXiv.2004.03032>.
- Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2020a. A Graph Auto-encoder Model of Derivational Morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1127–1138, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.acl-main.106.pdf]
- Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020b. DagoBERT: Generating Derivational Morphology with a Pretrained Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.emnlp-main.316.pdf]
- Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre Is Not Superb: Derivational Morphology Improves BERT’s Interpretation of Complex Words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics. [https://aclanthology.org/2021.acl-long.279.pdf]
- Ljiljana Jojić, Nada Vajs Vinja, Vesna Zečević, Anuška Nakić, Ivan Ott, Jelena Cvitanušić Tvico, Ranka Đurđević, Igor Marko Gligorić, Aida Korajac, Ines Kotarac, Ivana Krajačić, Ivan Ott, Katja Peruško, Nika Štriga, Dijana Vlatković. 2015. *Veliki rječnik hrvatskoga standardnoga jezika*. Zagreb: Školska knjiga.
- Rihards Krišlauks, and Mārcis Pinnis. 2020. Tilde at WMT 2020: News Task Systems. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 175–180, Online. Association for Computational Linguistics. [https://aclanthology.org/2020.wmt-1.15.pdf]
- Marko Tadić. 2000. Building the Croatian-English Parallel Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*, pages 523-530, Athens, Greece. European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2000/pdf/119.pdf]
- Marko Tadić. 2005. The Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29(1/2): 206-217.
- Artūrs Vasiļevskis, Jānis Ziediņš, Marko Tadić, Željka Motika, Mark Fishel, Bjarni Barkarson, Claudia Borg, Keith Aquilina, and Donatienne Spiteri. 2023. National Language Technology Platform (NLTP): The Final Stage. In *Proceedings of the International Conference HiT-IT2023*, pages 203-208, Naples, Italy. INCOMA Ltd., Šoumen, Bulgaria. doi:10.26615/issn.2683-0078.2023_019.

Morphosemantic analysis of Croatian nouns formed with the prefix *pred-*

Marta Petrak

Faculty of Humanities and Social Sciences
University of Zagreb (Croatia)

mpetrak@ffzg.unizg.hr

Abstract

In this paper, we present a morphosemantic study of Croatian nouns formed with the prefix *pred-* ‘in front of’. The nouns were retrieved from the new CLASSLA-web corpus, the largest extant one for Croatian, and subsequently analysed both formally (word-formation wise) and semantically from a cognitive linguistic point of view. Cognitive linguistics considers prefixes to be polysemic units which exhibit behaviour similar to that of the traditional “lexical” categories such as nouns or verbs. Our analysis has demonstrated that the derived nouns demonstrate interesting regularities: prefixation frequently yields a temporal reading of the prefix *pred-*, prefix-suffix combination frequently results in nouns denoting places situated in front of something, and suffixation creates nouns with a number of extended meanings based on metaphor and/or metonymy. Detailed analysis of the morphosemantic potential of a specific word-formation element are rare in the literature, but contribute to a better understanding of the mechanisms of the construction of the lexicon.

Keywords: word formation, semantic motivation, Croatian, *pred-*

1 Introduction

The goal of this paper is to present an exhaustive and comprehensive morphosemantic analysis of nouns formed with the Croatian prefix *pred-* (or its allomorph *pret-*).¹ *Pred-* is a native Croatian prefix which can form nouns, verbs and adjectives (Babić, 2002).

Such analyses are quite rare in the extant literature, but are highly important because they explore the morphosemantic potential of one specific derivational element, in this case a prefix, and analyse how it is used in the construal of the lexicon. By analysing both the formal level (i.e. the various word formation mechanisms it enters into) and the semantic level (i.e. the meanings that the prefix *pred-* realizes in its derivatives, and the mechanisms on which these meanings are based), we gain a detailed insight into its morphosemantic potential.

This study is rooted in the cognitive linguistic framework, which has been of the leading currents in linguistics in the past several decades. When it comes to word-formation, however, cognitive linguistic tenets have only recently started to be applied to it (Onysko and Michel, 2010). The question was raised by some authors whether cognitive linguistics is well-equipped to deal with word-formation issues. Ungerer (2010) believes that it “has the potential to stimulate word-formation research” because it “can provide both the theoretical background and the empirical tools” necessary to do so. The underlying goal of all cognitive linguistic efforts related to word formation would be, as Ungerer (2010) puts it, “the semanticization of word-formation analysis”. This process, it needs to be noted, is not a novelty introduced by cognitive linguistics, but it was largely emphasized within this theoretical framework. In a similar vein, it needs to be said that cognitive linguistics is compatible with some basic tenets of the structuralist approach to word formation (Onysko and Michel, 2010; Štekauer and Lieber, 2005; cf. Raffaelli, 2004).

A cognitive linguistic approach to word formation, which is also adopted in this study, is in line with the basic cognitive linguistic tenets, among which the most important ones are prototype theory (e.g. Rosch, 1975), radial categories (Lakoff, 1987) and the theory of conceptual metaphor and metonymy (e.g. Lakoff and Johnson, 1980). All these facts could be subsumed under the idea that the formation of

¹ In our study, we have disregarded the standard spelling rules related to the assimilation according to voicing (cf. Babić et al. 2007), due to which some words are spelled with a *t* rather than with a *d*. We have eliminated from our lists all double entries and have retained only one spelling version of the noun, the more frequent one.

words “is in essence a cognitive phenomenon” (Štekauer and Lieber, 2005, Onysko and Michel, 2010). In other words, word-formation is not merely a linguistic issue, but as all language faculties, it is inextricably linked to human cognitive capacities and the principles of conceptual organization.

As we have previously said, the central tenet of cognitive linguistics is that all aspects of word-formation are treated as meaningful (Ungerer, 2010), which is in line with the axiom of the centrality of meaning (Langacker, 1987). As such, this framework considers prepositions and prefixes to be polysemic radial categories (e.g. Lakoff, 1987; Tyler and Evans, 2003; Šarić, 2008; Matovac, 2013; Petrak, 2021), which is quite different from the homonymous approach to these linguistic elements that had prevailed in traditional approaches to language (Peytard, 1975; Belaj, 2008).

The remainder of the paper is structured as follows. In Section 2 we showcase studies related to the prefix *pred-*. Section 3 presents our methodology, which is corpus-based. In Section 4 we analyse and discuss the results we have obtained, and Section 5 brings some concluding remarks.

2 Previous work

The prefix *pred-* features in several works: Šipka (1989) mentions it as a means of forming antonymic pairs of motivated words, in pairs with the prefixes *poslije-* ‘after’ and *post-* ‘after’. Vulić (2020), in her paper on motivated words in contemporary literary works of Burgenland Croats, mentions that *pred-* is used to form masculine and feminine nouns, without further details.

Apart from these papers, *pred-* is the main subject of Belaj’s (2008) study which deals with verbs formed with that prefix. The author concludes that these verbs form a set unified by the *pre-locativity superschema*, which explains both the prototypical and less prototypical meanings and meaning extensions of the prefix.

In his dissertation, Matovac (2013) analyses the preposition *pred*, from which the prefix *pred-* originated, and claims that together with the prepositions *nad*, *pod* and *za*, *pred* makes the basis of the study of Croatian orientational prepositions. Furthermore, Matovac (2013) remarks that, unlike the prepositions *nad* ‘above’ and *pod* ‘under’, which describe relations on the vertical axis, and which have been abundantly analysed within cognitive linguistics, prepositions referring to the horizontal axis have received much less attention or have, as in the case of *pred*, even been completely neglected. The fourth preposition, *za* ‘behind; for’, has received a lot of attention in traditional grammars due to its specific usage (*ibid.*).

The goal of this paper is to deal with nouns containing the prefix *pred-* in order to shed more light on this previously rather unexplored set of formally unified words. In addition, building upon Ullmann’s (1966) ideas that lexicon is a product, essentially, of morphological and semantic motivation, and especially the hypothesis that these two types of motivation are not and should be not regarded separately (Koch and Marzo, 2007; Raffaelli, 2013, 2015), but that formal (morphological) motivation always accompanies semantic (conceptual) motivation, we set out to explore the mechanisms which gave rise to the part of the Croatian lexicon construed with the prefix *pred-*.

3 Methodology

In line with the usage-based approach, which is predominant in the cognitive linguistic framework, this study is based on real language data as found in large corpora. More precisely, the data for our research was retrieved from the brand new CLASSLA-web.hr corpus (Kuzman and Ljubešić, 2023), which is currently the largest extant Croatian corpus. It contains 2.7 billion tokens crawled primarily from the national top-level internet domain *.hr*.

In order to retrieve all *pred-* prefixed nouns, we did a graphic search for nouns starting with the graphemes *pred* and *pret*. The reason for this lies in the fact that the CLASSLA corpora are not derivationally segmented.² The minimal frequency threshold was set at two occurrences in order to retrieve all the extant nouns (except for hapaxes), which enabled us to reach both high- and low-frequency types. Low-frequency words are known to sometimes possess properties different from

² Derivational segmentation is seldom conducted in NLP because it has not been widely used so far, and due to the fact that it is rather difficult to perform (Nikola Ljubešić, private communication).

higher-frequency words, which may be linked to a different degree of lexicalisation, new meanings, newly developed properties, etc.

The corpus search resulted in two initial lists, one containing *pred-* (3 424) and the other containing *pret-* nouns (1 062). The lists contained a lot of noise. For the types below 10 occurrences, we have first examined whether these were indeed words or simply typos (spelling of prepositional phrases together with the noun). Also, several adjectives and verbs, as well as typos, were found in the lists. There were a few words which belong to closely related Southern Slavic languages, such as Serbian and Slovene (e.g. *preduzetništvo*, *predsednik*), as well as a quite large number of double entries differentiated by the use of hyphen³. All such words were excluded from the final list.

In addition, the largest part of the manual work consisted in checking the semantics of the *pred-* and *pret-* elements, as there were words in the lists which did not actually contain the prefix, but its homograph (e.g. *pretvorba* < *pre-* + *tvoriti* > ‘conversion’, *pretjerivanje* < *pre-* + *tjerati* > ‘exaggeration’, *predikacija* < *predikat*, Lat. *praedicatum* > ‘predicate’, *predavanje* ‘lecture’ < *pre-* + *dati* >, etc.). All the words were checked in the dictionary available at the Croatian Language Portal’s (Hrvatski jezični portal) website,⁴ while those that are not present in it were checked both in the corpus and online.

Having in mind all the tedious work that needs to be done in order to clear corpus lists in studies involving derivational morphemes, it would prove extremely useful to have corpora tagged at the morphological level. One of the possible solutions would be to take derivational morphology training data and train a model for morphological segmentation. In such a case, new transformer models such as BERTi⁵ could prove to be quite useful, especially for semantic recognition (Nikola Ljubešić, private communication).

Once all the manual work has been done, our search yielded a total of 1 006 nouns. When compared to the initial 4 486 types yielded by our corpus search, we can see that only about 22% of this initial bulk quantity of words were actually *pred-* prefixed nouns, and we can observe the large amount of noise and semantic errors which had to be manually checked.

Below is a list of the 50 most frequent types, accompanied by the number of their occurrences in the corpus and frequency per million:

1. *predsjednik* ‘chair, director (m.)’, 1054257, 388.38086
2. *predmet* ‘object’, 468947, 172.75677
3. *predstavnik* ‘representative (m.)’, 414332, 152.63699
4. *prednost* ‘advantage’, 335072, 123.43817
5. *predstava* ‘theater piece’, 333850, 122.98799
6. *predsjednica* ‘chair, president (f.)’, 210374, 77.50030
7. *predstavljanje* ‘presentation’, 164495, 60.59880
8. *preduvjet* ‘precondition’, 64727, 23.84497
9. *predsjedništvo* ‘presidency’, 57161, 21.05771
10. *predrasuda* ‘prejudice’, 41029, 15.11479
11. *predložak* ‘template’, 35058, 12.91512
12. *predstavnica* ‘representative (f.)’, 34965, 12.88086
13. *predak* ‘ancestor’, 30167, 11.11331
14. *predviđanje* ‘prediction’, 28200, 10.38868
15. *predsjedanje* ‘presiding’, 17285, 6.36767
16. *predstojnik* ‘director, head (m.)’, 16638, 6.12932
17. *predgovor* ‘foreword’, 16533, 6.09064
18. *predstavništvo* ‘representative body’, 15737, 5.79740
19. *predlagatelj* ‘proposer’, 15716, 5.78966
20. *predodžba* ‘idea, conception’, 15578, 5.73883

³ The hyphenation problem could easily be solved automatically, as an idea for the development of derivational tools.

⁴ <https://hjp.znanje.hr/>

⁵ <https://huggingface.co/classla/bcms-bertic-ner>

21. *predgrađe* ‘suburbs’, 14587, 5.37375
22. *predvorje* ‘vestibule’, 13894, 5.11845
23. *predvodnik* ‘leader’, 13743, 5.06282
24. *predznak* ‘sign, omen’, 13467, 4.96115
25. *predujam* ‘advance’, 11316, 4.16873
26. *predlaganje* ‘proposition’, 10936, 4.02875
27. *predočenje* ‘presentation’, 8406, 3.09671
28. *predsoblje* ‘anteroom’, 7906, 2.91251
29. *predznanje* ‘prior knowledge’, 7522, 2.77105
30. *predstojnica* ‘director, head (f.)’, 7299, 2.68890
31. *predškola* ‘preschooler’, 6515, 2.40008
32. *predgrupa* ‘band playing before the main one’, 6354, 2.34077
33. *predjelo* ‘appetizer’, 6016, 2.21625
34. *predstavka* ‘petition’, 5685, 2.09431
35. *predškola* ‘preschool’, 5543, 2.04200
36. *predigra* ‘foreplay’, 4994, 1.83975
37. *predvečerje* ‘eve’, 4918, 1.81176
38. *predbilježba* ‘pre-reservation’, 4849, 1.78634
39. *predračun* ‘invoice’, 4694, 1.72924
40. *predsezona* ‘preseason’, 4659, 1.71634
41. *predostrožnost* ‘precaution’, 4340, 1.59883
42. *predsjedatelj* ‘chairman (m.)’, 3980, 1.46620
43. *predugovor* ‘precontract’, 3661, 1.34869
44. *predlagač* ‘proponent’, 2788, 1.02708
45. *predvodnica* ‘leader (f.)’, 2533, 0.93314
46. *prednarudžba* ‘pre-order’, 2525, 0.93019
47. *predmetnica* ‘subject’, 2354, 0.86720
48. *predvoditelj* ‘leader (m.)’, 2353, 0.86683
49. *predbroj* ‘telephone prefix’, 2221, 0.81820
50. *predpojačalo* ‘preamplifier’, 2119, 0.78062

The number of 1 006 nouns formed with *pred-* (*pret-*) demonstrates that *pred-* has quite a large potential in the construction of the Croatian lexicon. Moreover, the prefix seems to be (much) more used in the formation of nouns than in that of verbs, which have been found to be much less numerous in some previous studies. For instance, Petrak (2021) found only 32 *pred-* verbs upon an analysis of the three Croatian most relevant Croatian corpora at the time: hrWaC, previously the largest Croatian web corpus, HNK, the Croatian national corpus, and the Riznica corpus. However, the frequency limit in her study was set at 10 occurrences, so the number of low-frequency *pred-* verbs remains unknown, and has probably influenced this large disparity in the results obtained.

Once we have obtained the list of *pred-* nouns from the CLASSLA corpus, we set out to study the following: 1) which word-formation processes *pred-* is found in, 2) what semantic mechanisms such formations are based on, 3) is there a correlation, i.e. are there regularities in the pairings of word-formation processes and semantic mechanisms, 4) is there any difference between high- and low-frequency nouns. We shall explore all these questions in the following section.

4 Results and analysis

In this section, we shall first take a detailed look into the word-formation types found in the corpus, and shall then turn to their semantic analysis.

4.1. Word-formation types

Below is a table demonstrating the part of the four word-formation types found in our corpus: prefixation, suffixation, prefix-suffix combination and backformation.

	Word-formation mechanism	No.	%
1	prefixation	846	83,5
2	suffixation	129	13
3	prefix-suffix combination	27	3
4	backformation	5	0,5
	Total	1 006	100%

Table 1: Word-formation types found in the CLASSLA-web.hr corpus

a) Prefixation

Prefixation is the prevalent word-formation type in our corpus, with more than 80% of all nouns stemming from it (e.g. *predpranje* ‘pewash’, *predromanika* ‘pre-Romanesque period’, *predznanje* ‘prior knowledge’, *predobrada* ‘preprocessing’, *predgovor* ‘preface’). All these nouns were formed by simply attaching the prefix *pred-* to a nominal base: *pranje* ‘wash’, *romanika* ‘Romanesque period or style’, *znanje* ‘knowledge’ and *obrada* ‘processing’ respectively. At the semantic level, a great majority of these nouns realise the temporal meaning, which points to a regularity we could label as following: $[pred-[N]]_N$ ‘before N’. We shall explain in detail the mechanisms underlying every meaning in the semantic part below (See 4.2).

b) Suffixation

Suffixation, which is otherwise the most productive and prevalent word-formation process both generally and in nouns in Croatian (cf. Babić, 2002), amounts to 13% of all formations (e.g. *predsjednik* ‘president’, *prethodnik* ‘predecessor’, *predmetak* ‘prefix’, *predsjedatelj* ‘chairman (m.)’, *predvoditeljica* ‘leader (f.)’, *predradnica* ‘foreman (f.)’, etc.).

Semantics-wise, most of the suffixed words are such that the suffixed nouns get a new, metaphorical reading, unlike the prefixation cases, in which the meaning of the prefix frequently and regularly extends the meaning of the nominal base towards the temporal domain. Here is an example: *predsjednik* ‘president’ is formed from *pred-* ‘before’ + *-sjed-* ‘sit’ and the agentive suffix *-nik*, where ‘sitting in front of other people’ stands for ‘presiding’ according to the LEADING IS BEING AHEAD metaphor (cf. Petrak, 2021). Some of the prefixed derivatives are exceptions to this rule, such as *predmetak* ‘prefix’, formed from *pred-*, *-met-* ‘put’ and the suffix *-ak*, which literally reads as ‘that which is put in front of (something)’. This meaning also includes the part for whole metonymy because a single property, i.e. standing in front of something else, is taken to designate the whole entity, i.e. a prefix.

c) Prefix-suffix combination

Prefixation and suffixation representing around 93% of all formations, the rest are, clearly, rarely used formation types. Such a finding is in line with the extant literature, according to which suffixation is the most frequent mechanism in the formation of Croatian nouns, while back-formation and prefix-suffix combinations are rather rare (Babić, 2002). In our corpus, prefix-suffix combinations represent 3% of all derivatives, with examples such as *predslovlje* ‘preface’, *predsoblje* ‘anteroom, antechamber’, *predgrađe* ‘suburbs’, *predgorje* ‘foothills’, *predvečerje* ‘early evening’.

A majority of these derivatives realise a concrete spatial meaning (*predsoblje* ‘antechamber’, *predgorje* ‘foothills’, etc.), but some have a temporal meaning (*predvečerje* ‘eve’ < *pred-* + *večer* ‘evening’ + *-je*>, *predzorje* < *pred-* + *zor(a)* ‘dawn’ + *-je*> ‘dusk’,). Most of these nouns receive the suffix *-je*, while only a few appear with other suffixes as well: *predoltarnik* < *pred-* + *oltar* ‘altar’ + *-*

nik > ‘altar frontal’, *predkolumbijanac* <*pred-* + *Kolumbo* ‘Columbus’ + *-ijanac* > ‘pre-Columbian’, *pred-bolonjac* <*pred-* + *Bolonja* + *-ac* > ‘student before the Bologna Process’.

In this group, two proper nouns are present: *Predkavkazje* ‘fore-Caucasus’ and *Predalpe* ‘Pre-Alps’, in which the prefix retains its concrete meaning ‘in front of’, and is attached to an oronym (*Kavkaz* ‘the Caucasus’ and *Alpe* ‘the Alps’).

d) Back-formation

The last type of derivatives is formed via back-formation, which is quite rare with only 0,5% of the total number of nouns (e.g. *predstava* ‘theatre piece’, *pretklon* ‘the forward lean’, *predosjećaj* ‘presentiment’, *predrasuda* ‘prejudice’). These nouns can have both concrete (*pretklon* <*pretkloniti* <*pred-* + *kloniti* > ‘to lean forward’ >) and abstract meanings (*predstava* <*pred-* + *-stav-* ‘put’ + *-a* >). The noun *predstava* is a result of a metonymy in which the act of putting something in front of someone, i.e. of “putting” a literary piece on stage, is taken to designate the whole theatre piece. In the noun *predrasuda* <*pred-* + *rasuditi* ‘judge’ + *-a* >, the prefix has a temporal reading ‘before’, and the meaning of the whole noun is based on encyclopaedic knowledge that prejudice are ideas made before facts are taken into consideration, in which the IDEAS ARE OBJECTS metaphor⁶ is also present, according to which prejudice is conceptualized as objects that are placed in front of the act of judgement.

4.2. Semantic analysis

When it comes to the semantic part of our study, we have first concluded that the analysed nouns can be divided into three large groups: 1) those in which the prefix retains its prototypical spatial meaning (e.g. *predbroj* ‘lit. in front-number; telephone prefix’), 2) the group in which only the prefix gets an extended meaning, which is temporal in the largest number of cases (e.g. *predugovor* ‘precontract’), and 3) the group in which the noun as a whole gets an extended meaning (e.g. *predlagač* ‘proponent’ is formed from *pred-* ‘in front’ and *ložiti* ‘arrange’). These three meanings correspond to the concrete-spatial-metaphorical & metonymical labelling as demonstrated in Table 2.

	<i>Meaning type</i>	<i>No.</i>	<i>%</i>
1	concrete (spatial)	110	11
2	temporal	801	80
3	metaphorical & metonymical	95	9,4
Total		1 006	100%

Table 2: Meaning types of the prefix *pred-* in the analysed derivatives

Even though the first and prototypical meaning of the prefix *pred-* is concrete and spatial, more precisely ‘in front of something’,⁷ as the meaning of the corresponding preposition *pred* from which the prefix has originated, in our corpus the prefix is prevalently used with a temporal meaning ‘before’. In other words, in derived nouns, *pred-* shows a clear semantic shift toward more abstract meanings, which is obviously a result of prefixation.

4.2.1. Concrete meaning

Nouns in which *pred-* retains its concrete meaning ‘in front of’ are, for example, the following: *predgrađe* <*pred-* + *grad* ‘town’ + *-je* > ‘suburbs’, *predvorje* <*pred-* + *dvor* ‘court; courtyard’ + *-je* > ‘vestibule’ and *predbroj* <*pred-* + *broj* ‘number’ > ‘telephone prefix’. Most of these nouns are a result of the prefix-suffix combination, but there are also some examples of prefixation (*predbroj*).

⁶ https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:IDEAS_ARE_OBJECTS

⁷ https://hjp.znanje.hr/index.php?show=search_by_id&id=eVhiWxc%3D&keyword=pred

4.2.2. Temporal meaning

The largest semantic group is formed of nouns in which the prefix has the temporal meaning ‘(occurring) before (something)’. This meaning is a result of the PAST IS BEHIND metaphor (Lakoff and Johnson, 1980), according to which events happening in the future are conceptualized as being in front of the speaker. For instance, in examples such as *predjelo* ‘appetizer’, *predradnja* ‘activity preceding another one’ or *predtestiranje* ‘pre-testing’, appetizers, previous activities and pre-testings are conceptualized as objects placed before a later meal, activity or testing. This metaphor is a rather frequent one in Indo-European languages, and it is an entailment of the very common TIME IS SPACE metaphor (e.g. Radden, 2003).

In addition, we would tentatively say that *pred-* formations in which the prefix has a temporal meaning are probably quite recent formations, which might be the result of calquing from English (cf. *pre-selection*, *pre-finance*, *precontract*, *foreplay*, etc.), but this question should be further studied.

4.2.3. Metaphorical and metonymical

The third group in Table 2 is labelled metaphorical and metonymical because it results from various metaphors and metonymies. We have used this designation for all other meanings which are extensions of the prototypical one, and which we have grouped under a single label due to the fact that they represent less than 10% of all meanings. The leading meaning in this group is a result of the LEADING IS BEING AHEAD metaphor, which is an entailment of the GOOD IS IN FRONT metaphor (Belaj, 2008)⁸ found in English. This metaphor has a metonymical basis because the property of being in front is used to refer to ‘leading’ (SALIENT PROPERTY FOR CATEGORY, Littlemore, 2010). Examples in which these metaphor and metonymy occur are the following: *predmolitelj* < *pred-* + *moliti* ‘pray’ > ‘leading celebrant’, *predradnik* < *pred-* + *radnik* ‘worker’ > ‘foreman’, *predšasnik* ‘predecessor’, *predvođenje* < *pred-* + *voditi* ‘lead’ + *-nje* > ‘leading’, *predsjedništvo* < *predsjedati* ‘preside’ < *pred-* + *sjediti* ‘sit’ > ‘presidency’.

Another relatively important subgroup in this group are nouns such as *predokus* < *pred-* + *okus* ‘taste’ > ‘foretaste’, *predosjećanje* < *pred-* + *osjećati* ‘feel’ + *-nje* > ‘presentiment’, *predskazivanje* < *pred-* + *kazati* ‘tell’ + *-nje* > ‘foretelling’ or *predskazatelj* < *pred-* + *kazati* ‘tell’ + *-telj* > ‘foreteller’. The meaning of these nouns is based on the PART FOR WHOLE metonymy, according to which foretelling or presentiments are conceptualised as saying, tasting or feeling something before something else. The metaphor PAST IS BEHIND is also present in the meaning of these nouns. In the noun *predokus* ‘foretaste’, the metaphor COGNIZING IS EATING⁹ is also present, according to which mental activities are conceptualized as tasting.

A third subgroup that should be mentioned here is the one consisting of the nouns *predočenje* ‘presentation’, *predodžba* ‘idea, conception’, *predočivost* ‘imageability’ and the like. These lexemes were all formed from the verb *predočiti* ‘display, put forward’, which comes from the root *oči* ‘eyes’, and their meaning is the result of the common metaphor KNOWING IS SEEING¹⁰, on the basis of which mental activities are conceptualised as the act of seeing. In these nouns, presenting, displaying and imaging something is conceptualised as putting objects in front of someone’s eyes.

In the conclusion to this part, we can say that in nominal derivatives, the prefix *pred-* sometimes retains its prototypical spatial meaning ‘before’, but in the majority of cases it is extended into the temporal domain via the TIME IS SPACE metaphor. In this second type of derivatives, only the prefix gets a metaphorical reading, while the meaning of the base remains the same. There is a third group of derivatives, in which the whole construction [prefix + noun] gets an extended meaning, motivated by metaphor and/or metonymy. Such insights are in line with the data from the literature on the preposition *pred*, according to which it extends its meaning to the temporal domain, but other than that, it does not develop many non-spatial meanings (Matovac, 2013).

⁸ It is very close to the very similar the LEADING IS A FORCE MOVING AN OBJECT FORWARD metaphor, cf. https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:LEADING_IS_A_FORCE_MOVING_AN_OBJECT_FORWARD.

⁹ Cf. https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:COGNIZING_IS_EATING.

¹⁰ Cf. https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:KNOWING_IS_SEEING.

Similar conclusions were signalled in a previous study (Petrak, 2021) with respect to the semantic structure of *pred-* verbs. More precisely, while the study did not find any *pred-* verbs in which the prefix retains its concrete spatial meaning, *pred-* verbs could be grouped into two groups identical to the second (temporal) and third (metaphorical and metonymical) groups in our study. In the first group, the prefix has a temporal reading, while the base retains its concrete meaning (e.g. *predugrijati* < *pred-* + *ugrijati* ‘heat’ > ‘preheat’), and in the second group, verbs develop an extended meaning as a whole, in which the two most frequent meanings are ‘leadership’ and ‘foretelling’ (e.g. *predvoditi* < *pred-* + *voditi* ‘lead’ ‘leading’ >; *predvidjeti* < *pred-* + *vidjeti* ‘see’ ‘foretell’ >).

4.3. Word-formation – semantic pairings

In this part of our analysis, we examine whether there are specific word-formation – semantic pairings or regularities in the formation of the Croatian lexicon with the prefix *pred-*.

