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

DERIMO 2025

Fribourg, Switzerland

September 4–5, 2025

EDITED BY

Richard Huyghe
Megan Prudent
Justine Salvadori

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Fonds national suisse

DeriMo 2025

Proceedings of the

Fifth International Workshop on Resources and Tools for Derivational Morphology

Editors

Richard Huyghe

Megan Prudent

Justine Salvadori

4–5 September 2025

University of Fribourg, Switzerland

https://events.unifr.ch/derimo2025/en/

Published by:

University of Fribourg,
Avenue de l'Europe 20
CH-1700 Fribourg
Switzerland

ISBN 978-2-8399-4786-2

Preface

This volume brings together the papers accepted for presentation at *DeriMo 2025: The Fifth International Workshop on Resources and Tools for Derivational Morphology*, held in Fribourg, Switzerland, on 4-5 September 2025. The 2025 edition of the workshop continues the series of discussions on language resources and tools for research on word formation, initiated at DeriMo 2017 (Milan, Italy) and carried forward at DeriMo 2019 (Prague, Czechia), DeriMo 2021 (Nancy, France), and DeriMo 2023 (Dubrovnik, Croatia).

The proceedings comprise 12 papers selected through a thorough reviewing process, with each submission evaluated by three program committee members. They also include the contributions of two invited speakers, Magda Ševčíková and Sabine Arndt-Lappe. We gratefully acknowledge the financial support of the University of Fribourg and the Swiss National Science Foundation for the organization of the workshop.

Richard Huyghe
Megan Prudent
Justine Salvadori

Program Committee Chairs

Richard Huyghe	University of Fribourg (Switzerland)
Megan Prudent	University of Fribourg (Switzerland)
Justine Salvadori	University of Fribourg (Switzerland)
Matea Filko	FFZG, University of Zagreb (Croatia)
Krešimir Šojat	FFZG, University of Zagreb (Croatia)

Program Committee Members

Alexandra Bagasheva	Sofia University St. Kliment Ohridski (Bulgaria)
Olivier Bonami	Université Paris Cité (France)
Maria Copot	Surrey Morphology Group, University of Surrey (England)
Cristina Fernández-Alcaina	Charles University (Czech Republic)
Jesús Fernández-Dominguez	University of Granada (Spain)
Livio Gaeta	University of Turin (Italy)
Nabil Hathout	Université Toulouse - Jean Jaurès (France)
Martin Hilpert	University of Neuchâtel (Switzerland)
Gianina Iordăchioaia	University of Graz (Austria)
Lívia Körtvélyessy	Pavol Jozef Šafárik University (Slovakia)
Eleonora Litta	Università Cattolica del Sacro Cuore (Italy)
Claudia Marzi	Institute for Computational Linguistics (Italy)
Fabio Montermini	Université Toulouse - Jean Jaurès (France)
Akiko Nagano	University of Shizuoka (Japan)
Fiammetta Namer	Université de Lorraine (France)
Renáta Panocová	Pavol Jozef Šafárik University (Slovakia)
Marco Passarotti	Università Cattolica del Sacro Cuore (Italy)
Ingo Plag	Heinrich-Heine-Universität Düsseldorf (Germany)
Jan Radimský	University of South Bohemia (Czech Republic)
Andrea Sims	The Ohio State University (USA)
Pavol Štekauer	Pavol Jozef Šafárik University (Slovakia)
Pavel Štichauer	Charles University (Czech Republic)
Marko Tadić	University of Zagreb (Croatia)
Salvador Valera Hernández	University of Granada (Spain)
Zdeněk Žabokrtský	Charles University (Czech Republic)

Local Organizing Committee

Richard Huyghe	University of Fribourg (Switzerland)
Megan Prudent	University of Fribourg (Switzerland)
Justine Salvadori	University of Fribourg (Switzerland)
Raphaël Cornaz	University of Fribourg (Switzerland)

Contents

Non-native roots in derivational morphology – any specifics?

Magda Ševčíková
Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
sevcikova@ufal.mff.cuni.cz

Abstract

Words borrowed from other languages are integrated into the morphological system of the recipient language by adopting inflectional markers and participating in derivation and other word-formation processes. By comparing 400 morphological families with non-native roots against the inherited vocabulary in Czech, this paper demonstrates that despite numerous commonalities, families with non-native roots differ from those with inherited roots notably in the role played by verbs. Non-native verbs resemble Czech denominal verbs and are typically dispreferred in favor of nouns, even when conveying action meanings. This tendency towards nominal expression, documented in the non-native vocabulary through qualitative and quantitative features, provides derivational evidence in support of the difficulty of borrowing verbs as verbs, which has been pointed out in linguistic typology.

1 Introduction

Lexical borrowing and word-formation are intimately intertwined in languages, making it difficult or even impossible to separate them sharply (ten Hacken and Panocová, 2020). In Czech, nominal loanwords exhibit various degrees of morphological adaptation, while borrowing of verbs implies the adoption of a thematic suffix and inflectional markers (cf. the nouns and verbs in 1 and 2). This means that a full integration into the conjugation system must be completed before a word is used as a verb. Since this integration employs native verbal markers that are also used in the formation of verbs in Czech, verbs borrowed from other languages do not differ from those derived directly in Czech (cf. Mravinacová (2005) on Czech and Wohlgemuth (2009) for similar cases in other languages).

(1) *start* – *start-ova-t*
start(M)[1] start-IPFV-INF
'start' 'to start'

(2) *koket-a* – *koket-ova-t*
coquet(F)-NOM.SG coquet-IPFV-INF
'coquette' 'to coquet'

Adhering to the principles of the study of foreign word-formation, as outlined, for instance, by Eins (2015), the present paper treats words with non-native (also, loan or foreign) roots[2] as an integral part of the Czech lexicon. The analysis is confined to a strictly synchronic perspective, examining the formal and semantic features of the relevant words and excluding diachronic and etymological factors. This makes it possible to avoid the difficulties encountered in previous descriptions of borrowed vocabulary, such as

[1] Gender (M for masculine, F for feminine, and N for neuter) is indicated in brackets immediately after the root as an inherent category of nouns in Czech. The following abbreviations also appear in the glosses: INF infinitive, IPFV imperfective (aspect), NOM nominative, PFV perfective (aspect), PREF prefix, SG singular.

[2] I speak of non-native roots (and, later, also suffixes), although – as pointed out by Dietz (2015, p. 1642), Waszakowa (2015, p. 1681) and others – morphemes are not borrowed separately, but come into a language as parts of loanwords and are perceived as such only subsequently, usually after a larger set of words with the particular element has been integrated into the language.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 1–10, Fribourg, Switzerland, 4-5 September 2025.

determining opposite directions of derivation in pairs of words with analogous structure and meaning based on different dates of first occurrence (Martincová, 2005). In line with this approach, the paper aims to address the following questions:

- Do noun/verb pairs with non-native roots differ in their properties from native pairs with analogous structures?
- What role do verbs play within morphological families containing non-native roots?
- Is the integration of non-native vocabulary into Czech, in the aspects under investigation, distinct from that in other languages?

The paper is structured as follows. Section 2 describes the compilation of the dataset on which the present study is based. In the data, the word-formation relations within 400 partial morphological families with non-native roots are established along the same lines as in families with inherited (native) roots. A comparison of morphological families with non-native and inherited roots in Section 3 shows that non-native verbs share features with native denominal verbs. This fits in with the cross-linguistic and typological debates, where, as briefly summarized in Section 4, it has been pointed out that verbs often come into other languages as nouns and are only subsequently turned into verbs. In this paper, however, I go further and show the denominal behavior of verbs with non-native roots as part of a broader tendency to prefer nominals over verbs in this part of the Czech vocabulary. Some summarizing remarks close the paper in Section 5.

2 Compilation of the dataset

2.1 Extraction of noun/verb pairs

The study is based on a dataset of more than 2,000 partial morphological families extracted from the SYN2015 corpus of written Czech, which contains 100 million words (Křen et al., 2015). The core of the dataset are pairs of suffixless nouns and corresponding verbs that share the root and possibly the prefix with the noun, but do not contain any derivational suffix or any other prefix. Suffixless nouns both without overt inflectional endings in the citation forms (cf. the bare-root noun *start* in 1) and with overt endings in the citation forms (cf. *koketa* with the *-a* ending in 2) are included in the data.

The relationship between the noun and the verb in individual pairs can be seen as the word-formation process of conversion (Ševčíková, in press; Manova, 2011). However, unlike the canonical conversion in English (Valera, 2015), this process in Czech does not occur without formal changes to the citation forms. A characteristic feature is the replacement of nominal inflection by verbal markers when a noun is converted into a verb, and conversely, the substitution of verbal markers with nominal ones when the direction is reversed. The noun/verb pairs were automatically extracted using the *Morfio* tool (Cvrček and Vondřička, 2013) and manually verified to ensure that the conversion pairs are linked by a synchronically available relationship. The final dataset contains 2,058 noun/verb pairs.

2.2 Annotation of the semantic relationship between the noun and the verb

The semantic change, which is brought about through the conversion, was determined based on sentences extracted from the SYN2015 corpus, which document the actual use of the noun and verb. For each conversion pair, a random sample of 50 sentences containing the noun and the same-size sample for the verb was analyzed. If the pair members were documented with less than 50 hits in the corpus, all respective sentences were taken for the annotation.

The relation between the noun and the verb, as evidenced in the sentence contexts, was manually tagged using a set of 9 semantic categories, which was compiled on the basis of existing discussions on the semantics of conversion in English (in particular, Cetnarowska, 1993, Plag, 1999, and Bauer et al., 2013) and is presented in Table 1. To avoid the issue of determining the direction of conversion, which, due to the absence of overt derivational markers in the Czech data, presents the same challenges as in English, the categories discussed separately for the denominal and deverbal direction in the previous

approaches were merged into a single category, which was then attributed to the noun in the conversion pair.[3]

SEMANTIC CATEGORY (Meaning of the noun wrt the verb)	Example conversion pair
ACTION (noun = action of V-ing)	*atak* 'attack' – *atakovat* 'to attack'
AGENT (noun = someone who performs V-ing)	*rebel* 'rebel' – *rebelovat* 'to rebel'
INSTRUMENT (noun = something used for V-ing)	*telefon* 'phone' – *telefonovat* 'to call'
OBJ/QUAL-ADDED (noun = something added through V-ing)	*zinek* 'zinc' – *zinkovat* 'to coat with zinc'
OBJ/QUAL-REMOVED (noun = something removed through V-ing)	*skalp* 'scalp' – *skalpovat* 'to scalp'
PLACE (noun = a place where something is V-ed)	*garáž* 'garage' – *garážovat* 'to garage'
RESULT (noun = result of V-ing)	*kompost* 'compost' – *kompostovat* 'to compost'
STATE (noun = the state of being V-ed)	*šok* 'shock' – *šokovat* 'to shock'
TIME (noun = the time spent V-ing)	*noc* 'night' – *nocovat* 'to stay the night'

Table 1: Nine categories capturing the semantic relationship between the suffixless noun and the corresponding verb. Each category is followed by a gloss of the noun's meaning with respect to the verb and by an example of noun/verb pair.

For 800 of the conversion pairs, semantic annotation was carried out by two annotators in parallel in order to assess whether the set of semantic categories has an appropriate level of granularity and to evaluate the inter-annotator agreement (inter-rater reliability). The agreement was calculated in percentage and by using Cohen's kappa coefficient, which is designed to consider the effect of chance agreement (Cohen, 1960). The annotators agreed on labels with 22,461 out of 30,277 sentences analyzed for the 800 conversion pairs, i.e., on 74.2 % of the annotated sentences. Expressed by Cohen's kappa, the inter-annotator agreement on this dataset reached the value 0.659, i.e., substantial agreement. The annotation of the remaining data was carried out by a single annotator. For more details on the annotation, see Ševčíková (in press).

A relationship between one sense of the noun and one sense of the verb was identified for 1,619 of 2,058 pairs analyzed, but two or three sense-sense relations were identified for the remaining 439 pairs (totaling 919 relations) – cf. *analýza* 'analysis' documented both as the ACTION and RESULT of the activity referred to by the verb *analyzovat* 'to analyze', or *sonda* 'probe' as the ACTION and

[3]For instance, Bauer et al.'s category RESULT, which is listed among the meanings of deverbal nouns (with the paraphrase "the outcome of V-ing" and the example *divorce* = "the result of divorcing"), and their category RESULTATIVE, which subsumes denominal verbs corresponding to the paraphrase "to make into X" (e.g., *to bundle* = "to make into bundles"), were merged into the category RESULT in the present analysis. Consequently, the meaning of the noun in both denominal and deverbal cases is rephrased in the same way, namely as a "result of V-ing" in the RESULT category (cf. Table 1), i.e., *divorce* = "the result of divorcing" and *bundle* = "the result of bundling". With the latter example, it has to be stressed that the gloss is not an explanation of the meaning, which is what the dictionary aims to do, but it specifies the role the noun plays with respect to the meaning of the verb.

INSTRUMENT used for the activity expressed by the respective verb (*sondovat* 'to probe') in (3) to (5).

(3) *Provedli jsme základní **sondy**.*ACTION *do podlahového souvrství v přízemí s **analýzou**.*ACTION *úrovně založení obvodových stěn.*
'We carried out basic **probes**.ACTION into the floor layer on the ground floor with **analysis**.ACTION of the foundation level of the perimeter walls.'

(4) *Čáp se ale brzo od konstruktérky Lidušky dozví, že takovou **analýzu**.*RESULT *už před časem vypracoval v podniku inženýr Křížek.*
'However, Čáp soon learns from Liduška, the engineer, that such an **analysis**.RESULT was produced by engineer Křížek some time ago.'

(5) *Tiskla mi **sondu**.*INSTRUMENT *na žaludek a pozorovala obrazovku ultrazvuku, která byla obrácená směrem k ní.*
'She pressed the **probe**.INSTRUMENT to my stomach and watched the ultrasound screen facing her.'

After processing the noun/verb pairs, additional members of the morphological families were added to the dataset, in particular, aspectual counterparts of the verbs, inflectional nominals and derivational nominals of the verbs. Each member of these partial morphological families was assigned the cumulative frequency of all forms of the respective lexeme (lemma frequency).

2.3 Identification of loanwords

As the last step, nouns and verbs with foreign roots were identified automatically by a pretrained classifier on the basis of the presence of particular graphemes or grapheme combinations felt as foreign. The results of the automatic annotation were checked manually. Words that have both formally and semantically close counterparts in multiple foreign languages (and thus comply with the definition of internationalisms; cf. Buzássyová, 2010, among others) were confirmed as non-native.

Based on this last annotation step, the sample was divided into 401 noun/verb pairs with non-native roots and 1,657 pairs with inherited roots. The dataset was published in the LINDAT/CLARIAH-CZ repository at `http://hdl.handle.net/11234/1-5142`.

3 Words with non-native roots in contrast with the inherited vocabulary

3.1 Verbs with non-native roots as denominals

Noun/verb pairs with non-native roots made up a smaller part of the data and they were present as a minor group in all **semantic categories**, except for the TIME category with no non-native instances at all. They also differed from the inherited subset in the extent of polysemy: A single sense-sense relation was determined with 350 out of 401 conversion pairs with non-native roots, while two or three sense-sense relations were assigned with 51, i.e., one eighth, of them; the average was 1.14 relations per pair. In the native part, 1,269 out of 1,657 conversion pairs were linked by a single sense-sense relation whereas 388, i.e., almost a quarter, with more than one relation (1.26 relations per pair on average). Table 2 lists the 9 semantic categories by the number of sense-sense relations separately for the non-native part and for the native part of the data. The categories OBJ/QUAL-ADDED and INSTRUMENT in the non-native data are larger at the expense of the pairs with ACTION and RESULT nouns, as compared to the inherited data.

A significant difference was in the expression of **grammatical aspect** in the verbs. Only 9 verbs out of 401 noun/verb pairs with non-native roots were capable of substituting their thematic suffixes for different ones to change the aspect (6). The verbs in the remaining non-native pairs did not allow substitution of the thematic suffix and instead added a prefix to mark aspectual change (7). In the native part, in contrast, verbs with thematic suffix substitution (1,082 verbs) outnumbered those which do not form aspectual counterparts by replacing the theme (575 verbs).

The strategy of aspectual change merits attention because, as shown in previous studies (Ševčíková, 2021a,b, 2022), the manner in which this grammatical category is expressed correlates with the direction of conversion between suffixless nouns and verbs in Czech. These studies, all limited to smaller samples with inherited roots, have revealed that compatibility with two different thematic suffixes is typical for

	Non-native		Inherited	
	# of paradigms	Percent	# of paradigms	Percent
ACTION	164	35.8	886	42.6
RESULT	110	24.0	596	28.7
INSTRUMENT	88	19.2	291	14.0
OBJ/QUAL-ADDED	55	12.0	105	5.0
AGENT	21	4.6	92	4.4
STATE	10	2.2	78	3.8
PLACE	9	2.0	23	1.1
OBJ/QUAL-REMOVED	1	0.2	7	0.3
TIME	0	0.0	2	0.1
Total	458	100.0	2,080	100.0

Table 2: Semantic categories ranked by the number of instances in the non-native and inherited part of the dataset.

verbs converting to suffixless nouns, while restriction to a single theme seems to be characteristic of verbs converted from nouns. If this observation is projected from native vocabulary onto the non-native data, this would indicate the almost exclusively denominal nature of the verbs in the pairs with non-native roots.

(6) *risk-ova-t* : *risk-nou-t*
risk-IPFV-INF risk-PFV-INF
'to risk (IPFV)' 'to risk (PFV)'

(7) *kontrol-ova-t* : *z-kontrol-ova-t*
control-IPFV-INF PREF-control-IPFV-INF
'to control (IPFV)' 'to control (PFV)'

The observation about the denominality of verbs in the non-native sample also finds support in **quantitative evidence**, whose utility for determining the direction in conversion was already pointed out by Marchand (1963, 1964) and discussed from various theoretical and empirical perspectives (cf. among others, Kisselew et al. 2016 and references therein). According to Marchand, when comparing the frequency of two overtly unmarked words, a higher frequency is indicative of a motivating (input) word because it is semantically broader and thus more usable than the motivated (output) word with a more specific meaning and thus more restricted use.

In the dataset under analysis, the comparison of the noun's lemma frequency count to the lemma frequency count of the corresponding verb (or of the more frequent verb if a noun corresponds to two verbs with different themes) in individual pairs with non-native roots shows that the verb is more frequent than the corresponding suffixless noun in only 35 out of 401 conversion pairs, while in the remaining 366 pairs the noun has a higher frequency than the verb. In the native part, on the other hand, the frequency of the verb was higher than the frequency of the noun in 870 pairs, and lower in 787 pairs.

For a more precise comparison, the distribution of noun-to-verb frequency ratios – where values greater than 1 correspond to pairs in which the noun outnumbers the verb, and values less than 1 to those with more frequent verbs – is shown in the box plots in Figure 1. In the non-native box plot (left), frequency ratio values are mainly distributed between 3 and 55, with the median at 12.27. In the native part of the data (box plot on the right), most values are located between 0.22 and 6.60, with the median slightly above 1. The box plots document that nouns tend to have a higher frequency than verbs in conversion pairs with non-native roots, or if taken from the speakers' perspective, that nouns are preferred over verbs in this segment of the lexicon. For the native sample, by contrast, no such clear picture emerges; both

pairs with the more frequent noun and pairs with the more frequent verb are present.

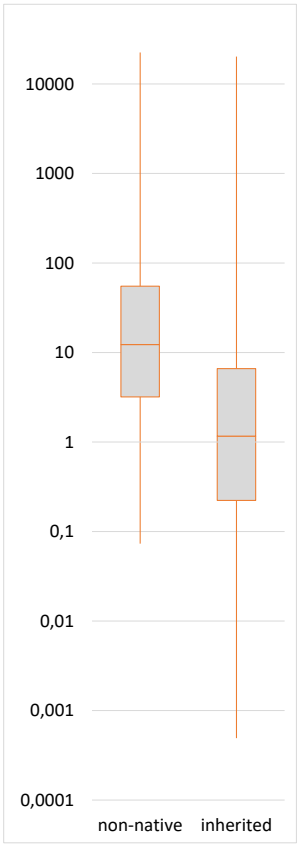


Figure 1: Distribution of frequency ratio values (calculated from the lemma frequency count of a suffixless noun divided by the lemma frequency count of the corresponding verb) in the non-native part vs. inherited part of the data. Y-axis is set to logarithmic scale.

3.2 Nouns preferred over verbs in morphological families with non-native roots

Even if the study is restricted to partial morphological families, the data support Waszakowa's (2003) observation that it is typical of internationalisms in West-Slavic languages that they share roots with many other words. Based on a dataset of 4,500 neologisms registered in Polish between 1985 and 2004, Waszakowa (2005) reports that about 70 % of the whole sample are motivated words, one-third of which are derivatives and two-thirds compounds, including neoclassical formations.

The data analyzed in this study document that verbs with non-native roots participate in **prefixation** similarly to native verbs. However, unlike the native vocabulary, these prefixed verbs do not enter conversion. Whereas in the native lexicon, unprefixed noun/verb pairs reoccur with a prefix in one or even several conversion pairs (cf. the noun *skok* from the pair in 8a reoccurring in the prefixed noun in 8b and with 12 other prefixes in the corpus data), the dataset with non-native roots includes only a single instance of this kind (9). In the native data, the repetition with prefixes appears with verbs that form their aspectual counterpart by substituting their thematic suffix and are interpreted as input to (deverbal) conversion. Noun/verb pairs with non-native roots, which in the vast majority of cases do not permit such repetition, thus behave – also in this respect – like native instances of denominal conversion.

(8) a. *skok* – *skák-a-t* : *skoč-i-t*
jump jump-IPFV-INF jump-PFV-INF
'jump' 'to jump (IPFV)' 'to jump (PFV)'

b. *vý-skok* – *vy-skak-ova-t* : *vy-skoč-i-t*
PREF-jump PREF-jump-IPFV-INF PREF-jump-PFV-INF
'upward jump' 'to jump up (IPFV)' 'to jump up (PFV)'

(9) a. *klik* – *klik-a-t* : *klik-nou-t*
click click-IPFV-INF click-PFV-INF
'click' 'to click (IPFV)' 'to click (PFV)'

b. *pro-klik* – *pro-klik-áva-t* : *pro-klik-nou-t*
PREF-click PREF-click-IPFV-INF PREF-click-PFV-INF
'click through' 'to click through (IPFV)' 'to click through (PFV)'

Similar to native vocabulary, **inflectional nominalizations** (ending in *-ní* in Czech and corresponding in many respects to *-ing* nominals in English) are attested for most verbs with non-native roots as well. What distinguishes the non-native data from the native, however, is the attestation of **derivational nominalizations**. These occur only rarely in morphological families with native roots, but are found in nearly 40 percent of the analyzed families with non-native roots. Moreover, in most of these cases, the derivational nominals exhibit a higher lemma frequency than the corresponding verbs, suggesting that speakers choose them more frequently than the verbs to refer to the respective actions. Table 3 provides examples of the conversion pairs and corresponding nominals with non-native suffixes along with their frequency counts from the SYN2015 corpus.

Noun	Lemma freq.	Verb	Lemma freq.	Nominalization	Lemma freq.
filtr	2,300	*filtrovat*	295	*filtrace*	434
'filter'		'to filter'		'filtration'	
extrakt	406	*extrahovat*	107	*extrakce*	190
'extract'		'to extract'		'extraction'	
archiv	4,235	*archivovat*	208	*archivace*	247
'archive'		'to archive'		'archiving'	
parfém	1,188	*parfémovat*	5	*parfemace*	22
'perfume'		'to perfume'		'perfuming'	
telefon	18,107	*telefonovat*	1,480	*telefonát*	1,271
'phone'		'to call'		'phone call'	

Table 3: Noun/verb conversion pairs and related derivational nominals (with lemma frequency counts extracted from the SYN2015 corpus).

It is also noteworthy that some of these nouns – despite combining a non-native root with a non-native suffix – do not have a direct counterpart in English, German, or French, which are the most common source languages for borrowings into Czech. The formation of the nouns *archivace* 'archiving' (but no **archivation* in English or **Archivation* in German), *parfemace* 'perfuming', *telefonát* 'phone call' in Table 3 and others in the dataset seems to have taken place in Czech without direct support from equivalent forms in other languages. It can be speculated that this may result from a systemic pressure to supply an action noun within the given morphological families. The tendency of favoring nouns over verbs when using non-native vocabulary can be further supported by morphological families that were not part of the analyzed dataset, where derivational nominals are the only means of expressing action meanings, because no corresponding verb is attested at all (cf. the nouns *kremace* 'cremation', *demise* 'demission', or *transfuze* 'transfusion' with no verbal counterparts that would correspond to the English verbs *to cremate*, *to demit*, and *to transfuse*, respectively).

Additional examples from the domain of neologisms show that, in many cases in Czech, it is the noun that appears first in usage, with the verb only emerging in the data once the noun has become established. For example, the noun *viktimizace* 'victimization' is attested in the SYN2000 corpus, which documents Czech from the final decade of the 20th century (Čermák et al., 2000), as well as in corpora covering subsequent periods, whereas the verb *viktimizovat* 'to victimize' does not appear until the SYN2020 corpus, which includes data from 2015–2020 (Křen et al., 2020). A similar pattern can be observed with the noun *prokrastinace* 'procrastination', which is attested in the SYN2005 corpus (Čermák et al., 2005), while the corresponding verb *prokrastinovat* 'to procrastinate' emerges only in the SYN2020 corpus.

4 Verbs with non-native roots and the typological debate

The observations regarding the denominal features of Czech verbs with non-native roots and their position in the respective morphological families, outlined in Sections 3.1 and 3.2, dovetail with the debate over borrowing of verbal lexemes in contrastive linguistics and linguistic typology. More than 100 years ago, Meillet (1921) noted that among lexical borrowings in French, nouns predominate, while verbs are borrowed only with difficulty. According to Meillet's explanation, this is due to the complex verbal morphology of the recipient language. Weinreich (1970), and similarly van Hout and Muysken (1994), offer a more general explanation: Of all word classes, nouns are the most easily transferred across languages, as they serve the communicative need to expand a language's referential capacity, which is considered a key motivation for lexical borrowing.

In her study on the structure of borrowed verbs in recipient languages and, subsequently, in her typological chapter on language contact, Moravcsik (1975, 1978) argued that verbs are not borrowed directly as verbs; instead, they are initially categorized as nouns in the recipient languages and thereafter integrated into the word class of verbs using native morphological means, resulting in at least a two-part structure combining a borrowed element with a native verbalizer. Although this claim has been rejected as untenable by some scholars (e.g., Campbell, 1993), Wichmann and Wohlgemuth (2008) and Wohlgemuth (2009), based on an analysis of dozens and hundreds source–recipient language pairs, confirm that this is a significant – though not exclusive – strategy for borrowing verbal concepts.

If the data analyzed in the present study are viewed through the prism of the typological discussion just outlined, the non-native verbs do exhibit the binary morphological structure. In the verbs, the non-native root, which is identical with or very close to the source verb in the foreign language, is combined with a native theme, whose form is mostly *-ova-*. The theme is a bound morpheme whose primary distribution is restricted to verbs, and whose occurrence in non-verbal word classes only results from word-class-changing word-formation. The analyzed dataset also provides a direct evidence for the nounhood of the non-native constituent: When the native theme is removed from the verb (along with the infinitive marker), the remaining string is the noun. Moreover, the corpus data provide evidence of cases where a string identical to a foreign verb functions as a noun in Czech (cf. the feminine noun *transfuze* 'transfusion' mentioned above), while no corresponding verb is attested in the language.

5 Concluding remarks

In the present study, which was based on a dataset of over 2,000 noun/verb pairs and morphologically related words, the denominal nature of verbs with non-native roots was inferred from their aspectual behavior – specifically, the verbs' compatibility with a single theme and the use of prefixation to form aspectual counterparts – and further supported by the higher frequency counts of the corresponding nouns relative to the verbs. This structural pattern and quantitative features are characteristic, in the native vocabulary, of verbs converted from nouns.

These observations regarding the denominal behavior of non-native verbs were further elaborated by examining additional members of the respective morphological families. While non-native verbs do not differ from native ones in prefixation or the formation of inflectional nominalizations, they stand out in that derivational nominals frequently compete with them in expressing action meanings within their families. These nominals appear to be created in Czech without the need for a foreign model, and in some families, they even constitute the sole means of expressing the relevant action. Taken together,

the analysis suggests that in morphological families with foreign roots, Czech demonstrates a tendency toward nominal expression and verbs are second choice even in expressing action meanings, for which verbs are preferred in the native segment of the data covered by the present study.

Acknowledgments

The research reported on in the present paper was supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press, Oxford.

Klára Buzássyová. 2010. Vzťah internacionálnych a domácich slov v premenách času. *Jazykovedný časopis* 61:113–130.

Lyle Campbell. 1993. On proposed universals of grammatical borrowing. In Henk Aertsen and Robert Jeffers, editors, *Historical Linguistics 1989. Papers from the 9th International Conference on Historical Linguistics*, John Benjamins, Amsterdam – Philadelphia, pages 91–109.

František Čermák, Renata Blatná, Jaroslava Hlaváčová, Jana Klímová, Jan Kocek, Marie Kopřivová, Michal Křen, Vladimír Petkevič, Věra Schmiedtová, and Michal Šulc. 2000. *SYN2000: A Balanced Corpus of Written Czech*. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; http://www.korpus.cz.

František Čermák, Drahomíra Doležalová-Spoustová, Jaroslava Hlaváčová, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Marie Kopřivová, Michal Křen, Renata Novotná, Vladimír Petkevič, Věra Schmiedtová, Hana Skoumalová, Michal Šulc, and Zdeněk Velíšek. 2005. *SYN2005: A Balanced Corpus of Written Czech*. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; http://www.korpus.cz.

Bożena Cetnarowska. 1993. *The Syntax, Semantics and Derivation of Bare Nominalisations in English*. Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

Václav Cvrček and Pavel Vondřička. 2013. Nástroj pro slovotvornou analýzu jazykového korpusu. In *Grammar and Corpora / Gramatika a Korpus 2012*, Gaudeamus, Hradec Králové, pages 1–10.

Klaus Dietz. 2015. Foreign word-formation in English. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, Mouton de Gruyter, Berlin, volume 3, pages 1637–1660.

Wieland Eins. 2015. Types of foreign word-formation. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, Mouton de Gruyter, Berlin, volume 3, pages 1531–1579.

Pius ten Hacken and Renáta Panocová. 2020. *The Interaction of Borrowing and Word Formation*. Edinburgh University Press, Edinburgh.

Roeland van Hout and Pieter Muysken. 1994. Modeling lexical borrowability. *Language Variation and Change* 6:39–62.

Max Kisselew, Laura Rimell, Alexis Palmer, and Sebastian Padó. 2016. Predicting the direction of derivation in English conversion. In Micha Elsner and Sandra Kuebler, editors, *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Berlin, pages 93–98.

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. 2020. *SYN2020: A Representative Corpus of Written Czech*. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; http://www.korpus.cz.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Jan Zasina. 2015. *SYN2015: A Representative Corpus of Written Czech*. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; http://www.korpus.cz.

Stela Manova. 2011. *Understanding Morphological Rules. With Special Emphasis on Conversion and Subtraction in Bulgarian, Russian and Serbo-Croatian*. Springer, Dordrecht and New York.

Hans Marchand. 1963. On a question of contrary analysis with derivationally connected but morphologically uncharacterized words. *English Studies* 44:176–187.

Hans Marchand. 1964. A set of criteria for the establishing of derivational relationship between words unmarked by derivational morphemes. *Indogermanische Forschungen* 69:10–19.

Olga Martincová. 2005. Nová slovesná pojmenování. In Olga Martincová, editor, *Neologizmy v Dnešní Češtině*, ÚJČ AV ČR, Praha, pages 119–133.

Antoine Meillet. 1921. Le problème de la parenté des langues. In Antoine Meillet, editor, *Linguistique Historique et Linguistique Générale*, Champion, Paris, pages 76–101.

Edith Moravcsik. 1975. Borrowed verbs. *Wiener Linguistische Gazette* 8:3–30.

Edith Moravcsik. 1978. Universals of language contact. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik, editors, *Universals of Human Language. Vol. 1, Method and Theory*, Stanford University Press, Stanford, CA, pages 93–122.

Jitka Mravinacová. 2005. Přejímání cizích lexémů. In Olga Martincová, editor, *Neologizmy v Dnešní Češtině*, ÚJČ AV ČR, Praha, pages 187–211.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.

Magda Ševčíková. 2021a. Action nouns vs. nouns as bases for denominal verbs in Czech: A case study on directionality in derivation. *Word Structure* 14:97–128.

Magda Ševčíková. 2021b. Bezpříponová substantiva a vyjadřování vidového protikladu u příbuzných sloves. *Slovo a slovesnost* 82:263–288.

Magda Ševčíková. 2022. Action meanings in noun/verb conversion: native and foreign word-formation in Czech. *Linguistica Pragensia* 2:173–197.

Magda Ševčíková. in press. *A Paradigmatic Approach to Conversion: Conversion between Nouns and Verbs in Czech*. Springer, Cham.

Salvador Valera. 2015. Conversion. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, Mouton de Gruyter, Berlin, volume 1, pages 322–339.

Krystyna Waszakowa. 2003. Internacionalizacija: Zapadnoslavianskije jazyki. Przejawy tendencji do internacjonalizaji w systemach słowotwórczych języków zachodnosłowiańskich. In Ingeborg Ohnheiser, editor, *Komparacja Systemów i Funkcjonowanie Współczesnych Języków Słowiańskich*, Universität Innsbruck – Uniwersytet Opolski, Opole, volume 1, pages 78–102.

Krystyna Waszakowa. 2005. *Przejawy Internacjonalizacji w Słowotwórstwie Współczesnej Polszczyzny*. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

Krystyna Waszakowa. 2015. Foreign word-formation in Polish. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, Mouton de Gruyter, Berlin, volume 3, pages 1679–1696.

Uriel Weinreich. 1970. *Languages in contact. Findings and problems*. Mouton, The Hague and New York, 7th edition.

Søren Wichmann and Jan Wohlgemuth. 2008. Loan verbs in a typological perspective. In Thomas Stolz, Dik Bakker, and Rosa Salas Palomo, editors, *Aspects of Language Contact. New Theoretical, Methodological and Empirical Findings with Special Focus on Romancisation Processes*, Mouton de Gruyter, Berlin, pages 89–121.

Jan Wohlgemuth. 2009. *A Typology of Verbal Borrowings*. Mouton de Gruyter, Berlin.

Explaining analogy in word-formation: The role of lexical network structure

Sabine Arndt-Lappe
Trier University
arndtlappe@uni-trier.de

Tammy Ganster
–
ganster.tammy@web.de

Aaron Seiler
Trier University
seiler@uni-trier.de

Abstract

Analogy-based theories of word-formation assume that similarity relations between words in the Mental Lexicon play a crucial role in the computation and processing of the form and meaning of complex words. This predicts that the analogies formed should be related to the network structure of this lexicon. Focussing on the form of complex words, this paper presents two case studies from morpho-phonology, specifically stress assignment in English, to explore this link. Study 1 uses Analogical Modeling (Skousen et al., 2013) and a Network Analysis of the similarity space used by the model to explain how effects of opaque morphology emerge in English stress. Study 2 uses experimentally elicited measures of vocabulary size and online lexical processing to explain how inter-speaker variability of stress position arises from individual differences between language users' lexicons.

1 Introduction

According to usage-based theories of linguistic morphology, analogy is a key mechanism of linguistic generalisation. It is often loosely defined as a linguistic form or meaning that is determined by the properties of similar forms or meanings in the language user's Mental Lexicon. Contrary to abstractionist models that emphasise rules or schemas as their main mechanisms, analogy-based models assume that analogies operate mostly 'on the fly', i.e., on the basis of relations between individual exemplars in the Mental Lexicon. Under a view of the Mental Lexicon as a highly interconnected network involving both distributed storage and shared structure, analogy is based on the co-activation of related forms within this lexical network structure. This network structure therefore is central to explaining analogy in word-formation, but is only rarely explored in detail in empirical research.

The present paper sets out to explore the relation between analogy and network structure further, and to make a case for the relevance of the latter as an explanatory force. Two studies will be presented, which exemplify the use of quantitative methods to study how derivational morphology is modulated by lexical network structure. Both studies will deal with morphologically governed stress assignment in English complex words. In the first study we ask how the apparent sensitivity of English verb stress to (opaque) morphological structure emerges from network structure in the Mental Lexicon. In the second study we ask how individual differences in network structure have an influence on paradigmatic co-activation, resulting in variation in stress position. Before the two case studies will be presented, Section 2 will set the stage, introducing the phenomena of interest.

2 Analogy, lexical networks, and morpho-phonology

Most usage-based work on analogy in derivation focusses on affix rivalry or competition. Thus, many studies have shown that in situations in which more than one derivational affix is available to express a given meaning, language users' choices depend on the degree of paradigmatic support that the available options receive in the lexicon (Huyghe and Varvara, 2023; Bonami and Strnadová, 2019; Rainer, 2018; Arndt-Lappe, 2014; Skousen, 1989). The selection of a particular affix among a set of alternatives is, however, not the only kind of linguistic form that is subject to analogical effects. Analogy has also been

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 11–20, Fribourg, Switzerland, 4-5 September 2025.

claimed to be relevant at the morphology-phonology interface, explaining how morphologically complex words are pronounced (Rebrus et al., 2024; Sano, 2015; Pierrehumbert, 2006; Eddington, 2000).

English lexical stress is well-known to be semi-regular, and existing patterns are usually described with reference to phonological, morphological, and morphosyntactic factors. Generally, stress tends to fall within a three-syllable window at the right word edge, where its specific position is influenced by the structure (especially the weight) of the penultimate syllable (for nouns) and the final syllable (for verbs). Adjective stress seems to show traces of both nominal and verbal patterns. Derivational morphology modulates the aforementioned structural regularities. The simplest case of such modulation has traditionally been described in terms of the difference between so-called 'stress-preserving' and 'stress-shifting' derivational affixes. Complex words with stress-preserving affixes are stressed on the same syllable as their bases; complex words with stress-shifting affixes are stressed on a different syllable. The two nominalising suffixes *-ness* and *-ity* may illustrate this point. *-ness* is stress-preserving: cf. e.g., *háppiness* (♦ *háppy*), *árbitrariness* (♦ *árbitrary*), *alértness* (♦ *alért*). Stress in *-ity* derivatives, by contrast, is always on the pre-suffixal syllable of the derivative, independent of stress in the base: cf. e.g., *productívity* (♦ *prodúctive*), *numerósity* (♦ *númerous*), *obésity* (♦ *obése*). This is commonly referred to as 'shifted' stresses. Beyond such simple cases, however, there are several phenomena which do not neatly fit the dichotomy of stress-preserving and stress-shifting affixation. One group of phenomena concerns formatives that are found bearing on stress position, but that are not typical affixes, in the sense that they have recognisable forms, but no recognisable meaning. The most relevant of these for English stress assignment are the large group of etymological prefixes found on English verbs, as illustrated in the disyllabic examples in (1a). The generalisation is that in such verbs the etymological prefix is almost never stressed, even if the rime of the word-final syllable contains a short vowel and a simple coda, a configuration that disfavours word-final stress in words without etymological prefixes (as in 1b, cf. Dabouis and Fournier, 2025 for discussion and a review of the literature).

(1) a. attách ?a-tach
commít ?con-mit
admít ?ad-mit

b. pólish
vómit
lísten

Another group of phenomena concerns variation between stress-shifting and stress-preserving behaviour. There are several affixes which can be both stress-shifting and stress-preserving, sometimes even with the same bases. Examples are provided in (2).

(2)

Base	**Preserving pronunciation**	**Shifting pronunciation**
ánalyse	ánalysable	analýsable
nécessary	nécessarily	necessárily
artículate	artículatory	articulátory

Under an analogy-based view of stress assignment, we expect the two phenomena illustrated in 1 and 2 to be a reflex of the contents of the Mental Lexicon. Thus, the effect of etymological prefixes should emerge from the lexicon containing frequently recurring word-initial sequences, which should be distributed in the lexicon in a different way than other word-initial sequences. In Study 1 in this paper we will use a computational analogical model and Network Analysis to test this assumption. Variation between preserving and shifting stresses with some suffixes, in turn, should be a reflex of the fact that the analogical mechanism is probabilistic, thus allowing for different stresses to emerge in usage. 'Preserving' pronunciations should emerge if paradigmatic links between bases and derivatives prevail; 'shifting' pronunciations should emerge if links across the morphological category prevail, causing uniform stress across that category (Arndt-Lappe et al., 2023). In Study 2 in this paper we will test how the use of these two strategies is related to the individual properties of speakers' Mental Lexicons. This

study uses an experimental methodology that correlates participants' elicited stress productions with a set of measures tapping into their individual lexical processing: vocabulary size, morphological sensitivity, and print exposure. The discussion of the two studies in this paper will focus on the methodologies used; for a detailed presentation of the findings please see Seiler (in prep.) for Study 1 and Ganster (2025) for Study 2.

3 Study 1: Explaining effects of opaque morphology in English verb stress

Regarding morphological determinants of stress, most theories distinguish between monomorphemic and morphologically complex (i.e., affixed) words. However, they disagree in how they partition the lexicon into words whose stress is influenced by morphology and words whose stress is not subject to morphological constraints. Consequently, there is also disagreement about how the nature of morphological effects is defined, with proposals ranging from those considering only transparent, productive form-meaning mappings as relevant to those including recurrent, but semantically opaque formatives. There is, however, leakage, no matter how morphology is defined. Approaches that include so-called 'opaque' morphological constituents have pointed to strong statistical correlations between stress position and the presence especially of opaque, i.e., etymological prefixes in the English lexicon (Dabouis and Fournier, 2025). The two verbs *commít* and *digréss* may illustrate the issue. Both verbs have final stress even though their final syllable counts as light by standard accounts. It is unclear why effects of opaque morphology should be productive synchronically.

Adopting an analogy-based view of linguistic generalisation, we hypothesised that English stress can be accounted for without recourse to an a-priori definition of morphological transparency. To test this hypothesis, a simulation experiment was conducted with a computational Analogical Model ('AML', Skousen et al., 2013). We used the TrAML interface to the algorithm (Arndt-Lappe et al., 2018), which provides access to a variety of model measures beyond accuracy of prediction. In particular, TrAML allows users to reconstruct the structure of the similarity space as relevant for the model, making it possible to see which exemplars in the lexicon are used as analogues in the classification of test items, and how they are weighted by the algorithm.

The database for this simulation experiment (the 'lexicon') comprises all polysyllabic verbs in the *Cambridge English Pronouncing Dictionary* (Jones, 2006), with some modification eliminating especially irrelevant pronunciation variants, instances of noun-verb conversion, and very rare words (N = 3,015, after data cleaning). All data were presented to the model as machine-readable phonological transcriptions. Vowels, however, were represented orthographically, as the phonological transcriptions of a word's vowels would have provided the model with information about which vowels are reduced, hence stressless. AML was set the task of predicting stress position by comparing the phonological string of the target verb to the phonological strings of all other English verbs in the lexicon (in 'leave-one-out' mode). One finding of this simulation experiment is that stress in English verbs is predictable from recurrence only, i.e., without semantic information. AML accurately predicts stress position in 91% of all items in our database with observed stress on the final syllable, 84% of all items with observed stress on the penultimate syllable, and 94% of all items with observed stress on the antepenultimate syllable.

A statistical analysis of the relational structure of the lexicon as established by the algorithm shows how individual 'morphological' elements emerge as relevant for analogical generalisation. Thus, inspection of the set of analogues used by the algorithm for classification reveals that classification is heavily influenced by 'hubs' of highly similar lexical items, suggesting that beyond claims in the literature about opaque prefixes, the English lexicon is generally heavily skewed in terms of disproportionately frequent, recurring bound elements, which can successfully be used for stress prediction. To quantify the degree of connectivity between different hubs of analogues as established by the computational model, we used measures from Network Analysis (e.g., Barabási, 2016). In this analysis, all exemplars in the test set function as nodes; edges in the network link exemplars classified with those exemplars that are used as analogues for the classification. In AML exemplars are usually classified on the basis of more than one lexicon item, hence each test item is normally linked to several analogues (termed the exemplar's 'Analogical Set' in the literature). Also, relations between exemplars are not always reciprocal in AML:

The fact that exemplar A is a member of the analogical set for the classification of exemplar B does not automatically entail that the reverse is the case as well. We systematically analysed how exemplars are integrated in the overall network, using degree centrality, modularity, and community structure (Bohlin et al., 2014) as quantitative measures.

The analysis reveals that the overall network is structured into three distinct subgraphs representing the three major stress categories in English verbs: final stress (e.g., *concéde*, *desérve*), penultimate stress (e.g., *shúdder*, *háppen*, *encúmber*), and antepenultimate stress (e.g., *alkálify*, *encápsulate*). The examples given are among the top five members of their 'communities', which means that they are particularly influential in stress classification, in the sense that they reoccur in several analogical sets in their subgraphs (and thus have relatively high degree centrality values). This tripartite structure of the overall network largely corresponds to different lexical strata typically identified in the literature (e.g., Guierre, 1979, 2000; Trevian, 2003, 2015; Fournier, 2007). Crucially, all three strata display opaque morphological effects – effects that are correlated with specific phonological shapes of either word-initial ('prefixes') or word-final ('suffixes') sequences. Thus, verbs with final stress are typically associated with the Romance stratum. Our analysis shows that this feature corresponds to a set of distinct phonological shapes: Such verbs are typically disyllabic, with their first and second syllable usually corresponding to a finite set of phonological forms ('etymological prefixes' and 'etymological roots') which reoccur in several exemplars. By contrast, verbs with penultimate stress are typically associated with the Germanic part of the English word stock; however, there is also a sizeable number of Romance words in this category. We see that such verbs typically have two or three syllables, and that many such verbs share one of a small set of phonological shapes at the end of the word (esp. [l̩], spelled *-le*; *-en*; *-er*; *-ish*). Word-final *-ish* (in words like *polish*, *admonish*, *embellish*) is a particularly interesting case, as this shape is etymologically Romance (cf. e.g., *Oxford English Dictionary*, "-ish (suffix2)," December 2024, https://doi.org/10.1093/OED/3997102342), but patterns with the Germanic words in terms of phonological shape, thus providing evidence that what is relevant for morphophonology is phonological shape, not etymological stratum. Finally, verbs with antepenultimate stress are often trisyllabic (or even longer), but differ from verbs with penultimate stress in the phonological shape of the end of the word, which is predominantly *-ate* or *-ify*. Both *-ate* and *-ify* are typically classified as verbal suffixes, but are not expected to be stress-shifting under standard accounts, given that they also attach to bound bases.

The three subgraphs for the three stress categories differ significantly in their internal structure. Figure 1 provides a comparison of two informative measures: network density and modularity.

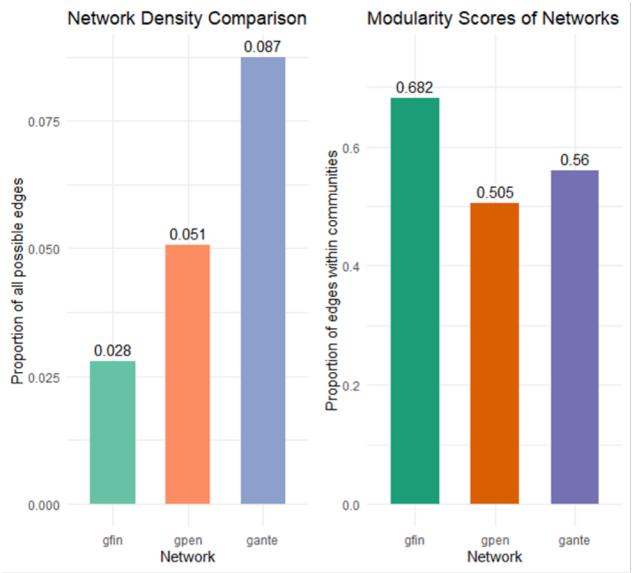


Figure 1: Network density and modularity scores for the three graphs: 'gfin' (i.e., final stress verbs, N = 1,558), 'gpen' (i.e., penultimate stress verbs, N = 795), 'gante' (i.e., antepenultimate stress verbs, N = 653).

Network density measures the number of used edges within a graph, compared against the total number of possible edges between lexical items. The higher the network density, the more are exemplars linked to other exemplars in the graph, which means in our case that more exemplars tend to reappear in analogical sets of several other exemplars. Modularity, in turn, provides the number of within-community edges, compared against the total number of used edges in the network. The higher the modularity score, the more do exemplars form clusters of interlinked nodes within their network, which means in our case that the network consists of separate communities of exemplars which reappear in several analogical sets within their communities, but tend not to reappear in the analogical sets outside their communities. 'Communities' are defined as clusters of nodes that are better connected amongst each other than with nodes outside the cluster. By comparing the left and right panels of Figure 1 we see that verbs that are stressed on the final syllable ('gfin') form a network that has the lowest overall density, but highest modularity. This means that in terms of their phonological shape, such verbs form smaller, distinct, well-separated clusters, with little to no overlap between them. Example words from two typical communities in the final-stress subgraph are provided in (3).

(3) Community 6: degrade, decide, explode, extrude
Community 9: redress, undress, possess, caress

Communities are characterised by family resemblance relationships. For example, members of Community 6 all end in a tense vowel or diphthong followed by [d], and they begin with an etymological prefix, which can, however, have different shapes. Members of Community 9, by contrast, end in [ɛs], and the first syllable of these words may or may not be a prefix. Crucially, the low overall density value of the final-stress network suggests that the shapes characterising these communities are distinct in the lexicon, meaning that they have few phonological neighbours.

The structure of the network of verbs with penultimate stress is different from that of verbs with final stress, with a higher density score and lower modularity. This means that clusters are still distinct, but a few shared endings blur the otherwise clear community boundaries. These are mostly *-le*, *-er*, and *-en*, as well as, to a lesser extent, *-y* and *-ish*. Example words from two communities are provided in (4).

(4) Community 2: grabble, brabble, scramble, scrabble
Community 8: juggle, struggle, snuggle, smuggle

Finally, the network of verbs with antepenultimate stress is characterised by the highest density value of the three subgraphs, and only moderate modularity. This group of verbs is more homogeneous than verbs with penultimate or final stress: All verbs with antepenultimate stress end in one of three verbal suffixes: *-ate* (by far the largest group), *-ify*, and *-ise*. There is considerable overlap between communities, and weak group boundaries. Examples are provided in (5).