Prefixed nouns exhibit a clear and straightforward preference for the temporal meaning. In other words, there is a formal-semantic regularity according to which [*pred-* + Noun] frequently results in nouns with a temporal meaning ‘before N’.

The suffixed nouns group is the most heterogenous one, in which various meanings appear: concrete (*predmetak* ‘prefix’), temporal (*predškola* ‘preschooler (f.)’) and metaphorical/metonymical (*predvodnik* ‘leader’), based on different metaphors and/or metonymies.

In prefix-suffix formation, *pred-* is typically used in its prototypical, spatial meaning to denote spaces that are located in front of another spatial landmark (e.g. *predvorje* ‘antechamber, entrance hall’ < *pred-* ‘before’ + *dvor-* ‘court’ + *-je* > denotes a space located in front of a large(r) front door), and in such formations the suffix *-je* is typically used. In other words, [*pred-* + N + *-je*] frequently results in nouns with a spatial meaning ‘the space in front of N’.

Backformations are very heterogenous and exhibit no clear regularities.

4.4. Comparison of high- and low-frequency types

Lastly, we need to address the question of whether there is any difference in the behaviour of high- and lower-frequency nouns. Below is a table summarizing word-formation data for low-frequency nouns:

	<i>Word-formation type</i>	<i>No.</i>	<i>%</i>
1	prefixation	491	96
2	suffixation	11	2,1
3	prefix-suffix combination	10	1,9
4	backformation	0	0
Total		512	100

Table 3: Word-formation types in low-frequency nouns

Low-frequency nouns exhibit the highest percentage of prefixation (96%, whereas it was 83,5% for high-frequency nouns), followed by suffixation, 2,1% (as opposed to 13% in high-frequency nouns), prefix-suffix combination 1,9% (as opposed to 3%). Backformations have not been found in the low-frequency range, while about 0,5% of high-frequency types exhibit that word-formation mechanism.

It can be concluded that low-frequency nouns formed with *pred-* are predominantly and almost systematically formed through prefixation, while other word-formation types are rather rare. Suffixation and prefix-suffix combination represent almost equal shares (2,1 and 1,9%, respectively), while there was a larger difference in the high-frequency group (13% and 3%, respectively).

On the semantic level, in the largest majority of the lower-frequency nouns, *pred-* has a temporal meaning (e.g. *predčistište* ‘ante-Purgatory’, *predijagnoza* ‘pre-diagnosis’, *predglaćanje* ‘pre-ironing’) according to the TIME IS SPACE metaphor mentioned before.

Data provided in this sub-section points to some differences between the high- and lower-frequency range, and explain why lower-frequency types (in this case, all nouns except for hapaxes) should be included in research on derivational processes.

5 Concluding remarks

Grammar-lexicon continuum

Our detailed study of the morphosemantic potential of the prefix *pred-* has demonstrated that it is found in numerous nouns whose morphosemantic structures exhibit interesting regularities. On the formal level, the prefix enters into four types of word-formation processes: prefixation, which is the prevalent one, suffixation, prefix-suffix formation and back-formation, the latter two being rather rare. Semantics-wise, the prefix in *pred-* nouns seldom retains its prototypical spatial meaning, and such nouns are usually the result of prefix-suffix combination. In the largest number of nouns, which are the result of prefixation, *pred-* gets an extended, temporal meaning according to the TIME IS SPACE metaphor. Suffixation usually produces nouns which get an extended meaning as a whole, based on metaphor, metonymy or both of these mechanisms. These are interesting morphosemantic regularities that have been found in the part of the Croatian nominal lexicon constructed with the prefix *pred-*. Our study of *pred-* nouns adds up to research proving that formal and semantic motivation are indeed intertwined (cf. Koch and Marzo, 2007), and sheds some more light on the semantic structure of motivated words, which remains a rather unexplored area of morphology (cf. Onysko and Michel, 2010).

References

- Stjepan Babić, Dalibor Brozović, Ivo Škarić, and Stjepko Težak. 2007. *Glasovi i oblici hrvatskoga književnoga jezika*. Nakladni zavod Globus, Zagreb.
- Stjepan Babić. 2002. *Tvorba riječi u hrvatskome književnome jeziku*. Globus and HAZU, Zagreb.
- Branimir Belaj. 2008. Pre-locativity as the schematic meaning of the Croatian verbal prefix *pred-*. *Jezikoslovlje* 9 (1-2): 123–140.
- Peter Koch and Daniela Marzo. 2007. A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. *Studies in Language* 31 (2): 259–291.
- Tanja Kuzman and Nikola Ljubešić. 2023. CLASSLA-web: Bigger and better web corpora for Croatian, Serbian and Slovenian on Clarin.si concordancers. <https://www.clarin.si/info/k-centre/classla-web-bigger-and-better-web-corpora-for-croatian-serbian-and-slovenian-on-clarin-si-concordancers/>
- George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The Journal of Philosophy* 77 (8): 453–486.
- George Lakoff. 1987. *Women, Fire and Dangerous Things. What Categories Reveal About the Mind*. University of Chicago Press, London, Washington, DC.
- Janet Littlemore. 2010. *Metonymy: Hidden Shortcuts in Language, Thought and Communication*. Cambridge University Press, Cambridge.
- Darko Matovac. 2013. *Semantika hrvatskih prijedloga*. PhD dissertation. Sveučilište J. J. Strossmayera, Osijek, Filozofski fakultet.
- Alexander Onysko and Sascha Michel. 2010. Introduction: unravelling the cognitive in word formation. In *Cognitive Perspectives on Word Formation*. De Gruyter Mouton, pages 1–28.
- Marta Petrak. 2021. Complex interplay of metaphor and metonymy: the case of *pred-* prefixed verbs. In *2nd International Conference for Young Researchers in Cognitive Linguistics (YRCL) Book of Abstracts*. Universidad de Alcalá, page 37.
- Jean Peytard. 1975. *Recherches sur la préfixation en français contemporain*. Librairie Honoré Champion, Paris.
- Günter Radden. 2003. The Metaphor TIME AS SPACE across Languages. In *Übersetzen, Interkulturelle Kommunikation, Spracherwerb und Sprachvermittlung - das Leben mit mehreren Sprachen. Festschrift für Juliane House zum 60. Geburtstag. Zeitschrift für Interkulturellen Fremdsprachenunterricht*. AKS-Verl, pages 226–239.
- Ida Raffaelli. 2004. Odnos strukturalne semantike prema kognitivnoj. *Suvremena lingvistika* 57-58(1-2):67–92.

- Ida Raffaelli. 2013. The model of morphosemantic patterns in the description of lexical architecture. *Lingue e linguaggio* 1:47–72.
- Ida Raffaelli. 2015. *O značenju. Uvod u semantiku*. Matica hrvatska, Zagreb.
- Eleanor Rosch. 1975. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General* 104: 193–233.
- Ljiljana Šarić. 2008. *Spatial Concepts in Slavic. A Cognitive Linguistic Study of Prepositions and Cases*. Harrasowitz Verlag, Wiesbaden.
- Danko Šipka. 1989. Tvorbena sredstva antonimizacije. *Jezik: časopis za kulturu hrvatskoga književnog jezika* 37(5): 139–145.
- Pavol Štekauer and Rochelle Lieber. 2005. *Handbook of Word-Formation*. Springer, Dordrecht.
- Andrea Tyler and Vyvyan Evans. 2003. *Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition*. Cambridge University Press, Cambridge.
- Stephen Ullmann. 1966. Semantic universals. In *Universals of language*. MIT Press, Cambridge MA, pages 217–262.
- Friedrich Ungerer. 2010. Word-Formation. In *The Oxford Handbook of Cognitive Linguistics*. Oxford University Press, Oxford.
- Sanja Vulić. 2020. Tvorbeno motivirane riječi u suvremenim gradišćanskohrvatskim književnim djelima. *Čakavska rič: Polugodišnjak za proučavanje čakavske riječi* XLVIII (1-2): 9–31.

Understanding Borrowing through Derivational Morphology: A Case Study of Czech Verbs

Abishek Stephen

Charles University,
Faculty of Mathematics and Physics,
stephen@ufal.mff.cuni.cz

Zdeněk Žabokrtský

Charles University,
Faculty of Mathematics and Physics,
zabokrtsky@ufal.mff.cuni.cz

Abstract

The transfer of morphemes across languages in language contact situations may lead to an alteration in the morphology of the recipient language. One of the possible outcomes can be the introduction of newer word forms or a formation of newer morphological variants of the existing word forms in the recipient language. Languages often borrow nominal roots and morphologically derive them into verbs and are thus integrated into respective derivational classes. This corpus-based analysis for Czech tries to show how synchronic derivational resources can be used to probabilistically analyze the effects of borrowing in language evolution by focusing on morphological integration of the borrowed nominal roots in verb formations.

1 Introduction

In Czech, the use of verbs like *studovat* ‘to study’, *rezervovat* ‘to reserve’, *fixovat* ‘to fix’, *blogovat* ‘to blog’, and so on is very common. On taking a closer look at such verbs we find that the verbal roots are of foreign origin and not native to Czech. Due to the rich derivational morphology, such verbs in Czech with foreign or borrowed roots take a number of prefixes as well. For example, *reagovat* ‘to react’ with the addition of the prefix *pře-* becomes *přereagovat* ‘to overreact’. Most of these verbs appear within the conjugation class *-ovat*.

According to Blaha (2022), the verbal conjugation with the affix *-ovat* is highly productive for the borrowed (nominal) roots, for especially those that denote an action done using an instrument like *scanovat* ‘to scan’ or an action defined after the concept denoted by the root like in *investovat* ‘to invest’ and this pattern has seen an increase in productivity in the last decade. In the presence of multiple conjugation classes in Czech, it looks like for the borrowed roots the affix *-ovat* is the most productive.

The process of integration or allocation of a derivational class happens once the foreign linguistic item is borrowed. In this study, we investigate this phenomenon in more detail to identify the underlying processes of such integration strategies that might have enabled the language to evolve special morphological machinery to deal with the incoming foreign material. Based on the integration strategies, we try to show that the existence of such internal mechanisms renders the language more flexible and competent to accept loanwords.

2 Background and Motivation

Based on sociolinguistic accounts, languages borrow primarily because of a need for a new concept or because of socio-attitudinal reasons (Campbell, 2020). For example, if a language does not have the concept ‘fax’ it will most probably be borrowed from a language where this concept exists and this borrowing process will be influenced by multiple linguistic and extra-linguistic factors. When a word for a particular concept is already a part of the language’s lexicon and still borrowing happens it is for prestige among the other social factors involved. For example, the dominance of Norman French led to the borrowings of culinary vocabulary from French to English even though English had words denoting those concepts (Campbell, 2020).

Languages thus borrow because of the social or attitudinal factors and also for grammatical reasons (Haspelmath, 2009). Ottawa-Hull French speakers might borrow from English because of their preference

for morphologically simple lexical items over more complex ones in French to express the same referent (Poplack, 2018). In these contact-induced changes, we can find the existence of certain asymmetries in the borrowability of linguistic items (Matras and Sakel, 2007) and these asymmetries reveal the properties of the human language faculty in terms of the stability of linguistic subsystems (Seifart, 2019).

Contact-induced borrowing can occur at variable rates during evolution due to bilingualism, the extent of contact between languages, the typological relatedness of languages or a combination of all of these factors (Thomason, 2001; Nelson-Sathi and List, 2011). According to Mufwene (2001) “Linguistic features are passed on primarily horizontally, more or less on the pattern of features of parasites, through speakers’ interactions with members of the same communicative network or of the same speech community. The default condition of linguistic transmission is with modification, however slight this may be. Horizontal and polyploidic transmission independent of generations makes it possible for a new feature to spread fairly rapidly”. This transfer of linguistic information can be visualized in parallel to gene transfer in molecular biology. The prokaryotic and eukaryotic evolution shows that the processes through which the gene families are created vary considerably based on the way the genetic material is transferred.

According to Hall et al. (2020) “Horizontal gene transfer (HGT) is particularly prevalent in prokaryotes, where it is one of the main mechanisms contributing to genetic variation and thus evolution”. If we were to look into how similar language and genome evolution are then the language evolution may resemble prokaryotic evolution (List et al., 2014). The horizontal or lateral gene transfer begins with the transfer of the foreign DNA in the cytoplasm followed by the recombination into the chromosome and integration with the gene regulatory circuits of the host (Skippington and Ragan, 2013).

Speaking of evolutionary changes, language evolution is usually looked upon in terms of family trees but it has been established that the horizontal components through lexical borrowing also contribute in evolution (Nelson-Sathi and List, 2011; List et al., 2014). Lexical borrowing can replace an existing word, introduce a new word that may co-exist with a native word having the same meaning or it can insert a new word referring to a concept that previously didn’t exist in the language (Monaghan and Roberts, 2019).

The incoming lexicon as a result of the lateral transfer or borrowing also needs to be adapted and integrated into the recipient language. This is very similar to the integration of the laterally acquired foreign genetic material into a host cell. According to Filipović (1981), the adaptation of loanwords on the morphological level is concerned primarily with the formation of its citation form. And this analysis is made based on the transmorphemization¹. Other claims suggest that speakers integrate verbs merely as lexical labels while others use them, to various degrees, as predicate-initiating devices (Matras and Adamou, 2020).

In case of the presence of multiple conjugation classes, for example, in Czech the citation forms of the verbs can take the affixes² *-ovat*, *-it*, *-at*, *-nout*, and so on. In other languages too, we see similar patterns. There has to be one form that is easily accessible or has a higher combinatory potential and thus will get attached to the incoming foreign root readily. In other words, a language could have possibly evolved or devised a mechanism for handling the morphology of the foreign linguistic materials by spreading the existing morphological processes to the borrowed vocabulary. In Croatian³, for the loanverbs, the English root gets attached to the infinitive affixes *-irati*, *-avati*, *-ivati*, *-ovati*, *-ati* according to the Croatian morphology and we find verbs like *intervjuirati* ‘to interview’, *flertovati* ‘to flirt’, etc. In Poplack (2018) we observe that the English bare infinitive itself serves as the root for conjugation when incorporated into Quebec French. English-origin verbs are assimilated into the *-er* group and conjugated according to French morphology like the verbs *mover* ‘to move’, *runner* ‘to run’, *shopper* ‘to shop’ and *skipper* ‘to skip’ to name a few.

Such evidences show that a language *somehow* assigns a conjugation class to the verbs formed with borrowed roots (Figure 1). Out of the total collection of roots or the *bag of roots* in the lexicon, verb

¹According to Filipović (1980), transmorphemization is one of the forms of substitution that comprises all the changes appearing in the adaptation of bound morphemes as they pass from the donor language to the recipient language.

²For simplicity we speak about affixes, but in fact the presented strings contain also endings (or also other affixes).

³As informed by Matea Filko through personal communication.

formation happens based on some underlying combinatorial mechanism let alone the conditions placed by different components of grammar like phonology, syntax, semantics, etc. With the current study, we empirically explore the possible reasons why in Czech the verbs with a borrowed root almost always fall into the *-ovat* conjugation class. We reason as to why only this particular affix is preferred over the others and how a language decides upon such a selection. We assume that these derivational processes make it possible for a language to accept foreign linguistic units and try to explore the reasons of morphological integration based on corpus analysis.

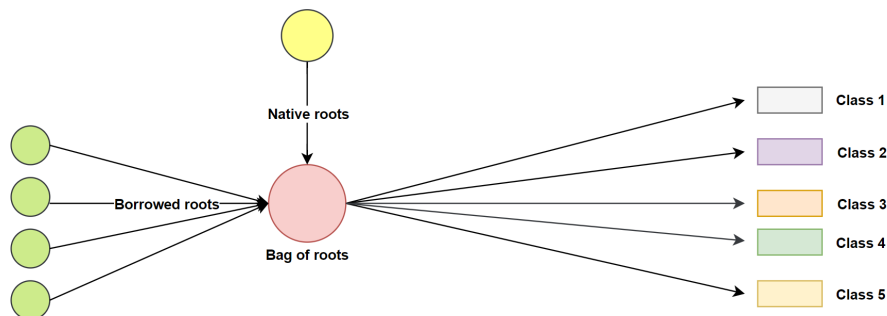


Figure 1: A potential flow of the borrowed and native roots into different conjugation classes

3 Approach

For our analysis we use DeriNet (Vidra et al., 2019). DeriNet⁴ is a lexical network that models word-formation relations in the lexicon of Czech. We only take into consideration the loanverbs i.e. verbs with a borrowed stem such as in Figure 2 based on the conjugation classes. There are tags within DeriNet for loanwords which were extracted in a supervised manner from multiple corpora based on language specific rewrite rules. Out of all the loanverbs we only consider those that are not idiosyncratic and are attested in the corpora. We present the unigram frequencies and the relative frequencies of native verbs and loanverbs belonging to different conjugation classes. We also calculate the conditional probabilities and entropies for the distribution of verbs with native and borrowed roots. We additionally calculate Dice coefficient and some other relevant statistic measures. The major topic of investigation is that when the verbs are borrowed into a language, there are certain affixes sensitive to get attached to the incoming foreign root. This when viewed through the lens of loanword integration and adaption seems like a rather probabilistic process than a discrete one. One of the possible reasons could be that the languages, in our case Czech could have possibly evolved special mechanisms to incorporate foreign linguistic material.

Frequency effects in this regard have also gained quite an attention. Pagel et al. (2007) report that the higher-frequency words are more stable and resistant to change or evolution. Such a word form is less likely to be replaced and it also won't admit co-existence with a semantically congruent counterpart but if the word form is represented in a less robust fashion then it is more likely to be replaced or to admit co-existence with a borrowed word form (Monaghan and Roberts, 2019). Hence, we find it worthwhile to analyze frequency effects in this regard i.e. how likely is it that a highly frequent word form would be a borrowed word, and so on. We also compare the derivational rate for the verbs with native and borrowed stems. The frequency of derivational nodes could possibly shed light on the difference in morphological productivity of both the classes of verbs under investigation.

4 Evaluation and Results

For our experiments, we consider only the verbs in DeriNet. We take the native and loanverbs based on corpus attestations (Table 1) i.e. the unigram corpus frequencies of the verbs belonging to both groups must be greater than 1 in the corpora. The absolute frequencies in DeriNet are taken from the Czech National Corpus, SYNv4 (Křen et al., 2016). As we focus on the loanverbs, we analyzed the derivational

⁴<https://ufal.mff.cuni.cz/derinet>

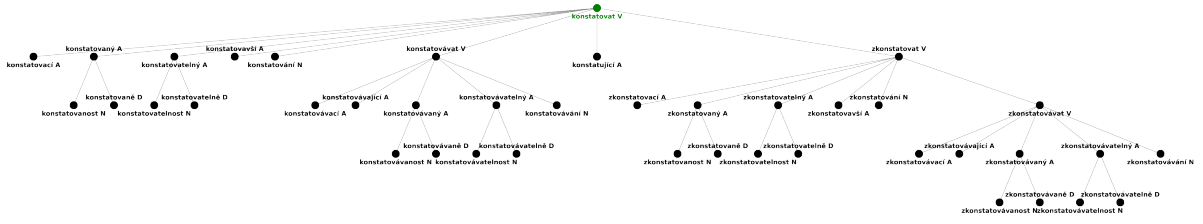


Figure 2: Word-formation relations for the most frequent loanverb *konstatovat* in DeriNet

affixes only for verbs and not for other parts-of-speech. On calculating the unigram frequencies for the

Type	Total verbs	Corpus attested
Native root	42930	19854
Borrowed root	13378	3972

Table 1: Frequencies of verbs in DeriNet

corpus-attested verbs in the DeriNet data we found that most of the verbs with a native root have the derivational affix *-at* followed by the affixes *-it* and *-ovat*. But we also find that a limited set of verbs with the affixes *-ít* and *-et* occur almost as frequently as the verbs with affixes *-at*, *-it* and *-ovat*. For the verbs with a borrowed root, almost all the verbs have the derivational affix *-ovat* followed by *-ovávat* and *-it*. (Table 2). Since our focus is on the loanverbs, we compare the frequencies only with those affixes with the native roots for which there are adequate counterparts with the borrowed roots. The frequencies point out that the conjugation classes of verbs have different *choices* of roots. We also compute the entropies for the distributions of the verbs with native and borrowed roots. Following the standard notion of entropy, we compute entropy H of a particular affix X as the negative summation of the log of relative frequencies of the affixes x , within the group of verbs with borrowed or native roots, $P(x)$.

$$H(X) = - \sum_{x \in X} P(x) \log_2 P(x) \quad (1)$$

To check for the strength of attraction between the conjugation affixes and the type of roots Dice coefficient was used as a statistical measure. It is one of the most common association measures used to detect collocations. Dice coefficient outperforms other association measures like mutual information, etc in the task of collocation detection (Kolesnikova, 2016). But for our analysis, we assume the combination of the root and affix is equivalent to a collocation.

Affixes	Native root	Tokens	Borrowed root	Tokens
<i>-at</i>	6481	19225633	62	10724
<i>-it</i>	4492	25400609	141	130751
<i>-ovat</i>	4132	10791009	3377	3749598
<i>-nout</i>	1780	6024365	41	22712
<i>-ovávat</i>	413	128614	307	3307
<i>-et</i>	778	12355416	3	347
<i>-ět</i>	519	6481870	28	860
<i>-át</i>	67	5676169	0	0
<i>-ít</i>	195	11439639	0	0
<i>-ýt</i>	28	228658	0	0

Table 2: Frequencies of derivational affixes

Origin of root	Affix	Lexicon frequencies		Corpus frequencies	
		P(Suffix Origin)	Entropy	P(Suffix Origin)	Entropy
Native	<i>-at</i>	0.326	2.330	0.200	2.803
	<i>-it</i>	0.230		0.254	
	<i>-ovat</i>	0.208		0.108	
	<i>-nout</i>	0.090		0.060	
	<i>-ovávat</i>	0.020		0.001	
	<i>-et</i>	0.040		0.123	
	<i>-ět</i>	0.026		0.064	
	<i>-át</i>	0.003		0.057	
	<i>-ít</i>	0.009		0.114	
	<i>-ýt</i>	0.001		0.002	
Borrowed	<i>-at</i>	0.015	0.873	0.023	0.401
	<i>-it</i>	0.035		0.033	
	<i>-ovat</i>	0.850		0.958	
	<i>-nout</i>	0.010		0.006	
	<i>-ovávat</i>	0.078		0.001	
	<i>-et</i>	0.001		0.000	
	<i>-ět</i>	0.007		0.000	
	<i>-át</i>	0.000		0.000	
	<i>-ít</i>	0.000		0.000	
	<i>-ýt</i>	0.000		0.000	

Table 3: Probabilities and entropies of derivational affixes

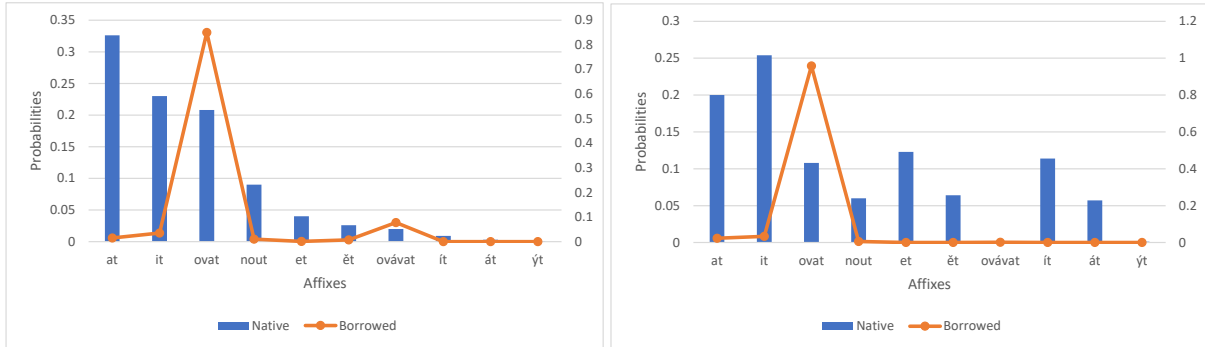


Figure 3: Probabilities based on lexicon frequencies Figure 4: Probabilities based on corpus frequencies

	<i>-at</i>	<i>-it</i>	<i>-ovat</i>	<i>-nout</i>	<i>-ovávat</i>	<i>-et</i>	<i>-ět</i>	<i>-át</i>	<i>-ít</i>	<i>-ýt</i>
Native root	0.322	0.404	0.188	0.114	0.002	0.220	0.122	0.107	0.205	0.005
Borrowed root	0.001	0.009	0.406	0.005	0.001	0	0	0	0	0

Table 4: Dice coefficient based on corpus frequencies

In Table 3, we observe that the entropy for the conjugation classes of verbs with native roots is much higher than the distribution of verbs with borrowed roots. The trend that the entropy of affixes is lower with borrowed roots is even stronger in a corpus with running text frequencies, as opposed to lexicon frequencies. One of the reasons is that the borrowed root almost always occurs with the affix *-ovat* (Figure 3 and Figure 4). This can be viewed as an analogy of the form $a : b = c : x$. In historical

	<i>do-</i>	<i>roz-</i>	<i>o-</i>	<i>po-</i>	<i>pod-</i>	<i>od-</i>	<i>u-</i>	<i>v-</i>	<i>vy-</i>	<i>z-</i>	<i>za-</i>	<i>pře-</i>	<i>před-</i>
Native root	616	649	568	661	134	506	652	623	1228	1263	1226	454	117
Borrowed root	108	38	70	44	7	104	23	47	187	338	214	140	14

Table 5: Lexicon frequencies of the prefixed verbs with native and borrowed roots

linguistics, an analogical change can be defined as a process whereby one form of language becomes more like another with which it is somehow associated (Arlotto, 1972). The analogy is also referred to as internal borrowing where a language *borrow*s some of its own patterns to change other patterns (Campbell, 2020). The conditional probabilities do indicate that one of the patterns of the derivation of verbs with a native root has been applied to derive verbs with a borrowed root i.e. the conjugation class *-ovat*.

Furthermore, the Dice coefficient scores in Table 4 also support that the affix most *sensitive* to borrowed roots is *-ovat*. This behaviour could be attributed to a quicker processing of verbs with this affix. Assuming the lexical units that have a higher information load are more costly to process, the lexical processing cost becomes directly proportional to the amount of information. The conditional probabilities in Table 3 indicate that the verbs with the affix *-ovat* carry the least amount of information⁵ and hence they are easier to process as compared to the other verbs with different affixes.

The argument around lexical processing itself requires its own space of discussion which is beyond the scope of this paper. But we would like to examine if the length of the affix plays any role behind the specific selection of *-ovat* for the borrowed roots. Most the affixes are of length 2 i.e. *-at*, *-it*, *-et*, and so on followed by the affixes *-ovat* and *-ovávat* with lengths 4 and 6 respectively. In word recognition and recall tasks, immediate memory span is better with short than with long words (Baddeley et al., 1975). The weighted average of the length of the affix and the conditional probabilities based on corpus frequencies were calculated (see Table 3) and it was found that the average length of the conjugation class affix is 2.3 for the verbs with the native roots and for the verbs with the borrowed roots it is 3.9. This again falls in accordance with the most preferred affixes by the both the type of roots. It is difficult to say if the borrowed roots fall into the *-ovat* class and hence a longer affix is preferred or it is the other way round. In Croatian, it can be speculated that the borrowed stems take the conjugation class with a longer affix like in *intervjuirati* ‘to interview’. In Slovak, we find examples like *fotografovať* ‘to take pictures’ and also Polish *komentować* ‘to comment’. Based on these examples, we might reach a probable conclusion that the verbs derived using a borrowed stem is *marked* with a longer suffix in the presence of multiple conjugation classes where the affix lengths vary. It might also indicate that the speakers of these languages label the loanverbs with a longer affix almost always but since we only deal with Czech primarily in this study, we do not make any concrete claims about other languages.