(5) Community 9: dedicate, medicate, triplicate, predicate
Community 13: derogate, congregate, interrogate, relegate

In sum, our simulation experiment and subsequent analyses show that effects of opaque morphology on stress assignment in English verbs can indeed be predicted successfully on the basis of the phonological shape of verbs only, without recourse to etymological morphological elements or abstract phonological representations (like, e.g., syllable weight). The network analysis reveals how such effects emerge. The phonological shapes of English verbs are not made up of a random combination of the sounds, but massively involve combinations that reoccur in many words and at the same time are distinct, i.e., substantially different, from others. We saw that all three stress categories in English verbs (final, penultimate, and antepenultimate stress) are characterised by such reoccurring shapes, but also that attested shapes are most distinct and modular in verbs with final stress, forming internal clusters within the subgraph of final-stressed verbs. Stress assignment, then, is an effect of the network structure of existing words in the Mental Lexicon. The analysis presented raises the question whether such effect is best classified as phonological or morphological. What we saw is that there is gradience between the two.

4 Study 2: Explaining individual differences in stressing English complex words

The idea that morpho-phonology is based on analogy, i.e., on co-activation of words in the Mental Lexicon, predicts that it should also reflect individual differences in processing. Individual differences relating to the structure of the Mental Lexicon should hence correspond to individual differences in the phonological realisation of complex words. One area where this has recently been shown to be the case is variable stress position in derived adjectives in English (Ganster, 2025). Ganster studied stress in derived adjectives with four different suffixes: *-able*, *-ant*, *-(at)ive*, and *-(at)ory* in British English. The derivational categories characterised by these four suffixes are well-known to show stress variation, in the sense both stress-shifting and stress-preserving patterns can be observed and, crucially for this study, both patterns can be observed for the same derivatives (for overviews cf. Bauer et al., 2013: chapter 14.2; Trevian, 2007). Examples of test words used in Ganster's study are provided in (6).

(6)

Derivative	**Base**	**Preserving stress**	**Shifted stress**
identifiable	idéntify	idéntifiable	identifíable
quantifiable	quántify	quántifiable	quantifíable
conversant	convérse	convérsant	cónversant
triumphant	tríumph	tríumphant	triúmphant
imaginative	imágine	imáginative	imaginátive
provocative	provóke	provócative	provocátive
anticipatory	antícipate	antícipatory	anticipátory
participatory	partícipate	partícipatory	participátory

In a production experiment, 153 participants read 30 test sentences containing derivatives of interest. These were presented in naturalistic sentences (based on scripted dialogue sentences from the Corpus of American Soap Operas, SOAP, Davies, 2011), in which they were embedded in rhythmically controlled contexts. In addition, participants completed three tasks tapping into various aspects of Mental Lexicon structure: a standardised vocabulary size test (Nation and Beglar, 2007), a masked priming experiment testing for morphological sensitivity (modelled on Hasenäcker et al., 2020), and an author recognition test, testing for print exposure (Acheson et al., 2008). Furthermore, a variety of sociodemographic data was collected by means of a questionnaire. The experiment was conducted as a remote experiment on the Labvanced platform (`www.labvanced.com`); participants were recruited via the Prolific service (`www.prolific.com`).

Variation in stress position was found for all 30 derivatives tested in the study. Contrary to claims in the literature that most of the morphological categories tested are stress shifting, stress preservation was the clear majority option overall: In 72% of all cases, primary stress in the base was preserved in the derivative (N = 3,442). Individual derivatives differed in the extent to which stress varied. Specifically, the probability of preserving stress was found to be correlated with structural linguistic factors (especially the weight of the pre-suffixal syllable) as well as the stress pattern of the base word. The most frequent type of stress shift that occurred in the study resulted in derivative stress on the stem-final heavy syllable (in the examples in 6: *quantifíable*, *triúmphant*, *imaginátive*, *provocátive*, *anticipátory*, *participátory*).

All participants varied in the extent to which they preserved primary stress in the derivative, but did so to different degrees. Participants preserved stress in between 13% and 92% of all derivatives (mean preservation rate per participant: 23%; half of the participants do not preserve stress in between 22% and 33% of all 30 derivatives). All three measures of individual differences in lexical processing (vocabulary size, morphological sensitivity, and print exposure) turned out to be predictive of participants' probability to preserve stress in the derivative. The three variables were correlated amongst each other, but the effect found was independent of age and other sociodemographic variables available in the study. We report on the vocabulary size effect here, as this measure turned out to be the strongest of the three measures, showing consistently robust effects in all statistical analyses conducted.

Nation and Beglar's (2007) vocabulary size test ('VST') is one of the most widely used, validated tests of (receptive) lexical knowledge, which also works well for native speakers. It systematically tests words

from different frequency bands in the British National Corpus (BNC Consortium, 2007) in a multiple choice design in which participants have to select the correct definition of the word. From the score obtained, the test extrapolates an estimate of the participant's vocabulary size (counted as the estimated number of word families). Figure 2 visualises the correlation between the preservation of primary stress in the data and vocabulary size in the raw data.

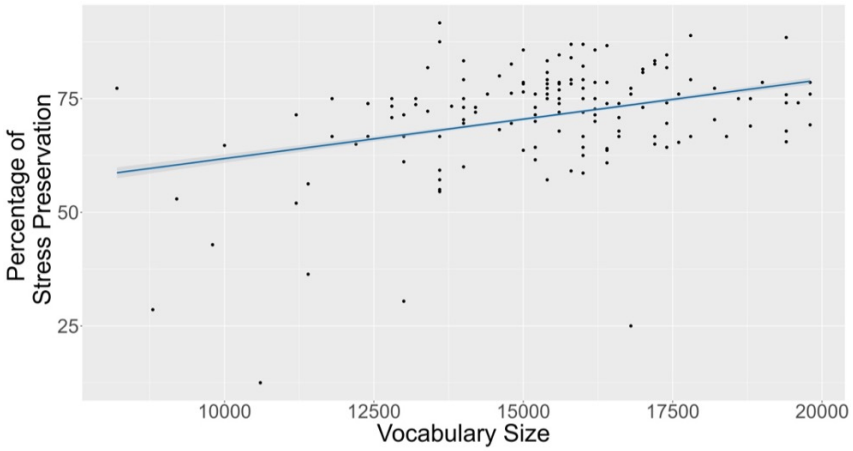


Figure 2: Relationship between participants' individual proportions of stress preservation and VST score (N = 3,442). Kendall's τ = 0.15, z = 13.17, $p$one-tailed $<$ 2.2e-16. The figure is a reproduction of Figure 34 in Ganster (2025, p. 122).

The effect of vocabulary size also survives as the most robust predictor of main stress preservation in a multivariate mixed effects regression analysis, which was used to explore how the probability of stress preservation can be predicted from all lexical processing variables (vocabulary size, morphological sensitivity, print exposure) as well as other variables of interest, which had emerged as promising in preliminary analyses. In particular, these were frequency-related variables (derivative frequency, relative frequency of the derivative with respect to its base). Also, the model included a variable coding whether the participant's first language was English ('L1'); the reason is that even though native-speaker status had been used as an exclusion criterion when recruiting participants for the experiment, a small group of 24 of all 153 participants indicated in the sociodemographic questionnaire that English was not their first language.[1] In addition, the analysis included a random intercept for derivative. The regression analysis was complemented by a random-forest analysis to explore the relationships between correlated predictors; the random forest analysis also confirmed the explanatory power of vocabulary size. The score from the vocabulary size test and the L1 variable were the only fixed effects that survived in the best regression model after simplification. Figure 3 visualises the partial effect of vocabulary size in our best model.

We see that the probability of primary stress preservation increases with the participant's vocabulary size. In addition, there is a small effect of native language: The probability of stress preservation is slightly higher for speakers whose first language is English than for the non-native speakers. The effect of vocabulary size is mirrored also in the measures gauging morphological sensitivity and print exposure (not included in the statistical model), although effects are generally much weaker (cf. Ganster, 2025, Chapter 6.6 for detailed analysis and discussion). The greater the priming effect that participants showed in the morphological sensitivity task, the more likely are they to produce preserving stresses. Likewise, greater print exposure (i.e., a higher score in the author recognition task) is associated with a higher probability of stress preservation. All three lexical processing measures are correlated, which is why it was only possible to include one of them in the regression analysis.

In sum, Ganster's study not only confirms earlier observations in the literature that the English adjectival

[1] Indeed, the variable 'L1' turned out to be a significant predictor in some of the analyses of the data in Ganster (2025), but differences between the two groups (English L1, English L2) were always gradual, never categorical, so that there was no independent justification for excluding the small L2 group from the analysis.

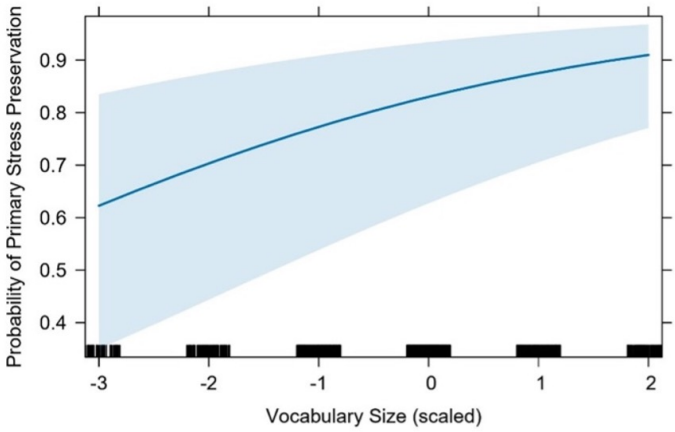


Figure 3: Partial effect of vocabulary size (top panel) on primary stress preservation. Model formula: `PrimaryStressPreservation ~ VSTScore_scaled (i.e., vocabulary size) + FirstLangBinary + (1| Derivative)`, N = 2,870. Interpretation aid: 0 indicates an average vocabulary size. Numbers below 0 represent lower than average vocabulary sizes, numbers over 0 represent higher than average vocabulary sizes. The figure is a reproduction of Figure 48 in Ganster (2025, p. 142).

morphological categories tested exhibit variable morpho-phonological stress. It also shows that a large portion of this variability is due to individual differences between speakers and is reflected in measures tapping into the structure of these speakers' lexicons. This has important implications for theories of morpho-phonology. On the one hand, it adds to the growing body of evidence that morpho-phonology depends on lexical processing (see also e.g., Collie, 2008; Breiss, 2024). On the other hand, it sheds new light on the question how exactly morphophonology should depend on lexical processing. Thus, previous accounts have largely claimed that effects of lexical processing depend on the resting activation of relevant words in the Mental Lexicon, which are correlated with lexical frequency. The study by Ganster (2025), by contrast, suggests that overall lexicon size and morphological sensitivity in processing are relevant, too. These measures have been associated in the literature not just with the accessibility of individual items, but with the strength and connectivity of the lexical network (e.g., Mainz et al., 2017). It seems that speakers with a large and strong such network tend to rely on this network to mark transparent morphological relations between words, by means of stress preservation. The result is phonologically nonuniform stress: e.g., two syllables before the suffix *-ive* in *provócative* and three syllables before the suffix in *imáginative*. Speakers with smaller networks, by contrast, seem to rely more on morphophonological rules, tending to produce stress that is more uniform within the morphological category (e.g., *identifíable*, *quantifíable*), and that conform to phonologically unmarked patterns like the tendency for heavy syllables to bear stress in English.

5 Summary and conclusion

This paper has discussed two case studies about how morpho-phonological stress is related to the network structure of the Mental Lexicon. In Study 1 we saw that English verb stress is sensitive to opaque morphological formatives, and that this sensitivity can be explained if we take into account how such formatives are distributed in the lexicon. For example, the strong correlation of the presence of an opaque prefix and final stress in disyllabic verbs was seen to be related to final stressed verbs forming a network with particularly high modularity, i.e., clusters of phonologically highly similar verbs. We used Analogical Modeling as a computational algorithm to test how predictable stress is on the basis of phonological similarity, and Network Analysis to explore the community structure of resultant lexical networks.

Study 2 focussed on transparent derivational morphology, specifically on complex adjectives in English.

We presented an empirical study to test the hypothesis that, if morpho-phonological stress is related to the network structure of the Mental Lexicon, then differences between speakers in terms of this structure should be reflected in their stress productions, leading to variation. Indeed, the stress patterns elicited in a reading experiment were found to be correlated with measures gauging lexicon structure, i.e., the overall size of speakers' Mental Lexicons and their morphological sensitivity in a masked priming task. Higher scores on these measures is associated with stress preservation, leading to greater morpho-phonological transparency. Lower scores, by contrast, are associated with more uniform stress across morphological categories.

Both studies are in line with an analogy-based view of word-formation in general, and morpho-phonology in particular. Thus, they have demonstrated the relevance of co-activation of individual, similar lexemes in the Mental Lexicon for morpho-phonological stress. For example, Study 1 demonstrates the relevance of relations between lexemes on a low level of abstractions or schematisation. This can be seen in the similarity relations between individual lexical items, which form networks that differ in structure. Study 2 also demonstrates that the relationship between morpho-phonology (i.e., stress position in complex words) and the contents of the Mental Lexicon is an immediate one. The former directly reflects individual differences in lexical processing which are related to the size and connectivity of the network, and hence the accessibility of relevant lexical items. On a methodological level, the two studies have provided examples of how Network Analysis (Study 1) and measures gauging vocabulary size and morphological sensitivity (Study 2) can be used to to explore the connection between analogy in morpho-phonology, co-activation, and the structure of the Mental Lexicon.

Acknowledgments

We are grateful to the *Deutsche Forschungsgemeinschaft* for supporting the research reported in Study 1 (project no. 469696746).

References

Daniel J. Acheson, Justine B. Wells, and Maryellen C. MacDonald. 2008. New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods* 40:278–289. https://doi.org/10.3758/BRM.40.1.278.

Sabine Arndt-Lappe. 2014. Analogy in suffix rivalry: The case of English *-ity* and *-ness*. *English Language and Linguistics* 18:497–548. https://doi.org/10.1017/S136067431400015X.

Sabine Arndt-Lappe, Ingo Plag, Kai Koch, and Mikalai Krott. 2018. Transparent Analogical Modelling (TrAML). https://github.com/SabineArndtLappe/TrAML.

Sabine Arndt-Lappe, Robin Schrecklinger, and Fabian Tomaschek. 2023. Stratification effects without morphological strata, syllable counting effects without counts – modelling English stress assignment with Naive Discriminative Learning. *Morphology* 33:433–457. https://doi.org/10.1007/s11525-022-09399-9.

Albert-László Barabási. 2016. *Network Science*. Cambridge University Press, Cambridge.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780198747062.001.0001.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). http://www.natcorp.ox.ac.uk/.

Ludvig Bohlin, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. 2014. Community detection and visualization of networks with the map equation framework. In Ying Ding, Ronald Rousseau, and Dietmar Wolfram, editors, *Measuring Scholarly Impact*, Springer International Publishing, Cham, pages 3–34. https://doi.org/10.1007/978-3-319-10377-8_1.

Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2):167–197. https://doi.org/10.1007/s11525-018-9322-6.

Canaan Breiss. 2024. When bases compete: A voting model of lexical conservatism. *Language* 100(2):308–358. https://doi.org/10.1353/lan.2024.a929738.

Sarah Collie. 2008. English stress preservation: The case for "fake cyclicity". *English Language and Linguistics* 12:505–532. https://doi.org/10.1017/S1360674308002736.

Quentin Dabouis and Jean-Michel Fournier. 2025. Opaque morphology and phonology: Historical prefixes in English. *Journal of Linguistics* 61:231–63. https://doi.org/10.1017/S002222672400015X.

Mark Davies. 2011. Corpus of American Soap Operas. https://www.english-corpora.org/soap/.

David Eddington. 2000. Spanish stress assignment within the Analogical Modeling of Language. *Language* 76:92–109. https://doi.org/10.2307/417394.

Jean-Michel Fournier. 2007. From a Latin syllable-driven stress system to a Romance versus Germanic morphology-driven dynamics: In honour of Lionel Guierre. *Language Sciences* 29:218–236. https://doi.org/10.1016/j.langsci.2006.12.010.

Tammy Ganster. 2025. *Variable stress patterns in English complex adjectives: Inter-individual differences and the nature of the phonology-morphology interface*. Ph.D. thesis, Trier University, Trier.

Lionel Guierre. 1979. *Essai sur l'accentuation en anglais contemporain : éléments pour une synthèse*. Ph.D. thesis, Université Paris-VII, Paris.

Lionel Guierre. 2000. Pourquoi la morpho-phonologie. In Pierre Busuttil, editor, *Points d'interrogation. Phonétique et phonologie de l'anglais*, Publications de l'Université de Pau, Pau, pages 32–46.

Jana Hasenäcker, Elisabeth Beyersmann, and Sascha Schroeder. 2020. Morphological priming in children: Disentangling the effects of school-grade and reading skill. *Scientific Studies of Reading* 24:484–499. https://doi.org/10.1080/10888438.2020.1729768.

Richard Huyghe and Rossella Varvara. 2023. Affix rivalry: Theoretical and methodological challenges. *Word Structure* 16(1):1–23. https://doi.org/10.3366/word.2023.0218.

Daniel Jones. 2006. *Cambridge English Pronouncing Dictionary*. Cambridge University Press, Cambridge.

Nina Mainz, Zeshu Shao, Marc Brysbaert, and Antje S. Meyer. 2017. Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Frontiers in Psychology* 8:1164. https://doi.org/10.3389/fpsyg.2017.01164.

Paul Nation and David Beglar. 2007. A vocabulary size test. *The Language Teacher* 31:9–13. https://openaccess.wgtn.ac.nz/articles/journal_contribution/A_vocabulary_size_test/12552197?file=23375153.

Janet B. Pierrehumbert. 2006. The statistical basis of an unnatural alternation. In Louis Goldstein, D. H. Whalen, and Catherine T. Best, editors, *Laboratory Phonology 8*, De Gruyter, Berlin, pages 81–106. https://doi.org/110.1515/9783110197211.1.81.

Franz Rainer. 2018. Patterns and niches in diachronic word formation: The fate of the suffix -MEN from Latin to Romance. *Morphology* 18(3):397–465. https://doi.org/10.1007/s11525-018-9333-3.

Péter Rebrus, Péter Szigetvári, and Miklós Törkenczy. 2024. No lowering, only paradigms: A paradigm-based account of linking vowels in Hungarian. *Acta Linguistica Academica* 71:137–170. https://doi.org/10.1556/2062.2023.00674.

Shin-Ichiro Sano. 2015. The role of exemplars and lexical frequency in Rendaku. *Open Linguistics* 1:329–344. https://doi.org/10.1515/opli-2015-0005.

Aaron Seiler. in prep. *Variability in English verb stress: Lexical distributions, analogy, and the nature of lexical representations*. Ph.D. thesis, Trier University, Trier.

Royal Skousen. 1989. *Analogical Modeling of Language*. Kluwer, Dordrecht.

Royal Skousen, Thereon Stanford, and Nathan Glenn. 2013. Algorithm::AM. https://github.com/garfieldnate/Algorithm-AM.

Ives Trevian. 2003. *Morphoaccentologie et processus d'affixation de l'anglais*. Peter Lang, Bern. https://www.peterlang.com/document/1096779.

Ives Trevian. 2007. Stress-neutral endings in contemporary British English: An updated overview. *Language Sciences* 29:426–450. https://doi.org/10.1016/j.langsci.2006.12.016.

Ives Trevian. 2015. *English Suffixes: Stress-assignment Properties, Productivity, Selection and Combinatorial Processes*. Peter Lang, Bern. https://doi.org/10.3726/978-3-0351-0761-6.

Additional lemmatization and measures of derivational productivity: The case of Lithuanian denominal suffixal nouns

Jurgis Pakerys
Vilnius University
jurgis.pakerys@flf.vu.lt

Agnė Navickaitė-Klišauskienė
Vilnius University
agne.navickaite@flf.vu.lt

Virginijus Dadurkevičius
Vytautas Magnus University
virginijus.dadurkevicius@vdu.lt

Abstract

We discuss how additional semi-automatic lemmatization and derivational annotation improve corpus data used to measure the derivational productivity of Lithuanian denominal suffixal nouns. A corpus of 1.3 billion tokens is used to estimate realized, potential, and expanding productivity based on types, hapaxes, and total frequencies. We exemplify our results by discussing the categories of quality, status, personal, and diminutive nouns expressed by 10 suffixes.

1 Introduction

For the estimation of derivational productivity in corpora, the measures of realized, potential, and expanding productivity were introduced based on the types, hapaxes, and total frequencies of the derivatives (Baayen, 1992, 1993; see overviews in Baayen, 2009; Gaeta and Ricca, 2015; Dal and Namer, 2016). To obtain credible results, reliable lemmatization tools are needed to capture rare lexemes in large corpora, along with meticulous manual review to exclude derivationally non-analyzable lemmas and those containing inner derivational cycles (Evert et al., 2000; Evert and Lüdeling, 2001; Dal et al., 2008; Gaeta and Ricca, 2006).

In our study, we focus on the expansion of the range of lemmas that are not automatically lemmatized and can be extracted by the application of a semi-automatic procedure. We also present the results of subsequent manual review and resolution of some homographic word-forms that may distort lemma frequencies.

2 Our data and methods

We used word and lemma lists of the Joint Corpus of Lithuanian containing 1.3 billion tokens (Dadurkevičius, 2020a,b; Dadurkevičius and Petrauskaitė, 2020). The initial fully automatic lemmatization was performed with the help of a Hunspell-type lemmatizer operating on the basis of a fixed dictionary and a set of inflectional rules (Dadurkevičius, 2017).

Upon discovering that the lemmatizer misses a significant number of derivatives that are not found in the inbuilt dictionary, we performed the following procedure of additional semi-automatic lemmatization. The word-forms of the corpus were automatically filtered according to the pattern SUFFIX + (all possible) ENDINGS for all relevant suffixes; the extracted forms were then morphologically annotated and grouped into potential lemmas. The resulting lists were manually reviewed, and derivationally non-transparent nouns and those with inner derivational cycles were excluded. The revisions were mostly done by one annotator – either Agnė Navick-aitė-Klišauskienė or Jurgis Pakerys – with the exception of some dubious cases that were analyzed and discussed by both annotators.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 21–25, Fribourg, Switzerland, 4-5 September 2025.

3 Results

In this study, we investigated categories of quality, status, personal, and diminutive nouns expressed by ten productive suffixes according to the evaluation of major grammars of Lithuanian (Ulvydas, 1965; Ambrazas, 1994).

3.1 Quality and status nouns

We included the following four suffixes of quality and status nouns in our study: three native (*-um-as*, *-yb-ė*, *-yst-ė*) and one borrowed (*-izm-as*), e.g., *saug-um-as* 'safety' ← *saug-us* 'safe', *kantr-yb-ė* 'patience' ← *kantr-us* 'patient', *krikščion-yb-ė* 'Christianity' ← *krikščion-is* 'Christian (noun)', *nar-yst-ė* 'membership' ← *nar-ys* 'member', *plokščiapad-yst-ė* 'flat-footedness' ← *plokščiapad-is* 'flat-footed', *individual-izm-as* 'individualism' ← *individual-us* 'individual (adjective)', *kapital-izm-as* 'capitalism' ← *kapital-as* 'capital'.

The results of the additional semi-automatic lemmatization and manual review are presented in Table 1. For the sake of brevity, we provide here and below the total frequencies and the potential productivity calculations (hapax counts divided by total frequencies) only for the final lemma lists obtained after the manual review.

Suffix	Automatic lemmatization		Additional lemmatization		Additional lemmatization and manual review			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-um-as	2,518	68	22,895	9,222	6,687	1,758	4,027,571	0.4365
-izm-as	484	8	4,280	1,856	772	214	7,845,016	0.0273
-yst-ė	269	2	2,355	1,017	1,136	365	826,408	0.4417
-yb-ė	276	1	3,802	1,715	1,054	332	196,472	1.6898

Table 1: Productivity data of quality and status noun suffixes (V – types, V_1 – hapaxes, P – potential productivity)

3.2 Personal nouns

We studied three personal noun suffixes: two native (*-inink-as*, *-ė*; *-uol-is*, *-ė*) and one borrowed (*-ist-as*, *-ė*). Personal nouns in Lithuanian can be masculine or feminine, and their gender correlates with certain declensions; in our case, the nominative singular *-as* and *-is* represent masculine nouns, while *-ė* is feminine, e.g., *men-inink-as*, *-ė* 'artist' ← *men-as* 'art', *blaiv-inink-as*, *-ė* 'abstainer' ← *blaiv-us* 'sober', *jaun-uol-is*, *-ė* 'young person' ← *jaun-as* 'young', *turt-uol-is*, *-ė* 'wealthy person' ← *turt-as* 'wealth', *gitar-ist-as*, *-ė* 'guitar player' ← *gitar-a* 'guitar', *real-ist-as*, *-ė* 'realist' ← *real-us* 'real'.

Unlike the traditional approach found in Lithuanian grammars and other word-formation studies, we treated masculine and feminine formations separately to highlight their differing degrees of productivity, as shown in Table 2.

In our earlier study of deverbal agent nouns in Lithuanian (Pakerys et al., 2024), it was noted that the frequencies of some derivatives are distorted due to a significant number of homographic cells found in the paradigms of masculine and feminine formations with certain suffixes. In the case of personal nouns, such homographic forms are less common: for the suffixes *inink-as*, *-ė* and *-ist-as*, *-ė*, the homographic forms are the locative and the vocative singular of masculine nouns (*-e*) and the instrumental and the vocative singular of feminine nouns (*-e*); for the suffix *-uol-is*, *-ė*, the genitive plural is the same for the nouns of both genders (*-ių*).

We performed manual disambiguation only for hapaxes, as they are used to estimate both potential and expanding productivity and require relatively little time for review. Disambiguating more frequent forms would be too time-consuming; however, it is important to keep in mind that non-disambiguated homographic forms distort the total frequencies to some extent, affecting the estimates of potential

Suffix	Automatic lemmatization		Additional lemmatization		Additional lemmatization and manual review			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-ininkas	760	18	7,412	3,138	2,657	693	4,461,517	0.1553
-ininkė	421	31	1,599	613	700	138	481,476	0.2866
-istas	268	1	6,665	2,980	996	341	638,651	0.5339
-istė	151	3	1,527	683	254	69	75,879	0.9093
-uolis	97	0	1,244	545	271	67	281,323	0.2382
-uolė	85	1	804	342	155	34	115,766	0.2937

Table 2: Productivity data of personal noun suffixes (V – types, V_1 – hapaxes, P – potential productivity)

productivity. The results of hapax disambiguation are as follows: for *-ininkas*, *-ė*, one feminine lemma was corrected to masculine and four cases could not be disambiguated due to a lack of context; for *istas*, *-ė*, three cases lemmatized as feminine were corrected to masculine and two cases could not be disambiguated; for *-uolis*, *-ė*, one feminine lemma was corrected to masculine and two masculine lemmas were corrected to feminine, while seven lemmas could not be disambiguated.

3.3 Diminutives

Lithuanian is rich in diminutive suffixes, but we were only able to evaluate three of them due to the large number of potential lemmas that had to be reviewed manually. Two of the chosen suffixes stand in complementary distribution with regard to the number of syllables of the base stem: the suffix *-el-is*, *-ė* attaches to bases that are one syllable long, while *-ėl-is*, *-ė* is used for longer bases. The suffix *-(i)uk-as*, *-ė* is insensitive to the base length. Derivatives with the nominative singular *-as* and *-is* are masculine, while those with *-ė* are feminine, e.g., *vaik-el-is* 'small child' ← *vaik-as* 'child' (monosyllabic base with a diphthong ⟨ai⟩ [ɐɪ]), *rank-el-ė* 'small hand' ← *rank-a* 'hand' (monosyllabic base), *ežer-ėl-is* 'small lake' ← *ežer-as* 'lake' (disyllabic base), *parduotuv-ėl-ė* 'small shop' ← *parduotuv-ė* 'shop' (trisyllabic base), *rat-uk-as* 'small wheel' ← *rat-as* 'wheel', *kavin-uk-ė* 'small cafe' ← *kavin-ė* 'cafe'. The results of this additional lemmatization and manual review are presented in Table 3.

Suffix	Automatic lemmatization		Additional lemmatization		Additional lemmatization and manual review			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-(i)uk-as	1,182	7	15,646	6,530	3,598	950	687,339	1.3821
-(i)uk-ė	147	0	3,807	1,521	686	209	24,372	8.5754
-ėl-is	483	3	6,467	2,500	1,900	566	376,211	1.5045
-ėl-ė	302	2	4,813	1,785	1,026	257	203,848	1.2607
-el-is	481	2	14,364	5,885	755	74	1,354,266	0.0546
-el-ė	319	5	6,862	2,673	645	65	1,098,362	0.0592

Table 3: Productivity data of diminutive noun suffixes (V – types, V_1 – hapaxes, P – potential productivity)

Just as in the case of personal nouns discussed above, diminutives of different genders also have homographic forms. For the suffixes *-ėl-is*, *-ė* and *-el-is*, *-ė*, the genitive plural is the same for nouns of both genders (*-ių*), and we reviewed and manually disambiguated their hapaxes. The results are as follows: for the suffix *-ėl-is*, *-ė*, one feminine lemma was corrected to masculine and ten hapaxes could not be disambiguated; for the suffix *-el-is*, *-ė*, one masculine lemma was corrected to feminine and one

feminine lemma was corrected to masculine. For the suffix *-(i)uk-as*, *-ė*, the locative and the vocative singular of masculine nouns (*-e*) coincides with the instrumental and the vocative singular of feminine nouns (*-e*). Here, one lemma was resolved as feminine and seven cases were disambiguated as masculine.

4 Discussion

It is evident that additional lemmatization based on simple search strings in word-form lists significantly increases the counts of types and hapaxes. However, it also introduces a substantial amount of random data, making manual review essential to exclude pseudo-lemmas, derivationally non-transparent words, and those with internal derivational cycles.

Lemma frequencies in both automatically- and semi-automatically-generated lemma lists may also be affected by non-disambiguated homographic forms. However, even for large corpora like ours, it is possible to disambiguate the hapaxes that are used for the calculation of potential and expanding productivity to arrive at more precise estimates. Our disambiguation of the homographic forms of hapaxes with regard to gender did not lead to significant changes compared to the disambiguation of hapaxes of agent nouns performed in our earlier study (Pakerys et al., 2024). This is because the number of homographic cells in the paradigms of personal and diminutive nouns examined in the present study is much lower than that in the paradigms of agent nouns. Further research is needed to determine the extent to which the total frequencies (and estimates of potential productivity) may be distorted. This can be achieved by manually reviewing and disambiguating homographic forms in selected data samples.

It should also be noted that our method of additional lemmatization may be very time-consuming due to large amounts of data requiring manual review. The procedure needs to be tested using data from smaller corpora of up to 100 million tokens to determine whether large corpora, like the one used in the present study, offer significant advantages for measuring derivational productivity.

Acknowledgments

The data are drawn from the project “Derivational productivity of Lithuanian suffixed nouns in the Joint Corpus of Lithuanian”, funded by the Research Council of Lithuania (LMTLT) under agreement No. S-LIP-22-61. We sincerely thank the anonymous reviewers of the paper for their remarks and Cristina Aggazzotti for editing the English of the article.

References

Vytautas Ambrazas, editor. 1994. *Dabartinės lietuvių kalbos gramatika*. Mokslo ir enciklopedijų leidykla, Vilnius.

Harald R. Baayen. 1992. Quantitative aspects of morphological productivity. In Geert E. Booij and Jaap van Marle, editors, *Yearbook of Morphology 1991*, Kluwer, Dordrecht, pages 109–149. https://doi.org/10.1007/978-94-011-2516-1_8.

Harald R. Baayen. 1993. On frequency, transparency and productivity. In Geert E. Booij and Jaap van Marle, editors, *Yearbook of Morphology 1992*, Kluwer, Dordrecht, pages 181–208. https://doi.org/10.1007/978-94-017-3710-4_7.

Harald R. Baayen. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, Mouton de Gruyter, Berlin and New York, volume 2, pages 899–919. https://doi.org/10.1515/9783110213881.2.899.

Virginijus Dadurkevičius. 2017. Lietuvių kalbos morfologija atvirojo kodo „hunspell“ platformoje. *Bendrinė kalba* 90:1–17. https://journals.lki.lt/bendrinekalba/article/view/156.

Virginijus Dadurkevičius. 2020a. Assessment data of the dictionary of modern Lithuanian versus joint corpora. CLARIN-LT digital library in the Republic of Lithuania. https://clarin.vdu.lt/xmlui/handle/20.500.11821/36.

Virginijus Dadurkevičius. 2020b. Wordlist of lemmas from the joint corpus of Lithuanian. CLARIN-LT digital library in the Republic of Lithuania. http://hdl.handle.net/20.500.11821/41.

Virginijus Dadurkevičius and Rūta Petrauskaitė. 2020. Corpus-based methods for assessment of traditional dictionaries. In Andrius Utka, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, and Danguolė Kalinauskaitė, editors, *Human Language Technologies – The Baltic Perspective*, IOS Press, Frontiers in Artificial Intelligence and Applications, pages 123–126.

Georgette Dal, Bernard Fradin, Natalia Grabar, Fiammetta Namer, Stéphanie Lignon, Clément Plancq, Pierre Zweigenbaum, and François Yvon. 2008. Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats. In Jacques Durand, Benoît Habert, and Bernard Laks, editors, *Actes du Premier Congrès Mondial de Linguistique Française, Paris, 9–12 juillet 2008*, Institut de Linguistique Française, Paris, pages 1587–1599.

Georgette Dal and Fiammetta Namer. 2016. Productivity. In Andrew Hippisley and Gregory Stump, editors, *The Cambridge Handbook of Morphology*, Cambridge University Press, Cambridge, pages 70–90. https://doi.org/10.1017/9781139814720.004.

Stephanie Evert, Ulrich Heid, and Anke Lüdeling. 2000. On measuring morphological productivity. In Werner Zühlke and Ernst Günter Schukat-Talamazzini, editors, *KONVENS-2000 Sprachkommunikation*. VDE-Verlag, Ilmenau, pages 57–61.

Stephanie Evert and Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja, editors, *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster University, Lancaster, pages 167–175.

Livio Gaeta and Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1):57–89. https://doi.org/10.1515/LING.2006.003.

Livio Gaeta and Davide Ricca. 2015. Productivity. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe*, De Gruyter, Berlin and Boston, volume 2, pages 842–858. https://doi.org/10.1515/9783110246278-003.

Jurgis Pakerys, Virginijus Dadurkevičius, and Agnė Navickaitė-Klišauskienė. 2024. How lemmatisation and derivational annotation affect productivity measures: The case of deverbal agent nouns in the joint corpus of Lithuanian. *Valoda: nozīme un forma / Language: Meaning and Form* 15:138–151. https://doi.org/10.22364/vnf.15.09.

Kazys Ulvydas, editor. 1965. *Lietuvių kalbos gramatika*, volume 1. Mintis, Vilnius.

Multilingual base word recognition in derivation

Vojtěch John and **Zdeněk Žabokrtský**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
{john, zabokrtsky}@ufal.mff.cuni.cz

Abstract

Given the scarcity of reliable derivational resources and the time- and work-consuming nature of their manual creation, reliable automatic recognition of base words is an important task. We have conducted to our knowledge the most extensive multilingual experiment on the task, using 28 datasets from the Universal Derivations project for a total of 20 languages. Our neural networks (mostly using the Transformer architecture) have achieved state-of-the-art results in both base word generation and base word recognition.

1 Introduction

Manual creation or improvement of derivational resources is laborious and time-consuming. As a result, the available manually created derivational resources are usually quite small. There have already been several attempts to improve the situation by generating word-formation networks using automatic or semi-automatic methods (Lango et al., 2018; Svoboda and Ševčíková, 2024). Nevertheless, automated word-formation analysis remains largely unexplored.

In our experiments, we have concentrated on base word generation and base word induction, as systems solving this subtask are typically the central component of word-formation network induction methods. We have trained models evaluating parent-child relations between lexemes (word relation classifiers) as well as models that generate parent lexemes for the input (parent generators) for 28 datasets across a total of 20 languages, achieving state-of-the-art performance.

2 Related work

While there is a considerable and growing number of derivational resources, most of the resources are monolingual and use resource-specific annotation formats. The largest currently available unified multilingual collection of word-formation networks is Universal Derivations (Kyjánek et al., 2019) – UDer, which in its latest release contains 28 word-formation networks across a total of 20 languages. Although UDer offers a unified format of resources, the data still retain resource-specific characteristics, including vast differences in scope and quality of resources. There are also differences in generality. For example, CroDeriV (Šojat et al., 2014), although a high-quality resource, covers only Croatian verbs, which makes it less useful as training data for generalized base word identification.

Some researchers have already attempted to remedy the lack of resources through automatic or semi-automatic generation, mostly using rule-based and language-specific methods (see, e.g., Vodolazsky (2020) for Russian or Knigawka (2022) (combined with automatic morpheme segmentation) for English). Baranes and Sagot (2014) use inflectional lexicons and automatically extracted transformation rules to generate word-formation networks for four languages, although it is expected that the method is largely language-independent. To our knowledge, recent advances in supervised morphological segmentation and morpheme classification(e.g., Bolshakova and Sapin 2022), have not been used for word-formation networks. This is reasonable since these approaches depend on very rare data: morphologically segmented corpora with classified morphemes. Methods based on supervised or semi-supervised machine learning are still rather rare, but usually quite successful. For example, Lango et al. (2018) presented a method for

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 27–36, Fribourg, Switzerland, 4-5 September 2025.

semi-automatic induction of derivational networks using only a lexicon and a small training set, whereas Vidra and Žabokrtský (2023) used supervised transfer learning to construct word-formation networks for under-resourced languages. More recently, Svoboda and Ševčíková (2024) introduced PaReNT, a collection of joint derivational parent retrieval and derivational edge classification models. This tool was trained on large datasets and is available in seven languages.

Svoboda and Ševčíková (2024) use word embeddings to capture meanings of words. As derivational affixes tend to correspond to relatively regular shifts in embedding spaces (Musil et al., 2019), we expect that the inclusion of word embeddings will increase the accuracy of the model. We will use two sets of embeddings, the traditional FastText (Bojanowski et al., 2017), as it takes into account the subword information, and BPEmb (Heinzerling and Strube, 2018) as it is largely multilingual and is used in Svoboda and Ševčíková (2024), so we will be able to compare our methods more accurately.

3 Data

3.1 Overview

We used data from the resources included in the latest release of UDer (Kyjánek et al., 2019). The datasets adhere to the tree-based (or rather DAG-based) model of derivation, with nodes representing lexemes and oriented edges representing derivational relations. As the datasets differ in quality and generality, so will our models. Since we have no gold data, we will test our models on selected subsets of the resources. Therefore, high accuracy might not actually correspond to high linguistic adequacy. This decision also prevents us from combining the resources. Hence, each model will have the same limitations as its respective resource. On the other hand, we will be able to approximate the specific (possibly mutually incompatible) linguistic decisions taken whilst creating these resources. We have, however, combined the training data (but not the test data) for fine-tuning pretrained multilingual models.

3.2 Data preparation

As we want to test our models' ability to generalize, we will divide our data so that the overlap between word-formation networks in train and test sets will be minimized. This means that the overlap of roots between the training and test sets will be as small as possible (although it can still be non-zero because of compounds).

For base word generation, we randomly select 500 trees for the test set and 100 other trees for the validation set. If the collection includes fewer than 600 trees, we include one half of the trees in the test set and one tenth of the trees in the validation set. Then, we remove the subtrees that appear also in the test set from the training set. For fine-tuning the models, we combine the resources, which might cause overlaps between train and test data for languages with multiple resources. Therefore, for fine-tuning, we excluded instances that appear in any test set. We, however, did not attempt to detect or even resolve inconsistencies across resources, nor did we combine the test sets.

For the classification of word formation relations, we take 5% of the training data instead. We have also sampled negative examples from words present in the same derivational tree, with approximately the same amount of positive and negative examples in each dataset.

We have embedded the source words using the BPEmb multilingual embeddings and FastText language-specific embeddings. The overview of input data is in the second column of Table 2.

4 Word relation classification

In the first experiment, we train neural models on the task of word relation classification. For two words a and b, the models have to decide whether a is a parent of b. In practice, the resulting models would be used for scoring candidate edges in derivational trees or networks. It is not guaranteed that the results will be consistent (for example, the model might produce a cycle) and they need the candidate words as input. Nevertheless, as the quality of these classifiers sets an upper boundary on the quality of resulting derivational databases and since there are different strategies for their use in building derivational resources, we will treat them separately, not as components of a large framework.

4.1 Preliminary ablation study

We have performed a preliminary ablation study on a subset of the Universal Derivations collection (including 16 resources covering 13 languages), using four model architectures. Words are fed to the models as padded lists of characters including also reverse forms of the words. For example, the input for the word *black* would look something like *b, l, a, c, k, 0, 0, 0, k, c, a, l, b*. We have exclusively used pretrained monolingual FastText embeddings of 300 dimensions.

In this preliminary study, we trained the models jointly on two related tasks – the models decided both whether *a* is a parent of *b* (*Parent*) and whether *a* and *b* are relatives, i.e., if they are in the same derivational tree (*Relative*). The architectures expect different inputs and have different initial layers, but the two classification heads are always similar. The *Parent* head consisted of three dense layers (sized 2048, 2048, and 512 respectively) with dropout on the last one and the final sigmoid classification layer; the *Relative* head is similar, but without the first two wide dense layers. As the *Relative* task proved too easy (due to the nature of the automatically generated negative examples), it was dropped in the second round of training, which included only the best-performing architecture – in practice, it meant dropping the second classification head.

Firstly, the baseline *Simple* architecture is a simple two-headed classifier with three hidden dense layers in the *Parent* head, which gets dot product of embeddings of the child and the candidate parent and Levenshtein edit distance between them.

Secondly, in the *Cosine* architecture, the input consists of the two words and cosine distance between the embeddings and the two words. The two words are embedded and processed using one ResNet block each. The outputs of the ResNet blocks are concatenated and processed by a bidirectional LSTM.

The *Full* architecture accepts the two words and their FastText embeddings; the words are embedded and processed by ResNet blocks, while the embeddings are run through by a dense layer with dropout, multiplied and then processed by a convolutional layer.

Finally, the *Subtract* architecture expects as input the difference between the FastText embeddings of the words and both candidate words. Otherwise, the process applied to words and embeddings is the same as in the *Full* architecture.

As seen in Table 1, the *Subtract* architecture was by far the most successful. This is interesting, as it has access to less information than the *Full* architecture. Hence, it seems that the difference between the parent and child vectors does capture important information about the derivational process which generated the child from the parent. This corresponds to the observation by Musil et al. (2019) that meanings of derivational affixes can be captured by the difference between the base word and the derived word.

Setting	Binary accuracy	Precision	Recall
Cosine	88.81%	74.42%	77.43%
Simple	78.91%	46.92%	51.24%
Subtract	93.05%	88.24%	83.87%
Full	87.45%	73.26%	79.05%

Table 1: Ablation study of word relation classifiers performed on a subset of the datasets. The results are macro-averaged across resources.

4.2 Methods

In the final version of the *Subtract* architecture, we have introduced some modifications. We feed the difference of token-and-positional embeddings of the candidate words to a Transformer decoder block, using the difference between the FastText embeddings as encoder input and the difference between word embeddings as decoder input. For classification, we use three dense layers of decreasing size (1024, 256, and 1) on the decoder output. We train for at most 20 epochs using early stopping on non-decreasing loss.

4.3 Results

We have evaluated the models using precision and recall. The scores achieved are very high, with average precision of 91.2% and average recall of 90.8%. It is to be noted that changes in the architecture (e.g., introducing the Transformers decoder block) seem to have helped quite significantly. It appears that the results do not depend much on the size of the data but rather on the quality and diversity of the datasets. The results are presented in Table 2.

Resource	Size	Subtract	
		Precision	Recall
et-EstWordNet	988	89.5	96.0
pt-EtymWordNetPT	2797	91.8	90.1
ru-EtymWordNetRU	4005	86.2	94.6
ru-GoldenComp...	4570	97.2	99.0
hr-CroDeriV	5092	83.5	90.0
pt-NomLexPT	7020	98.0	99.3
sv-EtymWordNetSV	7333	93.8	96.7
ca-EtymWordNetCA	7496	94.1	97.2
gd-EtymWordNetGD	7524	79.5	93.0
cs-EtymWordNetCS	7633	89.0	98.0
tr-EtymWordNetTR	7775	92.7	96.9
sh-EtymWordNetSH	8033	94.5	94.6
it-DerIvaTario	8265	86.9	95.1
en-WordNet	13810	98.4	97.3
fi-FinnWordNet	20032	97.9	99.8
fr-Demonette	22057	90.8	92.6
pl-EtymWordNetPL	27797	93.8	99.0
la-WFL	36204	85.6	95.0
fa-DeriNetFA	42258	87.5	86.0
sl-Sloleks	48054	96.5	99.1
en-CatVar	82667	90.6	55.8
hr-DerivBaseHR	99587	83.5	90.0
es-DeriNetES	151173	87.9	93.7
pl-PolishWFN	262887	95.9	97.4
ru-DerivBaseRU	270473	94.2	91.8
de-DErivBase	280775	86.8	61.9
ru-DeriNetRU	337584	79.1	64.5
cs-DeriNet2	1038991	91.1	87.7
Macro-average		91.2	90.8

Table 2: Overview of data resources and word relation classification results. The resources are sorted by size.

5 Base word generation

5.1 Methods

We have experimented with two methods of base word generation. Firstly, we use small Transformer models provided by the OpenNMT package (Klein et al., 2017) with several ablations. The *Basic* model uses the *TransformerTiny* architecture with no embeddings and no custom settings. We test the following ablations: *Big* model with increased capacity, embeddings as an additional input, FastText in the *FastText* models, and BPEmb in the *BPEmb* ones. Again, we train with early stopping (until the perplexity ceases to increase).

Secondly, we perform fine-tuning of pretrained multilingual models. Tokenization algorithms in a multilingual setting tend to favor languages with large datasets. We have therefore decided to work with ByT5 transformer models, which operate directly on UTF-8 bytes. These models achieved competitive results against comparable transformer models using subword tokenization. The models were pretrained on 101 languages (Xue et al., 2022). We have used the *small* and *base* versions, fine-tuning both for 5 epochs on all the datasets together.

5.2 Evaluation metrics

For base word generation, we have used word-level accuracy, i.e., the count of correctly predicted base words divided by the number of test instances. We take our resources as gold data including the missing links – the words that have no base word in the resources are regarded as unmotivated. This is obviously not always the case, so caution should be exercised in interpreting the results, especially with regard to the most incomplete databases. We have decided to include the unmotivated words anyway in order to avoid overgeneration – retrieving base words where there really are none.

5.3 Results

The results are presented in Table 3. Among the models trained from scratch, the worst results were achieved by *BPEmb* models, while *FastText* embeddings generally achieved the best results. It is quite unexpected that the inclusion of BPE word embeddings would worsen the performance. However, it might be caused by low quality of BPEmb embeddings or by the fact that they do not take into account subwords to the extent FastText embeddings do. The second explanation would mean, slightly counterintuitively, that the models mostly use the FastText embeddings for information about words' inner structure (as opposed to semantics). Also, as enlarging the model did not improve the performance even on large datasets, model size does not appear to be a limiting factor. Because the improvement caused by FastText embeddings is only slight, the same might hold for inclusion of semantic representations.

The fine-tuned ByT5 models performed better on average compared to the models trained from scratch. These models were trained on all the data jointly, using only language tags, so the results across resources from the same languages can be influenced by each other. Nevertheless, it seems that this is not always the case or that incompatibilities between resources might even cause a decrease in performance (e.g., for Portuguese or Russian). Interestingly, the largest improvement to the OpenNMT models is observed on EtymWordNets and other very small resources. This is especially interesting for languages with no other resource included in the training data.

We have compared our results with those of PaReNT (Svoboda and Ševčíková, 2024), with two PaReNT search variants – *greedy* and *best*. Since PaReNT was trained on multiple resources not present in UDer, it does not approximate the UDer resources and so it may generate a base word which is missing in any given resource (correctly or incorrectly). Therefore, we evaluate the methods for the sake of comparison only on motivated words (unlike in Table 3). PaReNT is trained for 7 languages. However, since we have no data for Dutch, we have compared the performance of our taggers on the test set for 6 languages only – Czech, German, English, Spanish, French, and Russian. As seen in Table 4, our models outperformed PaReNT on our test set in all cases for *best* and all cases except English for *greedy*. The small fine-tuned model achieved the best performance for all languages except Russian. However, this may be explained by the generally poor quality of the Russian data.

Resource	Basic	Big	FastText	BPEmb	Fine-tuned	
					ByT5-small	ByT5-base
et-EstWordNet	42.1	45.5	**46.7**	38.2	67.3	67.7
pt-EtymWordNetPT	63.9	61.5	**67.5**	62.0	52.7	59.3
ru-EtymWordNetRU	11.0	14.8	**16.3**	13.2	20.6	21.3
ru-GoldenComp...	40.6	37.5	**42.7**	0.0	8.6	7.8
hr-CroDeriV	11.6	9.5	**18.9**	9.5	25.0	20.6
pt-NomLexPT	**87.3**	85.2	83.9	84.5	90.9	90.2
sv-EtymWordNetSV	**71.1**	68.4	68.6	66.7	84.2	83.8
ca-EtymWordNetCA	73.5	74.3	**75.9**	71.8	83.7	83.1
gd-EtymWordNetGD	58.5	55.2	**59.8**	56.9	66.7	67.3
cs-EtymWordNetCS	38.1	37.6	**40.4**	34.1	58.1	58.0
tr-EtymWordNetTR	79.2	77.8	**83.7**	75.5	88.2	86.6
sh-EtymWordNetSH	63.9	61.3	**65.4**	60.4	76.9	78.9
it-DerIvaTario	**79.6**	78.2	75.7	77.9	82.3	83.2
en-WordNet	**52.4**	50.9	52.0	50.1	67.9	65.1
fi-FinnWordNet	52.5	51.0	**55.6**	49.2	60.6	61.6
fr-Demonette	**70.8**	70.3	69.4	69.3	93.0	93.8
pl-EtymWordNetPL	79.6	75.5	**83.7**	75.3	85.5	85.2
la-WFL	**41.4**	38.3	40.3	37.3	50.1	52.1
fa-DeriNetFA	68.4	65.7	**70.8**	67.1	73.4	72.0
sl-Sloleks	**91.5**	91.2	90.3	91.1	93.3	94.2
en-CatVar	75.0	73.9	**77.8**	75.6	77.4	77.0
hr-DerivBaseHR	65.4	68.6	**69.6**	68.9	71.4	73.3
es-DeriNetES	80.7	**81.8**	80.1	79.8	84.8	85.3
pl-PolishWFN	**75.5**	71.3	73.2	72.9	76.4	75.0
ru-DerivBaseRU	73.7	73.9	**75.9**	74.9	59.0	58.0
de-DErivBase	86.6	87.1	87.0	**87.5**	91.0	90.7
ru-DeriNetRU	28.8	29.1	**38.5**	26.4	33.5	33.9
cs-DeriNet2	81.1	80.4	**81.2**	80.9	80.9	80.5
Macro-average	62,3	61.3	**64.0**	59.2	68.0	68.0

Table 3: Accuracy of base word generation. The resources are sorted by size.