We also investigate if the prefixes play any role in the integration strategies of the borrowed roots. In Table 5, we present the lexicon frequencies of the prefixed verbs with native and borrowed roots. The difference in both the classes do not present any striking contrast. Moreover, based on the derivational trees (see Appendix) we can infer that the formation of verbs begins with the combination or selection of a root and a conjugation affix which is then followed by derivations by the addition of the prefixes and the roots do not seem to play any significant role in the selection of the prefixes. But in any case, we did calculate some frequency measures (Table 6) for the verbs with the conjugation affix *-ovat* and found that nearly 70% of the verbs with a native root are prefixed and only 45% of the verbs with a borrowed root are prefixed. To analyze the morphological productivity⁶, we calculated the average number of derivational nodes in the derivational tree for verbs with native and borrowed roots present in DeriNet (Figure 5). The results show that on an average a verb with a native root has 34.7 derived word forms whereas a verb with a borrowed root has 35.1 derived word forms. This indicates that most of the derivational processes are similar for the native and borrowed words.

⁵The amount of information carried is the negative logarithm of the probability.

⁶In a narrow sense the conditional probabilities in Table 3 can also serve as an indicator of morphological productivity.

	<i>do-</i>	<i>roz-</i>	<i>o-</i>	<i>po-</i>	<i>pod-</i>	<i>od-</i>	<i>u-</i>	<i>v-</i>	<i>vy-</i>	<i>z-</i>	<i>za-</i>	<i>pře-</i>	<i>před-</i>
Native root	112	123	141	129	30	126	129	106	306	270	293	119	26
Borrowed root	69	23	51	37	4	66	25	23	142	291	160	91	11

Table 6: Lexicon frequencies of the prefixed verbs with native and borrowed roots with affix *-ovat*

```

if word.is_loanword != "None" and word.absolute_count > 1:
    if word.is_loanword == "True":
        count_loan++
        count_children_loan += word.get_all_children()
        if word.pos == "VERB":
            count_verb_loan++
            count_children_verb_loan += word.get_all_children()
    else:
        count_native++
        count_children_native += word.get_all_children()
        if word.pos == "VERB":
            count_verb_native++
            count_children_verb_native += word.get_all_children()

avg_children_verb_loan: float = count_children_verb_loan / count_verb_loan
avg_children_verb_native: float = count_children_verb_native / count_verb_native

```

Figure 5: Pseudocode for extracting the number of derivations per word from DeriNet

For English as a donor language, [Monaghan and Roberts \(2019\)](#) report that for the mid- to high-frequency words in English the likelihood of borrowing drops but for mid- to low-frequency words (with frequencies less than one per ten thousand) the relationship is positive and monotonic i.e. the likelihood of borrowing increases. For analyzing the distribution of the verbs with borrowed and native roots, considering Czech as a recipient language, we calculated the probability of finding a loanword and compared it with the probability of finding a verb with a borrowed stem (Figure 6 and Figure 7). We observe that locating a loanword or a loanverb increases with an increase in the corpus frequencies i.e. a loanword can be as highly frequent as a native word.

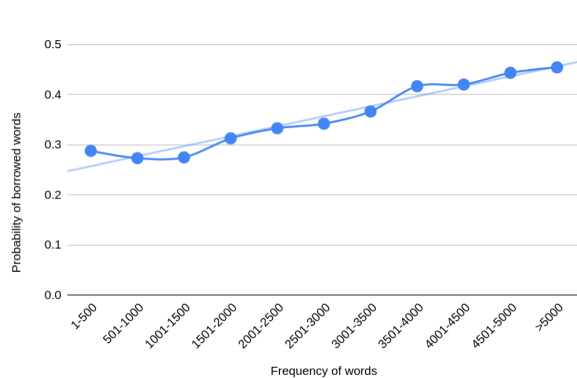


Figure 6: Probability of locating a loanword

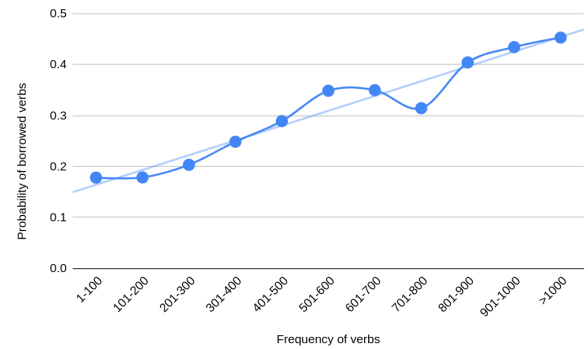


Figure 7: Probability of locating a verb with borrowed root

Based on all of the above statistical measures, we can conclude that the integration strategies in Czech treat the borrowed linguistic material in a very similar manner like the native vocabulary. The

derivations proceed in a very similar direction as the average number of derived word-forms indicates. It seems like due to the presence of multiple conjugation classes, loanverbs are preferably conjugated using the class with the longer affix. The observations based on the corpus frequencies indicate the *-ovat* conjugation class is neither the most frequent nor the least frequent choice of the native verbs. The aim of a morphological system is not to increase chaos and thus it identifies and expands the recurrent patterns to the borrowed words. This process also characterizes the cognitive capacity in a narrow sense.

For handling the integration, we assume that Czech chooses the pattern that has a central tendency given that a corpus is statistically dispersed. The measures of central tendency can be used to summarize the profile of verbs with either type of roots. We already know the probabilities and the corpus frequencies show that the native verbs falling within the *-ovat* conjugation class are neither the most frequent nor the least frequent. They lie somewhere in the middle of the distribution. The median of the corpus frequencies of verbs with a native root in Table 2 happens to be 8.6 million which is close to the corpus frequency of the conjugation class *-ovat*. There seems to be a higher probability that the choice of the conjugation class for loanverbs should fit around the median so as to keep the morphological system out of chaos. It is difficult to generalize this behaviour due to the lack of comparative corpus analysis across a good number of languages but it does seem to be prospective. The findings are purely empirical. There is a possibility that some extra-linguistic factor initiated the assimilation of loanverbs into the *-ovat* class. Language contact situations are complex and hence we cannot rule out the possibility that the integration strategies can be influenced by other factors as well.

5 Conclusion

This study analyzes the loanverbs in Czech based on DeriNet. The corpus analysis showed that the loanverbs almost always fall into the conjugation class *-ovat*. This can be seen as a strategy to mark the loanverbs with a longer suffix to indicate that the root is borrowed. Other statistical measures indicate that to keep the morphological system out of chaos that can be caused due to the incoming borrowed words, the central derivational process is extended towards handling the morphology of loanwords in the presence of multiple verb conjugation classes. These underlying mechanisms act as a positive pressure for accepting borrowings and thus contribute to the evolution of language in terms of its vocabulary range and morphological specializations to name a few among the various other modifications. Thus, the verb integration strategies or in this study the derivational processes led by the conjugation classes play a vital role in language change and evolution over time.

Acknowledgements

This work has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101). We thank the anonymous reviewers for their valuable comments and Michal Olbrich for his technical support with DeriNet.

References

- Anthony T. Arlotta. 1972. *Introduction to Historical Linguistics*. Houghton Mifflin, Boston.
- Alan D. Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* 14(6):575–589.
- Ondrej Blaha. 2022. Dynamics of Conjugation Pattern “Kupuje” in Contemporary Czech (On Material of Journalistic Texts, 1990–2019). *Bohemica Olomucensia* 14(2):40–54.
- Lyle Campbell. 2020. *Historical Linguistics: An Introduction*. Edinburgh University Press.
- Rudolf Filipović. 1980. *Transmorphemization: Substitution on the Morphological Level Reinterpreted*, Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l’Université de Zagreb, volume 25 (-), pages 1–8.
- Rudolf Filipović. 1981. *Morphological Categories in Linguistic Borrowing*, Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l’Université de Zagreb, volume 26 (-), pages 197–207.
- Rebecca J. Hall, Fiona J. Whelan, James O. McInerney, Yaqing Ou, and Maria Rosa Domingo-Sananes. 2020. Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes. *Frontiers in Microbiology* 11.
- Martin Haspelmath. 2009. *II. Lexical borrowing: Concepts and issues*, De Gruyter Mouton, Berlin, New York, pages 35–54.
- Olga Kolesnikova. 2016. Survey of Word Co-occurrence Measures for Collocation Detection. *Computacion y Sistemas* 20:327–344.
- Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kovářková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondříčka, and Adrian Zasina. 2016. SYN v4: large corpus of written czech. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Johann-Mattis List, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. 2014. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *BioEssays* 36(2):141–150.
- Yaron Matras and Evangelia Adamou. 2020. *Borrowing*, Routledge, pages 237–251.
- Yaron Matras and Janette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language* 31(4):829–865.
- Padraic Monaghan and Seán G. Roberts. 2019. Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition* 186:147–158.
- Salikoko S. Mufwene. 2001. *Language contact, evolution, and death: how ecology rolls the dice*, Cambridge University Press, page 145–166. Cambridge Approaches to Language Contact.
- Shijulal Nelson-Sathi and Johann-Mattis List. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society* pages 1794–1803.
- Mark Pagel, Quentin Atkinson, and Andrew Meade. 2007. Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History. *Nature* 449:717–720.
- Shana Poplack. 2018. Borrowing in the speech community. In *Borrowing: Loanwords in the Speech Community and in the Grammar*, Oxford University Press.
- Frank Seifart. 2019. *2. Contact-induced change*, De Gruyter Mouton, Berlin, Boston, pages 13–23.
- Elizabeth Skippington and Mark A. Ragan. 2013. *Lateral Genetic Transfer and Cellular Networks*, Springer New York, New York, NY, pages 123–135.
- S.G. Thomason. 2001. *Language Contact*. Edinburgh University Press.
- Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. 2019. DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, pages 81–89.

6 Appendix

Figure 8 and Figure 9 show the word-formation relations for the loanverbs *parkovat* ‘to park’ and *komentovat* ‘to comment’. The derivational trees presented here only contain few sub-branches and nodes that indicate the prefixation of the verbs. The complete visualization of the trees can be viewed using the DeriNet online viewer available at: <https://ufal.mff.cuni.cz/derinet/derinet-viewer>. Table 7 contains the corpus frequencies of the top 50 most frequent native verbs and loanverbs.

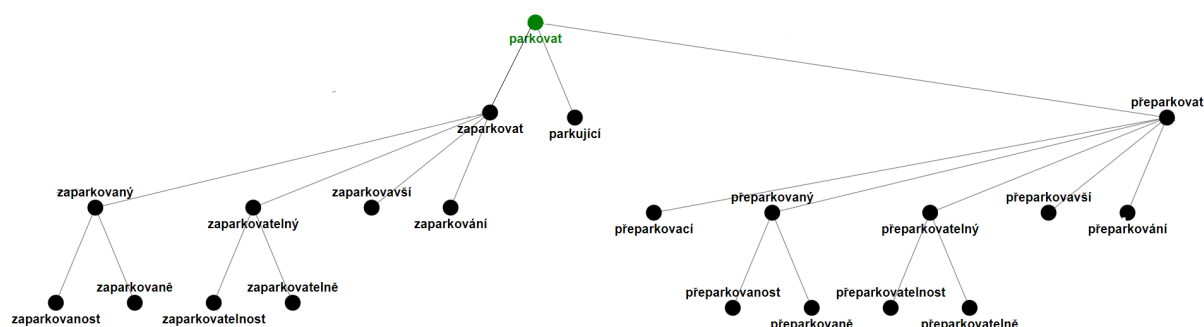


Figure 8: Word-formation relations for the loanverb *parkovat* in Derinet

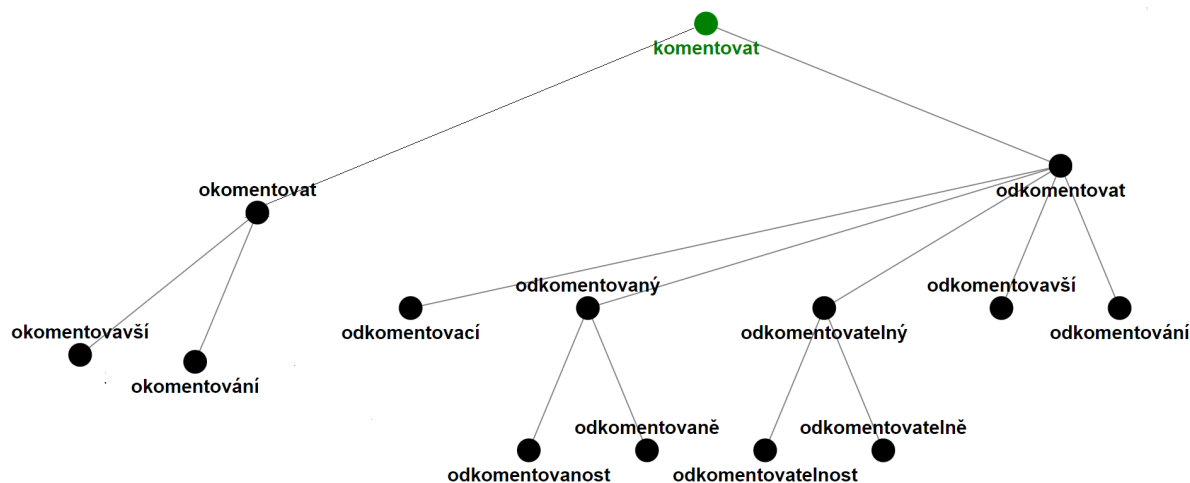


Figure 9: Word-formation relations for the loanverb *komentovat* in Derinet

Table 7: Corpus frequencies of 50 most frequent verbs with native and borrowed roots in DeriNet

Native	Corpus frequencies	Borrowed	Corpus frequencies
mušet	6758539	konstatovat	452229
chtít	6478359	investovat	272892
hrát	3334886	komentovat	248282
vědět	2973377	fandit	86867
potvrdit	996252	kontaktovat	83964
mluvit	924879	nominovat	78456
dávat	883624	instalovat	77583
jezdit	882479	rezignovat	72944
využít	875927	charakterizovat	72458
hledat	842844	kombinovat	67995
věnovat	828345	argumentovat	67867
řešit	825438	zareagovat	66842
příjet	798967	testovat	66225
vybrat	726762	zaregistrovat	65575
dosáhnout	724422	angažovat	59529
umět	674523	produkovat	58693
držet	664526	definovat	58582
nabídnout	663885	konkurovat	56473
psát	654944	akceptovat	53563
zajistit	653733	aplikovat	49925
představovat	648567	rekonstruovat	48446
připravovat	645968	parkovat	47738
koupit	628786	blokovat	46552
prohrát	592736	identifikovat	42783
odejít	578272	kopírovat	39466
bývat	577669	režírovat	38935
pořádat	553959	stabilizovat	38297
podívat	529262	nastudovat	37494
ztratit	525653	sportovat	37454
zúčastnit	522742	zrekonstruovat	36972
projít	522254	fotografovat	35896
odehrát	498576	eliminovat	35856
pohybovat	488446	iniciovat	34688
oznámit	488398	evakuovat	33585
zajímat	479852	avizovat	29564
uvidět	477494	formulovat	27853
vystoupit	469273	deklarovat	27839
upozornit	468482	kompenzovat	27287
bránit	466535	stagnovat	26643
sedět	452492	zkolabovat	25469
navštívit	448659	emigrovat	24536
vycházet	436446	interpretovat	23824
odpovídat	435243	finišovat	22958
popsat	435226	natankovat	9835
sejít	434936	pogratulovat	9742
připomínat	434872	proinvestovat	9726
připomenout	433386	marodit	9488
plánovat	433345	konkretizovat	9436
vzpomínat	428383	restaurovat	9355
určit	427396	zkorigovat	8988

Processing Croatian Morphology: Roots, Segmentation and Derivational Families

Krešimir Šojat

Faculty of Humanities
and Social Sciences
University of Zagreb
Croatia

kresimir.sojat@ffzg.unizg.hr

Matea Filko

Faculty of Humanities
and Social Sciences
University of Zagreb
Croatia

matea.filko@ffzg.unizg.hr

Abstract

This paper deals with the development of the Croatian derivational lexicon – CroDeriv. It is a 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. The database is available for online search according to various parameters. In this paper, 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, is the accurate recognition of lexical morphemes.

1 Introduction

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 of the lexicon 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 (Šojat et al., 2013). 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 (Filko et al., 2020).

In this paper, we present further development and enrichment of the existing version of CroDeriv. The paper is structured as follows: in Section 2.1 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. In section 2.2, the main derivational processes are presented as well as some that are not described or that are only marginally described in the existing literature. Each derivation process we describe is accompanied by examples. In section 2.3, we illustrate the structure of derivation families and lexical entries in CroDeriv. In section 3, we discuss some problems we have encountered in our work and outline possible solutions. We finish the paper with the Conclusion and the outline of future work.

2 Morphological analysis

2.1 Segmentation

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

Morpheme is the basic morphological unit. Usually, it is defined as the smallest language sign, i.e. the smallest language unit that can be associated both with the expression on one side and the content on the other (Marković, 2012; Silić and Pranjković, 2005; Barić et al., 1995). In other words, morphemes are the smallest units in the linguistic analysis with their meaning (Haspelmath and Sims, 2010; Booij, 2005). It is important to emphasize that morphemes are abstract units whereas morphs are their physical realization.

Types of morphemes recognized in lexemes are prefixes, lexical morphemes (roots), derivational suffixes, inflectional suffixes, and interfixes for compounds. Each type of morpheme can occur more than once in the morphological structure of lexemes.¹ The following example illustrates multiple prefixation and suffixation in derivation:

s-po-raz-um-je-ti se 'come to an agreement';

s = prefix; *po* = prefix; *raz* = prefix; *um* = root; *je* = suffix; *ti* = suffix; *se* = reflexive particle

As presented by Filko et al. (2019, 2020), 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.

For example, the surface layer segmentation of the verb *raščišćavati* 'to clean up_{IPF}' can be represented as: *raš-čišč-av-a-ti*; *raš* = prefix; *čišč* = root; *av* = derivational (aspectual) suffix; *a* = derivational (thematic) suffix; *ti* = inflectional (infinitive) suffix.

At the deep layer of presentation, the prefixal allomorph *raš* is connected to its representative morph *raz*, the root allomorph *čišč* to its representative morph *čist*, and the suffixal morph *av* to its representative morph *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*.

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

2.2 Derivational Processes

Two major word-formation processes in Croatian are derivation and compounding. The main difference between them is that word-formation processes based on derivation involve lexemes with one lexical morpheme, i.e. derivatives share the same lexical morpheme, whereas word-formation processes based on compounding involve lexemes with two or more lexical morphemes. In other words, compounds have usually two or possibly more different lexical morphemes.

Further in this work, we focus exclusively on derivation and discuss relations between lexemes that share the same root. Generally, derivation can be described as a word-formation process that is based on adding one or more affixes to lexical morphemes. Types of affixes recognized in Croatian lexemes are prefixes, suffixes, and interfixes for compounds. That means that the 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. In other words, 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. Generally, suffixation is the most productive derivational process. In the development of CroDeriv the following derivational processes were recognized:

¹There 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.

1. **suffixation** – addition of single or multiple suffixes or substitution of suffixes

- *bac(ati)* 'to throw' + *-ač* = *bacač* 'thrower, pitcher'
- *kazališt(e)* 'theater' + *-ar* + *-ac* = *kazalištarac* 'theater artist'
- *bac(iti)* _{PF} 'to throw' + *-ati* = *bacati* _{IPF} 'to throw'

2. **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'

3. **simultaneous prefixation and suffixation**

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

4. **back-formation + zero suffixation** - subtraction of stems

- *upis(ati)* 'to enroll' + \emptyset = *upis* 'enrollment'
- *uvid(jeti)* 'to see, to realize' + \emptyset = *uvid* 'insight'
- *dokaz(ati)* 'to prove' + \emptyset = *dokaz* 'proof'

5. **SE** - addition of the reflexive particle *se*²

- *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'

6. **ablaut** - a systematic variation of vowels in the same root, usually combined with various types of affixation

- *sagledati* _{PF} 'to perceive' = *saglédati* _{IPF} 'perceive'
- *pomoći* _{PF} 'to help' = *pomagati* _{IPF} 'to help'
- *smrdjeti* 'to stink' = *smrad* 'smell, stench'

7. **conversion / zero derivation** - derivation without any change in form of the stem

- *mlada* 'young (adjective)' = *mlada* 'bride (noun)'
- *nečist* 'impure (adjective)' = *nečist* 'dirt (noun)'
- *leteći* 'flying (participle, verbal adverb)' = *leteći* 'flying (adjective)'

These are major processes used in the derivation of Croatian lexemes. However, 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 most of these 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:

1. **ablaut + suffixation**

- *prigovor(iti)* _{PF} + *-ati* 'to complain' = *prigovarati* _{IPF} 'to complain'
- *bra(ti)* 'to pick' + *-ba* = *berba* 'harvest'

²The 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*.

2. prefixation + ablaut

- *pre-* + *zvati se* 'have a name' = *prezivati se* 'have a surname'

3. prefixation + ablaut + suffixation

- *o-* + *govor(iti)* 'to speak' + *-ati* = *ogovarati* 'to slander'
- *na-* + *vod(i-ti)* 'to lead_{IPF}' + *-ti* = *navesti* 'to lead_{PF}'

4. 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'

5. prefixation + se

- *na-* + *jesti* 'to eat' + *se* = *najesti se* 'to eat one's fill'
- *za-* + *trčati* 'to run' + *se* = *zatrčati se* 'to start running'

6. prefixation - se (dropping out of se)

- *u-* + *suglasiti (se)* 'to agree' = *usuglasiti* 'to agree, to get along'

7. ablaut + back-formation

- *iz(a)bra(ti)* 'to pick' + \emptyset = *izbor* 'choice'
- *razves(ti se)* 'to divorce' + \emptyset = *razvod* 'divorce'
- *opozva(ti)* 'to recall' + \emptyset = *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. We discuss the structure of derivational families and derivational relations between lexical entries in more detail in the next section.

2.3 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.³ 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_{PF}' from the root *bac*, *ruka* 'hand' from the root *ruk*, and *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 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, e.g.:

- *baciti* 'to throw' - *izbaciti* 'to throw out', *odbaciti* 'to reject', *ubaciti* 'to throw into' etc. All second-degree derivatives in this derivational families are derived via prefixation.
- *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.
- *nov* 'new' - *novac* 'money', *novak* 'rookie', *novost* 'news' (suffixation), *obnoviti* 'to renew', *ponoviti* 'to repeat' (prefixation + suffixation) etc.

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, e.g.: *ruče* 'gymnastic arms', *rukav* 'sleeve', *ručerda*, *ručetina* 'hand, (augmentative)', *ručica*, *ručka* 'handle', *naručje* 'bosom', *narukvica*

³In rare cases where we cannot base a family on only one lexeme, two lexemes are found at the center of the derivational family.

'bracelet' are the second-degree derivatives of the stem *ruka* that do not motivate any other lexeme. 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. This is the maximum number of derivatives in derivational chains recorded so far. It is possible that this number will increase with the further expansion of CroDeriv. For example:

- 1. *govor* 'speech' - 2. *govoriti* 'to speak' - 3. *odgovoriti* - 'to answer, to respond' - 4. *odgovarati* 'to answer, to match, to account for, to be responsible for' - 5. *odgovoran* 'responsible' - 6. *neodgovoran* 'irresponsible' - 7. *neodgovornost* 'irresponsibility'.
- 1. *glas* 'voice, tone, vote' - 2. *glasiti* 'to be addressed to, to read' - 3. *suglasiti se* 'to agree' - 4. *usuglasiti* 'to agree_{PF}' - 5. *usuglašavati* 'to agree_{IPF}' - 6. *usuglašavan* 'agreed upon (participle)' - 7. *usuglašavanje* 'harmonization'.

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*. The full structure of lexical entries in CroDeriv is presented in Filko et al. (2019, 2021).

In Table 1 below, we show how the lexical material is processed and prepared for input into CroDeriv. The examples are from the derivational family structured around the root *SĚK*. Its meaning is associated with cutting and dismembering. The first-degree derivative is the verb *sjeći* 'to cut'.⁴ We use the symbol *ě* for the reflexes of Proto-Slavic *jat* in the contemporary Croatian language. In this way, we solve the problem of numerous surface allomorphy and connect all reflexes to the representative *ě* at the deep layer. Note that there are four allomorphs at the surface layer of the same root in only nine examples in Table 1 below (SL column). At the deep layer there is only one representative morph - *sěk*, except in the last example. We will discuss this and similar cases in the next section.

I	II	SL	DL
sjeći, V - sjek + ti (S)		sje-ći	sěk-ø-ti
	sjecište, N - sjek(ti) + ište (S)	sjec-išt-e	sěk-išt-e
	sjekotina, N - sjek(ti) + otina (S)	sjek-ot-in-a	sěk-ot-in-a
	sječa, N - sjek(ti) + ja (S)	sječ-a	sěk-j-a
	siječanj, N - sjek(ti) + anj (S)	siječ-anj-ø	sěk-nj-ø
	sječivo, N - sjek(ti) + ivo (S)	sječ-iv-o	sěk-iv-o
	sjekira, N - sjek(ti) + ira (S)	sjek-ir-a	sěk-ir-a
	sjekutić, N - sjek(ti) + utić (S)	sjek-ut-ić-ø	sěk-ut-ić-ø
	sjekati, V - sjek(ti) + kati (S)	sjec-k-a-ti	sěc-k-a-ti

Table 1: An example from the derivational family of the root *SĚK* 'to cut'⁵

3 Discussion

In the previous sections, we indicated that each lexeme in CroDeriv is morphologically segmented and that the segmentation is performed at two layers - surface and deep. We also mentioned that in CroDeriv we combine two types of morphological data, i.e. in addition to the morphological segmentation for each lexeme, we record the word-formation relations with other lexemes as well as word-formation processes by which the lexemes were created. In Figure 1, we present a part of the derivational family of the root *VID* 'sight, to see'. For each lexeme, we provide information on the word class (*N* = noun, *V* = verb, *GPR* = active past participle, *GPT* = passive past participle etc.), stem, affixes that participate in

⁴I = first-degree derivatives, II = second-degree derivatives, SL = segmentation at the surface layer, DL = segmentation at the deep layer (cf. Section 2.1), V = verb, N = noun, (S) = suffixation.

the derivational process, and the type of the derivational process (S = *suffixation*, $SB+S$ = *subtraction + zero suffixation*, P = *prefixation* etc.). Columns I, II, III, etc. indicate whether a lexeme is a first-, second- or third-degree derivative. The final two columns refer to morphological segmentation at surface (PP) and deep (DP) layer. We have indicated that in CroDeriv's lexical entries, we do not provide information about full derivational chains. Instead, we provide information about the stem that served for the derivation of that lexeme. Information about the full derivation chain can be found using the visualization tool available to users of the lexicon.

I	II	III	IV	V	VI	VII	PP	DP
KORIJEN VID-								
vid, N < vid + \emptyset (S)							vid	vid- \emptyset
vidik, N < vid + ik (S)							vid-ik	vid-ik- \emptyset
vidikovac, N < vidik + ovac (S)							vid-ik-ov-ac	vid-ik-ov-c- \emptyset
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	vid-ov-it- \emptyset
vidovnjak, N < vidov(it) + njak (S)							vid-ov-njak	vid-ov-njak- \emptyset
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(tti) + i (S)							vid-i-o	vid-ě-l
vidjelac, N < vidjel + ac (S)							vid-je-l-ac	vid-ě-l-c- \emptyset
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)							vid-e-n	vid-je-n- \emptyset
viđenje, N < viđen + je (S)							vid-e-n-j-e	vid-je-n-j-e
vidan, A < vid(jeti) + an (S)							vid-an	vid-n- \emptyset
vidnost, N < vid(a)n + ost (S)							vid-n-ost	vid-n-ost- \emptyset
vidljiv, A < vid(jeti) + lživ (S)							vid-lživ	vid-lživ- \emptyset
vidljivost, N < vidljiv + ost (S)							vid-lživ-ost	vid-lživ-ost- \emptyset
nevidljiv, A < ne + vidljiv (P)							ne-vid-lživ	ne-vid-lživ- \emptyset
nevidljivost, N < nevidljiv + ost (S)							ne-vid-lživ-ost	ne-vid-lživ-ost- \emptyset
vidati, V < vid(jeti) + jati (S)							vid-a-ti	vid-ja-ti
izvidjeti, V < iz + vidjeti (P)							iz-vid-je-ti	iz-vid-ě-ti
izvid, N < izvid(jeti) + \emptyset (SB+S)							iz-vid	iz-vid- \emptyset
izvidni, A < izvid + ni (S)							iz-vid-n-i	iz-vid-n-i
izvidnica, N < izvidn(i) + ica (S)							iz-vid-n-ic-a	iz-vid-n-ic-a
izvidnik, N < izvidn(i) + ik (S)							iz-vid-n-ik	iz-vid-n-ik- \emptyset
izvidnički, A < izvidnik + ski (S)							iz-vid-n-ič-k-i	iz-vid-n-ik-sk-i
izvidaj, N < izvid + jaj (S)							iz-vid-aj	iz-vid-jaj- \emptyset
izvidajni, A < izvidaj + ni (S)							iz-vid-aj-n-i	iz-vid-jaj-n-i


Figure 1: The excerpt of the derivational family for the root *VID-*

In Figure 2, we give an example of how the entry in CroDeriv is structured. We also show a visualization tool used in the new CroDeriv's online search interface that shows the full derivation chain for the lexeme *zapisničarka* 'scorer, clerk (female)'. The full derivational chain is as follows:

- *pisati* 'to write' - *zapisati* 'to write down' - *zapisan* 'written down (participle)' - *zapisnik* 'record, minutes' - *zapisničar* 'scorer, clerk (male)' - *zapisničarka* 'scorer, clerk (female)'.

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. Such an approach to word formation in Croatian is new since it does not assume the existence of stems in which certain morpho-phonological processes have already been carried out before a certain word-formation process began. 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



Details
LEMMA zapisničarka
PART OF SPEECH noun
MORPHOLOGICAL STRUCTURE - SURFACE LAYER <div> za - pis - n - ič - ar - k - a </div>
MORPHOLOGICAL STRUCTURE - DEEP LAYER <div> za - pis - n - ik - ar - k - a </div>
WORD-FORMATION PATTERN <div> zapisničar - ka </div>
WORD-FORMATION PROCESS suffixation (noun > noun)
STEM zapisničar

Figure 2: The lexical entry *zapisničarka* in the new CroDeriv search interface

implemented in stems.⁶ Unlike the approach to Croatian morphology in CroDeriv, such an approach is represented in many works on word formation in Croatian and related Slavic languages (Babić, 2002; Klajn, 2002, 2003).

Although there are many advantages to the approach we advocate, there are certain cases that raise questions. We have stated that in Table 1 above, in the last example, the root at the deep layer is not connected to the morph that is representative of other root allomorphs. This also applies to examples 3. and 4. below;

1. *sjěci* 'to cut' - *sje*-*ći* / *sěk*-*ø*-*ti*
2. *sjeknuti* 'to cut (deminutive)' - *sjek*-*n*-*u*-*ti* / *sěk*-*n*-*u*-*ti*
3. *sjeckati* 'to cut (deminutive)' - *sjec*-*k*-*a*-*ti* / *sěc*-*k*-*a*-*ti*
4. *sjecnuti* 'to cut (deminutive)' - *sjec*-*n*-*u*-*ti* / *sěc*-*n*-*u*-*ti*

We will give a few more examples from another derivational family:

1. *pucati* 'to crack, to fire' - *puc*-*a*-*ti* / *puk*-*a*-*ti*
2. *puckati* 'to crack (deminutive)' - *puc*-*k*-*a*-*ti* / *puc*-*k*-*a*-*ti*
3. *pucnuti* 'to crack (deminutive)' - *puc*-*n*-*u*-*ti* / *puc*-*n*-*u*-*ti*

⁶In terms of morpho-phonological rules, we largely follow Marković (2013), a modern, precise, and extensive account of Croatian morpho-phonology.

The reason why we here list different root morphs in the deep structure is that there is no morpho-phonological rule that could explain the change of the root *puk* to *puc* before the diminutive suffix *-k* in the contemporary Croatian language. The same holds for the root *sěk* in the examples above. In addition, in example 2 for the root *sěk*, the deep-layer segmentation is *sěk-n-u-ti*. In example 4, the deep-layer segmentation is *sěc-n-u-ti*. In other words, we have two different root allomorphs in the same phonological environment. The same holds for example 3 for the root *puk*. Here again, the deep-layer segmentation is *puc-n-u-ti*, although there is a lexeme *puknuti* 'to crack, to fire' which is at the deep layer segmented as *puk-n-u-ti*.

Marković (2013, p. 140, 146) considers such examples to be "pre-sibilarized" or "pre-iotized". He states that in many similar examples "we have a possible sibilarization, however, it is probably more elegant to connect them with a pre-sibilarized verb root" (Marković, 2013, p. 140). The author does not provide an additional explanation, and we interpret this to mean that pre-sibilarization or pre-iotation were carried out in the earlier stages of language development and cannot be explained by the rules that apply in the contemporary language (cf. Mihaljević, 1991). 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.

We encountered a similar problem with lexemes derived by ablaut. For example:

1. *brati* 'to pick' - *berba* 'harvest' - *birati* 'to choose' - *izbor* 'choice',
2. *teći* 'to flow' - *protjecati* 'to flow' - *protok* 'flow',

Here, the question also arises as to which of the root allomorphs to take as the representative one, since morpho-phonological rules cannot justify the selection of only one. In CroDeriv we use a similar solution as in the examples above. The segmentation at the deep layer is as stated in the above examples for the roots *sěk* and *puk*, but one of the root allomorphs is taken to be representative and listed in parentheses. In this way, we can present such lexemes as members of the same derivational family.

The next problem we encountered relates to the homographic roots. For example, the verbs *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'. Both first-degree lexemes and many derivatives in their derivational families are semantically very similar. We solve the problem with homographic roots by marking them with a different number: *kup₁* for lexemes semantically associated with buying, *kup₂* for lexemes associated with gathering, *kup₃* for lexemes associated with bathing, *kup₄* for lexemes associated with docking etc. As for their semantic distinction, checking in etymological dictionaries (Skok, 1971, 1972; Matasović et al., 2016, 2021; Snoj, 2003) is the only way to solve such problems.

The last issue we will discuss here refers to the structuring of derivational families composed of suppletive stems. The problem we have not tackled yet concerns one of the biggest families in terms of the number of its members - the one which contains verbs like *ići* 'to go' and *otići* 'to leave', as well as *doći* 'to come_{PF}' and *dolaziti* 'to come_{IPF}'.

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 shown as follows:

- *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 *do-* and the suffix *-ći*, which is an allomorph of the infinitive ending *-ti*:

- *doći* 'to come' - *do* (prefix)-*XXX-ći* (suffix).

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_{PF}' and *dolaziti* 'to come_{IPF}', *naći* 'to find_{PF}' and *nalaziti* 'to find_{IPF}', *ući* 'to enter_{PF}' and *ulaziti* 'to enter_{IPF}' 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.

4 Conclusion

In this paper, we have briefly presented the structure of the CroDeriv, the derivational lexicon for Croatian which provides information about the morphological structure of words and about derivational links with other words, thus forming the derivational families. Since the structure of the lexical entries in CroDeriv has been explained in more detail in previous work (e.g. (Filko et al., 2020)), here, we have focused on the derivational processes in Croatian that have not yet been recognized in the existing literature. These processes emerged when the Croatian words were analyzed in the format used in CroDeriv. Moreover, such a formal analysis has forced us to find both computationally applicable and theoretically plausible solutions for unsolved (and even theoretically untackled) problems in Croatian morphology and word formation in order to include very frequent, but irregular lexemes in CroDeriv. Only a handful of the most interesting ones were presented here due to the limitations of this paper, but we can foresee that even more problems of this kind will emerge with further analysis of the data. We hope that the procedure and general rules applied in the examples presented here could be, with or without the modifications, applied to future issues, as well.

References

- Stjepan Babić. 2002. *Tvorba riječi u hrvatskome književnome jeziku*. Hrvatska akademija znanosti i umjetnosti : Globus, Zagreb.
- Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, and Marija Znika. 1995. *Hrvatska gramatika*. Školska knjiga, Zagreb.
- Geert Booij. 2005. *The Grammar of Words. An introduction to Linguistic Morphology*. Oxford Textbooks in Linguistics. Oxford University Press, New York.
- Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2019. Redesign of the Croatian derivational lexicon. In Zdeněk Žabokrtský, Magda Ševčíková, Eleonora Litta, and Marco Passarotti, editors, *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Karlovo sveučilište, Prag, pages 71–80.
- Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2020. *The Design of Croderiv 2.0*. *The Prague Bulletin of Mathematical Linguistics* 115:83–104. <https://doi.org/10.14712/00326585.006>.
- Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2021. Deriving the graph: Using affixal senses for building semantic graphs. In Fiammetta Namer, Nabil Hathout, Stéphanie Lignon, Magda Ševčíková, and Zdeněk Žabokrtský, editors, *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology*. Karlovo sveučilište, Prag, pages 120–128.
- Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*. Understanding Language. Hodder Education, London, 2nd edition.
- Ivan Klajn. 2002. *Tvorba reči u savremenom srpskom jeziku: prvi deo: slaganje i prefiksacija*. Zavod za udžbenike i nastavna sredstva : Institut za srpski jezik SANU ; Matica srpska, Beograd: Novi Sad.
- Ivan Klajn. 2003. *Tvorba reči u savremenom srpskom jeziku: drugi deo: sufiksacija i konverzija*. Zavod za udžbenike i nastavna sredstva : Institut za srpski jezik SANU ; Matica srpska, Beograd: Novi Sad.
- Ivan Marković. 2012. *Uvod u jezičnu morfologiju*. Number 6 in Biblioteka Thesaurus. Disput, Zagreb. OCLC: 815718585.

- Ivan Marković. 2013. *Hrvatska morfonologija*. Number 7 in Biblioteka Thesaurus. Disput, Zagreb.
- Ranko Matasović, Dubravka Ivšić Majić, and Tijmen Pronk. 2021. *Etimološki rječnik hrvatskoga jezika*, volume Sv. 2, O-Ž. Institut za hrvatski jezik i jezikoslovlje, Zagreb.
- Ranko Matasović, Tijmen Pronk, Dubravka Ivšić, and Dunja Brozović-Rončević. 2016. *Etimološki rječnik hrvatskoga jezika*, volume Sv. 1, A-Nj. Institut za hrvatski jezik i jezikoslovlje, Zagreb.
- Milan Mihaljević. 1991. *Generativna i leksička fonologija*. Školska knjiga, Zagreb.
- Josip Silić and Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta*. Školska knj, Zagreb. OCLC: ocm70847560.
- Petar Skok. 1971. *Etimologijski rječnik hrvatskoga ili srpskoga jezika*, volume Knjiga 1. Jugoslavenska akademija znanosti i umjetnosti, Zagreb.
- Petar Skok. 1972. *Etimologijski rječnik hrvatskoga ili srpskoga jezika*, volume Knjiga 2. Jugoslavenska akademija znanosti i umjetnosti, Zagreb.
- Marko Snoj. 2003. *Slovenski etimološki slovar*. Modrijan, Ljubljana.
- Krešimir Šojat, Matea Srebačić, and Vanja Štefanec. 2013. CroDeriV i morfološka raščlamba hrvatskoga glagola. *Suvremena lingvistika* 75:75–96.

Ontological modeling of morphological entities, allomorphy and representation in Modern Greek derivation

**Nikos
Vasilogamvris**
Ionian University
Corfu, Greece
l20vasi@ionio.gr

**Michalis
Sfakakis**
Ionian University
Corfu, Greece
sfakakis@ionio.gr

**Giannoula
Giannouloupoulou**
NKUA
Athens, Greece
giannouloup@ill.uoa.gr

**Maria
Koliopoulou**
NKUA
Athens, Greece
mkoliopoulou@gs.uoa.gr

Abstract

In the present article, we ontologically explore the entities of Modern Greek (MG) morphology as well as the variety of their allomorphic and representational relationships. The aim of this modeling is to fully enable the representation of lexical data in the MMoOn ontology and to propose an interactive allomorphy framework for MG derivation. According to this, interconnected allomorphy paradigms and derivational rules are placed inside the ontology, engulfing both the Permanent and Dynamic lexicon so that lexical data can be generated automatically and be morphologically justified. In respect of the morphological entities representation, different examples are presented to elaborate how allomorphy or morphological semantics affect them, as they show different or identical phonetic, morphemic and orthographic forms.

1. Introduction

Modern Greek (MG) is a synthetic inflectional language that presents a variety of morph types participating in complex morphological structures. Moreover, a significant characteristic is that it engulfs several non-transparent or phonologically unjustified allomorphic forms partly originated from Ancient Greek (AG) or based on AG roots. In order to explore language derivational processes, it is necessary to identify the different types of morphs, especially the stem and affix concepts and their subcategories. But it is equally important to look into these entities under the phenomenon of allomorphy involved in MG derivation and place it within suitable derivational environments (Melissaropoulou & Ralli, 2009) for creating a framework towards the generation of new forms.

In what follows, in section 2, we explore the different morphological entities of MG participating in derivation and then we focus on the types of allomorphy and propose a framework in which it can operate and be modeled. Then, we present the different representational aspects of these entities that justify the MMoOn ontology conceptual analysis. Finally, in section 3, we conclude on the topic.

2. Morpho-Ontological analysis

2.1. MG morphological typology

Morphemes or more precisely their realizations, *morphs*, are divided into two broad categories: *free* and *bound* (Booij, 2012; Ralli, 2005; Spencer, 2017). Free morphs are mono-morphemic words, either of *grammatical* or *lexical* nature, while bound cannot stand alone as free words and can be either roots, stems, affixes, confixes (Giannouloupoulou, 1999) or bound stems (Ralli, 2005).

Roots are the keystones of a lexeme but as Ralli postulates (Ralli, 2005), a root concept in MG cannot easily be located because roots are traced back in AG lexical forms. It would be more sensible, then, to use a *Stem* concept that may be either a *Base* (an initial stem) (e.g. *χορ-* (*xor-*) > *χορός* (*xorós*) ‘dance’) or an *Affixed Base* (e.g. *χορεύ-* (*xorév-*) > *xorévo* ‘to dance’).

Affixes are bound morphs that append to bases, operating as “satellites”, to form new affixed bases according to their *categorial signature* (Ralli, 2005). Affixes are divided into *Prefixes* when they precede

(e.g. *δια-* in *δια-δρασ-* (*δια-δρας-*)), or *Suffixes*, when they follow stems. Prefixes may also precede words, thus forming new words (e.g. *δρω* (*δρο*) ‘to act’ > *δια-δρω* (*δια-δρό*) ‘to interact’). Suffixes may in turn be of *Derivational* (e.g. *-εϋ-* in *χορ-εϋ-ω* (*χορ-έϋ-ω*) ‘to dance’) or *Inflectional* (*-ω* in *χορεύ-ω*) nature.

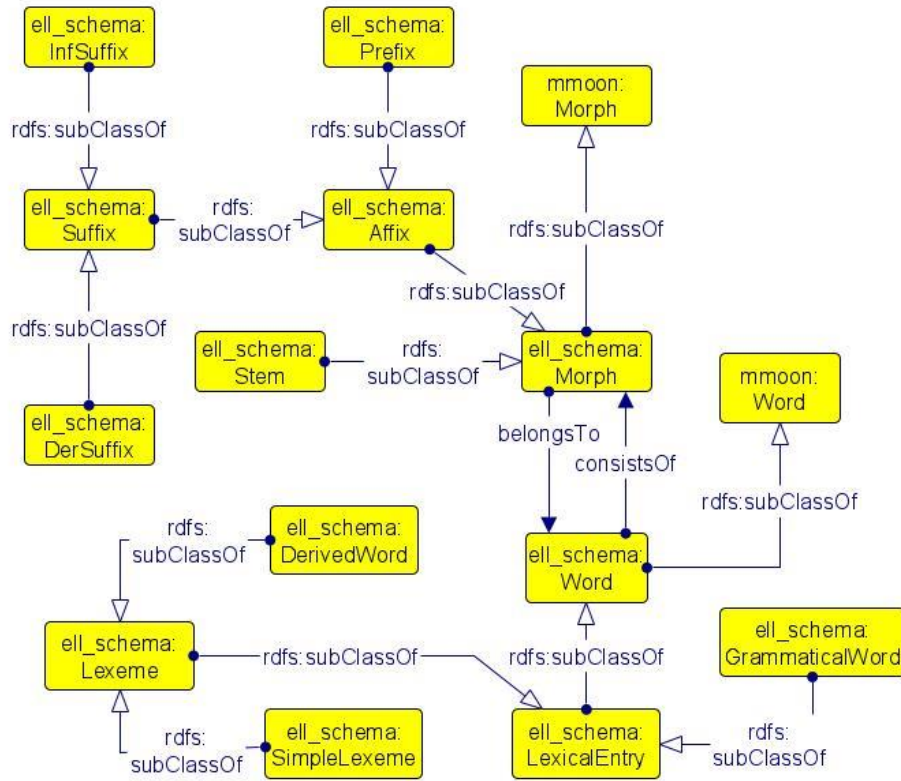


Figure 1. ell_schema morphological entities embedded into the MMoOn model

Confixes (Anastasiadi-Symeonidi, 1986; Giannouloupoulou, 1999), *Bound stems* (Ralli, 2005, 2007, 2012) or bound morphs of *neo-classical compounds* (Booij, 2012), as they are named, are a special group of morphs found as constituents in dual-structured forms of scientific or other vocabularies, usually coming from AG or Latin (e.g. *δολ-ο-πλόκος* ‘schemer’, *γloss-ο-λογία* ‘linguistics’, *meta-mondernismós* ‘post-modernism’ etc.). However, because these are rather placed between derivation and composition areas and because of their functional and semantic peculiarities, they are not analyzed or represented here as they will be considered at a later stage of analysis when decisions on data processing are to be made.

*Words*¹ can be either composed by a series of morphs (multi-morphemic) or consist of just a single morph (mono-morphemic) with no further morphological analysis. Mono-morphemic words can be *Grammatical* (e.g. conjunctions *όταν* (*όtan*) ‘when’, *και* (*κε*) ‘and’) or *Lexical* (usually loan words from foreign languages (e.g. *taxi* > *ταξι* (*taksí*)). Multi-morphemic words are always finalized by an

¹ Compounds are also regarded as word types but they are not part of this research.

inflectional suffix, even an unrealized one (\emptyset) (e.g. μητέρα- (mitéra-) > μητέρα (mitéra) ‘mother’) and can be *Simple Lexemes* (e.g. χορ-ός (xor-ós) ‘dance’) or *Derived Words* (e.g. χορ-ός (xor-ós) > χορ-εύ-ω (xor-év-o) ‘to dance’). The former uses a base and the latter an affixed base, which in both cases are finalized by an inflectional suffix.

Based on the previous conceptual analysis, in Figure 1, we identify the related classes in the MMoOn ontology (Klimek et al., 2020) and develop the specific *ell_schema*² embedded into it. We further add two classes: *ell_schema:DerSuffix* and *ell_schema:InfSuffix* as subclasses of *ell_schema:Suffix*. For the moment, we leave the *Stem* concept as it is, considering its subdivision in due time. We have chosen MMoOn, as already done before (Vasilogamvrakis et al., 2022; Vasilogamvrakis & Sfakakis, 2022), because it has been a comprehensive domain ontology for the representation of morphological language data (Klimek et al., 2019) and because it has been used as a template for the development of the Ontolex Morphology Module³.

2.2. Allomorphy

Allomorphy is the morphological phenomenon according to which a morpheme that is realized by a morph has more than one form with the same meaning. This morph variant⁴ is found in different morphological environments, that is why allomorphs stand in complementary distribution within words. Allomorphy can be basically of two types: a) *morpho-phonological*, when the change depends on some still-existent morpho-phonological rule⁵ (e.g. κλέβ- (klev-) ~ κλεφ- (klef-) ~ κλεψ- (kleps-) of the simple lexeme κλέβ-ω (klev-o) ‘to steal’) and b) *morphological* or *grammatical*, when the occurring allomorph is grammatically dependent and unpredictable (e.g. σώμα- (sóma-) ~ σωματ- (somat-) of the noun σώμα (sóma) ‘body’ or the AG form κλοπ- (klop-), an additional allomorph to κλεβ- ~ κλεφ- ~ κλεψ-) and it engulfs either bases or affixes alone or their combinations as affixed bases. An excessive type of allomorphy can also occur in forms, which substitute absent *lexical* realizations in inflection (e.g. είδ-α (íd-a) ‘I saw, which is the aorist word form of βλέπ-ω (vlép-o) ‘I see’). These forms are usually considered as instances of *suppletion* and, therefore, not true allomorphs as they do not show any phonological or semantic similarity (Ralli, 2005).

The representation of allomorphy in MG derivation is central because it triggers the creation of new derivatives (Karasimos, 2011) and offers connectivity between them. This is evident, in Figure 2, in the morpheme-based analysis⁶ of αγαπ-ώ (agap-ó) ‘to love’ and its derivative αγαπη-τός (agapi-t-ós) ‘beloved’, where their bases αγαπ- (agap-) and αγαπη- (agapi-) are allomorphs to each other.

² The *ell_schema* current version can be reached at: https://github.com/nvasilogamvrakis/nmoon_project/blob/main/ell_schema/ell_schema_03.owl.

³ <https://github.com/ontolex/morph/>.

⁴ Allomorphs are related to each other with appropriate morpholexical rules, which normally depict the morphological environment in which an allomorph occurs (Karasimos, 2011; Ralli, 2005).

⁵ For Ralli (2005), true allomorphs are synchronically unjustified and unpredictable forms and not those derived by phonological rules.

⁶ The MG morpheme-based analysis is elaborated in Vasilogamvrakis & Sfakakis (2022).

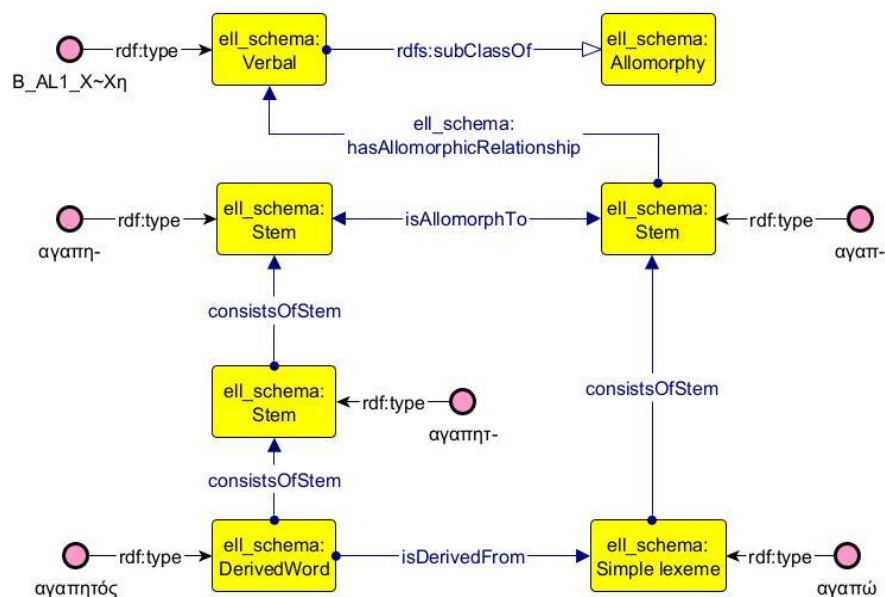


Figure 2. Interconnection between words through the allomorphs $\alpha\gamma\alpha\pi\sim\alpha\gamma\alpha\pi\eta$ -, belonging to paradigm $B \text{ } ALI \text{ } X\sim X\eta$

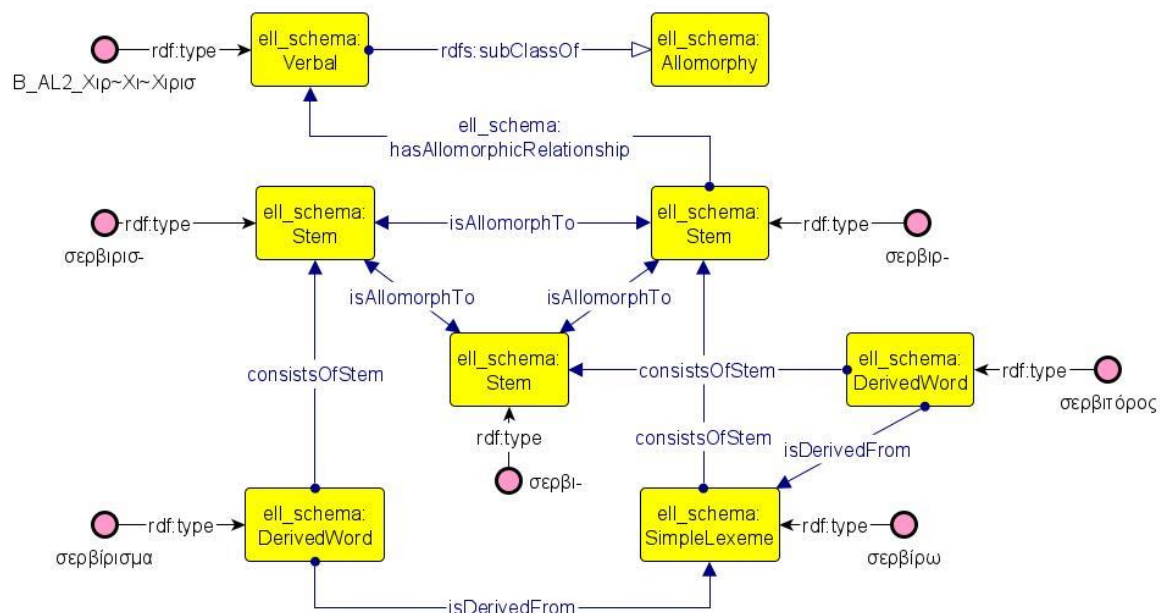


Figure 3. Allomorph instances $\sigma\epsilon\rho\beta i\rho\sim\sigma\epsilon\rho\beta i\sim\sigma\epsilon\rho\beta i\rho i\sigma$ - adapted to MG, belonging to paradigm $B_AL2_X i\rho\sim X i\sim X i\rho i\sigma$

Allomorphy can also occur in cases of loans from foreign languages. For example, in Figure 3, we show that the base *σερβιρ-* (*servir-*) of *σερβίρ-ω* (*servir-o*) ‘to serve’ (*servir* from French) is allomorph

to *σερβι-* (*servi-*) of *σερβι-τόρος* (*servi-tóros*) ‘waiter’ (*servi-tore* from Italian) and to *σερβιρις-* (*serviris-*) of *σερβίρις-μα* (*servíris-ma*) ‘serving’ (Karasimos, 2011; Ralli, 2005).

Furthermore, since allomorphs stand in complementary distribution, forms like *αγαπώ* (*αγαρό*) / *αγαπάω* (*αγαράω*) (Present, 1st Person, Singular) of Figure 4, emerged by *Reanalysis* of the AG contracted forms, are rather considered free variants (Ralli, 2005) and not true allomorphs. In the same figure, we also observe that the stem variant *αγαπα-* (*αγαπα-*) is specifically combined with the variant inflectional suffix *-γα* in *αγάπα-γα* (*αγάπα-γα*) whereas *αγαπ-* (*αγαπ-*) with the variant *-ούσα* (*-usa*) in *αγαπ-ούσα* (*αγαπ-úsa*) in Imperfect. We, therefore, create a new *ell_schema:hasFreeVariant* object property (OP) to connect the two morph entities, which we extend to also connect the two word lemma forms (*ell_schema:Morph* or *ell_schema:Word* as domain and range of the OP *ell_schema:hasFreeVariant*).