Resource	ByT-small-finetuned	Emb	PaReNT best	PaReNT greedy
cs-DeriNet2	80.1%	81.7%	38.8%	43.7%
de-DErivBase	71.7%	41.4%	22.2%	19.2%
en-CatVar	54.7%	47.0%	46.6%	53.8%
es-DeriNetES	71.8%	59.0%	42.6%	45.2%
fr-Demonette	93.0%	54.7%	35.9%	35.5%
ru-DerivBaseRU	21.5%	64.0%	29.0%	32.8%
Average	65.2%	58.0%	35,9%	38.4%

Table 4: Comparison with PaReNT results: accuracy for motivated lemmata.

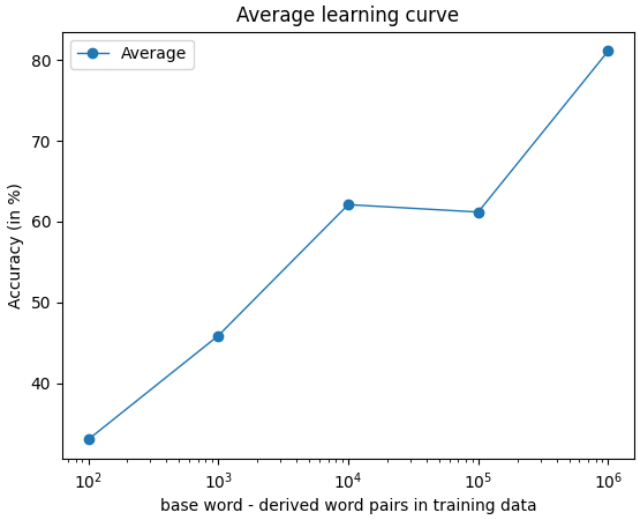


Figure 1: Learning curve. Averaged performance across resources. The x-axis stands for the size of training data on a logarithmic scale, while the y-axis corresponds to accuracy.

We have plotted learning curves for the *FastText* models. The learning curves for individual resources differ considerably in starting point, steepness and even direction (Figure 2), although generally the fastest growth can be observed between 100 and 10,000 training examples. Although the learning curves appear messy, they allow some instructive observations. For automatically generated resources (All DerivBases, Spanish and Russian DeriNets), the increases are very slow or even nonexistent – 100 examples is almost as good as 100,000, sometimes better. This is probably caused by the fact that the automatic methods used for building these resources detect only easily recognizable derivational relations. Accordingly, for these resources, we also find the lowest scores for the least consistent resources (e.g., the Russian DeriNet). This goes a long way to explain low scores on languages with big data (like Russian).

Generally, the scores achieved by models trained only on 100 examples seem to depend on the complexity of derivational relations captured by the resource and on the quality of the resource. Hence, large, manually or semi-automatically created resources tend to score lower than automatically generated resources or resources capturing only simple derivational relations. There may be some relation to language typology, as witnessed by different results on EtymWordNets, where Slavic languages consistently perform worse and improve more slowly than the rest of the languages. Unfortunately, as we know of no reliable metrics of word-formational complexity, we were unable to verify this suspicion. Nevertheless, the general impression is that comparatively small data are enough for achieving reasonable results on word-formationally simple languages, while for more complex languages, several hundred thousand training examples might be required (as seems to be the case for Czech DeriNet).

5.4 Error analysis

While the results for word-formation relation classification are consistently high, the accuracy of generated base words varies across the resources. One factor that appears to play a rather large role is the size of training data, as seen in Figure 1. However, this is not the catch-all explanation, as for some small datasets, the results are surprisingly good. For example, the Portuguese NomLexPT has the second highest accuracy even though it is among the smallest resources. The results also often vary between different resources for the same language. This can be explained by the difference in size of training data in the cases of Czech and Croatian, but not in the case of Russian, where results for similarly sized resources (DeriNet and DerivBase) differ by approximately 45%.

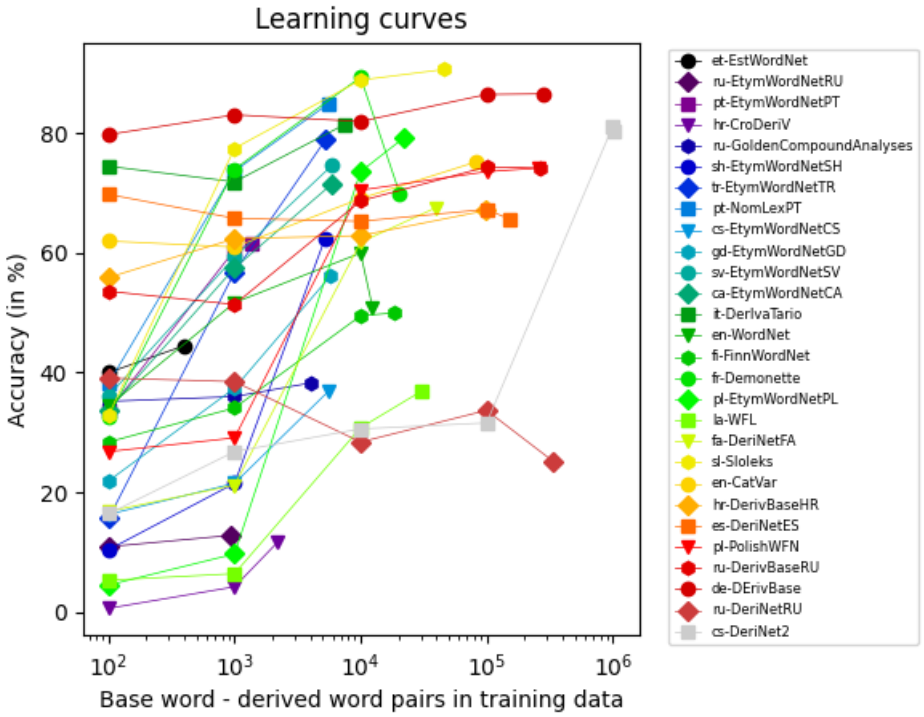


Figure 2: Learning curves by resource.

5.4.1 Selected error types

Although an overview of the most common errors in base word generation can be found in Svoboda and Ševčíková (2024), there are still some comparatively common errors not covered there which warrant discussion. Firstly, since the datasets we use also contain errors and, more critically, miss many edges, the models sometimes predict a correct derivational parent for a word which is either marked as unmotivated or wrongly motivated in the resource. This concerns mainly the automatically formed resources. For example, *studiously*, which has no parent in CatVar, has been correctly identified as a child of *studious*, and *revealing*, wrongly marked as a child of *revel*, was correctly assigned to *reveal*.

Some errors are caused by the models not being able to identify conversion, which nevertheless appears in the training and test data (e.g., *festering* can be regarded as a child of *fester* by derivation or *festering* by conversion); this might be resolved in future work either by incorporating part-of-speech tags or by avoiding conversion altogether.

Quite often, the models choose wrong order of word-formation operations; mostly, they tend to retain prefixes and change suffixes. As a result, *nečinný* ('inactive'), for example, is proposed as parent of *nečinnost* ('inaction') instead of *činnost* ('action'); while this case seems quite plausible, there are many completely implausible examples of the same (e.g., *západopennsylván* as derivational parent for *západopennsylvánský*, which is in fact a compound of *západ* and *pennsylvánský*).

Another common error occurs when the model segments off an affix, generating a more or less plausible but nonexistent base word. Sometimes (albeit rarely), the base word did exist but is defunct (e.g., *svat* as a parent of *svatba*, which is really derived from Proto-Slavic *svat*, see Rejzek 2015), while in the case of borrowings or neoclassical compounds it might exist in a different language (e.g., *bióza* as a parent of *antibióza*). Such cases raise an interesting theoretical question. Should derivational resources include words that no longer exist in the language, or should they be entirely synchronic? In practice, if we want to retain only synchronically present words, we could filter the outputs of the models with the help of dictionaries or large corpora. It should be noted that the models hallucinate much more often than they come up with a correct etymology.

Finally, there is the issue of word variants and their derivatives. The datasets resolve these in different

ways. Perhaps most notably, in the Czech DeriNet, the variants are children of the dictionary versions of given words, and nothing is derived from them. Accordingly, sometimes most or all descendants of the dictionary version of word *A* have variants that look as if they were derived from the same variant of *A*. While this decision is well-founded (e.g., we definitely want words with the same meanings and very similar forms close to each other in the tree), it confuses the models, which tend to generate parallel subtrees. Hence, for example, the correct parent for *oučinkování* is *účinkování*, instead of the predicted *oučinkovat*.

5.4.2 Fine-tuned models and multilinguality

The performance of both fine-tuned models on different datasets for the same language is dissimilar. As in the fine-tuning data the datasets are distinguished only by language, this might point to differences in either resource reliability or linguistic decisions taken during their creation. Some of the errors might also be introduced by edges missing in one and only one of the resources. Cutting off ends of words is a surprisingly common error, as well as hallucinations, adding or removing several letters without clear reason. Also, these models have harder time guessing which words are unmotivated in which resource, but perform significantly better on predicting parents that are present in the dataset – when evaluating only on parent-child pairs with parents present in datasets, where some parent was in fact predicted by the models, results of the small fine-tuned model are usually much better than the results of the models trained from scratch. They also tend to be better than results on the whole dataset, including unmotivated words. Therefore, even though nominally the results of fine-tuned models and models trained from scratch are quite similar, the fine-tuned models are more reliable.

6 Conclusion and future work

As the size of derivational resources tends to be rather limited, the problem of automatic induction of derivational relations is of great importance for morphological processing of large datasets. To address this issue, several approaches have been proposed, including parent-child relation labeling or derivational parent generation. Using data from Universal Derivations, we have trained state-of-the-art parent generators for 21 languages, achieving on average 64.0% accuracy. Furthermore, we have fine-tuned a multilingual model on the same task, with even better average performance (68.0% accuracy). We have also trained word-formation relation classifiers with 91.2 % precision and 90.8 % recall. The results, although state-of-the-art, leave much room for further research, especially in the case of languages with very small resources. It would also be interesting to investigate whether mostly accurate morphological segmentation (possibly with morpheme classification) could improve the performance of the generators.

Acknowledgements

This work has been supported by the Charles University Research Center program No. 24/SSH/009, by the project GA UK No. 101924, by the SVV project number 260 698. The authors thank anonymous reviewers for their insightful feedback.

References

Marion Baranes and Benoît Sagot. 2014. A language-independent approach to extracting derivational relations from an inflectional lexicon. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, pages 2793–2799. https://aclanthology.org/L14-1327/.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. In Lillian Lee, Mark Johnson, and Kristina Toutanova, editors, *Transactions of the Association for Computational Linguistics*. MIT Press, Cambridge, MA, volume 5, pages 135–146. https://doi.org/10.1162/tacl_a_00051.

Elena I. Bolshakova and Alexander S. Sapin. 2022. Building a combined morphological model for Russian word forms. In Evgeny Burnaev, Dmitry I. Ignatov, Sergei Ivanov, Michael Khachay, Olessia Koltsova, Andrei Kutuzov,

Sergei O. Kuznetsov, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Jari Saramäki, Andrey V. Savchenko, Evgenii Tsymbalov, and Elena Tutubalina, editors, *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*. Springer International Publishing, Cham, pages 45–55.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki. https://aclanthology.org/L18-1473/.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Mohit Bansal and Heng Ji, editors, *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, pages 67–72. https://aclanthology.org/P17-4012/.

Łukasz Knigawka. 2022. Constructing a derivational morphology resource with transformer morpheme segmentation. In Robin Schaefer, Xiaoyu Bai, Manfred Stede, and Torsten Zesch, editors, *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. KONVENS 2022 Organizers, Potsdam, pages 104–109. https://aclanthology.org/2022.konvens-1.12/.

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. 2019. Universal derivations kickoff: A collection of harmonized derivational resources for eleven languages. In Magda Ševčíková, Zdeněk Žabokrtský, Eleonora Litta, and Marco Passarotti, editors, *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, pages 101–110. https://aclanthology.org/W19-8512/.

Mateusz Lango, Magda Ševčíková, and Zdeněk Žabokrtský. 2018. Semi-automatic construction of word-formation networks (for Polish and Spanish). In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1291/.

Tomáš Musil, Jonáš Vidra, and David Mareček. 2019. Derivational morphological relations in word embeddings. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, pages 173–180. https://aclanthology.org/W19-4818/.

Jiří Rejzek. 2015. *Český Etymologický Slovník [Czech Etymological Dictionary]*. Leda, Voznice.

Krešimir Šojat, Matea Srebačić, Marko Tadić, and Tin Pavelić. 2014. CroDeriV: A new resource for processing Croatian morphology. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, pages 3366–3370. https://aclanthology.org/L14-1057/.

Emil Svoboda and Magda Ševčíková. 2024. PaReNT (parent retrieval neural tool): A deep dive into word formation across languages. In Nicoletta Calzolari et al., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Turin, pages 12611–12621. https://aclanthology.org/2024.lrec-main.1104/.

Jonáš Vidra and Zdeněk Žabokrtský. 2023. Transferring word-formation networks between languages. *The Prague Bulletin of Mathematical Linguistics* (120):47–71.

Daniil Vodolazsky. 2020. DerivBase.Ru: A derivational morphology resource for Russian. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), Marseille, pages 3937–3943. https://aclanthology.org/2020.lrec-1.485/.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. In Brian Roark and Ani Nenkova, editors, *Transactions of the Association for Computational Linguistics*. MIT Press, Cambridge, MA, volume 10, pages 291–306. https://aclanthology.org/2022.tacl-1.17/.

LexEco: Exploring how derivational morphology contributes to the semantics of French nouns

Lucie Barque
Université Sorbonne Paris Nord & LLF (Université Paris Cité, CNRS)
lucie.barque@univ-paris13.fr

Abstract

In this study, we introduce LexEco, a comprehensive lexical resource designed to support empirical research on the French nominal lexicon. It includes nearly 19,000 nouns, primarily drawn from existing databases, and selected to reflect a broad and representative sample of predominantly familiar vocabulary. Each entry is annotated with morphological, semantic, and corpus-based frequency information, allowing for detailed analyses of the relationships between form and meaning. As a case study, we examine the role of suffixation in meaning formation and assess how morphologically derived meanings compare to those in the simplex lexicon. The results indicate that while all major semantic categories are represented in both simplex and derived nouns, each group exhibits distinct semantic tendencies that allow for a certain degree of predictability in word meaning based on morphological structure. Simplex and suffixed nouns also differ in their ambiguity profiles: simplex nouns tend to have a greater number of meanings than suffixed nouns. This difference can be partly attributed to their semantic characteristics, as concrete meanings – proportionally more frequent among simplex nouns – tend to be more prone to semantic extension than abstract meanings, which are more common among suffixed nouns.

1 Introduction

Derivational morphology, alongside polysemy, is a key mechanism in the construction of lexical meaning. Some linguistic approaches attribute to it a complementary role, viewing it as a means of generating meanings different from those typically associated with unmarked elements of a given grammatical class. For instance, according to (Croft, 1991), nouns prototypically denote concrete objects; when they are used to express less typical meanings, such as actions or properties, they are generally morphologically marked across languages. In the case of French, several empirical studies have focused on the semantic properties of specific morphological categories of nouns. While most simplex nouns do denote concrete objects, a non-negligible proportion encodes other types of meanings (Tribout et al., 2014; Huyghe et al., 2017). The semantic behavior of suffixed nouns – particularly deverbal ones – has also been widely investigated, with research examining their aspectual and argumental properties (e.g. Balvet et al., 2011; Huyghe et al., Submitted), affix competition (e.g. Guzmán Naranjo and Bonami, 2023; Huyghe et al., 2023), and their ambiguity patterns (Salvadori, 2024).

However, to our knowledge, no previous study has offered a comprehensive overview of the nominal category that would allow for a systematic assessment of relationships between morphological structure and meaning within the lexicon. The recent availability of large-scale lexical resources for French, enriched with both morphological and semantic information, offers fertile ground for such empirical investigations. This is the motivation behind LexEco, a new morpho-semantic database developed to meet the need for resources that comprehensively reflect the structure of the French nominal lexicon, enabling in-depth analysis of its morpho-semantic properties. Inspired by Échantinom, a morphological resource created to offer a representative sample of French nouns (Bonami and Tribout, 2021), LexEco expands the scope of noun sampling with an emphasis on familiar nouns and, importantly, incorporates a detailed semantic categorization.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 37–47, Fribourg, Switzerland, 4-5 September 2025.

N	Freq	Fam	N	Freq	Fam
bétel 'betel'	1.54	30%	tendinite 'tendinitis'	0.12	100%
gandin 'dandy'	0.92	25%	peaufinage 'refinement'	0.1	100%
trèpe 'huddle'	0.74	19%	physionomiste 'face reader'	0.1	100%
vertex 'vertex'	0.61	17%	luxembourgeois 'Luxembourger'	0.1	100%
boutéon 'mess tin'	0.57	3%	fluor 'fluoride'	0.06	100%
voussure 'arch'	0.41	19%	déforestation 'deforestation'	0.02	100%

Table 1: Morphological information associated with nouns in LexEco.

In the remainder of this paper, we first detail the methodology used to compile a large sample of nearly 19,000 French nouns, annotated with morphological and semantic information primarily sourced from existing databases, along with key statistics describing this resource (Section 2). We then leverage this lexicon[1] to perform a comparative analysis of the semantic properties of simplex versus complex nouns, aiming to elucidate the role of derivational morphology in shaping nominal meaning (Section 3). Our findings reveal that although all major types of nominal meanings occur in both simplex and derived noun sets, distinct semantic tendencies within each group allow for a meaningful prediction of a noun's semantic profile based on its morphological structure. Additionally, we identify differences in the patterns of ambiguity characterizing nouns across the two morphological classes.

2 Building a morpho-semantic lexicon of frequent French nouns

The nouns included in LexEco are drawn from Lexique-3, a comprehensive lexical database that provides morpho-syntactic and frequency information derived from two extensive corpora of contemporary French: one literary corpus comprising 31 million words, and another composed of 52 million words from film subtitles (New et al., 2004, 2007). Altogether, Lexique-3 lists 32,066 unique nominal lemmas. Our primary objective was to define a large nominal set that is representative of everyday French, i.e., words that are familiar to most adult French speakers. Instead of relying on a fixed textual frequency threshold, we leveraged information from the "deflem" section available in Lexique-3. This information is not derived from a formal norming study. Instead, it is based on data collected from an informal online survey initiated by a user, with no details on its methodology. As such, it cannot be considered a reliable source on French speakers' lexical knowledge, but remains an effective method for filtering out most obscure nouns in Lexique-3, as illustrated by the 6 nouns randomly selected from those known by less than 33% of respondents (left column of Table 1). At the same time, it preserves commonly known nouns that might otherwise be excluded by the textual frequency strategy, such as those randomly selected nouns known by 100% of respondents despite having a textual frequency below 0.3 word per million in the corpora (right column of Table 1).

LexEco includes nouns from Lexique-3 that meet two criteria: *i)* they have a nominal entry in the French Wiktionnaire, and *ii)* they are known by at least 50% of the respondents. The first criterion ensures that semantic information – drawn from the Wiktionnaire (see Section 2.2) – is available and helps exclude words that do not correspond to standard common nouns. The second criterion guarantees that the sample remains "ecological", which means that it reflects vocabulary commonly known by French speakers. This process results in 18,979 nominal lemmas having both textual frequencies (Mean = 16.5, SD = 77.7) and familiarity ratings (Mean = 88.5, SD = 13.2).

2.1 Morphological description

To provide nominal lemmas with derivational information, we first made use of available data in existing morphological resources. In total, 14,898 nouns (approximately 78% of the dataset) were identified in

[1]LexEco is available at https://osf.io/cd5hf/?view_only=d58fb3f51a0c4afb881d8fb156613cf4.

existing resources, distributed as follows: 2,351 nouns from The Simplex Lexicon (Tribout et al., 2014), 3,274 from Échantinom (Bonami and Tribout, 2021), 1,513 from Sonde (Huyghe et al., Submitted), and 7,760 from Démonette-2 (Namer et al., 2023).[2] Morphological descriptions for the remaining 22% of nouns had to be created from scratch.

The morphological encoding adhered to the guidelines established for constructing Échantinom (Bonami and Tribout, 2021). Specifically, LexEco indicates whether a noun is simplex (e.g., *cou* 'neck') or derived, through one of the following morphological processes: suffixation (e.g., *embrassade* 'kissing' < *embrasser* 'to kiss'), prefixation (e.g., *irrespect* 'disrespect' < *respect* 'respect'), conversion (e.g., *réveil* 'wake-up, alarm clock' < *réveil* 'réveiller'), compounding (e.g., *cerf-volant* 'kite'), or non-concatenative processes (e.g., *resto* 'restaurant'). For affixation, the annotation specifies the suffix and/or prefix involved, along with the morphological base and its part of speech (e.g., *embrassade* is derived from the verb *embrasser* with the suffix *-ade*). In the case of complex nouns that have undergone multiple derivational processes, as in *malchanceux*$_N$ 'an unlucky person', the assumed final process is encoded (conversion in this case), along with the remaining components of the composition. Table 2 presents the morphological information for the seven nouns mentioned above, as encoded in LexEco.

	cstr	**suff**	**suff_norm**	**pref**	**conv**	**conv_pos**	**aff_base**	**aff_pos**
cou	simplex	0	0	0	0	0	0	0
embrassade	suffixed	ade	ade	0	0	0	embrasser	V
irrespect	prefixed	0	0	in	0	0	respect	N
réveil	convert	0	0	0	réveiller	V	0	0
cerf-volant	coumpound	0	0	0	0	0	0	0
boudeuse	suffixed	euse	eurM	0	boudeur	A	bouder	V
malchanceux	convert	eux	eux	mal	malchanceux	A	chance	N

Table 2: Morphological information associated with nouns in LexEco.

It is important to emphasize that neither the overall quality of the morphological annotation – particularly for the 22% of nouns absent from existing morphological resources and partially annotated using heuristic rules (e.g., treating hyphenated nouns as compounds, or labeling as "converted" those nouns also tagged as adjectives in Lexique-3) – nor its reliability has yet been formally assessed. By *reliability*, we refer to the internal consistency of the encoding, both within each individual resource and across these different primary sources. Indeed, although a substantial portion of the morphological encoding is drawn from manually curated resources, two main limitations remain. First, with the exception of Sonde (Huyghe et al., Submitted), the primary resources have not undergone inter-annotator agreement evaluation, leaving the reliability of their annotations uncertain. Second, analytical choices regarding the treatment of morphological phenomena may vary between projects, potentially leading to inconsistencies in how derivational information is encoded. For instance, the "convert" value in the *cstr* (construction) feature is intended to apply only to nouns that result from a morphological conversion process, which is particularly difficult to assess in most cases (Marchand, 1964; Balteiro, 2007; Tribout, 2020, a.o.). As mentioned above, regarding the occasional use of automatic rules for nouns not included in any of the four morphological resources, we know that this analytical choice has not been applied consistently. Other well-known morphological challenges, such as determining whether a noun is a compound or a prefixed form (e.g., *épiphénomène* 'epiphenomenon'), or whether it is suffixed when derived from a non-autonomous base (e.g., *ablation* 'ablation'), have likely resulted in inconsistencies in the LexEco encoding. To date, approximately one quarter of the nouns in the database have undergone manual verification of their morphological encoding.[3]

[2]Given that the four morphological resources employed share a common subset, data extraction was carried out following a predefined order, prioritizing resources that provide information on simplex nouns and those considered as primary sources. While Démonette-2 accounts for the majority of morphological descriptions, these stem from the aggregation of several morphological resources, rather than from independent, original annotation.

[3]The revised subset, which includes both suffixed and simplex nouns, was reviewed by the four authors of a study on

2.2 Semantic description

The semantic data associated with nouns in LexEco is derived from the SuperWikt-fr lexicon (Angleraud et al., 2025), a version of the French Wiktionary in which all nominal word senses have been automatically annotated with their general semantic class. This annotation is provided at two levels of granularity. At the *supersense* level, noun senses are grouped into 23 semantic classes, including 20 simple classes (e.g., Act, Person, Plant) and 3 complex classes (Act*Cognition, Artifact*Cognition, and Group×Person). At the broader *hypersense* level, senses are categorized into 9 overarching semantic categories (e.g., Inanimate_entity, Animate_entity, Dynamic_situation). For instance, the two senses of *lave-glace* ('windshield washer') are annotated with both a supersense and a hypersense, informed by their lexicographic definition and examples.