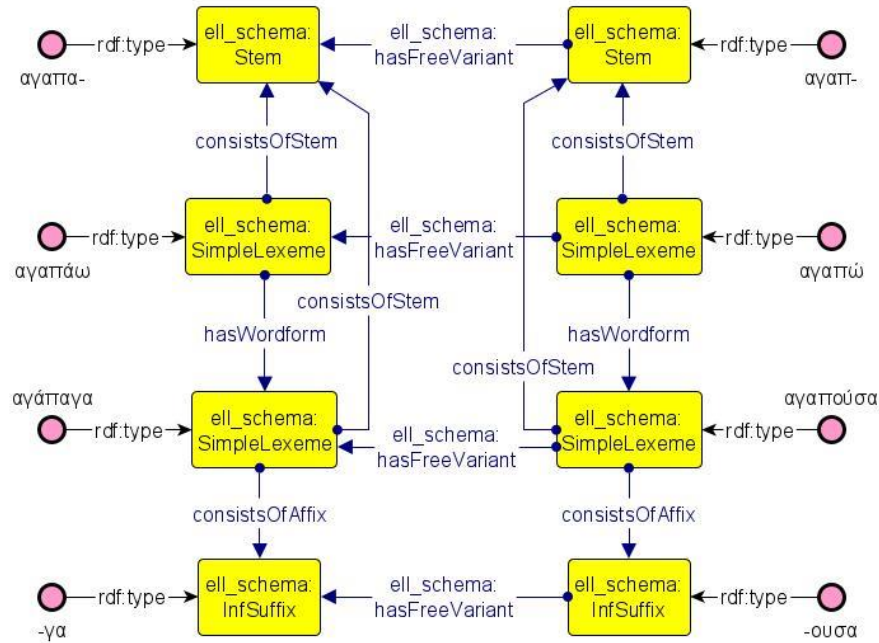


Figure 4. Interconnection between free variants via the *ell_schema:hasFreeVariant* OP

Allomorphy framework

The insertion of rules in the ontology does not contradict the assumption of some linguists that the Mental or Permanent Lexicon may include, next to morphological lemmas and non-transparable words, the dynamic area of word construction, i.e. the grammar or morphology (Kiparsky, 1982; Lieber, 1980; Selkirk, 1982). As presented in Vasilogamvrakis et al., 2022, the kind of morphological rules inserted in the ontology are rather descriptive, i.e. a top-down element that clusters similar lexical data. However,

these rules, as it will be shown next, can be leveraged for modulating an appropriate pipeline workflow for generating new forms.

In a computational-based approach, allomorphy is categorized into nominal, verbal and prefixal according to the affected lexico-grammatical category (Karasimos, 2011). Each of these categories encapsulates a series of allomorphy paradigms⁷, which are destined to operate as Regex patterns to bootstrap a morphological analyzer. These patterns are combined with appropriate computational rules placed within a specific morphological environment so as to predict the allomorphic change of a word.

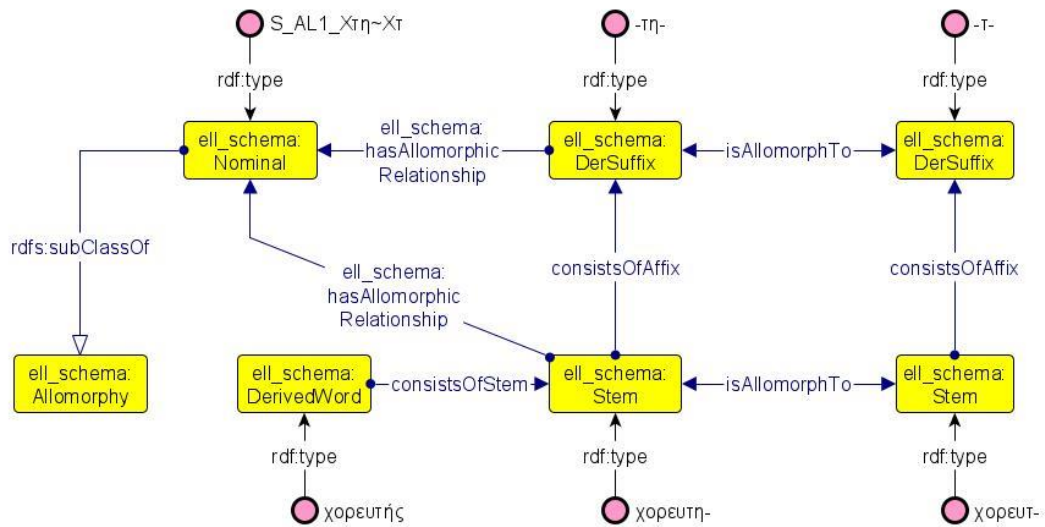


Figure 5. Interconnection between allomorph derivational suffixes $-τη-$ ~ $-τ-$ and between their attached stems, belonging to paradigm $S_AL1_Xτη~Xτ$

In a similar manner, we want to create allomorphy paradigms as morpholexical rules (Karasimos, 2011; Ralli, 2005) and relate them to specific derivational environments according to suffix-driven selectional restrictions (Melissaropoulou & Ralli, 2009). To host allomorphy paradigms, we introduce a new *ell_schema:Allomorphy* class in the core MMoOn schema, which, for the moment, we subdivide into *ell_schema:Verbal* and *ell_schema:Nominal* subclasses (Figures 2, 3 and 5). Although all variant forms are allomorphs to each other, which is represented in the ontology, the allomorphy paradigm is linked only to the basic morph lemma⁸ ($σερβιρ-$) and not to its alternative forms ($σερβι-$ ~ $σερβιρiσ$) (Booij, 2012; Karasimos, 2011). For doing so, we add an *ell_schema:allomorphic_relationship* OP, with *ell_schema:Morph* as domain and *ell_schema:Allomorphy* as range (Figures 2, 3 and 5). We represent this specific allomorphy paradigm starting with the base (B) paradigm number and an X character for the common lexical part, followed by each variant with the symbol ~ in between ($B_AL1_X~Xτη$ or $B_AL2_Xτ~Xτ~Xτiσ$)⁹. We choose this inclusive pattern, adhering to the common morphological representation of allomorphs (Ralli 2005) but alternative ways may be also considered in the course of

⁷ We chose the term ‘paradigm’ instead of ‘class’ so that it is distinguished from the ontological term ‘class’.

⁸ This forms the initial lexical entry of the derivational family ($σερβιρ- > σερβιρ-ω$).

⁹ The given paradigm numbers are arbitrary.

the research. Similar is the modelling for allomorph suffixed bases (S), in Figure 5 (e.g. $S_ALI_X\tau\eta\sim X\tau$ for the derivational suffix $-\tau\eta\sim -\tau-$, preceded by the common lexical part X).

In order for an allomorphy paradigm to operate as a data classification module, an additional built-on programming pipeline should be implemented, based on pattern matching queries, which are sent to a core Lexicon component. According to the modelling of Figure 6, a verbal allomorphy paradigm (B_ALI) finds matches by its instance ($X\sim X\eta$) inside the Lexicon of lemmas and clusters them according to the common lexical part X, making a unique set of related bases (e.g. $\alpha\gamma\alpha\pi\sim\alpha\gamma\alpha\pi\eta$: set1). Then, every term of the set replaces the placeholder AL of the derivational word-pairs based on suffixation rules (Melissaropoulou & Ralli, 2009; Vasilogamvrakis et al., 2022), which are simultaneously validated against the existent lemmas of the Lexicon.

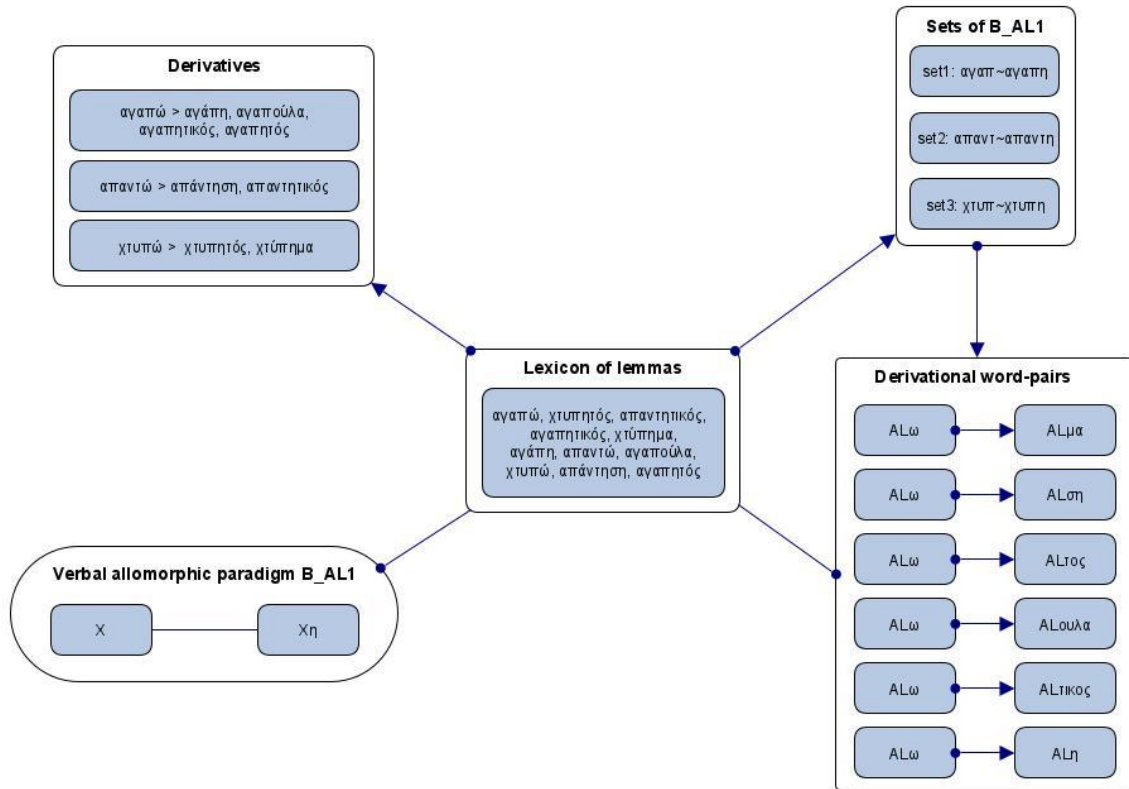


Figure 6. Provisional pipeline model for creating derivatives based on allomorphy paradigms

As a filtering rule, the placeholder of the source word always uses the common allomorph ($\alpha\gamma\alpha\pi$), whereas the placeholder of the target word may use all available allomorphs of the paradigm ($\alpha\gamma\alpha\pi\sim\alpha\gamma\alpha\pi\eta$). For example, for the $\alpha\gamma\alpha\pi\sim\alpha\gamma\alpha\pi\eta$ set1 of B_ALI paradigm, the derived words, $\alpha\gamma\alpha\pi\sim\eta$ ($\alpha\gamma\alpha\pi\sim i$) ‘love’, $\alpha\gamma\alpha\pi\sim\omicron\upsilon\lambda\alpha$ ($\alpha\gamma\alpha\pi\sim\acute{\iota}\lambda\alpha$) ‘sweetheart’, $\alpha\gamma\alpha\pi\eta\sim\tau\iota\kappa\omicron\varsigma$ ($\alpha\gamma\alpha\pi\eta\sim\tau\iota\kappa\omicron\varsigma$) ‘lover’ and $\alpha\gamma\alpha\pi\eta\sim\tau\omicron\varsigma$ ($\alpha\gamma\alpha\pi\eta\sim\tau\omicron\varsigma$) ‘beloved’ will be generated from the simple lexeme $\alpha\gamma\alpha\pi\sim\acute{\omega}$ ($\alpha\gamma\alpha\pi\sim\acute{\omicron}$) ‘to love’, within a

specific derivational environment of word-pairs, and after validated against the Lexicon of lemmas. Apparently, a derivational word-pair can be combined with more than one allomorphy paradigm, which makes the model particularly economical.

This modeling reproduces the theoretical assumption that the Dynamic lexicon (morphology) applies rules to the Permanent lexicon to generate or re-analyze derivational structures, placing the ontology at the centre of this operation. However, it would be wise, here, to stress that until we test the model's effectiveness upon real lexical data, it is likely that it will be modified to optimize performance and consistency and is, therefore, considered provisional.

2.3. Representation

With regard to the representation of form, the MMoOn provides the class *Representation* as domain of the data properties (DP): *morphological*, *phonetic* and *orthographic* representation. The usability of this class is evident mostly in cases of allomorphy or homonymy.

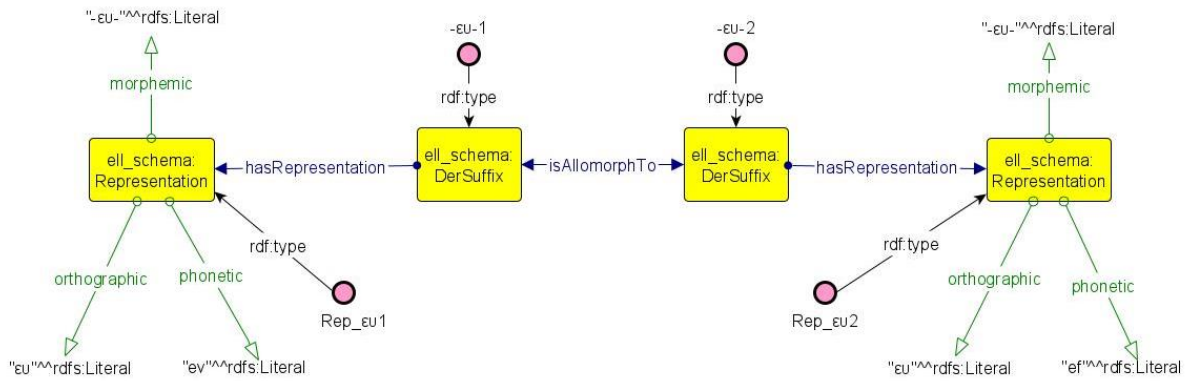


Figure 7. Representation of allomorphs (with allophones) with different *Representation* instances

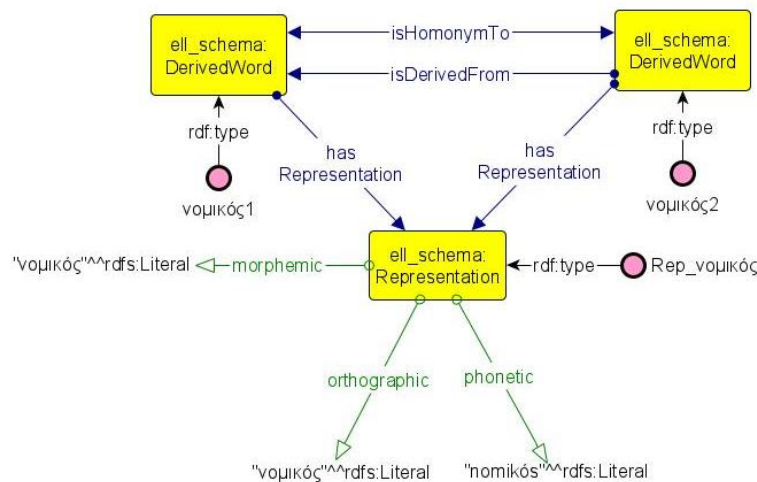


Figure 8. Representation of homonyms with the same *Representation* instance

Except for those cases explored previously, *allomorphy* can also occur when there are variant phonetic realizations of a phoneme (*allophones*) within a morph. Accordingly, in Figure 7, each of the derivational suffixes *-ev-1* and *-ev-2* retains different *Representation* instances, *Rep_ev1* and *Rep_ev2*, because of their different phonetic transcription (*ev* and *ef* respectively). This is better understood when both are seen as constituents of their belonging words within a very common MG derivational chain e.g. *χορ-εβ-ω* (*xor-év-o*) ‘to dance’ > *χορ-εβ-τή-ς* (*xor-ef-tí-s*) ‘dancer’. Their phonologically-based allomorphic interconnection is captured by the symmetrical OPs *is allomorph to*, having, at the same time, a common morphemic and orthographic representation *-ev-* value.

On the other hand, *homonymy* occurs when there are similarly spelled (homographs) and pronounced (homophones) morphs or words while having different lexical or grammatical meanings. For example, as shown in Figure 8, the two different words *νομικός1* (*nomikós1*) ‘juristic’ and *νομικός2* (*nomikós2*) ‘lawyer’ are also marked as Adjective and Noun respectively. Each word is connected to the other with a symmetrical *is homonym to* OP, while both of them have a common *Representation* instance *Rep_νομικός* and identical morphemic, orthographic and phonetic representation values. Furthermore, they have a derivational relation, as the second word *νομικός2* is derived from the first *νομικός1* by *Conversion*.

3. Conclusion

In the present article, we ontologically analyzed the types of MG morphological entities participating in derivational structures, justifying their presence in the MMoOn *ell_schema* ontology. In particular, we focused on the stem and affix concepts and their subclasses because we showed that these entities are affected by the phenomenon of allomorphy. We additionally provided evidence that the latter impacts significantly on derivational processes and, for that reason, we modeled and placed it within certain derivational environments so that it is functional and can generate new lexical forms. This framework is actually consistent with the postulation that the Lexicon can incorporate both morphological rules and lexical data and we assigned the ontology that role. Finally, we showed how morphological semantics or certain allomorphy types can affect the representational aspects of morphs or words.

4. Acknowledgements

This research was supported by the project: “Activities of the Laboratory on Digital Libraries and Electronic Publishing of the Department of Archives, Library Science and Museology”.

References

- Anastasiadi-Symeonidi, A. (1986). *Neology in Standard Greek (in Greek)*. Aristotle University of Thessaloniki.
- Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.
- Giannouloupoulou, G. (1999). *Morphosemantic comparison between affixes and confixes in Modern Greek and*

- Italian (in Greek)* [Ph.D. Diss., Aristotle University of Thessaloniki (AUTH). School of Italian Language and Literature]. <http://hdl.handle.net/10442/hedi/32432>
- Karasimos, A. (2011). *Computational processing of allomorphy in Modern Greek derivation (in Greek)* [Ph.D. Diss., University of Patras]. <http://hdl.handle.net/10442/hedi/32458>
- Kiparsky, P. (1982). *Lexical Morphology and Phonology*. <https://web.stanford.edu/~kiparsky/Papers/Lexical%20Morphology%20and%20Phonology.pdf>
- Klimek, B., Ackermann, M., Brümmer, M., & Hellmann, S. (2020). MMoOn Core—The Multilingual Morpheme Ontology. *Semantic Web*, 4, 1–30. <http://www.semantic-web-journal.net/system/files/swj2549.pdf>
- Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K., & Chiarcos, C. (2019). Challenges for the Representation of Morphology in Ontology Lexicons. *Proceedings of the eLex 2019 Conference. 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o.*, 570–591. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_33.pdf
- Lieber, R. (1980). *On the Organization of the Lexicon* [Ph.D. Diss.]. MIT.
- Melissaropoulou, D., & Ralli, A. (2009). Combinatoriality of Derivational Suffixes in Modern Greek: A First Approach (in Greek). *Studies in Greek Linguistics*, 29, 97–107. http://ins.web.auth.gr/images/MEG_PLIRI/MEG_29_97_107.pdf
- Ralli, A. (2005). *Morfologia (in Greek)*. Patakis.
- Ralli, A. (2007). *The composition of words (in Greek)*. Patakis.
- Ralli, A. (2012). Deverbal Compounds with Bound Stems. In *Compounding in Modern Greek* (pp. 201–220). <https://www.angelaralli.gr/sites/default/files/Neoclassical%20compounds%20with%20bound%20stems.pdf>
- Selkirk, E. (1982). *The syntax of Words*. MIT Press.
- Spencer, A. (2017). Morphology. In *The Handbook of Linguistics* (pp. 211–233). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119072256.ch11>
- Vasilogamvrakis, N., Koliopoulou, M., Sfakakis, M., & Giannouloupoulou, G. (2022). Testing the Word-Based Model in the Ontological Analysis of Modern Greek Derivational Morphology. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørnvåg, B. Catania, G. Vargas-Solar, & E. Zumpano (Eds.), *New Trends in Database and Information Systems* (pp. 572–584). Springer International Publishing. https://doi.org/10.1007/978-3-031-15743-1_52
- Vasilogamvrakis, N., & Sfakakis, M. (2022). A Morpheme-Based Paradigm for the Ontological Analysis of Modern Greek Derivational Morphology. In E. Garoufallou, M.-A. Ovalle-Perandones, & A. Vlachidis (Eds.), *Metadata and Semantic Research* (pp. 389–400). Springer International Publishing. https://doi.org/10.1007/978-3-030-98876-0_34

Of Families and Occurrences. Derivation and Word Usage in Latin

Marco Passarotti, Eleonora Litta
Università Cattolica del Sacro Cuore / Milano
marco.passarotti@unicatt.it
eleonoramaria.litta@unicatt.it

Abstract

In this paper we present the results of an investigation on the relation between derivational morphology, represented in terms of derivational families from a word formation lexicon for Latin, and the number of textual occurrences of their members in a large set of Latin corpora made interoperable in a Linked Data Knowledge Base.

1 Introduction

The current availability of several linguistic resources for the Latin language has raised the issue of their dispersion, which affects the full exploitation of the (meta)data they provide. This means that, even when they are published in common repositories or infrastructures (like, for instance, CLARIN),¹ resources still stay confined in separate silos that do not communicate with each other.

This situation impacts negatively on the use of data, because it prevents scholars from running federated queries across different resources, although this is a typical need when linguistic (meta)data are concerned. Particularly, this is the case when Classical and ancient languages are concerned, as scholars for centuries have been joining information from texts in different collections, as well as from lexical resources like dictionaries and glossaries.

To address the issue of dispersion and lack of interaction among the available linguistic data for Latin, the *LiLa: Linking Latin* ERC project (2018-2023)² has built a Knowledge Base of interoperable lexical and textual resources for Latin based on the principles of the Linked Data paradigm (Berners-Lee et al., 2001), by representing and publishing the (meta)data from these resources using common vocabularies (provided by ontologies) for knowledge description.

The resources currently made interoperable by the LiLa Knowledge Base include several corpora, which cover a wide chronological and typological span of Latin texts, and a number of lexical resources, like a bilingual dictionary, an etymological lexicon and a polarity lexicon.³ Among the lexical resources published in LiLa is *Word Formation Latin*, a derivational lexicon for Latin where derived words are assigned a word formation rule and a link to the lexical item (or items, in the case of compounds) from which they are derived. The interoperability between the derivational information provided by *Word Formation Latin* and the (meta)data of all the other resources published in LiLa makes it possible to collect lexical information and textual evidence to empirically test hypotheses (or assumptions) about the relation between word formation processes in the lexicon and the use of derived words in texts.

In this paper, we want to address and evaluate empirically the hypothesis that, given a derivational family (i.e., a set of words sharing the same ancestor, henceforth the ‘root’), the member with the highest number of occurrences in texts is derivationally simple (i.e., not featuring any affix), the root of the family being the most typical case. After providing the quantitative results taken from the corpora linked to the LiLa Knowledge Base, we focus on some cases that exceed the prototype, i.e., those where the root of a family is not also the most frequently attested member in texts. Finally, we compare the distribution of

¹<https://www.clarin.eu>

²<https://lila-erc.eu>

³For the full list of the resources currently linked to LiLa see <https://lila-erc.eu/data-page/>.

the root and the most frequent members of the derivational families common to two different corpora, showing that this helps to highlight some lexical properties of the texts included in the corpora concerned.

2 LiLa and Word Formation Latin

To make distributed linguistic resources interact following the Linked Data principles, the LiLa Knowledge Base adopts the data model of the Resource Description Framework (Lassila et al., 1998) (RDF). According to RDF, information is coded in terms of triples, that connect a subject – a labelled node – to an object – another labelled node or a literal – by means of a property – a labelled edge. The vocabulary used for properties and the criteria for their application are provided by a set of ontologies developed and widely adopted by the Linguistic Linked Open Data community and adopted in LiLa to describe linguistic annotation (OLiA, cf. Chiarcos and Sukhareva 2015), corpus annotation (NIF, cf. Hellmann et al. 2013; CoNLL-RDF, cf. Chiarcos and Fäth 2017) and lexical resources (Lemon, cf. Buitelaar et al. 2011; OntoLex, cf. McCrae et al. 2017).

LiLa connects resources building on the intuition that words play a central role in both lexical and textual resources, and that through words these can be interlinked and interact. Following this intuition, the core of the Knowledge Base is its Lemma Bank, an ever growing collection of around 215,000 canonical citation forms of Latin words. Through the Lemma Bank, the entries of the various lexical resources published in LiLa and the word occurrences in the corpora included therein are linked to their appropriate citation form in the Lemma Bank, thus achieving interoperability (Passarotti et al., 2020).

Word Formation Latin (WFL) is a derivational lexicon for Classical Latin that includes 41,977 entries connected by input-output relations, grouping all members of the same derivational family in a hierarchical structure taking root from the ancestor – the lexeme from which all the members of the family ultimately derive – and branching out to all derivatives by means of the successive application of individual word formation rules (Litta et al., 2020).

In building the LiLa Lemma Bank, derivational data were extracted from WFL to describe the word formation construction of the WFL entries in a flat way, i.e. without inferences on their derivational history, but only with details about the presence of affixes and affiliation to a derivational family. In the LiLa ontology, this information was encoded in two classes (sub-classes of the class *Morpheme*), namely *Affix* – divided into *Prefix* and *Suffix* – and *Base*. Bases are abstract connectors between lemmas that belong to the same family. These connectors are labelled with the lemma of the root word of the family concerned. In the Lemma Bank, a lemma is linked to the base to which it is related by means of the property *lila:hasBase*, and to the affixes it contains by means of the property *lila:hasPrefix* or *lila:hasSuffix* (Litta et al., 2019). Hence for example lemma *aduersaria* is connected through the property *hasPrefix* to the prefix *ad-*, through the property *hasSuffix* to suffix *-ari*, and through the property *hasBase* to the connector node labelled *uerto*. WFL was subsequently linked as a lexical resource to the LiLa Knowledge Base, in order to preserve precious data about more detailed, hierarchical information on the order of application of different word formation processes (Pellegrini et al., 2021).

3 Data and Discussion

Among the 4,769 derivational families provided by WFL, we select those that feature at least 10 members (1,086 families), which means that in the LiLa Knowledge Base the individual representing the base that connects all the members of a family has an in-degree via the property *lila:hasBase* ≥ 10 . Also, among such families, we select those where the total number of occurrences of the members in all the textual resources currently linked to the LiLa Knowledge Base⁴ is ≥ 100 , leading to 878 families.⁵

3.1 Derivation, Frequency and Lexicalisation

In 582 out of the 878 families under investigation, the root member is also the most frequent one in the corpora linked to the LiLa Knowledge Base (e.g., *pono* ‘to put’), while this is not the case for the

⁴The textual resources in LiLa contain more than 3 million occurrences in Latin texts of different period (from Classical era to Medieval times) and genre (including literary, documentary, historical and philosophical texts).

⁵The script providing the SPARQL queries that we used to collect the data described in this Section is available at <https://github.com/CIRCSE/DevAttFreq>.

remaining 296 families (e.g., *accipio* ‘to accept’ is the most frequently encountered word from the family rooted by *cipio* ‘to take’). In 89 out of these 296 families, the most frequent member is derivationally simple, i.e., it does not include any affix (either prefix, or suffix), as it formed by a conversion process, like in the case of *cursus* ‘course’, converted from the base of the supine of the root verb *curro* ‘to run’). Given these figures, we can confirm that in most families (671 out of 878) the member with the highest number of textual occurrences is derivationally simple and, very often (582 out of 671), it is also the root word. Conversely, in 207 (296 - 89) out of the 878 families concerned, the most frequently attested member is a derivationally complex word, formed with one, or more affixes.

Table 1 shows the 10 most attested affixes in the most frequent derived words of a family. For instance, the prefix *con-* appears in the most frequent word of 25 families, like in the case of the verb *cognosco* ‘to know’, which is the most frequent word (2,543 occurrences in the LiLa corpora) of the family whose root is the derivationally simple verb *nosco* ‘to know’ (1,497 occurrences).

Moreover, Table 1 shows the ranking of each of the 10 affixes in the LiLa Lemma Bank, resulting from the number of lemmas therein formed with that suffix. For instance, the prefix *con-* is present in 2,204 entries of the Lemma Bank, which makes it the third most attested affix in the Bank. The difference between the lexical ranking of an affix (i.e., the number of lemmas in the LiLa Bank formed with that affix) and its textual ranking (i.e., the number of occurrences of the lemmas formed with that affix in the LiLa corpora) is remarkably positive as for the suffixes *-i* (from 11 to 2), *-id* (from 36 to 3) and *-in* (from 19 to 5), and negative as for *-(t)io* (from 1 to 6). As for the latter, this means that, although in the Latin lexicon the number of available derived words featuring the suffix *-(t)io* (3,418) is higher than for any other affixed word, the number of families whose most frequent member features the suffix *-(t)io* is quite low (8). The opposite holds, for instance, for suffix *-i*: although the number of words in the Lemma Bank formed with this suffix is much lower (1,323) than those featuring *-(t)io*, 22 out of them are the most frequently occurring member in as many families, against only 8 formed with *-(t)io*.