(1) LAVE-GLACE ('windshield washer')

a. (Automobile) Dispositif qui envoie du liquide nettoyant sur le pare-brise et, en option, sur la lunette arrière ou les optiques. Artifact - Inanimate_entity
'(Automotive) Device that sprays cleaning fluid onto the windshield and, optionally, onto the rear window or headlights.'

b. (Par métonymie) Liquide lave-glace. ex. *Notre antigivre permet de réduire le gel du lave-glace sur le pare-brise, en hiver.* Substance - Inanimate_entity
'(By metonymy) Windshield washer fluid. e.g., *Our antifreeze reduces the freezing of the windshield washer on the windshield during winter.*'

The semantic annotation was performed using supervised classifiers trained and evaluated on a large set of manually curated data. The automatic classification achieved a precision of nearly 85% at the supersense level and nearly 92% at the hypersense level, with performance varying across semantic categories. The easiest senses to classify belong to the Animate_entity class (F-score: 97.7%), while the most challenging ones are from the Informational_object class (F-score: 77.5%).

2.3 Statistics

Table 3 summarizes the morphological properties of the 18,979 nouns, and their associated 56,590 senses, with semantic properties outlined. As for morphological properties (left), the data indicate that the vast majority of nouns in LexEco are morphologically derived, primarily through suffixation (47%), while approximately one quarter are morphologically simplex (26%). These proportions differ markedly from those observed in Échantinom (Bonami and Tribout, 2021), despite both being sampled from the same Lexique-3 resource, where simplex and suffixed nouns account for 41% and 37%, respectively. A likely explanation is that nouns excluded from LexEco due to the "familiarity" constraint are predominantly morphologically simplex (see Table 1, left column). As for semantic properties (right), the data indicate that nouns in LexEco mostly denote Inanimate entities (34%), and then Animate entities (19%), which is in line with what is generally presented as the prototypical semantic class of nouns (Lyons, 1977; Wierzbicka, 1986; Croft, 1991, a.o.). However, the hypersense distribution reveals a fairly even split between nouns denoting concrete entities (Inanimate + Animate, 53%) and those representing abstract entities (Dynamic_situation, Informational_object, etc., 47%).[4]

3 Case study: how does derivational morphology shape nominal meaning?

In this section, we illustrate how LexEco can be used to investigate the relationships between the morphological and semantic properties of nouns. Previous studies have offered valuable insights into the semantics of simplex French nouns and their contrasts with complex nouns, particularly deverbal forms (Tribout et al., 2014; Huyghe et al., 2017). The data provided by LexEco make it possible to extend these initial analyses to a broader and more morphologically diverse set of nouns.

pseudo-suffixed French nouns (Barque et al., Submitted).

[4]Unfrequent hypersenses (grouped in the Other class in Table 3) are: Time, Possession, Institution, Quantification, Dyn_sit*Info, Inanimate*Info, Quantity×Animate.

	Nb of lemmas	%
Suffix	8,939	47
Simplex	4,927	26
Conversion	3,677	19
Compounding	816	4
Prefix	308	2
Nonconcat	312	2
	18,979	100

	Nb of senses	%
Inanimate_entity	18,945	34
Animate_entity	10,825	19
Dynamic_situation	10,816	19
Stative_situation	5,470	10
Informational_object	5,422	10
Other	5,112	8
	56,590	100

Table 3: Distribution of nouns by types of morphological processes (left) and distribution of nominal senses by hypersenses (right) in the dataset. Hypersenses with a representation of less than 3% are grouped under the label `other`.

3.1 Focus on simplex vs. suffixed nouns

For these analyses, the data is restricted to clear-cut cases of simplex and suffixed nouns. A conservative approach to conversion is adopted, given the well-documented challenges in determining the directionality of conversion (Tribout, 2020). Consequently, any noun with a converted counterpart – whether simplex (e.g., *jeune*$_{\text{N}}$ – *jeune*$_{\text{A}}$ 'young') or complex (e.g., *arriviste*$_{\text{N}}$ – *arriviste*$_{\text{A}}$ 'social climber') – is excluded from the analysis, regardless of whether there is evidence that the noun is the primary form in the conversion relation.[5] General statistics of this reduced sample are provided in Table 4.

	Total N	Monosemous N	Ambiguous N	Senses	Mean Ambiguity Level	Freq
Simplex N	3,971	1,202	2,769	12,802	3.2	28.3
Suffixed N	8,007	2,887	5,120	21,380	2.6	7.1
Total	11,978	4,089	7,889	34,182	2.8	14.1

Table 4: Statistics of the dataset reduced to simplex and suffixed nouns.

The difference in ambiguity between the two groups (3.2 vs. 2.6) is significant, as revealed by a Mann–Whitney U test ($Z = 9.7$, $p < .001$), indicating that simplex nouns tend to be more ambiguous than suffixed nouns. This difference may partly be an artifact of lexicographic practices, which often describe derived forms relationally, reducing their apparent ambiguity by referencing an already ambiguous base. Alternatively, it may reflect a genuine semantic tendency: derived nouns could be inherently less susceptible to ambiguity than simplex nouns. At least two factors may help explain this tendency. First, ambiguity rates are well known to correlate with frequency (e.g., Zipf, 1945; Piantadosi et al., 2012), a pattern confirmed by our data: simplex nouns are, on average, significantly more frequent than suffixed nouns, as indicated by a Mann–Whitney U test on log-transformed frequencies ($Z = 23.2$, $p < .001$). However, it remains unclear whether frequency drives ambiguity (i.e., the more a word is used, the more meanings it acquires), or whether ambiguity drives frequency (i.e., the more meanings a word has, the more often it is used across contexts). Second, certain semantic types may be more susceptible to semantic extension than others, types that could be more frequently represented among simplex nouns than among derived ones.

Another aspect to consider is that the mechanisms underlying the creation of ambiguous forms differ between simplex and suffixed nouns. In simplex nouns, polysemy typically arises through sense extensions stemming directly or indirectly from the word's primary meaning. In contrast, ambiguity in suffixed nouns may result not only from similar sense extensions based on a morphologically constructed primary

[5]The number of excluded nouns is more important in the simplex group (956/4,927, 19%) than in the suffixed group (932/8,939, 10%).

meaning, but also from morphological co-derivation – i.e., the derivation of multiple forms from a common formal base, which may itself be semantically unique or ambiguous (Rainer, 2014; Bauer, 2017; Salvadori, 2024, a.o.). For instance, the noun *explorateur* 'explorer', derived from the verb *explorer* 'to explore', has one sense referring to a Person and another referring to a software application. However, it is theoretically difficult to determine whether one meaning is derived from the other – since metaphorical extensions between humans and Artifacts are attested in the simplex lexicon – or whether both senses are independently co-derived from the verb, given that *-eur* in French can yield both Animate and Inanimate interpretations.

In what follows, we begin by analyzing semantic tendencies among monosemous nouns. While this subset is more restrictive, it provides a clear and unambiguous foundation for assessing the meanings of suffixed nouns, which can reasonably be assumed to be morphologically derived (Section 3.2). We then move on to compare the ambiguity profiles of simplex and suffixed nouns (Section 3.3).

3.2 Semantic tendencies among monosemous nouns

Table 5 presents the distribution of senses within the two sets of monosemous nouns (4,165 lemmas), among which 29% are simplex and 71% are suffixed. Simplex nouns predominantly denote concrete entities (Animate and Inanimate, 69%) – as predicted by theoretical accounts (e.g., Croft, 1991) – but also exhibit a substantial proportion of abstract meanings (Information, Dynamic and Stative situations, 31%), as shown by previous empirical studies (e.g., Tribout et al., 2014). This trend is reversed for derived nouns, which primarily denote abstract concepts (59%), although a significant proportion of derivatives refers to concrete entities (41%).

Concrete meanings (detailed in the first eight rows of Table 5) show distinct distributional patterns across simplex and suffixed nouns. At the hypersense level, simplex nouns predominantly refer to Inanimate entities rather than Animate ones (48% vs. 21%), whereas suffixed nouns show the opposite trend, with Animate entities being more frequently denoted than Inanimate ones (29% vs. 12%). Among Animate referents, simplex nouns include both Persons (e.g., *samouraï* 'samurai', *gendre* 'son-in-law') and Animals (e.g., *poulpe* 'octopus', *alligator* 'alligator'), while suffixed nouns mostly denote Persons (e.g., *alpiniste* 'mountaineer', *armateur* 'shipowner') and more rarely refer to Animals (e.g., *coquelet* 'young rooster', *oursonne* 'female bear cub'). As for Inanimate referents, the ratio of natural objects (Body, Food, Object, Plant) to man-made objects (Artifact) is reversed between the two morphological groups. Simplex nouns more often denote natural objects (32%; e.g., *topaze* 'topaz', *toundra* 'tundra') than Artifacts (16%; e.g., *calèche* 'carriage', *véranda* 'veranda'), whereas in the suffixed group, the proportions are more balanced, with Artifacts (6%; e.g., *téléviseur* 'television set', *plantoir* 'dibble') only slightly outnumbering natural objects (5%; e.g., *abricotier* 'apricot tree', *neutron* 'neutron').

Abstract meanings (listed in the last nine rows of Table 5) are more prevalent among suffixed nouns than simplex nouns. Among abstract meanings, dynamic situations constitute the dominant class in both groups – accounting for 10% of simplex nouns (e.g., rixe 'fight', salto 'flip') and 32% of suffixed nouns (e.g., comptage 'counting', identification 'identification'). As for non-dynamic meanings, States are dominant among simplex nouns, mostly denoting diseases (e.g., *rhume* 'cold', *syphilis* 'syphilis') whereas Attributes are dominant among suffixed nouns (e.g., *prudence* 'prudence', *héroïsme* 'heroism'). Finally, nouns denoting informational objects are proportionally more frequent in the simplex group (e.g., *axiome* 'axiom', *calembour* 'pun') than in the derivatives group (e.g., *comptine* 'nursery rhyme', *devinette* 'riddle').

Overall, the semantic distinctions between simplex and suffixed nouns are statistically validated through a chi-square test, demonstrating a significant dependency between the morphological structure of the noun (simplex or suffixed) and its meaning (hypersense) ($\chi^2(5, N = 4{,}089) = 845.9, p < .001$), with a moderately strong effect size (Cramér's V = 0.45).

We can explore form–meaning relationships in more detail by examining the semantic contribution of suffixes within the set of monosemous derivatives (2,887 items). As expected, these relationships are rarely one-to-one. On the formal side, these nouns are morphologically derived using 98 distinct suffixes, which vary widely in their degree of polyfunctionality, i.e., their tendency to produce nouns belonging to multiple semantic classes (Mean = 3.3, SD = 3.1). Some suffixes are monofunctional, such as *-iste*, which

Supersense	Hypersense	Simplex		Suffixed		All	
Animal	Animate	8.7	21.3	1.2	29.3	3.4	26.9
Person		12.6		28.1		23.5	
Artifact	Inanimate	16.5	48.6	6.4	12.2	9.4	22.9
Body		5.1		0.7		2.0	
Food		13.1		1.1		4.6	
Object		4.7		1.1		2.2	
Plant		4.8		1.0		2.1	
Substance		4.5		1.9		2.7	
Cognition	Information	4.7	5.6	4.3	4.5	4.4	4.8
Communication		0.8		0.2		0.4	
Act	Dynamic_sit.	7.2	10.1	26.6	32.1	20.9	25.7
Event		1.7		4.5		3.7	
Phenomenon		1.2		1.0		1.1	
Attribute	Stative_sit.	1.4	4.4	10.1	18.1	7.6	14.1
Feeling		0.8		1.7		1.4	
State		2.2		6.3		5.1	
Other (6)	Other (6)	10.1		3.8		5.6	

Table 5: Distribution of senses, in percentage, among simplex monosemous nouns (1,202), suffixed monosemous nouns (2,887), and the total (4,089).

exclusively yields nouns referring to Persons (e.g., *dentiste* 'dentist'), while others display varying degrees of polyfunctionality, ranging from 2 to 15 supersenses. For instance, monosemous nouns formed with *-eur* are associated with several supersenses: they predominantly denote Persons (87%, e.g., *ingénieur* 'engineer'), but also *Artifacts* (12%, e.g., *carburateur* 'carburetor'), and, more rarely, Animals (<1%, *décomposeur* 'decomposer') or Substances (<1%, *durcisseur* 'hardener'). On the semantic side, most categories are realized through multiple suffixes (Mean = 14.3, SD = 8.6). For example, nouns denoting Persons appear with various suffixes, including *-eur* (50%, e.g., *ingénieur* 'engineer'), *-iste* (14%, e.g., *dentiste* 'dentist'), *-ier* (10%, e.g., *parolière* 'lyricist'), and *-ant* (6%, e.g., *fabricant* 'manufacturer'), among others. Table 6 presents the semantic distribution of monosemous nouns derived with one of the 13 most frequent suffixes, which together account for 81% of the monosemous subset. The distributions show that the semantic diversity of suffixes, within the monosemous lexicon, remains largely constrained by the concrete–abstract distinction. In fact, there is an almost clear-cut division between suffixes such as *-eur*, *-ier*, and *-et*, which typically produce nouns denoting concrete entities (first eight rows of the table), and suffixes like *-ment*, *-ion*, and *-isme*, which tend to form abstract entities or situations (following nine rows). These observations suggest that ambiguity in the suffixed lexicon involving abstract-to-concrete shifts (e.g., *construction* 'construction/structure', *armement* 'armament/weaponry') is more likely to arise from semantic extension than from true cases of morphological co-derivation.

3.3 Ambiguity profiles

Building on the semantic contrasts observed between simplex and suffixed monosemous nouns, this section examines how ambiguity patterns differ across the two groups – both in average number of senses and in sense alternation types. We first consider the reasons why simplex nouns show greater ambiguity, then assess how senses are distributed among ambiguous nouns in each group.

	-eur	*-ment*	*-ion*	*-age*	*-ité*	*-isme*	*-ier*	*-iste*	*-erie*	*-et*	*-ance*	*-ie*	*-ant*	**Other**	**All**
Animal	0.1	-	-	-	-	-	-	-	-	0.2	-	-	-	0.9	1.2
Person	14.0	-	-	-	-	-	3.0	4.1	-	0.4	-	-	1.9	4.6	28.0
Artifact	2.0	0.2	-	0.1	-	-	0.7	-	0.2	1.0	-	-	0.2	2.0	6.4
Body	-	0.1	-	-	-	-	-	-	-	0.1	-	-	-	0.4	0.7
Food	-	-	-	-	-	-	-	-	-	0.3	0.1	-	-	0.6	1.1
Object	-	-	-	-	0.1	-	0.1	-	-	0.1	-	-	-	0.9	1.1
Plant	-	-	-	-	-	-	0.7	-	-	-	-	-	-	0.2	1.0
Substance	-	-	-	-	-	-	-	-	-	-	-	-	0.1	1.8	1.9
Cognition	-	-	0.1	-	-	3.0	-	-	0.1	0.1	-	0.3	-	0.5	4.3
Com.	-	-	0.1	-	-	-	-	-	-	-	-	-	-	0.1	0.2
Act	-	7.2	8.6	6.8	-	0.4	-	-	1.0	0.1	0.4	0.4	0.1	1.6	26.6
Event	-	2.7	1.2	-	-	-	-	-	-	-	-	-	0.1	0.3	4.5
Phenom.	-	0.7	0.1	-	-	-	-	-	-	-	-	-	-	0.2	1.0
Attribute	-	0.1	0.2	-	5.5	1.6	-	-	0.6	-	0.8	0.2	-	1.0	10.1
Feeling	-	0.7	0.4	-	0.2	-	-	-	-	-	0.2	-	-	0.1	1.7
State	-	0.5	0.6	-	1.2	0.5	-	-	0.1	-	0.6	0.7	-	2.1	6.3
Other	-	0.2	0.5	0.2	0.2	-	0.1	-	0.8	0.2	0.1	0.2	-	1.3	3.8
Total	16.2	12.5	11.7	7.2	7.3	5.5	4.7	4.1	3.0	2.6	2.2	1.9	2.3	18.7	100.0

Table 6: Distribution of senses, in percentage, among suffixed monosemous nouns (2,887), for the 13 most frequent suffixes (81.3%). The remaining 85 suffixes, which account for 18.7% of the nouns, are grouped under the 'Other' label.

3.3.1 Level of ambiguity

The average level of ambiguity across the entire sample is higher for simplex nouns (see Table 4). As expected, simplex nouns also tend to be more frequent than derived ones. However, the causal relationship between these two collinear variables remains unclear. A less explored, yet potentially significant, hypothesis is that ambiguity levels may be influenced by the semantic type of a word's base meaning. Specifically, concrete meanings might be more prone to semantic extension than abstract ones (e.g., Lakoff and Johnson, 2008). To test this hypothesis, we used a Poisson regression, with the number of meanings as the dependent variable, and, as predictors, the log-transformed frequency of the noun and the concreteness of its source meaning. As shown in Table 7, concreteness significantly contributes to predicting the number of meanings a word has ($p < 0.01$), although its effect is smaller than that of frequency.

	Estimate	Std Error	z value	Pr($<|z|$)
(Intercept)	0.584326	0.010071	58.023	$<$ 2e-16
Concreteness-concrete	0.029294	0.010830	2.705	0.00683
Log_Freq	0.642598	0.007921	81.125	$<$ 2e-16

Table 7: Coefficients of predictors of ambiguity level.

Given that simplex nouns are more frequently associated with concrete referents than suffixed nouns – as observed among monosemous items – this semantic tendency may partly account for their higher average ambiguity. However, further investigation will be needed at a finer level of analysis and should take into account additional factors that might explain this difference. One such factor is the lexical longevity of the nouns, as simplex nouns may, on average, be older in the lexicon than derived forms.

3.3.2 Sense alternations

As previously mentioned, one can expect the properties of sense alternations to differ between the simplex and suffixed groups for at least two reasons. First, since the distribution of semantic types differs between these two groups – as observed in the monosemous lexicon (see Table 5) – it follows that the types of sense alternations they exhibit are also likely to differ. For example, because body-part nouns are predominantly simplex, ambiguities arising from metaphors such as Body→Artifact (e.g., *bouche* 'mouth/entry') or metonymies like Body→Person (e.g., *tête* 'head/intelligent person') are more characteristic of the simplex group than of the suffixed one. Second, the source of ambiguity also differs between the groups. In simplex nouns, ambiguity typically stems from semantic factors alone, such as meaning change or extension from an existing sense. In contrast, in suffixed nouns, ambiguity may arise from both semantic and morphological sources. This distinction is expected to influence the patterns and profiles of sense alternation observed in each group. However, a detailed analysis of these two aspects requires explicit encoding of the relationships between a noun's different senses. In the case of suffixed nouns, it also requires clarification of the relationship between the meaning of the derived noun and that of its morphological base. This type of information is not available in LexEco, as the sense inventories are sourced from Wiktionary, which only sporadically makes such relations explicit – typically through metalinguistic labels such as "par extension" ('by extension').

Nevertheless, the data available in LexEco allow for a distinction between two subtypes of ambiguous nouns, providing a coarse-grained account of their semantic dispersity. The first subtype, hereafter referred to as *monocategorical* ambiguous nouns, includes words whose senses all belong to the same semantic category – as exemplified by *suffragette* in (2), which consistently denotes a Person, regardless of the specific usage. The second subtype, hereafter referred to as *polycategorical* ambiguous nouns, includes words with at least two senses that fall under distinct supersenses – for instance, *cuisinière* ('female cook') in (3), which can denote either a Person or an Artifact.

(2) SUFFRAGETTE 'suffragette'

a. (Histoire, Politique) Femme qui militait pour obtenir le droit de voter. Person
'(History, Politics) Woman who campaigned for the right to vote.'

b. (Par extension) Féministe, femme qui milite pour l'admission des femmes dans une institution ou un ordre qui lui est jusqu'alors fermé. Person
'(By extension) Feminist; woman advocating for the inclusion of women in an institution or order previously closed to them.'

(3) CUISINIÈRE 'cook/cookstove'

a. (Cuisine) Celle qui cuisine, qui prépare, qui cuit la nourriture. Person
'(Cooking) A woman who cooks or prepares food.'

b. (Électroménager) Fourneau de cuisine servant à chauffer ou faire cuire les aliments, souvent muni d'éléments chauffants sur sa surface de travail. Artifact
'(Appliance) Kitchen stove used to heat or cook food, often equipped with heating elements on its surface.'

Table 8 presents the distribution of ambiguous nouns in the analyzed sample. A first observation is that monocategorical ambiguous nouns are more frequent in the suffixed group than in the simplex group (43% vs. 34%). This difference is statistically significant, $\chi^2(1, N = 7{,}889) = 61.5$, $p < .001$, although the effect size is negligible (Cramér's $V = .08$). A plausible explanation for this difference is that simplex nouns tend to exhibit higher overall ambiguity than suffixed nouns (4.1 vs. 3.6, within the set of ambiguous nouns), thereby increasing the likelihood that an ambiguous simplex noun spans multiple semantic classes. A more noteworthy observation, however, is that the mean ambiguity among monocategorical ambiguous nouns is no longer significantly different between the two groups (2.8 vs. 2.7), as shown by a Mann–Whitney test ($Z = 1.4$, $p = .1$). This holds despite a substantial difference in mean frequency (29.4 vs. 7.0), which mirrors the pattern observed in the full dataset, and despite a semantic class distribution in each group that resembles that of the monosemous group. A possible

explanation is the inflation in sense count due to morphological co-derivation, i.e., the generation of multiple senses from a single morphological base – but this hypothesis requires further investigation. It should also be considered in light of the fact that, within the polycategorial group of ambiguous words, the proportion of co-derived senses in suffixed nouns is apparently not sufficient to reach the level of ambiguity exhibited by simplex nouns.

	Lemmas	**Senses**	**Ambiguity**	**Freq**
Simplex-monocat	957	2,717	2.8	29.4
Simplex-polycat	1,812	8,883	4.9	42.9
Suffixed-monocat	2,236	6,107	2.7	7.0
Suffixed-polycat	2,884	12,386	4.2	12.1
Total	7,889	30,093	3.81	19.8

Table 8: Descriptive statistics for the subset of ambiguous nouns.

4 Conclusion

In this paper, we presented LexEco, a large-scale morpho-semantic database designed to serve as a representative lexicon of French nouns known by most adult native speakers. Morphological information was primarily sourced from several existing, manually curated derivational databases, while semantic information was obtained through supervised classification based on the French Wiktionary. Although the precision of the semantic annotation has been rigorously evaluated and deemed sufficient for quantitative analysis, the reliability of the morphological information has yet to be formally assessed. A key direction for future work is to evaluate the coherence of the morphological encoding – currently suboptimal due to the heterogeneous origins of the data – and to improve this coherence in future releases of the database.

As a case study, we used LexEco to explore how derivational morphology contributes to the semantics of French nouns. The comparative analysis of simplex and suffixed nouns revealed a partially complementary distribution of semantic types between the two groups. Furthermore, the results indicate that the ambiguity profiles of simplex and suffixed nouns differ both in their degree of ambiguity – with simplex nouns tending to be more ambiguous – and in the nature of their sense alternations.

The database and these initial findings open promising avenues for further research, including more fine-grained semantic investigations into the complementary role of morphologically constructed meanings and a deeper understanding of the respective roles of morphology and polysemy in the construction of nominal meaning.

References

Nicolas Angleraud, Lucie Barque, and Marie Candito. 2025. Annotating the French Wiktionary with supersenses for large scale lexical analysis: a use case to assess form-meaning relationships within the nominal lexicon. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, Abu Dhabi, UAE, pages 5321–5332. https://aclanthology.org/2025.coling-main.356/.

Isabel Balteiro. 2007. *The Directionality of Conversion in English: A dia-synchronic study*, volume 59. Peter Lang, Bern.

Antonio Balvet, Lucie Barque, Marie-Hélène Condette, Pauline Haas, Richard Huyghe, Rafael Marin, and Aurélie Merlo. 2011. La ressource Nomage : Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Revue TAL: Traitement Automatique des Langues* 52(3):129–152. https://shs.hal.science/halshs-01077819v1.

Lucie Barque, Pauline Haas, Richard Huyghe, and Delphine Tribout. Submitted. Les noms pseudo-suffixés en français : Quelle proximité sémantique avec leurs équivalents suffixés ?

Laurie Bauer. 2017. Metonymy and the semantics of word-formation. In *Mediterranean Morphology Meetings*. volume 11, pages 1–13. https://doi.org/10.26220/mmm.2868.

Olivier Bonami and Delphine Tribout. 2021. Échantinom: A hand-annotated morphological lexicon of French nouns. In *International Workshop on Resources and Tools for Derivational Morphology*. pages 42–51. https://shs.hal.science/halshs-03520602.

William Croft. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. University Press of Chicago, Chicago.

Matías Guzmán Naranjo and Olivier Bonami. 2023. A distributional assessment of rivalry in word formation. *Word Structure* 16(1):87–114.

Richard Huyghe, Lucie Barque, Pauline Haas, and Delphine Tribout. 2017. The semantics of underived event nouns in French. *Italian Journal of Linguistics* 29(1):117–142.

Richard Huyghe, Alizée Lombard, Justine Salvadori, and Sandra Schwab. 2023. Semantic rivalry between French deverbal neologisms in *-age*, *-ion* and *-ment*. In Sven Kotowski and Ingo Plag, editors, *The Semantics of Derivational Morphology: Theory, Methods, Evidence*, Berlin: De Gruyter, pages 143–175. https://doi.org/10.1515/9783111074917-006.