Table 2 shows the 5 words with the highest frequency in the *-i* and *-(t)io* sets. The words of the *-i* set have more occurrences than those of the *-(t)io* set, which is headed by one much frequent word (*ratio*), while the others show a lower number of occurrences. We notice that some of the non-root members of a derivational family that are the most frequently attested in corpora are cases of lexicalisation.⁶ According to Lehmann (2002, pp. 1-2), “grammar is concerned with those signs which are formed regularly and which are handled analytically, while the lexicon is concerned with those signs which are formed irregularly and which are handled holistically. [...] The analytic approach consists in considering each part of the object and the contribution that it makes to the assemblage by its nature and function, and thus to arrive at a mental representation of the whole by applying rules of composition to its parts. The holistic approach is to directly grasp the whole without consideration of the parts”. For instance, the first sense of the noun *substantia* provided by the Oxford Latin Dictionary (Glare, 2012) is “the quality of being real”. Other senses are “underlying, or essential nature”, “the material of which a thing is made”, “possessions” and “the basic unit of measurement (in any calculation)”. Clearly, this is a case of lexicalisation, as the meaning of the word does not result from the simple composition of the semantic contribution from each of its parts, but underwent a process of shift from the original spatial semantic field to a metaphorical meaning. As for the derivation process of *substantia*, the Oxford Latin Dictionary reports “substo+ia”. The verb *substo* means “to hold one’s ground”, still pertaining to the spatial semantic field that the lexicalisation process has made *substantia* loose.

Some interesting insights come from comparing the distribution of the parts of speech (PoS) of the root words with those of the most frequent words of the families.⁷ If we focus on adjectives, common nouns and verbs only (as the PoS with most words here concerned), Table 3 shows that adjectives and verbs are the root of a family more often than playing the role of the most frequent word. The opposite holds when common nouns are concerned: while the root word is a common noun in 364 families, the most frequently attested word of a family is a common noun in 415 cases. The great majority of these are shifts from adjectives or verbs as the root word to common nouns as the most frequent word in texts,

⁶According to Lehmann (2002, pp. 1-2), lexicalisation is a lexical semantic process “concerned with those signs which [...] are handled holistically”, which means “to directly grasp the whole without consideration of the parts”.

⁷The LiLa Lemma Bank adopts the Universal PoS tagset (Petrov et al., 2012).

Affix	Number of families	Lemma Bank ranking	Example
con-	25	3	cognosco
-i	22	11	substantia
-id	11	36	frigidus
-or	11	4	calor
de-	11	9	detrimentum
ad-	10	10	accipio
-in	9	19	dominus
ex-	9	5	exsulto
in(entering)-	9	8	instruo
-(t)io	8	1	oratio

Table 1: The 10 most attested affixes in the most frequent derived words of a family.

Ranking	-i set	-(t)io set
1	consilium (2,147)	ratio (3,513)
2	gratia (2,051)	oratio (1,250)
3	substantia (1,697)	opinio (504)
4	sententia (1,606)	fornicatio (179)
5	memoria (1,039)	satisfactio (175)

Table 2: The 5 most frequent words in the -i and -(t)io sets.

often due to conversion, like in the case of the family whose root word is the verb *lugeo* ‘to mourn’ (frequency: 174) and whose most frequent one is the common noun *luctus* ‘sorrow’ (274), which is derived by conversion from the perfect participle of *lugeo*. Moreover, if we compare the distribution of adjectives, common nouns and verbs playing either the role of root or most frequent word in a family with their number in a large collection of Latin words like the LiLa Lemma Bank, we notice the importance of verbs in derivational families. Indeed, if we consider that the total number of verbs in the Lemma Bank (16,618) is much lower than nouns (80,892) and adjectives (65,006), the figures in Table 3 show that verbs are either the root or the most frequent word in a family much more often ($351+291 = 642$) than nouns ($364+415 = 779$) and adjectives ($133+114 = 247$).

PoS	Root	Most frequent
adjective	133	114
common noun	364	415
verb	351	291

Table 3: PoS distribution of root words and most frequent words in derivational families.

3.2 Comparing Corpora

As mentioned, the Latin corpora currently interlinked through LiLa include very diverse texts, which belong to different periods and genres. If such a diversity makes the corpora of LiLa quite a representative set of data to draw conclusions about the Latin language, merging the data from all the corpora prevents from identifying the characteristics of the lexicon of one specific corpus.

To this aim, we compare two of the largest corpora in LiLa, namely the LASLA collection of Classical Latin texts (around 1.7 million words) (Fantoli et al., 2022) and the *Index Thomisticus* Treebank (ITTb), which includes the full text of *Summa contra Gentiles*, a Medieval Latin philosophical treatise by Thomas

Aquinas, for a total of approximately 350,000 words (Passarotti, 2019).

Table 4 shows the quantitative results of the LASLA-ITTB comparison. Out of the 878 families selected, 214 are common to the LASLA and the ITTB data sets, i.e. the total number of the occurrences of their members in the two corpora is ≥ 100 . 116 out of these 214 families have the same most frequent word in the two corpora, while for 98 families it is different. Focusing on the latter, we notice (1) that there are 34 families where the most frequent word is different in the data from the two corpora and, for both of them, it is not the root, and (b) that the number of cases where the most frequent word of a family is also the root in the LASLA corpus, while it is not in the ITTB, is much higher than the opposite (55 vs 9). This result is worth noticing. Indeed, in several such cases from the ITTB, the most frequent word of a family is a derivationally complex word, featuring one (e.g., *transeo* ‘to cross over’) or two affixes (e.g., *differentia* ‘difference’). The fact that the texts of Thomas Aquinas tend to make use of derived words more than those from the LASLA corpus might be due to the specific characteristics of the lexicon found in Aquinas’ works. As a matter of fact, often in the ITTB the most frequent word of a derivational family is a technical word of the philosophical terminology of Thomas Aquinas (and, overall, of Medieval Scholasticism), like *substantia*, which belongs to the family rooted by *sto* ‘to stay’ (the most frequent family member in the LASLA data). Other cases are *passio* ‘passion’ (from the family rooted by *patior* ‘to bear’), *sensibilis* ‘sensible’ (*sentio* ‘to feel’) and *virtus* ‘strength’ (*vir* ‘man’). Instead, looking at the 9 cases, where the LASLA most frequent word in a family does not correspond to its root while the opposite holds for the ITTB, we find terms related to the political area, like *rex* ‘king’ (from the family of *rego* ‘to guide’), *gubernator* ‘steersman’ (*guberno* ‘to steer’) and *libertas* ‘liberty’ (*liber* ‘free’).

	LI-r	LI-no-r	L-r I-no-r	L-no-r I-r	Total
Same most frequent word	89	27	NA	NA	116
Different most frequent word	NA	34	55	9	98
Common families					214

Table 4: Comparing the LASLA and ITTB corpora. L=LASLA. I=ITTB. r=root.

Also in order to verify this hypothesis, we focus on the 15 families (among the 878 selected) with the highest number of members. Table 5 shows for each of them its root word and the name of the most frequently attested one in the LASLA and ITTB corpora, respectively. Finally, it is worth noticing that a quite frequent family in LASLA like that of *gero* does not reach the minimum number of occurrences (100) in the ITTB.⁸

4 Conclusion and Future Work

In this paper we presented the results of an investigation about the relation holding between derivational morphology, represented in terms of derivational families, and the number of textual occurrences of their members in a large set of Latin corpora.

In the near future, we plan to exploit the evidence that we collected in order to explore some trends. For instance, for those families where the most frequent word is not the family root but another derivationally simple word (related to the root proper by conversion), we shall investigate whether the evidence suggests that the root-derivative relation should be reversed. Indeed, the criteria followed by dictionaries to identify the order in derivations are not always consistent and, in most cases, do not take into account the frequency of use of the words in texts. As an example, for the verb *nuntio* ‘to announce’, the Oxford Latin Dictionary reports that it derives from the noun *nuntium* ‘announcement’ (“nuntium+o”), while in our textual data *nuntio* is far more frequent than *nuntium* (411 vs. 28 occurrences). However, it is clear that frequency cannot be the only criterion to identify derivational roots when conversion takes place. This is especially true for a language like Latin, that shows a wide diachronic span of more than two millennia: we must investigate the role possibly played by a chronological shift of prominence between two derivationally

⁸The case of the family of *fluo* is less surprising, because the high number of occurrences of the noun *flumen* in LASLA makes alone the family suitable for selection of the experiment described here.

Root	Most frequent in LASLA	Most frequent in ITTB
facio	facio	facio
fero	fero	differentia
capió	accipio	principium
ago	ago	actus
verto	versus	universalis
gero	gero	NA
pes	pes	impedio
lego	legio	intellectus
eo	eo	transeo
fluo	flumen	NA
pario	pario	comparo
sto	sto	substantia
mitto	mitto	praemitto
loquor	loquor	loquor
duco	duco	produco

Table 5: Most frequent word of the 15 largest families in the LASLA and ITTB corpora.

simple items candidate to be the root of a family, considering, for instance, the possibility that the original root had become obsolete in Late, or even already in Classical Latin. Such investigation would shed light also on the thin line holding between etymology and derivation.

We collected the results discussed in the paper thanks to the interoperability among the lexical and textual resources for Latin made possible by the LiLa Knowledge Base, showing how making the (meta)data provided by different linguistic resources interact is helpful and much needed. This is particularly relevant when an ancient language is concerned, because the absence of native speakers requires any investigation on the lexicon to be grounded on a steady confrontation with the evidence provided by texts, as the only still surviving witnesses of the real use of the words of a dead language. However, having the possibility to make the wealth of lexical and textual data from several available resources interact would prove helpful also for living languages. In such respect, it is necessary that, in the near future, the research community makes an effort towards making real (and effective) the interoperability between distributed digital resources for as many languages as possible, with the outlook of making all their data finally interact in multi-lingual fashion. To this aim, Latin might play the important role of connection among, at least, romance languages.

5 Credits

The “LiLa - Linking Latin” project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

References

- Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284(5):34–43.
- Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. *Ontology Lexicalisation: The lemon Perspective*. In *9th International Conference on Terminology and Artificial Intelligence (TIA11) – Proceedings of the Workshops*. Paris, France, pages 33–36. <https://pub.uni-bielefeld.de/record/2486962>.
- Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*. Springer, Cham, Switzerland, pages 74–88.

- Christian Chiarcos and Maria Sukhareva. 2015. *OLiA – Ontologies of Linguistic Annotation*. *Semantic Web* 6(4):379–386. <https://www.semantic-web-journal.net/system/files/swj5180.pdf>.
- Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the lasla corpus in the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of the Linked Data in Linguistics Workshop@ LREC2022*. pages 26–34.
- Peter Geoffrey William Glare. 2012. *Oxford Latin Dictionary*. Oxford Languages. Oxford University Press, Oxford, UK, 2 edition. <https://global.oup.com/academic/product/oxford-latin-dictionary-9780199580316?cc=uslang=en>.
- Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *Proc. 12th International Semantic Web Conference, 21-25 October 2013*. Sydney, Australia. Also see <http://persistence.uni-leipzig.org/nlp2rdf/>.
- Ora Lassila, Ralph R. Swick, World Wide, and Web Consortium. 1998. *Resource Description Framework (RDF) Model and Syntax Specification*. <https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
- Christian Lehmann. 2002. New reflections on grammaticalization and lexicalization. *New reflections on grammaticalization* pages 1–18.
- Eleonora Litta, Marco Passarotti, Marco Budassi, and Marco Pappalepore. 2020. Of nodes and cells. two perspectives on (and from) word formation latin. *Lingue antiche e moderne* 9:131–155.
- Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. *The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin*. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, pages 35–43. <https://www.aclweb.org/anthology/W19-8505>.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. *The OntoLex-Lemon Model: Development and Applications*. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Lexical Computing CZ s.r.o., Brno, Czech Republic, pages 587–597. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.
- Marco Passarotti. 2019. *The Project of the Index Thomisticus Treebank*, De Gruyter Saur, Berlin, Germany; Boston, MA, USA, pages 299–320. Number 10 in Age of Access? Grundfragen der Informationsgesellschaft. <https://doi.org/10.1515/9783110599572-017>.
- Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. *Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin*. *Studi e Saggi Linguistici* LVIII(1):177–212. <https://doi.org/10.4454/ssl.v58i1.277>.
- Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2021. *The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources*. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*. ATILF, Nancy, France, pages 101–109. <https://doi.org/10.5281/zenodo.5532501>.
- Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. *A Universal Part-of-Speech Tagset*. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 2089–2096. http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_paper.pdf.

Morphological Resources for the Study of Turkish Derived Nouns

Yağmur Öztürk

CRIT,

Université de
Franche-Comté

yagmur.ozturk@edu.univ-fcomte.fr izabella.thomas@univ-fcomte.fr

Izabella Thomas

CRIT,

Université de
Franche-Comté

Snejana Gadjeva

CREE,

Institut National
des Langues et

Civilisations Orientales
snejana.gadjeva@inalco.fr

Abstract

In terms of morphological resources, Turkish turns out to be an underresourced language. In particular in the field of Natural Language Processing (NLP), there are not enough resources that sufficiently (and systematically) describe Turkish derivational morphology, especially concerning semantic aspects of the derivational process. The research aims to describe and use existing resources and studies to develop an NLP tool for Turkish nominal derivation. The first part of our study presents the current morphological analysers revealing a gap in derivational morphology of nominals. We then discuss how derivational morphemes, specifically nominal morphemes, are rendered in linguistic studies and the problems it poses for a systematic study. Finally, we introduce *Semantiürk*, which is an ontology of semantic categories, and *DerivBaseTR*, which is a morpheme database with specific features, as the formalised resources we created for a systematic study of noun-to-noun morphemes.

1 Introduction

The study we propose came into existence following research in a morphosemantic project for the processing of Turkish derived nouns. The primary goal is to create a morphosemantic analyser that describes the internal structure of derived nouns as well as explicits the semantic role of each detected morpheme in the derived noun. A similar tool, *DériF* (Namer, 2002) exists in French with a “pseudo-definition” output as shown in (1).

- (1) appauvrissement/NOM à [[a [pauvre ADJ] VERBE] ment NOM],
(appauvrissement/NOM, appauvrir/VERBE, pauvre ADJ),
“(Action – résultat de l’action) de appauvrir”
en. (Action – result of the action) of impoverish

Although Turkish is an agglutinative language, with a high degree of regularity and productivity in its derivational processes, as stated in Section 2, there is currently no morphosemantic analyser for Turkish. Additionally, there are no computerised resources, such as a morpheme database, that can be used for the development of a morphosemantic analyser. However, as discussed in Section 3.1, most of the existing analysers yield excellent results in inflectional morphology. Regarding the analysis of the derivatives, the analysers primarily focus on the derivation of verbs. In contrast, the analysis of nominal derivatives remains notably limited.

To enhance the formal and semantic analysis of nominal derivatives, it is necessary to build formalised resources. Our approach starts with the investigation of the representation of nominals in the descriptive linguistic studies and Turkish textbooks. Nonetheless, our analysis revealed another issue; a lack of formalised description of the nominal morphemes in Turkish derivational morphology. This matter is further discussed in Section 3.2.

The lack of a formalised description of derivational morphemes may also explain the lack of a morphosemantic analyser, especially for nominal derivatives. Therefore, we established a methodology to standardise the representation of nominal morphemes and their description, as presented in Section

4. This approach considers the formal, categorial and semantic aspects of the morphemes to enable their automatic processing. It then resulted in the development of two different resources, built in an Open Science perspective¹. These resources are *Semantürk*, an ontology of semantic categories and *DerivBaseTR*, a database of Noun-to-Noun (N-to-N) morphemes and their corresponding descriptions.

2 Turkish Derivational Morphology

2.1 Formal Level

Turkish is an agglutinative language where suffixation is the predominant morphological process. In N-to-N derivation, derivational morphemes are all bound morphemes attached to a free morpheme, which may be either a simple word (zero derivation) or a complex word (having one or more derivational morphemes). Exceptions aside, the root, whether complex or not, does not change. Typically, the suffix is concatenated directly to the root word, as in (2).

- (2) göz (en. eye)
 gözlük (en. eyeglasses)
 gözlükçü (en. optician)
 gözlükçülük (en. opticianry)

However, the majority of bound morphemes conform to the vowel and consonant harmony rules, leading to allomorphy. These rules do not alter the semantic or grammatical features of the morphemes. Instead, they trigger formal and phonological adaptations of the morpheme. Particular conventional writing rules are used to represent all allomorphs in a single form. Firstly, if the morpheme starts with a capital consonant, it denotes consonant harmony. The uppercase consonant corresponds to a voiceless and voiced consonant pair as shown in Table 1. In cases where the root word ends with a voiceless consonant, such as the word *kitap*, the suffix begins with the voiceless consonant of the pair, as in the suffix -Cİ (3a). Otherwise, the suffix begins with the corresponding voiced consonant (3b).

Voiceless	Voiced	Symbol
ç [tʃ]	c [ɟ]	C
t [t]	d [d]	D
k [k]	g [g]	G

Table 1: Consonant pairs in consonant harmony

- (3) a. kitap-ç**i**
 book-C**I**
 “bookseller”
 b. şark**i**-c**i**
 song-C**I**
 “singer”

Morphemes can undergo either simple or complex vowel harmony rules. Simple vowel harmony is usually represented by the symbol A². It applies to the two open vowels *a* and *e*. If the last syllable of the root word contains a front vowel, such as *e*, *i*, *ü* or *ö*, then the vowel in the suffix will be *e*, as in (4a). Otherwise, it will be the vowel *a* (4b).

- (4) a. Türk-ç**e**
 Türk-CA
 “Turkish language”

¹Our resources are to be accessible and usable for future works.

²In some instances, the letter E represents simple vowel harmony. Here, we use the symbol A.

- b. Fransız-ca
French-CA
“French language”

Complex vowel harmony is denoted by the letter $\dot{\text{I}}$ ³ with four possible closed vowels: *i*, *ü*, *ı* or *u*. If the last syllable contains a closed vowel, then the vowel in the suffix will be identical (5a). Else, the vowel in the suffix will be its closed vowel counterpart (5b). Table 2 displays the possible combinations that arise due to the complex vowel harmony rule.

Last vowel	Suffix vowel	Last vowel	Suffix vowel
a [a]	ı [ɯ]	e [e]	i [i]
ı [ɯ]	ı [ɯ]	i [i]	i [i]
u [u]	u [u]	ü [y]	ü [y]
o [o]	u [u]	ö [œ]	ü [y]

Table 2: Complex vowel harmony

- (5) a. ayakkabı-lık
shoe-lık
“shoe cupboard”
b. göz-lük
eye-lık
“eyeglasses”

2.2 Categorical Level

Derivational morphemes, unlike inflectional morphemes, allow for the creation of new lexemes, mainly characterised by a possible change in the word class. A significant number of morphemes come into play in Turkish nominalisation, such as N-to-N morphemes, Verb-to-Noun morphemes, Adjectives-to-Noun morphemes, and so on. However, our research focuses on N-to-N derivation which limits our scope to the semantics of nominals.

The distinction between word classes is very significant since the semantics of the morphemes closely correlates to the grammatical class of either the root or the derivative, as explained in Section 2.3. Turkish linguistic studies, particularly in morphology, offer a different perspective on word class distinction in comparison to the word class distinction put forth in Western linguistic studies. Derivational morphemes are classified into two separate categories, verbs (tr. *fil*) and nouns (tr. *ad* or *isim*⁴). The latter includes numerals, adjectives, adverbs and pronouns (further discussed in Section 3.2 and Section 4).

Furthermore, this classification of nominal morphemes reflects their polycategorical nature. This is because many morphemes classified as nominal morphemes can result in derivatives of various word classes (nouns, adjectives, adverbs, or sometimes pronouns). (6) clearly shows the polycategoriality of the morpheme -CA as it can, attached to the noun *kadın* (en. woman), derive a new noun (6a), adjective (6b) or adverb (6c).

- (6) a. kadın-ca → N-to-N
woman-CA
“the language of women”
b. kadın-ca → N-to-Adj.
woman-CA
“womanlike”

³In some instances, the letter I or H represents complex vowel harmony. Here, we use the symbol $\dot{\text{I}}$.

⁴These terms are synonymous and can be used interchangeably within the context of a nominal lexeme or a nominal class that covers different categories.

- c. kadın-ca → N-to-Adv.
 woman-CA
 “womanly”

Moreover, a change in the meaning of the lexemes in (6) can be noticed, indicating a direct link between morpheme meaning and grammatical category. To minimize ambiguity in the analysis of nominal morphosemantics in Turkish derivational morphology, we restrict our analysis to N-to-N derivation.

2.3 Semantic Level

A morpheme is traditionally defined as the smallest meaningful unit of a language. This approach is especially appropriate for the description of agglutinative languages. As mentioned earlier, derivational morphemes enable the formation of new lexemes. This leads to a change in the word class, but it can also lead to a change in the meaning, as shown in (6). Meaning can change significantly, which is the case between the nominal form (6a) which refers to an abstract entity and the adjectival form that denotes a more qualitative concept in (6b). However, it can also be more ambiguous as in (6b) and (6c), with both examples showcasing the qualitative aspect.

It is important to note that morpheme polysemy is not necessarily related to polycategoriality. In fact, a morpheme that creates N-to-N derivatives can produce entirely different meanings. In (7), the morpheme -lîk first produces a concrete material object designated by the noun (7a). However, it also creates an abstract noun (7b). The combination of the morphemes -Cî-lîk results in the abstraction of the lexeme, noted as a recurrent distributive pattern. Therefore, the meaning of the morpheme in question can also be context-dependent. This can be observed with various morphemes, cf. example (4) given previously, where the addition of the suffix -CA to a noun denoting nationality results in a noun denoting the language or the dialect spoken in that nation.

- (7) a. göz-lük
 eye-lîk
 “eyeglasses”
 b. gözlükçü-lük
 optician-lîk
 “opticianry or the occupation of an optician”

A semantic category can also be conveyed by different morphemes, resulting in synonymous or quasi-synonymous morphemes. Typically, the diminutive morphemes -Cîk and -cAğîz both convey a sense of a pity felt by the speaker towards the referred entity, as shown in (8).

- (8) a. kedi-cik
 cat-Cîk
 “the poor little cat”
 b. adam-cağiz
 man-cAğîz
 “the poor little man”

Therefore, the correlation between form and meaning can be qualified as a many-to-many relationship, that is a morpheme can be associated to one or more semantic categories, just as a semantic category can be associated with one or more morphemes. It can be either dependent on the category or its distribution.

Lexicalisation is also a phenomenon present in the Turkish language. Some derivatives can show a high degree of lexicalisation. Some morphemes can be synchronically difficult to detect and more root dependent where many others are completely distinct and are independent from the root word. Lexicalised derivatives are not taken into consideration in this research as the morpheme in these cases loses its semantic component and requires an etymological analysis.

3 Resources and Studies in Turkish Nominal Morphology

3.1 Nominal Morphemes in NLP Tools and Resources

A lot of research on Turkish language is currently being conducted in the fields of NLP (Ofłazer and Saraçlar, 2018; Çöltekin et al., 2023). One of the issues we met concerns the availability of existing resources as was highlighted in Çöltekin et al. (2023): “The locations of published resources are not always stable and/or permanent. The URLs indicating the location of the resources in papers or on the webpages of the authors or institutions are not always maintained and the resources often disappear after publication. Although our efforts to reach out to the authors/creators of the resources often yielded positive results, it is desirable to diminish these barriers to keep up with the fast-paced research community.”

While there are numerous studies available for the French language, e.g. Missud et al. (2020); Mailhot et al. (2020); Varvara et al. (2022); Hathout and Namer (2022), to our knowledge, very few focus on the derivation of Turkish nouns, and even less to the particular subject of N-to-N derivation. Among the most well-known NLP tools in Turkish, there is *Zemberek*⁵ (Akın and Akın, 2007), an open-source Java library (no longer updated). The morphology processing section offers various analyses, i.e. single word morphological analysis, stemming and lemmatisation, contextual ambiguity resolution, and word generation. However, the processing mainly results in inflectional analyses, with word generation producing an output of inflected forms of the entry word, as shown in the examples of outputs for the entry *ev* (en. house) in (9).

- (9) a. evime
ev-im-e
house-1SG.POSS-DAT
“to my house”
b. evimde
ev-im-de
house-1SG.POSS-LOC
“in my house”

Another well-known tool is *TRmorph*⁶ (Çöltekin, 2010), an open-source morphological analyser, written using a Foma Finite State Transducer (FST) compiler, which produces a list of possible analyses for an out-of-context lexeme. In addition to a complete inflectional analysis, it accurately identifies verbal derivational morphemes. However, it only identifies a short list of the most productive nominal morphemes. 17 derivational suffixes with nominal roots are described in the resource. Only seven of these (four of which have been regrouped) are annotated as N-to-N suffixes: *-lîk_(lîk)*, *-Cîk_(dim)*, *-cAk_(dim)*, *-(İ)cAk_(dim)*, *-cAğİz_(dim)*, *-Cî_(ci)*, *-gil_(gil)*, which is a rather small sample of nominal suffixes. However, the part-of-speech categorisation of the morphemes by *TRmorph* does not exactly match ours. For instance, unlike in our classification, *-CA* is not categorized as an N-to-N morpheme in this analyser.

A new open source Java library, *Turkish Morphological Analyzer*⁷ (Yıldız et al., 2019), was released in 2019. Again, only four N-to-N suffixes are identified: *-Cî*, *-Cîk*, *-(İ)ncî*, *lîk*. However, they added specific tags, AGT, DIM, ORD and NESS respectively, representing a possible semantic role of these derivational suffixes.

*Trnlp*⁸ (Bayol, 2018) is an ongoing project, an open source Python API. It has several components including lemmatisation, stemming, spellchecking and tokenisation. It identifies a more diverse set of nominal derivational suffixes. Although it gives good results, it still needs improvement: 1. the suffixes listed in the N-to-N section are not all correct (e.g. *-m* is included but actually corresponds to the first person possessive suffix); 2. among the 27 suffixes listed as N-to-N suffixes, several do not result in nominal derivatives (e.g. *-sî* results in adjectival derivatives); 3. the output of the analysis is not disambiguated. Nevertheless, it produces an analysis on 15 nominal suffixes, which is one of the best

⁵<https://github.com/ahmetaa/zemberek-nlp>

⁶<https://github.com/coltekin/TRmorph>

⁷<https://github.com/olcaytaner/TurkishMorphologicalAnalysis>

⁸<https://github.com/brolin59/trnlp>

results we have observed so far.

As our project is carried out in an Open Science perspective, we did not analyse publicly unavailable resources. Some examples are PC-KIMMO-based analyser (Oflazer, 1994), SakMP (Sak et al., 2008), ITU Turkish NLP Web Service (Eryiğit, 2014).

Not only are there very few tools available for Turkish derivation in nominal morphology, but there are also no available computerised morphological resources. To our knowledge, there are no accounts of:

- dictionaries with morphological descriptions,
- exhaustive inventories of morphemes, whether formalised or not.

For instance, while the French Wiktionary has 1,935,402 entries, the Turkish Wiktionary has only 3,958 entries⁹ and therefore does not provide a usable dataset for any morphological analysis. Moreover, it does not contain any information on derivatives. As shown in Figure 1¹⁰, there is only the “definition” (or a synonym of the word as given in this example) of the word whereas *fakirlik* (en. “poverty”) is a noun derived from *fakir* (en. “poor”) with a very productive suffix -lik.

Ad [değiştir]
fakirlik (belirtme hâli **fakirliği**, çoğulu **fakirlikler**) .ği
1. (toplum bilimi) yoksulluk

Figure 1: Example from Turkish Wiktionary

To overcome the scarcity of easily accessible and available resources in derivational morphology from an NLP perspective, we collected data from various linguistic studies in order to design and then implement new computerised resources. However, this is not a trivial task as we faced several difficulties originating from the descriptions proposed in these studies, as discussed in the following subsection.

3.2 Nominal Morphemes in Linguistic Studies

The linguistic books we examined were Turkish (Adalı, 2004; Korkmaz, 2014; Boz, 2015), French (Bazin, 1994) and English grammar books (Göksel and Kerslake, 2005) as well as a few Turkish textbooks for learners (Bozdémir, 1991; Erikan et al., 2008). We have also looked at the two other sources, an article by Akçataş and Taşdemir (2020), and a master’s thesis by Ozturk (2016), focusing on the morphosemantics of Turkish morphemes. However, new difficulties arose during the data collection. These difficulties were more or less common to all of the above-mentioned studies as listed below.

1. Lack of descriptions in alphabetically ordered lists

Descriptive linguistic studies of the Turkish language mainly consist of a set of morphemes listed alphabetically with instances of derived words without any explanation on the morphotactics or the semantic value of the morpheme, e.g. Adalı (2004).

2. Difference in morpheme categorisation

As introduced in Section 2.2, Turkish linguistic studies introduce a different word class categorisation, describing morphemes of different word classes in the section dedicated to nominal morphology. For example, (10), extracted from the section “Suffixes that attach to nominals to form nominals” in Göksel and Kerslake (2005), is a morpheme that produces an adjective. We can also find suffixes attaching to or deriving adverbs and pronouns in addition to nouns and adjectives.

(10) -(A)C Attaches to nouns to form adjectives: *anaç* ‘motherly’, *kıraç* ‘infertile’

This is a traditional categorisation of word classes in the literature of Turkish linguistics (and other Turkic linguistics in general). The inclusion of adjectives, adverbs, pronouns and numerals in a single nominal class reflects the close interaction of these classes and their ability to function as nominal

⁹<https://fr.wiktionary.org/wiki/Wiktionnaire:Statistiques> (last accessed: June 21st, 2023)

¹⁰<https://tr.wiktionary.org/wiki/fakirlik> (last accessed: June 21st, 2023)

elements, whether the lexeme is polycategorical or not. Syntax, in Turkish, has a relatively flexible lexeme order so that nouns, adjectives, adverbs and pronouns can occur in different positions, including a nominal position, i.e. adjectives can be used as nouns in a sentence, without any formal indication on the functional change apart from the syntactic position. In addition, they can easily function as nouns and take nominal inflectional suffixes.

3. Non-exhaustiveness

The number of morphemes described varies from study to study, as illustrated by a few examples in Figure 2. Introductory studies (Bazin, 1994) or pedagogical textbooks (Bozdémir, 1991; Erikan et al., 2008) for language learning do not have complete descriptions of derivational morphemes. They tend to focus on a few of the most productive ones. Among the remaining linguistic studies, Göksel and Kerslake (2005) and Korkmaz (2014) have the highest number of morphemes described¹¹. This variation is due to different approaches to morpheme description. Indeed, Korkmaz (2014) also describes dead affixes in lexicalised forms, which is on a borderline with a diachronic approach to morphemes. Göksel and Kerslake (2005) include many loaned morphemes mainly of Arabic or Persian origin. Incoherence in morpheme description between different sources also explains the difference in the number of morphemes described.

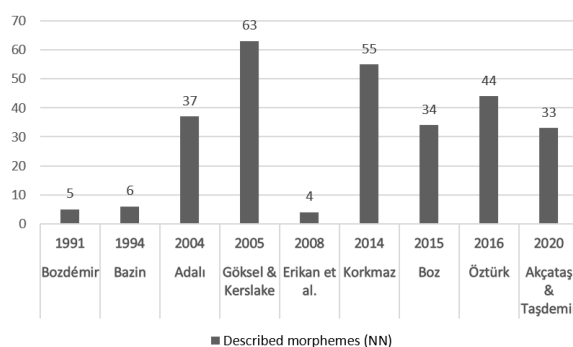


Figure 2: Number of morphemes per source

4. Incoherence in morpheme description

Different sets of references show discrepancies in different descriptive aspects. For example, there is a difference in the description of the suffix *-sal* in Göksel and Kerslake (2005) and Korkmaz (2014). On one hand, in Göksel and Kerslake (2005), this suffix is described as mainly a Noun-to-Adj. suffix which, in rare cases, also forms nouns: *kumsal* ‘sandy beach’. On the other hand, Korkmaz (2014) clearly states that the suffix is not related to the form *sal* in *kumsal*. We can also see incongruent morpheme representations across various sources, as for the morpheme *-cağız*. In Korkmaz (2014), we have *-Cağız*, whereas in Adalı (2004), the morpheme *-IZ* (*ız, iz, uz, üz*) is a separate morpheme entry attached to stems ending with the morpheme *-CAK* (*-cak, -cek, çak, çek*), including non-grammatical stems such as **çocukcak, *kızcak*, etc. Another difference we noted in most of the sources, is that each morpheme is described with a different set of information throughout the same source. The semantic function of a suffix is explained for some of the morphemes, as in (11) (Göksel and Kerslake, 2005). However, some suffixes are described only from a grammatical point of view (10). The description is therefore unsystematic and may be incoherent.

- (11) *-Daş/Deş* Added to nouns to form nouns denoting possessors of a shared attribute: *yandaş* ‘supporter’, *kardeş* ‘sibling’ (from *karın* ‘abdomen’), *meslektaş* ‘colleague (i.e. person of the same profession)’.

¹¹A few loaned prefixes are mentioned in several of the sources, but are not further studied or described.

4 Formalised Morpheme Description in Machine-readable Resources

As discussed earlier, we assume that the semantics of N-to-N morphemes can be identified using existing linguistic sources. In this section, we present the processing steps of our methodology for the development of two formalised resources, *DerivBaseTR* and *Semantürk*. Figure 3 illustrates the workflow for the creation of these two independent resources.

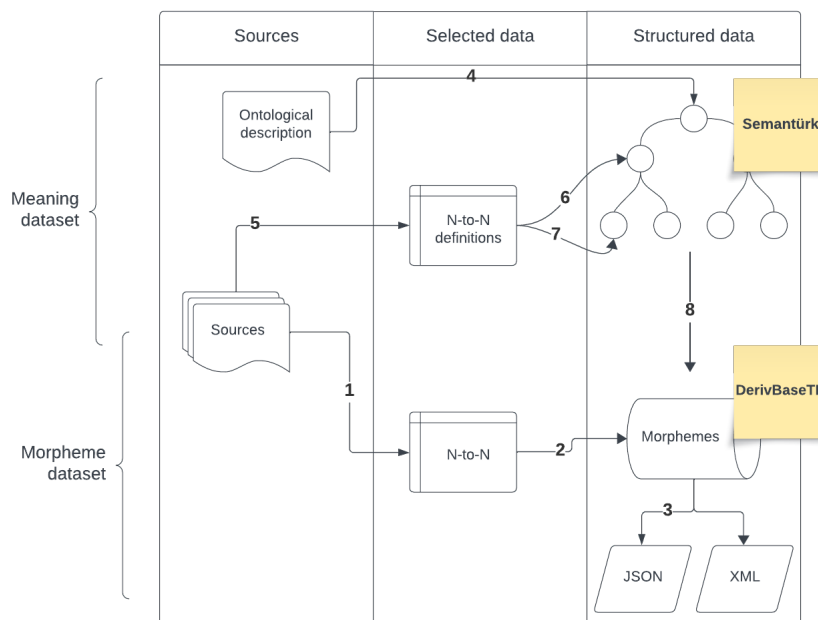


Figure 3: Processing workflow

1/ Following the examination of the existing sources in Turkish linguistics, we extracted the morphemes producing N-to-N derivation. This task was complicated by the different morpheme categorisation in Turkish linguistics¹². Some linguists claim that the categorial flexibility of the lexemes is a proof of a functional variation rather than a categorial variation. That is, the categorial function of a lexeme is syntax-dependent. However, it can also be argued that lexemes inherently carry categorial information, so that their category can be identified in the lexicon¹³. In fact, any given word in a dictionary, such as Türk Dil Kurumu Sözlükleri¹⁴ (TDK sözlükleri, *the dictionaries of the Turkish Language Association*), is assigned a grammatical category per meaning. After excluding all dead suffixes which result in lexicalised forms, and selecting suffixes from sources where the grammatical category (or categories) of the root and the derivative were already given as nouns, we studied the examples of derivatives for the unclassified ones. We proceeded to the selection by identifying the “primary function” (Göksel and Kerslake, 2005) of the examples of derivatives given in the morphemes description with the help of the TDK dictionaries. 2/ We then formalised and stored all the information given on the selected morphemes in an Excel file. In this way, we collected the morpheme representation¹⁵, its allomorphs, its origin, and examples of derivatives. We also added *Base category* and *Derived category* entries in the morphemes’ descriptive properties to ensure the possibility of adding other grammatical categories. We developed a first version of *DerivBaseTR* with a formalised description of the morphemes at both the formal and categorial levels, offering the possibility of filtering or ordering the morphemes by features. 3/ We plan to add the possibility to generate a json and/or xml file of the stored data. This would facilitate and enable its use in any NLP project. We have chosen two formats in order to make it accessible to a wider public.

¹²Mentioned in Section 2.2 and Section 3.2.

¹³Gorgülü (2012, Ch. 1) gives an insight of the different theories on the subject matter.

¹⁴<https://sozluk.gov.tr/>

¹⁵As aforementioned, we sometimes encountered discrepancies in morpheme representation. We chose the morpheme that best represented the actual allomorphs found in derivatives.

Semantürk, the second resource is an ontology of semantic categories encoding meanings. Therefore the semantic category refers to the meaning of the morpheme and is representative of it. We have built this resource, written in Web Ontology Language (OWL), using a hybrid methodology applying both a top-down and a bottom-up method. 4/ Firstly, the main structure of the ontology is adapted from an existing tagset for the description of nominal semantics in French (Huguin et al., 2022). This tagset is based on WordNet's¹⁶ top concepts called Unique Beginners (Fellbaum, 1998). Initially not defined for morpheme description, it proved adaptable as we applied the set to define the N-to-N morphemes at the semantic level as explained later. 5/ We then collected all the definitions and meanings found in the various sources and stored them in a single file, aligning them by morphemes and source. 6/ Once we had collected all the morpheme definitions, we matched them to the main structure of our ontology. 7/ If no match was found, or if the existing category was too broad to reflect the meaning of the morpheme, we created a new semantic category. As the semantic categories are hierarchically ordered, we could adapt the set and add new semantic categories specific to Turkish derivational morphemes, with the possibility of having different levels of granularity.

8/ In addition, we added a new *Semantic category* entry to *DerivBaseTR* and annotated each morpheme with the semantic categories of *Semantürk*, so that the morphemes are now described at the formal, categorial and semantic levels. Some morphemes present semantic transparency and are annotated with only one semantic category. Others are more ambiguous and have multiple semantic categories.

5 Conclusion

Prior to the construction of the morphosemantic analyser, the establishment of a formalised descriptive resource of derivational morphemes is necessary. The development of formalised resources requires the establishment of a specific framework for the description of Turkish morphemes. Therefore, we have created two different sets of resources: an ontology of semantic categories for the description of morphemes called *Semantürk* and *DerivBaseTR*, a database that formalises the description of morphemes at the formal, categorial and semantic levels. The resources are built with the perspective of possibly being used as additional components in various linguistic or NLP projects, and extended with other types of morphemes or new features. As the majority of published computerised resources are either not available or not easily accessible, this project is conducted in an Open Science perspective. We hope to provide extendible and interoperable resources to help improve the progress of the research in processing of the Turkish derivational morphology.

References

- Oya Adalı. 2004. *Türkiye Türkçesinde Biçimbirimler*. Papatya, Istanbul, Türkiye.
- Ahmet Akçataş and Serpil Taşdemir. 2020. Türkiye türkçesinde kök ya da gövdeye gelen ekler üzerine bir anlambilim incelemesi. *Avrasya Dil Eğitimi ve Araştırmaları Dergisi* 4(1):129–149.
- Ahmet Afşın Akın and Mehmet Dünder Akın. 2007. Zemberek, an open source NLP framework for Turkic languages. *Structure* 10:1–5.
- Esat Mahmut Bayol. 2018. *Türkçe Doğal Dil İşleme Macerası*. <https://turkceddi.blogspot.com/2018/08/turkce-dogal-dil-isleme-maceras-her.html>.
- Louis Bazin. 1994. *Introduction à l'étude pratique de la langue turque*. Librairie d'Amérique et d'Orient, Paris, 3rd edition.
- Erdoğan Boz. 2015. *Türkiye Türkçesi, Biçimbilimsel ve Anlamsal İşlevli Biçimbilgisi*. Gazi Kitabevi Tic. Ltd. Şti., Ankara, Türkiye, 4th edition.
- Michel Bozdémir. 1991. *Méthode de turc*, volume 1. L'Asiathèque, maison des langues du monde, Paris.
- Catherine Erikan, Ayhan Erdal, and Marie Koçoğlu. 2008. *Apprenons le turc ensemble / Beraber Türkçe Öğrenelim*, volume 1. Ataturque, Paris, 2nd edition.

¹⁶Wordnet is a lexical database, used for more than 200 languages, including Turkish.

- Gülşen Eryiğit. 2014. *ITU Turkish NLP Web Service*. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1–4. <https://doi.org/10.3115/v1/E14-2001>.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press, Cambridge, London.
- Emrah Gorgülü. 2012. *Semantics of nouns and the specification of number in Turkish*. Simon Fraser University.
- Aslı Göksel and Celia Kerslake. 2005. *Turkish: a comprehensive grammar*. Routledge comprehensive grammars. Routledge, London.
- Nabil Hathout and Fiammetta Namer. 2022. *ParaDis: a family and paradigm model*. *Morphology* 32(2):153–195. <https://doi.org/10.1007/s11525-021-09390-w>.
- Mathilde Huguin, Lucie Barque, Pauline Haas, Fiammetta Namer, and Delphine Tribout. 2022. *Guide d’annotation Demonext: typage lexical des noms du français*.
- Zeynep Korkmaz. 2014. *Türkiye Türkçesi Grameri, Şekil Bilgisi*. Türk Dil Kurumu Yayınları, Ankara, Türkiye, 4th edition.
- Hugo Mailhot, Maximiliano A. Wilson, Joël Macoir, S. Hélène Deacon, and Claudia Sánchez-Gutiérrez. 2020. *MorphoLex-FR: A derivational morphological database for 38,840 French words*. *Behavior Research Methods* 52(3):1008–1025. <https://doi.org/10.3758/s13428-019-01297-z>.
- Alice Missud, Pascal Amsili, and Florence Villoing. 2020. *VerNom : une base de paires morphologiques acquise sur très gros corpus (VerNom : a French derivational database acquired on a massive corpus)*. In Christophe Benoit, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla, and Stéphane Schneider, editors, *Actes de la 6e conférence conjointe Journées d’Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, Nancy, France, June 8-19, 2020. ATALA et AFCP, pages 305–313. <https://aclanthology.org/2020.jeptalnrecital-taln.30/>.
- Fiammetta Namer. 2002. *Acquisition automatique de sens à partir d’opérations morphologiques en français : études de cas*. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs, TALN 2002, Nancy, France, June 2002*. ATALA, pages 237–246. <https://aclanthology.org/2002.jeptalnrecital-long.21/>.
- Kemal Oflazer. 1994. *Two-level description of turkish morphology*. *Literary and Linguistic Computing* 9(2):137–148. <https://doi.org/10.1093/lc/9.2.137>.
- Kemal Oflazer and Murat Saraçlar. 2018. *Turkish Natural Language Processing*, volume 1 of *Theory and Applications of Natural Language Processing*. Springer, Cham, Switzerland.
- Seda Ozturk. 2016. *Création et reconnaissance de néologismes par méthode de suffixation*. Université de Franche-Comté, Besançon, France.
- Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. *Turkish language resource: Morphological parser, morphological disambiguator and web corpus*. In Bengt Nordström and Aarne Ranta, editors, *Advances in Natural Language Processing*. Springer, Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-540-85287-2_40.
- Rossella Varvara, Justine Salvadori, and Richard Huyghe. 2022. *Annotating complex words to investigate the semantics of derivational processes*. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*. European Language Resources Association, pages 133–141. <https://aclanthology.org/2022.isa-1.18>.
- Olçay Taner Yıldız, Begüm Avar, and Gökhan Ercan. 2019. *An open, extendible, and fast turkish morphological analyzer*. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. INCOMA Ltd., pages 1364–1372. https://doi.org/10.26615/978-954-452-056-4_56.
- Çağrı Çöltekin. 2010. *A freely available morphological analyzer for turkish*. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/109_paper.pdf.
- Çağrı Çöltekin, A. Seza Doğruöz, and Özlem Çetinoğlu. 2023. *Resources for Turkish natural language processing: A critical survey*. *Language Resources and Evaluation* 57(1):449–488. <https://doi.org/10.1007/s10579-022-09605-4>.

A keymorph analysis of Russian political news reporting

Thomas Samuelsson

Stockholm University

Stockholm, Sweden

thomas.samuelsson@slav.su.se

Abstract

This paper presents a diachronic study of Russian prefixes in the political news reporting. The analysis examines derivational prefixes in the Russian media discourse for each year in the time period 2012–2020. The prefixes are analyzed using derivational keymorphs. The data material consists of a corpus of political texts from more than 60 Russian online media resources. Key-morphs have previously been used to investigate Czech presidential discourse (Fidler and Cvrček, 2019), the Russian media resource Sputnik Czech Republic (Fidler and Cvrček, 2018; Cvrček and Fidler, 2019) and Putin’s speeches (Janda et al., 2023). The use of keymorphs enables one to focus on morphological features and to capture general characteristics of the textual content in a language corpus. The work uses Corpus-assisted discourse studies (CADS) as the main framework and is a contribution to the understanding of Russian political discourse.

1 Introduction

This paper studies Russian morphological derivation in political discourse. More precisely, it studies prefixes in the Russian political news reporting from a chronological perspective in time period 2012–2020 by applying the concept of keymorphs. This concept has been introduced by Fidler and Cvrček (2019) as an extension of keyword techniques. Keywords are a useful part of corpus-based discourse studies (Partington and Duguid, 2020) and are often used in the study of political discourse (Ädel, 2010). While an investigation of keywords highlights major key topics and stylistic features in the discourse, keymorphs identify to a greater extent general characteristics of the discourse (Fidler and Cvrček, 2019). The extension of the concept of keyness beyond keywords to other key items have previously also been introduced for other linguistic units than morphemes. One major approach that uses key part of speech tags (pos-tags) and semantic categories in the analysis of linguistic characteristics has been developed by Rayson (2004, 2008). The method of identifying key pos-tags and key semantic categories has been exploited in a variety of studies. Culpeper (2009) utilized key part of speech and semantic field analysis to analyze the characters in Shakespeare’s *Romeo and Juliet*. Archer et al. (2009) explored Shakespeare’s plays in terms of key semantic fields and Afida (2007) analyzed business magazines using the same technique. Some recent studies on key parts of speech have been conducted by Breeze (2019) on legal genres and Smith and Waters (2019) on a British radio show.

The studies have not focused on morphosyntactic features since the opportunity to do that in English is limited due to a quite high degree of analyticity (Cvrček and Fidler, 2019). Czech and Russian are on the other hand typologically more synthetic and have inventories of identifiable inflectional markers that are richer. Fidler and Cvrček (2019) show that an investigation of inflectional keymorphs (case, number, person, finiteness, verb negation) and part of speech keymorphs revealed representations of situations and speaker images in Czech presidential speeches. Ideological differences between the presidents were associated with parts of speech and stylistic variations with inflectional features. They have also explored the Russian media portal Sputnik Czech Republic and showed that inflectional keymorphs are a tool that provide information about the structure of the discourse (Fidler and Cvrček, 2018; Cvrček and Fidler, 2019). As an illustration, the analyses show that Russia is likely to be portrayed as a victim and Putin to

be represented as an actor with agency in the Russian-controlled media outlet. Janda et al. (2023) tested for the first time keymorph analysis on Russian data in a study of Putin’s speeches delivered around the time of Russia’s full-scale invasion of Ukraine. The study revealed the roles of the actors Russia, NATO and Ukraine in the narrative by exploring key grammatical cases. To mention some roles, the different actors could be portrayed as agents, victims, dynamic, static, places and states.

This study, however, focuses on the derivational morphological system in Russian political news reporting within the framework of Corpus-assisted discourse studies (CADS) (Partington et al., 2013). In particular, it is built on the perspective that discourse studies target semantic issues and consequently ascribe meanings to all linguistic elements in the contexts, including the smallest meaningful constituents in the language (Spitzmüller and Warnke, 2011). The study is also motivated from the approach of Cognitive Grammar, which “treats all linguistic units and categories as meaning-bearing, in all contexts” (Divjak and Janda, 2008). From empiricism, researchers have noticed that the Russian morphological derivational system reacts to extra-linguistic factors such as socio-political processes (Ratsiburskaya et al., 2015). Based on the Russian socio-political development during the investigated time period, diachronic variations are to be expected. Characteristic of the contemporary Russian language is the rise of the derivational nominal prefixes whose use extensively are of non-Slavic origin (Zemskaya, 2006). The prefixes originate alongside English from Greek and Latin (Koriakowcewa, 2009).

There have been advances in the creation of Russian derivational resources in the recent years that enables one to investigate the Russian discourses by exploiting derivational morphemes.¹ This paper examines co-occurrences of lexical constituents and discursive functions in the Russian political media discourse that belong to other aspects of the discourse that are not otherwise investigated such as key topics or typical linguistic items identified through the use of keywords. Since a keymorph analysis uses the same principles as other keyness statistical approaches, it also has the same advantages compared to a qualitative analysis. A keyword analysis can reveal features that are not obvious to an unaided eye and are hard for an observer to detect (Culpeper, 2009). Another benefit according to the author is that it also uncovers patterns without the use of intuition of an observer. The aim of this preliminary study is to probe the Russian political media reporting by identifying changes in the salience of derivational prefixes. The keyness of the prefixes in the study is estimated by using the derivational keymorph technique. Against this background, the study will present preliminary results based on the use of the derivational morphology resource DeriNet.RU 0.5 for Russian (Kyjánek et al., 2022).

2 Methodology

The diachronic study approaches Russian political news reporting through the evolution of keyness of derivational prefixes. The prominent prefixes for each year are identified by comparing the relative frequencies of the prefixes in a target corpus composed of texts published in the same single year with the relative frequencies of the same prefixes in a reference corpus in the time frame 2012–2020. The target corpora are created by partitioning the language data according to year. The corpora of interest represent Russian political online reporting for each year in the studied time period. The comparisons are made with the entire undivided reference corpus that represents Russian language presented on the web. Since Russian has a comparatively high degree of inflection and the target is semantic constituents, lemmas from both corpora are used. Every lemma in the corpora is automatically looked-up in the morphological database DeriNet.RU 0.5 to get derivational information (Kyjánek et al., 2022). If the lexicon does not contain information about the subparts of a prefixed lexeme for some reason, it is not integrated into the results.

The keymorph analysis is conducted in the same way as a keyword analysis. In both approaches, a target corpus is contrasted to a reference corpus. The prominence of the morphemes is calculated by comparing the frequencies using two types of calculations (Fidler and Cvrček, 2015, 2019). The first calculation makes sure that there are enough data evidence. For this purpose, the study uses the statistical test log-likelihood ratio. The null hypothesis is that there is no difference in the frequencies of a prefix. The differences in the uses of each of the most prominent prefixes are statistically significant at a level

¹<https://ufal.mff.cuni.cz/universal-derivations>

$p < 0.05$. The second calculation estimates the effect size through the metric Log Ratio introduced by Hardie (2014). The effect size estimator is calculated for each prefix according to equation 1, where f_{target} denotes the relative frequency of a prefix in a target corpus, $f_{reference}$ designates the relative frequency of the same prefix in the reference corpus and \log_2 is the binary logarithm:

$$\text{Log Ratio} = \log_2 \frac{f_{target}}{f_{reference}}. \quad (1)$$

Previously, the effect size of keymorphs has been measured using Difference Index (DIN) (Fidler and Cvrček, 2019). The DIN-index is essentially the difference between the relative frequency of an item in the target corpus and the relative frequency of the same item in the reference corpus divided by the mean of the relative frequencies; and then the index is normalized so that the range is ± 100 . Difference Index was introduced to handle words that are present in the target corpus but absent in the reference corpus (Fidler and Cvrček, 2015). However, the choice of Log Ratio appears more intuitive and less complicated to interpret. On the one hand the measures differ when it comes to magnitude, range and treatment of the absence of lexical items. On the other hand they both generate the same rank order (Gabrielatos, 2018), which is an important part of the analysis (Fidler and Cvrček, 2015). While Difference Index has a range between -100 and 100 , the range of Log Ratio is unbounded for both positive and negative values. Difference Index handles absence in corpora while Log Ratio on the other hand does not allow frequencies to be equal to zero. But in large corpora, the absence of prefixes is unlikely to be an issue. The interpretation of the value of the Log Ratio is as follows. A Log Ratio value of zero shows that the relative frequencies of the prefixes are equal in both corpora. A Log Ratio value of 1 means that the relative frequency of an item in the corpus of interest is twice the relative frequency of the same item in the reference corpus. A Log Ratio value of 2 corresponds to a relative frequency that is 4 times larger in the target corpus compared to the reference corpus. If we continue with the values 3, 4 and 5, the relative frequency will be 8, 16 and 32 times larger (Hardie, 2014).

3 Data

To be able to calculate the Log Ratio values for the Russian prefixes, both well annotated corpora and a high-quality derivational analysis tool are needed. Since Russian, as mentioned above, has rich morphology, compared to for example English, corpora annotated with lemmas are therefore required. This section presents the target corpus and the reference corpus used in the study, as well as the database containing derivational information about the Russian lemmas.

3.1 Corpora

The study requires, as previously mentioned, target corpora and a reference corpus to calculate the Log Ratio values of the prefixes. The target corpora consist of a collection of Russian political texts published online between 2012 and 2020. The data material is sampled from the most influential Russian online media resources according to a citation index provided by the leading Russian media monitoring company Medialogia (2023)² and the selection of texts is made on the basis of political classification. In the linguistic processing, deduplication of the texts has been applied. The raw texts are tokenized and lemmatized using the natural language analysis tool Stanza (Qi et al., 2020) trained on the SynTagRus treebank (Dyachenko et al., 2015; Droganova et al., 2018). The performance for Russian measured in F1 scores are for tokens 99.57 and for lemmatization 97.51.³ The size is more than 500 million tokens from more than 60 outlets. The extracted texts originate from a diversity of journalistic resources of different genres, geographical cover and political orientations. The dataset contains Russian political media texts annotated with publication date.

The acquisition of relevant keyed items is connected to the relation between the target corpus and the reference corpus (Culpeper and Demmen, 2015). A relation that is close between the corpora increases the likelihood to obtain keywords that are specific to the target corpus (Culpeper, 2009). Since the interest

²<https://www.mlg.ru/>

³<https://stanfordnlp.github.io/stanza/v100performance.html>

here is to uncover keymorphs that are characteristic to Russian political journalism, a general web corpus is preferred to a balanced general corpus. The target corpora are contrasted with a reference corpus that is much larger in size than any of the target corpora and is assessed to have achieved representativeness of the Russian web language use during the given time period. In this way, the political aspects of the textual content will be highlighted. Used as a reference corpus in the study is the internet corpus *Ara-neum Russicum III Maximum 19.03* (Benko, 2014a,b; Benko and Zakharov, 2016, 2021; Rychlý, 2007). It is based on web-crawled Russian language data that have been acquired by applying the strategy of including everything that is possible to come across. The size of the reference corpus is almost 20 billion tokens from the actual time frame. The reference corpus is tokenized and lemmatized using TreeTagger (Benko, 2014a; Benko and Zakharov, 2021). In an evaluation of the performance in lemmatization tasks, TreeTagger trained on the disambiguated subcorpus of the Russian National Corpus (RNC) performed an accuracy of 97.0% on RNC⁴ and 86.9% on RU-EVAL gold standard (Kuzmenko, 2016). Kotelnikov et al. (2017) tested the parser on three corpora. The accuracy is 95.21% on the RU-EVAL corpus, 97.31% on the disambiguated subcorpus of the Russian National Corpus (RNC) and 96.95% on the disambiguated subcorpus of OpenCorpora. The results were biased on the RNC since TreeTagger was trained on it. Compared to the Russian national corpus, the selected reference corpus contains more rare lexical items but may on the other hand be less balanced (Benko and Zakharov, 2016).