Richard Huyghe, Justine Salvadori, Rossella Varvara, Lucie Barque, Pauline Haas, Alizée Lombard, Matthieu Monney, Delphine Tribout, and Marine Wauquier. Submitted. SONDE: A database for exploring the semantics of nouns derived from verbs in French.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

John Lyons. 1977. *Semantics*, volume 2. Cambridge University Press, Cambridge.

Hans Marchand. 1964. A set of criteria for the establishing of derivational relationship between words unmarked by derivational morphemes. *Indogermanische Forschungen* 69:10. https://doi.org/10.1515/9783110243116.10.

Fiammetta Namer, Nabil Hathout, Dany Amiot, Lucie Barque, Olivier Bonami, Gilles Boyé, Basilio Calderone, Julie Cattini, Georgette Dal, Alexander Delaporte, Guillaume Duboisdindien, Achille Falaise, Natalia Grabar, Pauline Haas, Frédérique Henry, Mathilde Huguin, Juniarta Nyoman, Loïc Liégeois, Stéphanie Lignon, Lucie Macchi, Grigoriy Manucharian, Caroline Masson, Fabio Montermini, Nadejda Okinina, Frank Sajous, Daniele Sanacore, Mai Thi Tran, Juliette Thuilier, Yannick Toussaint, and Delphine Tribout. 2023. Démonette-2: A derivational database for French with broad lexical coverage and fine-grained morphological descriptions. *Lexique* 33:6–40. https://doi.org/10.17605/OSF.IO/DB2W8.

Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28(4):661–677. https://doi.org/10.1017/S014271640707035X.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers* 36(3):516–524. https://doi.org/10.3758/BF03195598.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition* 122(3):280–291. https://doi.org/10.1016/j.cognition.2011.10.004.

Franz Rainer. 2014. Polysemy in derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford University Press, Oxford, pages 338–353.

Justine Salvadori. 2024. *L'ambiguïté des noms déverbaux en français: Une étude quantitative du sens construit*. Ph.D. thesis, Université de Fribourg. https://doi.org/10.51363/unifr.lth.2025.044.

Delphine Tribout. 2020. Nominalization, verbalization or both? Insights from the directionality of noun-verb conversion in French. *Zeitschrift für Wortbildung/Journal of Word Formation* 4(2):187–207. https://doi.org/10.3726/zwjw.2020.02.10.

Delphine Tribout, Lucie Barque, Pauline Haas, and Richard Huyghe. 2014. De la simplicité en morphologie. In *SHS Web of Conferences*. EDP Sciences, volume 8, pages 1879–1890. https://doi.org/10.1051/shsconf/20140801182.

Anna Wierzbicka. 1986. What's in a noun? (or: How do nouns differ in meaning from adjectives?). *Studies in Language* 10(2):353–389. https://doi.org/10.1075/sl.10.2.05wie.

George K. Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology* 33(2):251–256. https://doi.org/10.1080/00221309.1945.10544509.

Morphophonological alternation patterns in the Phononette database: The role of families and series

Fiammetta Namer
Université de Lorraine &
UMR 7118 ATILF, CNRS
fiammetta.namer@
univ-lorraine.fr

Stéphanie Lignon
Université de Lorraine &
UMR 7118 ATILF, CNRS
stephanie.lignon@
univ-lorraine.fr

Nabil Hathout
Université Toulouse - Jean Jaurès & UMR 5263 CLLE, CNRS
nabil.hathout@univ-tlse2.fr

Abstract

Morphophonology generally addresses the question of stem alternation using the notions of allomorphy and suppletion. However, these descriptive concepts are difficult to define and apply, and they do not explain why a particular variation appears in a specific pair. We propose a variation description based on four criteria: stem integrity, formal distance, stem frequency within the family, and variation frequency within the lexicon. To validate these criteria empirically, we created the Phononette database. This database contains 19,032 derivationally related word pairs from French with morphophonological features. The analyses made possible by Phononette are grounded in a paradigmatic framework of word formation. Through case studies relying on 5,500 Phononette entries, we show how the proposed criteria offer a fresh perspective on allomorphy and suppletion and provides an alternative view of epenthesis.

1 Introduction

One of the many questions of interest to morphology is the formal (i.e., morphophonological) conditions on the formation of derived words.[1] For example, in French, how to predict that the relation between a noun and its relational adjective in *-ier* involves the epenthetic consonant [p] at the boundary between the noun stem and the exponent [je] in the case of *drap*$_N$*/drapier*$_A$ 'sheet/of sheet', whereas it is [d] with *cauchemar*$_N$*/cauchemardier*$_A$ 'nightmare/of nightmare', [s] with *peau*$_N$*/peaucier*$_A$ 'skin/of skin', [əl] with *chapeau*$_N$*/chapelier*$_A$ 'hat/of hat', or [t] with *clou*$_N$*/cloutier*$_A$ 'nail/of nail' and *caoutchouc*$_N$*/caoutchoutier*$_A$ 'rubber/of rubber'?

Alternations, that is, string adjunction like above or string modification, can affect the stems of related lexemes, for example [œ]∼[ɔ] apophony in *fleur*$_N$ [flœʁ] 'flower'−*flor-al*$_A$ [flɔʁ-al] 'floral'. They may also affect rule exponents, such as the negative prefix, which in the *pur*$_A$ [pyʁ] 'pure'−*im-pur* 'impure'$_A$ [ɛ̃-pyʁ] relation is realized [ɛ̃], whereas in *réel*$_A$ [ʁeɛl] 'real'−*ir-réel*$_A$ [i-ʁeɛl] 'unreal' it is realized [i].

The traditional analysis of alternation relies almost exclusively on the comparison of the phonetic representations of pairs of forms, and on the phonological and historical properties of the compared forms. For example, in research fields in line with OT, alternation is considered as a (morpho-)phonological phenomenon (Thornton, 1997): variations are treated primarily with phonological criteria. As a result, they are divided into two broad categories (see Section 2): allomorphy (falling within the domain of phonology) and suppletion (processed at the lexical level).

However, the evidence presented in this article, which focuses exclusively on stem alternation issues in French, suggests that measuring the similarity of morphologically related forms alone is not a reliable

[1]We are grateful to Richard Huyghe, Megan Prudent, Justine Salvadori, Matea Filko and Krešimir Šojat for the organization of DeriMo 2025. We thank the three anonymous reviewers for their constructive comments. All errors remain ours.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 49–61, Fribourg, Switzerland, 4-5 September 2025.

predictor of the alternating sequence. On the contrary, the level of saturation of a word family by a stem, and the frequency of morphophonological variations in derivational relations, seem to have an influence on the alternation mechanism. These factors have an impact on the distribution of alternations between allomorphy and suppletion.

It is essential to have access to the phonological properties of large and varied derivational relations to confirm the role of each of these factors in predicting the most likely formal variations, analyzing atypical cases, or explaining why speakers produce a suppletive form (or recover its meaning) with greater or lesser ease. More generally, an extensive approach based on the use of a derivational resource equipped with morphophonological annotations can shed further light on formal variations in derivational relations. To our knowledge, there is no such resource for French: the need to provide a broad-coverage derivational database, with complete, manually-checked phonetic transcriptions has motivated the development of Phononette, an extension to a large, publicly available derivational database for French (Namer et al., 2023).

The rest of the paper is organized as follows: after outlining, in Section 2, the issues at stake and the hypotheses we intend to explore, we present Démonette in Section 3, and explain how its contents have been used to create Phononette in Section 4.1. The development of Phononette involves several steps: (i) computation of the similarity measure for the stems involved in each derivational relation documented in the database, (ii) the generalization of similarity relations in the form of patterns (Section 4.2); (iii) manual validation and standardization of these annotations (Section 4.3); (iv) identification of the frequency of each stem in a word family (Section 4.4). In Section 3, finally, we show through several examples to what extent the hypotheses proposed in Section 2 are confirmed, and, consequently, how Phononette will be an important asset in derivation research.

2 What determines phonological variation in word formation?

Traditionally, the formal (i.e., phonological) comparison of pairs of derivationally related words leads to an organization that hierarchizes them according to the similarity of their stems. Two stems are either identical (Table 1-(a)), partially different (Table 1-(b-f)), or completely different (Table 1-(g-h)).

Variations that affect only part of the stem are cases of allomorphy, the others being cases of suppletion. According to Lieber (1982), **allomorphy** is concerned with "morpheme variants which share such lexical information as semantic representation and argument structure but which differ unpredictably and arbitrarily in the phonological forms and in the morphological environments in which they occur". Other scholars like Aronoff (1994), adopt a revised notion of stem allomorphy, where relations between stems range from identity through regular phonological alternation or arbitrary change to full suppletion (see also Aronoff 2012; Booij 2009). **Suppletion** is a notion mainly used in inflection (Carstairs, 1988; Boyé, 2006; Hippisley et al., 2004). However this concept is obviously also of interest in word formation (Dressler, 1985; Corbin, 1985; Plénat, 2008). Suppletion is generally contrasted with allomorphy; see Kiparsky (1996); Dressler (2015); Bonami and Beniamine (2021) for a discussion.

Similarity between two words can be approximated using the edit distance between strings, as shown in Table 1. Column **Stem1** (resp. **Stem2**) indicates which stem of $\mathbf{W_1}$ (resp. $\mathbf{W_2}$) is involved in the derivational relation. The distance between Stem1 and Stem2 (Column 5) is 0 in (a), accounting for the fact that the stems of $\mathbf{W_1}$ and $\mathbf{W_2}$ are identical; in (b) and (c), it is 1, the only difference being a vowel change (i.e., apophony). The same applies to (d), (e) and (f), where the stems differ only by phoneme adjunction (i.e., epenthesis). Finally, stems (g) and (h) have no characters in common, as indicated by 'NO'.

The data seem to suggest that the similarity criterion alone is insufficient. For instance, the epenthesis at the stem-suffix boundary in (d) does not affect its integrity, unlike the epenthesis in (e) and (f). Additionally, some alternations occur more frequently than others. For instance, when in contact with a vowel-initial suffix, the final sequence [bl] of the stem of *diable*$_{\text{N}}$ in (e) changes to [bol], whereas the same sequence [bl] in (f) changes to [bil]. This internal [i] epenthesis is observed not only in (f) with *stable*$_{\text{A}}$, but also for example in *possible*$_{\text{A}}$ 'possible' or *soluble*$_{\text{A}}$ 'soluble', when suffixed with *-iser* (*solubil-iser*$_{\text{V}}$ 'make soluble'), *-ité* (*stabil-ité*$_{\text{N}}$ 'stability', *possibil-ité*$_{\text{N}}$ 'possibility', *solubil-ité*$_{\text{N}}$ 'solubility'), *-isme*

(*stabil-isme*$_{N}$ 'stabilism', *possibil-isme*$_{N}$ 'possibilism'), among others. So it seems that [bl]/[bil] is more widespread in the lexicon than [bl]/[bol]. Therefore, (e) and (f) would not form a homogeneous group. The same applies to (b) and (c): unlike [a]/[i] in (c), at prefix-stem boundary, the alternation [ɛ]/[a] in (b) at stem-suffix boundary is easily observed in the lexicon: e.g., [pɛʁ]/[paʁ] in *pair*$_{A}$ 'pair', *parity*$_{N}$ 'parity', or [popylɛʁ]/[popylaʁ] in *popular*$_{A}$ 'popular', *populariser*$_{V}$ 'popularize'. Finally, unlike in (h), stem suppletion in (g) does not confuse speakers who spontaneously associate [ip] with the meaning 'horse'. Interestingly, [ip] is a well-represented stem in the *horse*$_{N}$ family, unlike [ev] in the *water*$_{N}$ family.

	W_1		W_2		$Stem_1$	$Stem_2$	Distance
(a)	*cendre*$_{N}$	'ash'	*cendr-ier*$_{N}$	'ashtray'	sɑ̃dʁ	sɑ̃dʁ	0
(b)	*clair*$_{A}$	'clear'	*clar-té*$_{N}$	'clarity'	klɛʁ	klɑʁ	1
(c)	*amitié*$_{N}$	'friendship'	*in-imitié*$_{N}$	'enmity'	amitje	imitje	1
(d)	*abricot*$_{N}$	'apricot'	*abricot-ier*$_{N}$	'apricot tree'	abʁiko	abʁikot	1
(e)	*diable*$_{N}$	'devil'	*diabol-ique*$_{N}$	'devilish'	djabl	djabol	1
(f)	*stable*$_{A}$	'stable'	*stabil-iser*$_{V}$	'stabilize'	stabl	stabil	1
(g)	*cheval*$_{N}$	'horse'	*hipp-isme*$_{N}$	'hippism'	ʃəval	ip	NO
(h)	*eau*$_{N}$	'water'	*év-ier*$_{N}$	'sink'	o	ev	NO

Table 1: Examples of stem alternations in (W_1,W_2) word pairs in a derivational relation

These cases suggest that several factors interact in the prediction of stem variations in derivation. These criteria are formulated in Table 2. Two of them (C_1 and C_2) appeal to the phonemic context of the stems being compared. The other two (C_3 and C_4) take place at the level of the paradigmatic organization of the morphologically complex lexicon (Bochner, 1993; Hathout and Namer, 2022; Bonami and Strnadová, 2019; Bauer, 1997; Štekauer, 2014): the first focuses on the content of families (Hathout, 2011), the second with derivational series (Fradin, 2018).

C_1	**Formal distance:** The closer two word stems are to each other, the easier they are to associate. This distance can be calculated in two ways: Levenshtein edit distance and the ratio of the length of the longer sequence to that of the shorter one.
C_2	**Stem integrity:** Modifications at the stem-affix boundary are less damaging to stem identification than those in other positions. They are more likely to preserve integrity.
C_3	**Stem reproductibility:** The more a sequence appears in the stems of the words in a family, the more it will be associated with that family, and consequently with the set of words it contains. This criterion therefore addresses the distribution of a stem within a word family.
C_4	**Alternation pattern frequency:** The more frequent the alternation in the lexicon, the stronger the relation between the two stems.

Table 2: Hypothesis: Criteria for predicting formal variations in a derivational relation

3 Démonette

Démonette2.0 (Namer and Hathout, 2020; Namer et al., 2023) is a French derivational resource designed and developed in line with a paradigmatic approach to derivation. It is organized in the form of a database with several co-indexed tables, including a table of relations (TR) and a table of lexemes (TL).[2] The data

[2]Démonette2.0 was produced as part of the ANR Demonext project (ANR-17-CE23-0005). See https://www.demonext.xyz/ for project description and results, and https://demonette.fr/demonext/vues/front_page.php to query and download the database.

that initially fed the TR presented in Section 3.1 are derivational lexicons, mostly elaborated and validated manually. The content of the TL, detailed in Section 3.2, comes from GLàFF (Hathout et al., 2014).

3.1 Table of relations (TR)

From a paradigmatic point of view, the identification of interpredictability relations in word families is essential. These relations delimit what Hathout and Namer (2022) call "paradigmatic families." In this context, each of the 222,118 TR's entries describes morphological, categorical and semantic properties of (relations existing between all) members of these families.

Specifically, each word in a ($\mathbf{W_1}$,$\mathbf{W_2}$) pair, as in Table 3, is labelled with its lemma (columns **W1** and **W2**) and its part-of-speech (columns **Cat1** and **Cat1**).[3] On the other hand, the relation between these words is characterized by its morphological **Complexity** and **Orientation**. The value of **Complexity** is "simple" for base-derived pairs, as in Table 3-(a), and for word pairs derived from the same base. The base may not be attested in the contemporary lexicon of French. For instance, while *exécutif* and *exécution* derive from the verb *exécuter* 'execute' (Table 3-(b)), there is no attested verb *addicter* 'addict' that may serve as base for *addiction* and *addictif* (Table 3-(c)) – at best, $addicter_V$ could be back-formed from *addiction* or *addictif* (Hathout and Namer, To appear). In all other cases, the value of **Complexity** is "complex". As far as **Orientation** is concerned, its value indicates which of the related words is the ancestor of the other. For example, the value "as2des" in Table 3-(a) indicates that *actif* is the ancestor of *inactif*, and the value "indirect", in Table 3-(b) indicates that neither *exécutif* nor *exécution* derives from the other – the same applies to Table 3-(c).

The ($\mathbf{W_1}$,$\mathbf{W_2}$) relation is generalized as a schema between derivational patterns we call Derivational Alternation Pattern (**DAP**). In Table 3-(a), (*actif*, *inactif*) is an instance of the **DAP** Z/inZ, where Z stands for the morphological structure shared by $\mathbf{W_1}$ and $\mathbf{W_2}$; likewise, in Table 3-(b) and (c), both (*exécutif*, *exécution*) and (*addictif*, *addiction*) are instances of Zif/Zion.

	$\mathbf{W_1}$	$\mathbf{W_2}$	$\mathbf{Cat_1}$	$\mathbf{Cat_2}$	**DAP**	**Complexity**	**Orientation**
(a)	*actif* 'active'	*inactif* 'inactive'	A	A	Z/inZ	simple	as2des
(b)	*exécutif* '$\text{drinker}_{\text{mas}}$'	*exécution* '$\text{drinker}_{\text{fem}}$'	A	Nf	Zif/Zion	simple	indirect
(c)	*addictif* 'addictive'	*addiction* 'addiction'	A	Nf	Zif/Zion	simple	indirect

Table 3: Démonette's table of relations (excerpt)

3.2 Table of lexemes (TL)

The TL contains 388,306 lexemes. Such a large coverage results from the vocabulary of all the resources included in Démonette2.0, completed by the lexemes of the electronic dictionary GLAWI (Sajous and Hathout, 2015), derived from Wiktionary. This coverage ensures consistency and stability in the database. Namely, the TL lexical coverage, which tends to be complete, allows the future addition of any new derivational relation or the modification of existing descriptions without affecting its current content.

Each TL entry assigns lexical properties to every recorded lexeme, independent of the morphological relations the lexeme may have in the TR. These properties include the graphic representation of the word's inflectional paradigm (Table 4, Column 2) and its complete transcription in IPA format (Table 4, Column 3) generated from the GLÀFF lexicon (Hathout et al., 2014). The combination of morphosyntactic features encoded by each wordform is encoded in the Multext format (Ide and Veronis, 1994) (for instance, in the second row of Table 4, the label Ncfs groups the following values: 'Noun', 'common', 'feminine', 'singular').

[3] Additionally, words are provided with their ontological category, following Huguin et al. (2022).

In French, the size of inflectional paradigms of adjectives is 4 cells, that of nouns, 2 cells, and that of verbs, 47 cells. In their phonetic transcription, verb forms (unlike noun and adjective forms) are provided with inflectional markers. For instance, for all French verbs (e.g., Vmii3s-:mãtɛ in Table 4) the IND.IPFV.3S is marked by the suffix [ɛ]. In order to avoid distorting stem comparison, we have to remove the inflectional material. Therefore verb paradigms are complemented by a structured list of stems, called "stem spaces" by Bonami and Boyé (2003). The number of cells in a stem space is determined by the distribution of the verb stems in order to cover the inflectional paradigm of all the verbs in a given language. In French, a verb stem space has 12 cells, and some are illustrated in the last raw of Table 4, Column 4. For nouns and adjectives, the stem space is identical to their inflectional paradigm.

Lemma	**Infl. Paradigm**	**Infl. Paradigm IPA Transcription**	**Stem Space**
inactif$_{\text{A}}$ 'inactive'	Afpms:inactif; Afpmp:inactifs; Afpfs:inactive; Afpfp:inactives	Afpms:inaktif; Afpmp:inaktif; Afpfs:inaktiv; Afpfp:inaktiv	
exécution$_{\text{Nf}}$ 'execution'	Ncfs:exécution; Ncfp:exécutions	Ncfs:egzekysjɔ̃; Ncfp:egzekysjɔ̃	
mentir$_{\text{V}}$ 'lie'	Vmip1s-:mens; Vmii3s-:mentait; Vmcp1s-:mentirais; Vmif2p-:mentirez; Vmis3p-:mentirent; [...]	Vmip1s-:mɑ̃; Vmii3s-:mɑ̃tɛ; Vmcp1s-:mɑ̃tiʁɛ; Vmif2p-:mɑ̃tiʁe; Vmis3p-:mɑ̃tiʁ; [...]	Vmii---:mɑ̃t; Vmip3p-:mɑ̃t; Vmip-s-:mɑ̃; [...] Vmif---:mɑ̃ti; Vmis---:mɑ̃ti

Table 4: Démonette's table of lexemes (excerpt)

4 Phononette

Phononette's goal is to record morphophonological variation in derivational families, and to provide a linguistic interpretation based on the combination of the four criteria mentioned in Table 2, whose relevance must be verified: $\mathbf{C_1}$ – string similarity, $\mathbf{C_2}$ – stem integrity, $\mathbf{C_3}$ – stem representativeness in each family, $\mathbf{C_4}$ – phonological pattern frequency. In other words, Phononette's purpose is to give users the means to explore on a large scale the mechanisms of formal alternation at work in derivational relations, identify and describe on a large scale formal alternations in French derivational families, rank them according to quantitative measures, and predict the conditions of these alternations. Database design and populating rely on the phonetic annotation of the words present in Démonette. The values needed to measure $\mathbf{C_1}$ and $\mathbf{C_2}$ are calculated at the level of each entry. For the other criteria, additional features are required to group (i) relations into morphophonological series ($\mathbf{C_4}$) and (ii) words that share the same stem in relations where they appear ($\mathbf{C_3}$).

4.1 From Démonette to Phononette

Ideally, Phononette's coverage should be identical to that of the TR. However, only a subset of the TR is covered at this stage, because not all the words involved in the derivational relations are fully documented in the TL. The subset of TRs where $\mathbf{W_1}$ and $\mathbf{W_2}$ have a full phonemic description in the TL contains 41% of the 222,118 relations present in the TR. These 90,845 ($\mathbf{W_1}$,$\mathbf{W_2}$) entries, fully documented in the TL, are retained in this version of Phononette.

4.2 Measuring stem similarity in ($\mathbf{W_1}$,$\mathbf{W_2}$) relations

Each stem included in the stem space $\mathbf{SS_1}$ of $\mathbf{W_1}$ is compared to each stem included in the stem space $\mathbf{SS_2}$ of $\mathbf{W_2}$ (Table 6). To do this, derivational affixes (if any) involved in ($\mathbf{W_1}$,$\mathbf{W_2}$) derivational alternation pattern (or **DAP**, see Table 3) have to be removed. As illustrated in Table 5, exponent values are stored

in a table of derivational affixes compiled from all 518 derivational patterns of Démonette, where the dot '.' stands for the affix-stem boundary.

Derivational Pattern	Exponent (IPA transcription)	Example		
Xisme	.ism	[kybism]	*cubisme*$_{\text{N}}$	'cubism'
inX	i.	[ilɔʒik]	*illogique*$_{\text{A}}$	'illogical'
inX	in.	[inanime]	*inanimé*$_{\text{A}}$	'inanimate'
inX	ɛ̃.	[ɛ̃pasjɑ̃]	*impatient*$_{\text{A}}$	'impatient'

Table 5: Derivational pattern/Exponent value matching table (excerpt)

After each form in **SS$_1$** and **SS$_2$** has been stripped of any derivational exponent, a similarity measure is used to identify the most similar stems (**MSS**) among these two sets. Table 6 provides some examples, illustrating stem identity (a), epenthesis (b) and (c), apophony (d) and suppletion (e). Currently, this task uses a Levenshtein edit distance (Yujian and Bo, 2007). We refer to Albright and Hayes (2006, 2002) for a discussion of learning systems for morphophonological variation patterns in inflection, and to Beniamine (2017) for a universal algorithm inferring these variation patterns.

The similarity program generalizes each **MSS** by means of a phonological alternation pattern or **PAP** (see Table 6). Each phonetic string shared by the two stems in the **MSS** (e.g., [pɛ] in [pɛ/pɛz], Table 6-(b)) is represented by the same symbol in the **PAP**, here X/Xz. No stem variation is represented by a **PAP** value of X/X, and suppletive stems, as in Table 6-(e), are labelled with X/Y.

Overall, the raw application of the similarity measure (based on Levenshtein edit distance) results in 19,032 pairs with variation, representing 21% of the 90,845 fully documented entries in Démonette. In other words, French word formation involves at least 79% of formally regular relations. This already very high proportion is revised upwards during the manual verification phase (Section 4.3).

4.3 Checking annotation and adding new features

Human validation is a process that involves several tasks, carried out in parallel by two annotators. A corrector's guide is currently being drawn up: in its current version, it is a working document that cannot yet be distributed. As they are encountered, all cases of inappropriate coding and errors are recorded, along with the corresponding correction choices.

The first task is to standardize phonetic characters (e.g., [g] and [ɡ]), and in particular to neutralize free phonetic variation, like [ɑ̃] and [ɑ̃] we found in [mamɑ̃] vs. [mamɑ̃] (*maman* 'mummy'), where the opposition is morphophonologically irrelevant. The consequence of this standardization work is that what appeared to be relations involving stem alternation needs to be reviewed in many entries. The rest of human validation, complemented by manual annotation of additional features, is carried out on each of the 19,032 word pairs involving a formal alternation as reported and encoded by the similarity measure program. The most important tasks are as follows:

Checking affix stripping. The program sometimes misplaces affix-stem boundaries and they have to be corrected. For example, for the (*inaccentué*$_{\text{A}}$ 'unaccentuated', *accentué*$_{\text{A}}$ 'accentuated') word pair, there is a wrong segmentation of the prefixed stem *inaccentué*$_{\text{A}}$ into [naksɑ̃tɥe] instead of [aksɑ̃tɥe], due to the erroneous choice by the program of the [i] variant (instead of [in]) of the inX pattern of negation (Table 5).