3.2 Database

The frequencies of the prefixes in each of the corpora are calculated using the derivational lexical resource DeriNet.RU 0.5 (Kyjánek et al., 2022). DeriNet.RU is an open license state-of-the-art derivational model that captures derivational processes for the Russian language and includes more than 300 thousand lexemes and 164 thousand binary derivational relations, including derivational prefixation. The database outperforms other Russian derivational resources. It is the resource that contains the most number of lexemes and derivational relations. Besides that, all the lexemes are corpus-attested. The database includes derivational relations within and between nouns, adjectives, verbs and adverbs, but no compounds. Of the lexemes in the database, a majority (58%) consists of nouns; verbs and adjectives have about the same share, 20% respective 19%. The most common derivational relations involve nouns. The far most common derivational relation is the one where both the base and the derivate are nouns (42%). The maximum oracle score for a set of derivational relations was calculated to be 87.3%.

4 Results

The morpheme-discursive probe of the Russian political media reporting for each year of the time period 2012–2020 is presented in table 1. Among the most prominent prefixes, one can observe some key prefixes that are consistently high-ranked as well as prefixes that show large diachronic changes. Three prefixes, *eks-* ‘ex-’, *vnutri-* ‘intra-’ and *vice-* ‘vice-’, have a Log Ratio value of more than 2 for every year. They can be referred to as the most prominent prefixes in the discourse for the whole time period. The international prefix *giper-* ‘hyper-’ shows the largest increase (Spirkin et al., 1982, 129; Ryazanova-Clarke and Wade, 1999, 197–198). Between 2017 and 2018, Log Ratio increased from 2.08 to 4.82 and the prefix took the top place for the years 2018 and 2019. Its native counterpart *sverch-* ‘over-, super-’ is also activated in the second half of the time period (Ryazanova-Clarke and Wade, 1999, 194–196). The prefix *ul’tra-* ‘ultra-’ displays a growing trend during the time period and lands on a Log Ratio value of 2.77 in 2020. The prefix *trans-* ‘trans-’ is another prefix that has gained an increase during the time period and peaks in 2016 with a Log Ratio of 3.33. The prefixes *anti-* ‘anti-’, *kontr-* ‘counter-’ and *mež-* ‘inter-’ decline throughout the investigated time frame.⁵

The study confirms several findings in the literature. The nominal prefixes are more important in the Russian political reporting than the verbal prefixes and the international prefixes are more prominent than the prefixes of Slavic origin, for instance the international derivational prefixes such as *anti-* ‘anti-’ (Spirkin et al., 1982, 41), *eks-* ‘ex-’ (Spirkin et al., 1982, 574), *giper-* ‘hyper-’ (Spirkin et al., 1982, 129),

⁴<https://ruscorpora.ru/>

⁵Transliteration according to Scando-Slavica is used (<https://www.tandfonline.com/journals/ssla20>).

kontr- ‘counter-’ (Spirkin et al., 1982, 249), *sub-* ‘sub-’ (Spirkin et al., 1982, 477), *trans-* ‘trans-’ (Spirkin et al., 1982, 502), *ul’tra-* ‘ultra-’ (Spirkin et al., 1982, 513), *vice-* ‘vice-’ (Spirkin et al., 1982, 104) are more dominant than the inherited derivational prefixes like *mež-* ‘inter-’ (Vasmer, 1955, 112; Gerd, 2008, 10; Kuznecov, 1998, 529) and *vne-* ‘extra-’ (Vasmer, 1953, 210; Gorbačevič, 2005, 641; Kuznecov, 1998, 137) and those in turn are more salient than prefixes typically associated with verbs like *po-* ‘a little’, *s-* ‘together, down’, *pro-* ‘through’ (Janda and Ljashevskaya, 2013), *vy-* ‘out of’ (Ožegov and Švedova, 1999, 108–109), *voz-* ‘up’ (Ožegov and Švedova, 1999, 108–109) and so on. An exception from this pattern is the salient nominal prefix *vnutri-* ‘intra-’ of Slavic origin (Ožegov and Švedova, 1999, 88; Vasmer, 1953, 211).

Year	Prefix	LR	Year	Prefix	LR	Year	Prefix	LR
2012	<i>eks-</i> ‘ex-’	5.23	2013	<i>eks-</i> ‘ex-’	5.07	2014	<i>eks-</i> ‘ex-’	4.89
	<i>vnutri-</i> ‘intra-’	4.07		<i>vnutri-</i> ‘intra-’	3.76		<i>vnutri-</i> ‘intra-’	4.14
	<i>vice-</i> ‘vice-’	3.05		<i>vice-</i> ‘vice-’	2.64		<i>vice-</i> ‘vice-’	3.26
	<i>mež-</i> ‘inter-’	2.22		<i>mež-</i> ‘inter-’	2.30		<i>sub-</i> ‘sub-’	2.64
	<i>anti-</i> ‘anti-’	1.68		<i>sub-</i> ‘sub-’	1.96		<i>anti-</i> ‘anti-’	2.05
	<i>sub-</i> ‘sub-’	1.52		<i>vne-</i> ‘extra-’	1.87		<i>trans-</i> ‘trans-’	2.03
	<i>kontr-</i> ‘counter-’	1.44		<i>anti-</i> ‘anti-’	1.71		<i>vne-</i> ‘extra-’	2.02
	<i>protivo-</i> ‘counter-’	1.43		<i>nedo-</i> ‘under-’	1.69		<i>protivo-</i> ‘counter-’	1.94
	<i>nedo-</i> ‘under-’	1.38		<i>trans-</i> ‘trans-’	1.67		<i>mež-</i> ‘inter-’	1.70
	<i>vne-</i> ‘extra-’	1.35		<i>kontr-</i> ‘counter-’	1.62		<i>kontr-</i> ‘counter-’	1.67
2015	<i>eks-</i> ‘ex-’	4.58	2016	<i>eks-</i> ‘ex-’	5.17	2017	<i>eks-</i> ‘ex-’	5.41
	<i>vnutri-</i> ‘intra-’	4.23		<i>vnutri-</i> ‘intra-’	4.43		<i>vnutri-</i> ‘intra-’	4.69
	<i>vice-</i> ‘vice-’	3.10		<i>vice-</i> ‘vice-’	4.00		<i>vice-</i> ‘vice-’	3.42
	<i>trans-</i> ‘trans-’	2.46		<i>trans-</i> ‘trans-’	3.33		<i>trans-</i> ‘trans-’	2.80
	<i>sub-</i> ‘sub-’	2.40		<i>sub-</i> ‘sub-’	2.50		<i>sub-</i> ‘sub-’	2.67
	<i>anti-</i> ‘anti-’	1.98		<i>anti-</i> ‘anti-’	2.02		<i>giper-</i> ‘hyper-’	2.08
	<i>obez-</i> ‘dis-’	1.86		<i>giper-</i> ‘hyper-’	1.88		<i>anti-</i> ‘anti-’	2.04
	<i>mež-</i> ‘inter-’	1.80		<i>sverch-</i> ‘over-’	1.77		<i>ul’tra-</i> ‘ultra-’	2.01
	<i>vne-</i> ‘extra-’	1.75		<i>ul’tra-</i> ‘ultra-’	1.74		<i>mež-</i> ‘inter-’	1.84
	<i>kontr-</i> ‘counter-’	1.67		<i>mež-</i> ‘inter-’	1.59		<i>sverch-</i> ‘over-’	1.68
2018	<i>giper-</i> ‘hyper-’	4.82	2019	<i>giper-</i> ‘hyper-’	5.12	2020	<i>eks-</i> ‘ex-’	5.02
	<i>eks-</i> ‘ex-’	4.77		<i>eks-</i> ‘ex-’	4.71		<i>giper-</i> ‘hyper-’	4.83
	<i>vnutri-</i> ‘intra-’	4.39		<i>vice-</i> ‘vice-’	4.23		<i>vnutri-</i> ‘intra-’	3.98
	<i>vice-</i> ‘vice-’	3.54		<i>vnutri-</i> ‘intra-’	4.22		<i>ul’tra-</i> ‘ultra-’	2.77
	<i>trans-</i> ‘trans-’	2.93		<i>sub-</i> ‘sub-’	2.51		<i>trans-</i> ‘trans-’	2.49
	<i>ul’tra-</i> ‘ultra-’	2.41		<i>ul’tra-</i> ‘ultra-’	2.38		<i>vice-</i> ‘vice-’	2.42
	<i>sub-</i> ‘sub-’	2.34		<i>trans-</i> ‘trans-’	2.14		<i>sub-</i> ‘sub-’	1.90
	<i>sverch-</i> ‘over-’	1.91		<i>mež-</i> ‘inter-’	1.76		<i>sverch-</i> ‘over-’	1.73
	<i>mež-</i> ‘inter-’	1.53		<i>sverch-</i> ‘over-’	1.63		<i>protivo-</i> ‘counter-’	1.65
	<i>protivo-</i> ‘counter-’	1.50		<i>vne-</i> ‘extra-’	1.58		<i>vne-</i> ‘extra-’	1.40

Table 1: The top-10 Log Ratio (LR) values of the prefixes in the corpus of Russian political news for each year in the time period 2012–2020

The Log Ratio values of the derivational prefixes point to a number of discourse properties in the Russian political news reporting. The prefixes *eks-* ‘ex-’, *vice-* ‘vice-’ and *vnutri-* ‘intra-’ belong to the most prominent prefixes in Russian political reporting. The most distinguishing prefix in the reporting is *eks-* ‘ex-’ with the meaning of former, for example:

- (1) *S drugoj storony, Belyj dom ne spešit so vtorym paketom sankcij protiv Moskvy v svjazi s otravle-*

*niem **ěks-polkovnika** GRU Sergeja Skripalja i ego dočeri Julii v britanskom Solsberi.* (https://www.gazeta.ru/politics/2019/05/16_a_12357661.shtml)

‘On the other side, the White House is in no hurry with a second package of sanctions against Moscow due to the poisoning of the **ex-colonel** of GRU Sergei Skripal and his daughter Yulia in the British Salisbury.’

Also salient in the political media discourse is the prefix *vice-* ‘vice-’ with a most likely focus on political deputies:

- (2) *Otmetim, čto ranee **vice-prem’er** Rossii Dmitrij Rogozin zajavil o tom, čto serijnoe proizvodstvo novjšego rossijskogo tanka T-14 "Armata" mozet načat’sja v 2019 godu.* (<https://rg.ru/2017/06/21/rossiia-ne-budet-postavliat-za-rubezh-tank-armata-i-sistemu-s-500.html>)

‘Let us note that the **Deputy Prime Minister** of Russia Dmitry Rogozin earlier said that the serial production of the newest Russian tank T-14 “Armata” could begin in 2019.’

The prominence of the prefix *vnutri-* ‘intra-’ suggests a focus on internal relations, for example when they correlate with Russia’s ambitions to weaken the Western countries from within:

- (3) *Pri ètom Putin v èkskljuzivnom interv’ju avstrijskomu telekanalu ORF zajavil, čto ego vstreča s Trampom do sich por ne sostojalas’ iz-za ožestočennoj **vnutripolitičeskoj** bor’by v SŠA.* (<https://www.vesti.ru/doc.html?id=3024574>)

‘Putin said in an exclusive interview with the Austrian TV channel ORF that his meeting with Trump had not yet taken place due to the fierce **intra-political** struggle in the USA.’

The prefix *giper-* ‘hyper-’ shows the strongest increase of all the prefixes, especially from 2017 to 2018. In parallel, the prefix *sverch-* ‘over-, super-’ with a similar meaning also increased its activity. They are associated with the militarization of Russia reflected in the political discourse, especially the development and introduction of powerful weapons in the Russian military arsenal:

- (4) *Putin upomjanul o rakete s jadernoj ènergoustanovkoj, okeanskoj sisteme s bespilotnymi podlodkami na jadernoj ustanovke i **giperzvukovyh** raketach "Kinžal".* (<https://www.newsru.com/russia/23mar2018/kremlinglad.html>)

‘Putin mentioned a nuclear-powered missile, an oceanic system with nuclear-powered unmanned underwater vehicles and the **hypersonic** missiles “Kinzhal”.’

- (5) *Uničtožat’ protivotankovyje orudija, bronetekniku i betonnye doty protivnika "Terminator" sposoben s pomošč’ju **sverchzvukovyh** raket "Ataka-T".* (<https://rg.ru/2017/09/07/rossijskie-terminatory-pokorili-voennyh-sirii-i-izrailia.html>)

‘The “Terminator” is capable of destroying anti-tank guns, armored vehicles and enemy concrete pillboxes with the help of the **supersonic** missiles “Ataka-T”.’

The prefix *ul’tra-* ‘ultra-’ is often used with bases denoting political orientations like left, right, liberal and nationalist, for instance referring to events in the West:

- (6) *V poslednie mesjacy v Germanii usililis’ **ul’trapravye** nastroenija.* (https://www.gazeta.ru/politics/2018/11/30_a_12078325.shtml)

‘**Ultra-right** sentiments have intensified in Germany in recent months.’

The prominence of the prefix *trans-* ‘trans-’ is connected to the crossing of geographical spaces, for example in contexts where Kremlin’s intention of splitting the West is expressed:

- (7) *I v Kieve, i v Brjussele, i v Vašingtone, i v absoljutnom bol’šinstve zapadnyh stolic net nikakich somnenij v podgotovke Kremlëm masštabnyh vmešatel’stv v izbiratel’nye processy, čtoby slomat’*

*opasnoe dlja agressora evropejskoe edinstvo i solidarnost' s Ukrainoj, vnesti raskol v **transatlantičeskij** al'jans i podderžat' populistskie, nacionalističeskie i evroskeptično-političeskie sily.* (<https://life.ru/p/1143537>)

'Both in Kiev and in Brussels and in Washington and in the vast majority of Western capitals, there is no doubt that the Kremlin is preparing large-scale interventions in the electoral processes in order to break the dangerous European unity and the solidarity with Ukraine, which is dangerous for the aggressor, to cause a split in the **transatlantic** alliance and to support populist, nationalist and Eurosceptic political forces.'

The prefixes *anti-* 'anti-' and *kontr-* 'counter-' with the meaning of opposition show a decreasing trend throughout the time period. It suggests less polarization in the reporting over time. In the beginning of the studied time period, Putin's return to the Kremlin caused street protests in Russia:

- (8) *Aktivisty prokremlevskich dviženij razdavali vsem želajuščim georgievskie lenty i kričali "Putin ljubit vsech!", a protivniki izbrannogo prezidenta skandirovali **antiputinskie** lozungi.* (<https://utro.ru/articles/2012/05/07/1045258.shtml>)

'Activists of the pro-Kremlin movements handed out St. George ribbons to everyone who wished it and shouted "Putin loves everyone!", but the opponents of the elected president chanted **anti-Putin** slogans.'

The early 2010s is also the time of armed conflict in the Caucasus:

- (9) *Obstrel školy proižošel v chode **kontrterrorističeskoj** operacii, kotoraja provoditsja v Bujnaskom rajone s 5 sentjabrja.* (https://www.vedomosti.ru/politics/news/2012/09/14/skr_shkoluinternat_v_bujnakske_obstrelyali_iz_minometa_po)

'The shelling of the school occurred in the course of an **anti-terrorist** operation that has been underway in the Buynaksky district since September 5.'

The prefix *mež-* 'inter-' also displays a decreasing keyness trend during the actual time frame and that points to less interconnectedness between different political forces as in this example from the beginning of the time period:

- (10) *Usiliya meždunarodnogo soobščestva dolžny byt' napravleny prežde vsego na dostiženie **mežsirijskogo** primirenija.* (<https://rg.ru/2012/02/27/putin-politika.html>)

'The efforts of the international community should be directed first and foremost towards achieving **inter-Syrian** reconciliation.'

5 Discussion

In this preliminary work, an approach to probe a discourse by using prefixal keymorphs in combination with a derivational resource within the framework of Corpus-assisted discourse studies has been described. The results suggest that a derivational keymorph analysis has the potential to reveal general properties of a discourse. A refinement of the methods awaits further research.

Acknowledgments

I thank Tora Hedin for proofreading this paper and the anonymous reviewers for their comments and suggestions on important aspects of the research.

References

- Annelie Ädel. 2010. How to use corpus linguistics in the study of political discourse. In Anne O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, Routledge, chapter 42, pages 591–604.
- Mohamad Ali Afida. 2007. *Semantic fields of problem in business English: Malaysian and British journalistic business texts.* *Corpora* 2(2):211–239. <https://doi.org/https://doi.org/10.3366/cor.2007.2.2.211>.

- Dawn Archer, Jonathan Culpeper, and Paul Rayson. 2009. *Love – ‘a familiar or a devil’? an exploration of key domains in Shakespeare’s comedies and tragedies*. In Dawn Archer, editor, *What’s in a Word-list?: Investigating Word Frequency and Keyword Extraction*, Routledge, London, pages 137–157. <https://doi.org/https://doi.org/10.4324/9781315547411>.
- Vladimír Benko. 2014a. *Aranea: Yet another family of (comparable) web corpora*. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue: 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014*. Springer, Cham, volume 8655 of *Lecture Notes in Computer Science*, pages 247–254. <https://doi.org/https://doi.org/10.1007/978-3-319-10816-2>.
- Vladimír Benko. 2014b. Compatible sketch grammars for comparable corpora. In Andrea Abel, Chiara Vettori, and Nataschia Ralli, editors, *Proceedings of the 16th EURALEX International Congress*. EURAC research, Bolzano, Italy, pages 417–430.
- Vladimír Benko and Victor P. Zakharov. 2016. Very large Russian corpora: new opportunities and new challenges. In *Computational linguistics and intellectual technologies*, Rossijskij gosudarstvennyj humanitarnyj universitet, pages 79–93.
- Vladimír Benko and Victor P. Zakharov. 2021. *Crowdsourcing for the Russian morphological lexicon*. In Radomir V. Bolgov, Nikolay V. Borisov, Andrei V. Chugunov, Dmitry E. Prokudin, Alexander E. Voiskounsky, and Victor P. Zakharov, editors, *Proceedings of the International Conference “Internet and Modern Society” (IMS-2021)*. St. Petersburg, Russia, pages 111–119. <http://ceur-ws.org/Vol-3090/>.
- Ruth Breeze. 2019. *Part-of-speech patterns in legal genres: Text-internal dynamics from a corpus-based perspective*. In Teresa Fanego and Paula Rodríguez-Puente, editors, *Corpus-based research on variation in English legal discourse*. John Benjamins, volume 91 of *Studies in Corpus Linguistics*, pages 79–103. <https://doi.org/https://doi.org/10.1075/sci.91>.
- Jonathan Culpeper. 2009. *Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare’s Romeo and Juliet*. *International Journal of Corpus Linguistics* 14(1):29–59. <https://doi.org/https://doi.org/10.1075/ijcl.14.1.03cul>.
- Jonathan Culpeper and Jane Demmen. 2015. *Keywords*. In Douglas Biber and Randi Reppen, editors, *The Cambridge Handbook of English Corpus Linguistics*, Cambridge University Press, Cambridge Handbooks in Language and Linguistics, pages 90–105. <https://doi.org/10.1017/CBO9781139764377.006>.
- Václav Cvrček and Masako Fidler. 2019. *More than keywords: Discourse prominence analysis of the Russian web portal Sputnik Czech Republic*. In Martina Berrocal and Aleksandra Salamurovič, editors, *Political Discourse in Central, Eastern and Balkan Europe*, John Benjamins, volume 84 of *Discourse Approaches to Politics, Society and Culture*. <https://doi.org/https://doi.org/10.1075/dapsac.84.05cvr>.
- Dagmar Divjak and Laura A. Janda. 2008. *Ways of attenuating agency in Russian*. *Transactions of the Philological Society* 106(2):138–179. <https://doi.org/https://doi.org/10.1111/j.1467-968X.2008.00207.x>.
- Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*. Linköping University Electronic Press, 155, pages 52–65.
- Pavel V. Dyachenko, Leonid L. Iomdin, Aleksandr V. Lazurskij, Leonid G. Mityushin, Olga Ju. Podlesskaya, Viktor G. Sizov, Tatyana I. Frolova, and Leonid L. Tsinman. 2015. *Sovremennoe sostojanie gluboko anotirovannogo korpusa tekstov russkogo jazyka (sintagrus)*. *Trudy Instituta russkogo jazyka im. V.V. Vinogradova* 6:272–300.
- Masako Fidler and Václav Cvrček. 2015. *A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis*. *Journal of Slavic Linguistics* 23(2):197–239. <https://www.jstor.org/stable/24602151>.
- Masako Fidler and Václav Cvrček. 2018. *Going beyond “aboutness”: A quantitative analysis of Sputnik Czech Republic*. In Masako Fidler and Václav Cvrček, editors, *Taming the Corpus: From Inflection and Lexis to Interpretation*, Springer, Quantitative Methods in the Humanities and Social Sciences, chapter 10. <https://doi.org/https://doi.org/10.1007/978-3-319-98017-1>.
- Masako Fidler and Václav Cvrček. 2019. *Keymorph analysis, or how morphosyntax informs discourse*. *Corpus Linguistics and Linguistic Theory* 15(1):39–70. <https://doi.org/doi:10.1515/cllt-2016-0073>.

- Costas Gabriellatos. 2018. *Keyness analysis: nature, metrics and techniques*. In Charlotte Taylor and Anna Marchi, editors, *Corpus Approaches to Discourse: A Critical Review*, Routledge, London, pages 225–258. 1 edition. <https://doi.org/https://doi-org.ezp.sub.su.se/10.4324/9781315179346>.
- Aleksandr Sergeevič Gerd, editor. 2008. *Bol'shoj akademičeskij slovar' russkogo jazyka*, volume 10. Nauka.
- Kirill Sergeevič Gorbačevič, editor. 2005. *Bol'shoj akademičeskij slovar' russkogo jazyka*, volume 2. Nauka.
- Andrew Hardie. 2014. *Log ratio: An informal introduction*. <http://cass.lancs.ac.uk/?p=1133>.
- Laura Alexis Janda, Masako Fidler, Václav Cvrček, and Anna Obukhova. 2023. *The case for case in Putin's speeches*. *Russian Linguistics* 47(1):15–40. <https://doi.org/10.1007/s11185-022-09269-2>.
- Laura Alexis Janda and Olga Ljashevskaya. 2013. *Semantic profiles of five Russian prefixes: po-, s-, za-, na-, pro-*. *Journal of Slavic Linguistics* 21(2):211–258. <https://doi.org/doi:10.1353/jsl.2013.0012>.
- Elena I. Koriakowcewa. 2009. *Internacional'noe vs. nacional'noe v slovoobrazovatel'noj sisteme: k postanovke voprosa*. In Elena Koriakowcewa, editor, *Przejawy internacjonalizacji w językach słowiańskich*, Wydawnictwo Akademii Podlaskiej, Siedlce, pages 179–198.
- Evgeny Kotelnikov, Elena Razova, and Irina Fishcheva. 2017. *A close look at Russian morphological parsers: Which one is the best?* In Andrey Filchenkov, Lidia Pivovarova, and Jan Žižka, editors, *Conference on Artificial Intelligence and Natural Language*. Springer International Publishing, Cham, pages 131–142. https://doi.org/https://doi.org/10.1007/978-3-319-71746-3_12.
- Elizaveta Kuzmenko. 2016. *Morphological analysis for Russian: integration and comparison of taggers*. In Dmitry I. Ignatov, Mikhail Yu. Khachay, Valeri G. Labunets, Natalia Loukachevitch, Sergey I. Nikolenko, Alexander Panchenko, Andrey V. Savchenko, and Konstantin Vorontsov, editors, *Proceedings of 5th International Conference on Analysis of Images, Social Networks and Texts*. Springer Cham, Communications in Computer and Information Science, pages 162–171. <https://doi.org/https://doi.org/10.1007/978-3-319-52920-2>.
- Sergej Aleksandrovič Kuznecov. 1998. *Bol'shoj tolkovyj slovar' russkogo jazyka*. Norint, Saint Petersburg.
- Lukáš Kyjánek, Olga Ljashevskaya, Anna Nedoluzhko, Daniil Vodolazsky, and Zdeněk Žabokrtský. 2022. *Constructing a lexical resource of Russian derivational morphology*. In *Proceedings of the 13th Conference on Language Resources and Evaluation Conference (LREC)*. European Language Resources Association, Marseille, France, pages 2788–2797. <https://aclanthology.org/2022.lrec-1.298>.
- Medialogia. 2023. <https://www.mlg.ru>.
- Sergej Ivanovič Ožegov and Natalija Jul'evna Švedova. 1999. *Tolkovyj slovar' russkogo jazyka*. Azbukovnik, Moscow.
- Alan Partington and Alison Duguid. 2020. *Political media discourses*. In Eric Friginal and Jack A. Hardy, editors, *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, Routledge, chapter 8, pages 116–135. 1 edition.
- Alan Partington, Alison Duguid, and Charlotte Taylor. 2013. *Patterns and meanings in discourse: theory and practice in corpus-assisted discourse studies (CADS)*. Number 55 in *Studies in corpus linguistics*. John Benjamins, Amsterdam/Philadelphia.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. *Stanza: A Python natural language processing toolkit for many human languages*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pages 101–108. <https://doi.org/https://doi.org/10.18653/v1/2020.acl-demos.14>.
- Larisa V. Ratsiburskaya, Nadezhda A. Samylicheva, and Anna V. Shumilova. 2015. *Novye tendencii v sovremenom medijnom slovotvorčestve*. In Larisa V. Ratsiburskaya, editor, *Novye tendencii v russkom jazyke načala XXI veka*, FLINTA, Moscow, chapter 3, pages 134–221. 2 edition.
- Paul Rayson. 2004. *Keywords are not enough*. Invited talk for JAECS (Japan Association for English Corpus Studies) at Chuo University, Tokyo, Japan. https://www.lancaster.ac.uk/staff/rayson/publications/jaecs_tokyo04.pdf.
- Paul Rayson. 2008. *From key words to key semantic domains*. *International journal of corpus linguistics* 13(4):519–549. <https://doi.org/https://doi.org/10.1075/ijcl.13.4.06ray>.
- Larissa Ryazanova-Clarke and Terence Wade. 1999. *The Russian language today*. Routledge.

- Pavel Rychlý. 2007. Manatee/Bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Masaryk University, Brno, pages 65–70.
- Nicholas Smith and Cathleen Waters. 2019. [Variation and change in a specialized register: a comparison of random and sociolinguistic sampling outcomes in desert island discs](https://doi.org/https://doi.org/10.1075/ijcl.17117.smi). *International Journal of Corpus Linguistics* 24(2):169–201. <https://doi.org/https://doi.org/10.1075/ijcl.17117.smi>.
- Aleksandr Georgievič Spirkin, Igor' Alekseevič Akčurin, and Regina Semenovna Karpinskaja. 1982. *Slovar' inostrannyh slov*. Russkij jazyk, Moscow, 9 edition.
- Jürgen Spitzmüller and Ingo H. Warnke. 2011. [Discourse as a 'linguistic object': methodical and methodological delimitations](https://doi.org/10.1080/17405904.2011.558680). *Critical Discourse Studies* 8(2):75–94. <https://doi.org/10.1080/17405904.2011.558680>.
- Max Vasmer. 1953. *Russisches etymologisches Wörterbuch*, volume 1. Carl Winter, Heidelberg.
- Max Vasmer. 1955. *Russisches etymologisches Wörterbuch*, volume 2. Carl Winter, Heidelberg.
- Elena Zemskaya. 2006. Aktivnye processy v russkom slovoobrazovanii našego vremeni. *Acta Neophilologica* VIII:9–21.

Croatian Language Technologies Society
Faculty of Humanities and Social Sciences,
University of Zagreb
HR-CLARIN
Zagreb, Croatia

<https://derimo.ffzg.unizg.hr/>

ISBN 978-953-55375-5-7