Reshaping patterns. Automatic labeling of **PAP**s yields patterns in which all identical sequences of the compared stems are represented by variables. For example, [ɛ/a] apophony is represented by XɛY/XaY. However, we have opted for a general strategy according to which a **PAP**'s value is refined in order to be as specific as possible. Therefore, in case of apophony as in Table 6-(d), the phonemes following the alternating vowel are reproduced as they are. Consequently, the original XɛY/XaY **PAP** value of *clair*/*clarté*, *pair*/*parité*, *populaire*/*populariser* (Section 2) is manually replaced with

Xɛʁ/Xaʁ. Similarly, the **PAP** of the pairs *fonctionnel*$_{\text{A}}$ 'functional' / *fonctionnalité*$_{\text{N}}$ 'functionality' and *rituel*$_{\text{A}}$ 'ritual' / *ritualiser*$_{\text{V}}$ 'ritualize' is Xɛl/Xal and that of the pairs *marocain*$_{\text{A}}$ 'Moroccan' / *marocanité*$_{\text{N}}$ 'Moroccanness', *humain*$_{\text{A}}$ 'human' / *humaniser*$_{\text{V}}$ 'humanize', is Xɛn/Xan.[4]

A further annotation step will be the adjunction of a 'meta-**PAP**' feature that groups together all patterns with the same variation. For example, XɛC/XaC may generalize the three **PAP**s above with a common alternation label indicating '[ɛ/a] apophony on the last, coda-ending stem syllable'. This will allow alternation to be analyzed at more than one level of generality.

Generalizing suppletion patterns. In addition to 'strong suppletive' relations, that is, entries with totally disjoint valued stems like [ʃəvo/ip] in Table 6-(e), other 'borderline' cases of suppletion, sometimes referred to as 'weak suppletion' (Dressler, 1985; Mel'çuk, 1994), are manually tagged with X/Y. The idea is to come back later to these complex alternations in order to refine their description. An example is the *argyrisme/argent* word pair in Table 6-(f): although they have a large part of their stems in common ([aʁʒiʁ] alternation), we assume a case of 'weak suppletion', because the stem ending ([iʁ]/[ã]) variation is morphophonologically unpredictable.

Identifying last shared phoneme. Phononette provides the indication of X's last phoneme of all its entries that are not suppletive pairs (Table 6, Column **X last Ph.**), that is, the left context to epenthesis and alternating phonemes (when relevant). This value can therefore help explain certain variations.

Another feature is currently being encoded in each entry. Its value is the graphic transcription of the last character of the stem corresponding to X in **PAP**. This value is encoded in all entries, excepted for relations involving suppletion (**PAP** = X/Y). It is useful for analyzing epenthesis (Section 5.2), since it allows to verify whether the epenthetic consonant of the derivative stem, as p in [drap] (Table 6-(c)) or z in [pɛz] (Table 6-(b)), matches the last letter of the base (that is, 'p' for *drap* and 'x' for *paix*) or not.

	W$_1$	**SS$_1$**	**W$_2$**	**SS$_2$**	**DAP**	**MSS**	**X last Ph.**	**PAP**
(a)	*actif*$_{\text{A}}$ 'active'	([aktif], [aktiv])	*activité*$_{\text{N}}$ 'activity'	([aktivite])	Z/Zité	[aktiv/aktiv]	v	X/X
(b)	*paix*$_{\text{N}}$ 'peace'	([pɛ])	*apaiser*$_{\text{V}}$ 'soothe'	([apɛz], ...)	Z/aZ	[pɛ/pɛz]	ɛ	X/Xz
(c)	*drap*$_{\text{N}}$ 'sheet'	([dʁa])	*drapier*$_{\text{N}}$ 'sheet maker'	([dʁapje])	Z/Zier	[dʁa/dʁap]	a	X/Xp
(d)	*clair*$_{\text{A}}$ 'clear'	([klɛʁ], [klɛʁ])	*clarté*$_{\text{N}}$ 'clarity'	([klaʁte])	Z/Zté	[klɛʁ/klaʁ]	l	Xɛʁ/Xaʁ
(e)	*cheval*$_{\text{N}}$ 'horse'	([ʃəval], [ʃəvo])	*hippique*$_{\text{A}}$ 'equestrian'	([ipik], [ipik])	Z/Zique	[ʃəvo/ip]	NO	X/Y
(f)	*argyrisme*$_{\text{N}}$ 'argyrism'	([aʁʒiʁizm])	*argent*$_{\text{N}}$ 'silver'	([aʁʒã])	Zisme/Z	[aʁʒiʁ/aʁʒã]	NO	X/Y

Table 6: Examples of alternations in (**W$_1$**,**W$_2$**) relations

4.4 A family affair

Once the formal properties of each relation have been encoded and verified, the number of occurrences of each stem in its family is calculated. To do this, all relations within a given family are grouped together using the family index provided by Démonette and shared by all family members. For example, the family of *eau* 'water' (f85070) contains 72 relations between 27 words, and the stems of these words are ranked by frequency in Table 7. The fact that [idʁat] occurs 13 times in its family and [ev] only once is consistent

[4]Here, the **PAP** value is justified by the fact that the adjectival stem used for feminine: [maʁokɛn], [ymɛn] is a better candidate for stem comparison than the stem used for masculine [maʁokɛ̃], [ymɛ̃].

with the fact that speakers unconsciously associate [idʁat] with the meaning 'water' but are unable to do so for [ev] (see below).

For this task, we need to sum the number of occurrences of each stem, as recorded in each entry (cf. Table 6, **MSS** column). This is a work in progress. Currently, stem distribution by family is calculated manually.

Example	*év-ier*$_{A}$ 'sink'	*aig-aire*$_{A}$ 'ewer'	*hydr-ique*$_{A}$ 'hydric'	*aqu-eux*$_{A}$ 'watery'	*aquat-ique*$_{A}$ 'aquatic'	*hydrat-er*$_{V}$ 'hydrate'
Stem	[ev]	[ɛg]	[idʁ]	[ak]	[akwat]	[idʁat]
Distribution	1	1	2	3	7	13

Table 7: Distribution of stems in the family of *eau* (excerpt)

4.5 Phononette current state

To date, 5,500 Phononette word pairs have been manually checked, accounting for nearly 30% of the database. The values of the 225 **PAP**s identified (Table 8) therefore cover a relatively significant sample and provide insight into the distribution of derivational relations according to the observed variations.

The first column lists the main **PAP**s identified in this portion of the base, grouped by major type of alternation: (a) stem identity represented by X/X; (b) suppletion (X/Y); (c) addition of phoneme sequence represented by Σ and including cases of epenthesis; (d) (de)nasalization; (e) consonant alternation at stem ending, where ϕ and ϕ' are the two alternating consonants; (f) sequence variation (Σ / Σ') within the two stems, including cases of apophony; (g) any other alternation. The total frequency of each type in the 5,500 analyzed entries is given in Table 8, Column 5. The second column of Table 8 indicates how many patterns are part of each type. Column 3 shows the type's most prevalent pattern, the number of word pairs involved (in brackets), and an example of alternation Ccolumn 4).

	Type of relations among stems	**Number of PAPs**	**Most represented PAP**	**Example**	**Number of word pairs**
(a)	stem identity: X/X	1			2,064
(b)	suppletion: X/Y	1			636
(c)	adjunction: X/XΣ	67	X/Xt (465)	[klu/klut]	1,509
(d)	(de)nasalization	6	Xɔ̃/Xɔn (497)	[tɔ̃/tɔn]	582
(e)	stem ending variation: Xϕ/Xϕ'	22	Xk/Xʃ (25)	[aʁk/aʁʃ]	143
(f)	stem internal variation: XΣC/XΣ'C	42	Xɛʁ/Xaʁ (46)	[klɛʁ/klaʁ]	229
(g)	other cases	86			337

Table 8: Distribution of phonological alternation patterns in 5,500 Phononette entries

Unsurprisingly, this synthesis confirms that morphological relations mostly involve stem identity (X/X). Another expected result, due to the annotation strategy explained in Section 4.3, regards the important representation of the X/Y pattern. Moreover, in line with the literature (Pagliano, 2004), the data confirm that [t] epenthesis is the most frequent pattern of sequence addition (Table 8-(c)). A closer look at epenthesis (Section 5.2) will allow us to refine this result. The other figures in Table 8 are also interesting but deserve to be confirmed by an account on the complete database.

First, the most represented (non suppletive) variation, both in terms of number of different patterns (Column 2) and number of different word pairs (Column 5), is the addition of a phonetic sequence (Table 8-(c)): 67 patterns, 24 of which being cases of epenthesis, 1,509 derivational relations. A further

result not displayed in the table is that epenthesis patterns apply to 73%, that is, 1,099/1,509 of these relations.

Second, Table 8-(d-e) seem to indicate that phoneme alternation, as in *thon*$_N$*/thonier*$_N$ 'tuna/tuna boat' in (d) or *arc*$_N$*/archet*$_N$ 'bow' in (e), is both less frequent (22 patterns vs. 42) and less represented in derivational relations (143 vs. 229) than internal stem alternation (Table 8-(f)). Additionally, in the current state of the database, stem alternation consists mainly of apophony – they form 31 of the 42 patterns, and apply to 206 of the 229 relations (these figures are not shown in the table) – as illustrated with *clair*$_A$*/clarté*$_N$ 'clear/clarity'.

5 Phononette use cases

The descriptions and measures summarized above can be used in a variety of ways in morphophonology (Kiparsky, 1996; Spencer, 2017; Tranel, 1981). Below we present two results: (i) a new insight into the debate about the allomorphy/suppletion distinction; (ii) a case study of epenthesis, where the aim is to identify the factors favoring the emergence of the different epentheses.

5.1 Somewhere in between allomorphy and suppletion

Formal series, that is, **PAP**s, and derivational families (Section 4.4), can provide a new explanation for the distinction between allomorphy and suppletion in derivation. Precisely, the data encoded in Phononette are used to assess the impact of each of the four criteria presented in Table 2 in order to shed new light on the classification of derivational phonological alternation. Since "local" comparison criteria are well-known, we focus on the specific impact of the two "global" criteria, that is, $\mathbf{C_3}$, based on the frequency of each stem, and $\mathbf{C_4}$, based on the **PAP** frequency.

To do so, in each example of Table 9, two **($\mathbf{W_1}$/$\mathbf{W_2}$)** relations are compared, where criteria $\mathbf{C_1}$ and $\mathbf{C_2}$ have the same value, so their effect is neutralized. We examine whether $\mathbf{C_3}$ and $\mathbf{C_4}$ play a decisive role in a more accurate ranking of alternations, reflecting their different predictability.

- In (a) two suppletive forms ($\mathbf{C_1}$ = 'NO') of *eau*$_{\text{N}}$ 'water' are compared. Here, $\mathbf{C_4}$ is not relevant (suppletion means a PAP frequency of 1). What makes the difference is $\mathbf{C_3}$: the stem [idʁat] is found in half (13/27) of the words of the family of *eau*, whereas [ev] occurs only once (1/27). The distance between these two values clearly indicates that [idʁat] is much more "regular" for speakers than [ev], and thus much less "suppletive".

- In (b), the two compared stem pairs have the same string edit distance ($\mathbf{C_1}$ = 3), and alternation preserves integrity ($\mathbf{C_2}$ = +) in both cases. In the *corps/corselet* 'body/corselet' word pair, the alternating stem [kɔʁsəl] occurs only once in its family – the stem used for the other derived words is [kɔʁs] as in *corsage*$_{\text{N}}$ 'bodice', *corser*$_{\text{V}}$ 'make spicy', and *corset*$_{\text{N}}$ 'corset' – and the frequence of **PAP** is 1. On the other hand, for *instruire/instructeur* 'teach/instructor', the allomorphic stem "saturates" the family because it is used in the formal representation of all its members (*instructif*$_{\text{A}}$ 'instructive', *instructrice*$_{\text{N}}$ 'female instructor', *instructivement*$_{\text{Adv}}$ 'instructively'), and the Xɥi/Xykt **PAP** is quite largely used in the lexicon since it applies to all the word pairs containing any *-uire* ending verbs (*conduire* 'drive', *déduire* 'deduce', *introduire* 'introduct', *produire* 'produce', etc.) as well as its masculine (*producteur* 'producer', etc.) or feminine (*productrice* 'female producer', etc.) agent noun, or when attested, an *-if* suffixed adjective (*productif* 'productive', etc.).

- In (c) two cases of internal vowel epenthesis are compared (therefore $\mathbf{C_2}$ = −). The stem [stabil] of *stable/stabilité* is both present in almost the whole family (17/19) and widespread in the lexicon, confirming our assumption in Section 2. The situation with *diable/diabolique* 'devil/devilish' is the opposite for both criteria.

5.2 Different types of epenthesis

In Section 5.1, we saw that $\mathbf{C_3}$ and $\mathbf{C_4}$ have a decisive power in the morphophonological classification of derivational relations. Their value seems consistent with the speakers' ability to produce the alternating

	W_1/W_2		MSS	C_1	C_2	C_3	C_4
(a)	*eau*$_N$/*hydrater*$_V$	'water/hydrate'	[o]/[idʁat]	NO	+	13/27	1
	eau$_N$/*évier*$_N$	'water/sink'	[o]/[ev]	NO	+	1/27	1
(b)	*corps*$_N$/*corselet*$_N$	'body/corselet'	[kɔʁ]/[kɔʁsəl]	3	+	1/8	1
	instruire$_V$/*instructeur*$_N$	'teach/instructor'	[ɛ̃stʁɥi]/[ɛ̃stʁykt]	3	+	saturated	>26
(c)	*stable*$_A$/*stabilité*$_N$	'stable/stability'	[stabl]/[stabil]	1	−	saturated	700
	diable$_N$/*diabolique*$_A$	'devil/devilish'	[djabl]/[djabol]	1	−	2/7	1

Table 9: Combination of criteria for classifying morphophonological alternations

form without hesitation. Let us turn now to examples of external consonant epenthesis (thus preserving integrity, and with $\mathbf{C_1} = 1$).

Often, epenthesis analysis draws on etymological, phonological motivations (Moradi, 2017), or on the grapheme-phoneme correspondence, as shown for example by Corbin (1987) or Pagliano (2004) for French.

Table 10 illustrates how the grapheme-phoneme relationship, when combined with criteria C3 and C4, can shed new light on this question. We assume (see Section 4.3) that we have access to the ending graphic consonant of the base word (i.e., $\mathbf{W_1}$). It takes the value 't' in Table 10-(a) and (d), 'p' in Table 10-(c), and remains undefined in Table 10-(b), since the base word (*clou*$_N$) ends with a vowel.

Example (a) corresponds to the ideal situation: the epenthetic consonant is [t], which as we know is the most represented epenthesis in the lexicon. Moreover, it realizes the (silent) graphic ending of the base and the epenthetic stem saturates the *abricot*$_N$ 'abricot' family. In (b), $\mathbf{C_4}$ has the same, high frequency value for epenthesis (see Table 8), but the epenthesis stem here does not saturate the family – half of its members are formed from the non alternating stem [klu], including a rival verb, *clouer* 'nail') – and the consonant is not the realization of the base graphic ending. In (c), the epenthetic stem is found in all its family and the epenthetic consonant matches the base final letter. On the other hand, $\mathbf{C_4}$ value is quite low, while X/Xp is systematically used for derivational relations where the base word graphically ends with 'p'. We have here a case of lexical saturation of X/Xp. Finally, in (d), $\mathbf{C_3} = 1$, and the base ending graphic character 't' is different from the epenthetic consonant [z]. These two properties suggest that [nɥiz] is not an expected epenthetic stem for [nɥi]. The value $\mathbf{C_4} = 29$ seems to contradict this conclusion, but a closer look reveals that in almost all the word pairs with a stem variation matching X/Xz, the base ending graphic character has never the value 't'.

	W_1/W_2		MSS	C_3	C_4	W_1 Gr. ch
(a)	*abricot*$_N$/*abricotier*$_N$	'apricot/apricot tree'	[abʁiko]/[abʁikot]	saturated	**X/Xt** = 465	't'
(b)	*clou*$_N$/*clouter*$_V$	'nail/nail'	[klu]/[klut]	half family	**X/Xt** = 465	–
(c)	*champ*$_N$/*champêtre*$_A$	'field/of field'	[ʃã]/[ʃãp]	saturated	**X/Xp** = 6	'p'
(d)	*nuit*$_N$/*nuisette*$_N$	'night/babydoll'	[nɥi]/[nɥiz]	1	**X/Xz** = 29	't'

Table 10: Epenthesis: family, series and graphic consonant

As we can see here, the value of the graphic consonant at the end of the base is another feature to be considered when estimating the predictability of epenthetic variants: the results show that when the epenthetic phoneme, whatever it may be, realizes the ending graphic consonant of the base, then the resulting epenthetic stem is mostly the one that is best represented in its family, which gives it a privileged formal status, that we could call "family stem".

At a more general level, it seems that epenthesis brings four dimensions into play, which interact in derivation in the relation between stems: (i) the **phonological properties** of the alternating sequence

(here, the predominance of [t] for epenthesis), (ii) the general **lexical organization**, resulting in the frequency of the alternation pattern, (iii) the **structure of word families**, that is, the level of saturation of word families by each of their ancestor's stem, which has to be considered in conjunction with (iv) the **relation** between the **phonological form** of derived stems and the **graphic form** of their base.

6 Conclusion

Global measures (stem frequency in the family, **PAP** frequency) shed new light on the question of morphophonological alternations in the morphologically complex lexicon of French, by introducing two new criteria that interact with the traditional factors of string similarity and stem integrity. Alternations can be classified according to a combination of these four criteria, which appear to provide superior accuracy in predicting alternations to models based solely on formal proximity. The study of epenthesis shows that alternations are not (only) a matter of derivational relations between lexemes, but also, and above all, a question concerning their families and the series of relations that verify the same variations. Morphophonology in word formation can thus already be considered from a paradigmatic point of view.

References

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*. Association for Computational Linguistics, pages 58–69. https://doi.org/10.3115/1118647.1118654.

Adam Albright and Bruce P. Hayes. 2006. Modelling productivity with the gradual learning algorithm: The problem of accidentally exceptionless generalizations. In Gisbert Fanselow Fanselow, Caroline Féry, Matthias Schlesewsky, and Ralph Vogel, editors, *Gradience in Grammar: Generative Perspectives*, Oxford University Press, Oxford, pages 185–204.

Mark Aronoff. 1994. *Morphology by Itself*. MIT Press, Cambridge.

Mark Aronoff. 2012. Morphological stems: what William of Ockham really said. *Word Structure* 5(1):28–51.

Laurie Bauer. 1997. Derivational paradigms. In *Yearbook of Morphology 1996*, Springer, Cham, pages 243–256.

Sacha Beniamine. 2017. Un algorithme universel pour l'abstraction automatique d'alternances morphophonologiques. In Iris Eshkol-Taravella and Jean-Yves Antoine, editors, *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*. ATALA, Orléans, France, pages 77–85.

Harry Bochner. 1993. *Simplicity in Generative Morphology*. Mouton de Gruyter, Berlin.

Olivier Bonami and Sacha Beniamine. 2021. Leaving the stem by itself. In Sedigheh Moradi, Marcia Haag, Janie Rees-Miller, and Andrija Petrovic, editors, *All Things Morphology: Its Independence and its Interfaces*, John Benjamins, Amsterdam, pages 81–98.

Olivier Bonami and Gilles Boyé. 2003. Supplétion et classes flexionnelles. *Langages* 152:103–126.

Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29(2):167–197.

Geert Booij. 2009. Allomorphy and the autonomy of morphology. *Folia Linguistica* 31(1-2):25–56.

Gilles Boyé. 2006. Suppletion. In Keith Brown, editor, *Encyclopedia of Language and Linguistics, 2nd Edition*, Elsevier, Oxford, volume 12, pages 297–299.

Andrew Carstairs. 1988. Some implications of phonologically conditioned suppletion. In *Yearbook of Morphology 1987*, Springer, Cham, pages 67–94.

Danielle Corbin. 1985. Les bases non-autonomes en français ou comment intégrer l'exception dans le modèle lexical. *Langue Française* 66:54–76.

Danielle Corbin. 1987. *Morphologie Dérivationnelle et Structuration du Lexique*. Presses Universitaires de Lille, Lille.

Wolfgang U. Dressler. 1985. Suppletion in word formation. In Jacek Fiziak, editor, *Historical Semantics, Historical Word Formation*, Mouton de Gruyter, Berlin, pages 97–112.

Wolfgang U. Dressler. 2015. Allomorphy. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe, Volume 1*, De Gruyter Mouton, Berlin, pages 500–516.

Bernard Fradin. 2018. Paradigms and the role of series in derivational morphology. *Lingue e linguaggio* 17(2):155–172.

Nabil Hathout. 2011. Une analyse unifiée de la préfixation en *anti-*. In Michel Roché, editor, *Des Unités Morphologiques au Lexique*, Hermès, Paris, pages 251–318.

Nabil Hathout and Fiammetta Namer. 2022. ParaDis: a family and paradigm model. *Morphology* 32(2):153–195.

Nabil Hathout and Fiammetta Namer. To appear. What do derivational paradigms tell us about back-formation and what does back-formation tell us about derivational paradigms? *Word Structure* .

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. GLÀFF, a large versatile French lexicon. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, pages 1007–1012. https://aclanthology.org/L14-1469/.

Andrew Hippisley, Marina Chumakina, Greville Corbett, and Dunstan Brown. 2004. Suppletion: frequency, categories and distribution of stems. *Studies in Language* 28(2):387–418.

Mathilde Huguin, Lucie Barque, Pauline Haas, Fiammetta Namer, and Delphine Tribout. 2022. Guide d'annotation Demonext - typage sémantique de lexèmes nominaux hors contexte. Report, CNRS ATILF.

Nancy Ide and Jean Veronis. 1994. MULTEXT: Multilingual text tools and corpora. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*. Kyoto, Japan. https://aclanthology.org/C94-1097/.

Paul Kiparsky. 1996. Allomorphy or morphophonology? In Rajendra Singh, editor, *Trubetzkoy's Orphan. Proceedings of the Montréal Roundtable "Morphonology: Contemporary Responses"*. John Benjamins, Amsterdam, pages 13–31.

Rochelle Lieber. 1982. Allomorphy. *Linguistic analysis* 10(1):27–52.

Igor Mel'çuk. 1994. Suppletion: Toward a logical analysis of the concept. *Studies in Language* 18(2):339–410.

Sedigheh Moradi. 2017. Non-canonical epenthesis: epenthetic quality and the role of morphonology. Manuscript, Stony Brook University.

Fiammetta Namer and Nabil Hathout. 2020. ParaDis and Démonette – From theory to resources for derivational paradigms. *The Prague Bulletin of Mathematical Linguistics* 114:5–33. https://doi.org/10.14712/00326585.001.

Fiammetta Namer, Nabil Hathout, Dany Amiot, Lucie Barque, Olivier Bonami, Gilles Boyé, Basilio Calderone, Julie Cattini, Georgette Dal, Alexander Delaporte, Guillaume Duboisdindien, Achille Falaise, Natalia Grabar, Pauline Haas, Frédérique Henry, Mathilde Huguin, Juniarta Nyoman, Loïc Liégeois, Stéphanie Lignon, Lucie Macchi, Grigoriy Manucharian, Caroline Masson, Fabio Montermini, Nadejda Okinina, Frank Sajous, Daniele Sanacore, Mai Thi Tran, Juliette Thuilier, Yannick Toussaint, and Delphine Tribout. 2023. Démonette-2, a derivational database for french with broad lexical coverage and fine-grained morphological descriptions. *Lexique* 33:6–40.

Claudine Pagliano. 2004. Elaboration d'un corpus morphophonologique : l'épenthèse consonantique à la frontière suffixale en français. *Corpus* 3:357–398.

Marc Plénat. 2008. Le thème l de l'adjectif et du nom. In Jacques Durand, Benoît Habert, and Bernard Laks, editors, *1er Congrès Mondial de Linguistique Française*. ILF, Paris, pages 1613–1626.

Frank Sajous and Nabil Hathout. 2015. Glawi, a free xml-encoded machine-readable dictionary built from the French Wiktionary. In Iztok Kosem, Miloš Jakubíček, Jelena Kallas, and Simon Krek, editors, *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, Ljubljana/Brighton, pages 405–426.

Andrew Spencer. 2017. Morphophonological operations. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, John Wiley & Sons, Hoboken, pages 123–143.

Anna Maria Thornton. 1997. Stem allomorphs, suffix allomorphs, interfixes or different suffixes? on italian derivatives with antesuffixal glides. In Geert Booij, Angela Ralli, and Sergio Scalise, editors, *Proceedings of the 1st Mediterranean Meeting of Morphology*. University of Mytilene, pages 86–98.

Bernard Tranel. 1981. *Concreteness in Generative Phonology: Evidence from French*. University of California Press, Berkeley.

Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6):1091–1095.

Pavol Štekauer. 2014. Derivational paradigms. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, Oxford, Oxford University Press, Oxford, pages 354–369.

A data-based analysis of the effect of prefixation on the syntactic-semantic characteristics of verbs in Czech

Hana Hledíková
Institute of Formal and Applied Linguistics
Charles University, Faculty of Mathematics and Physics
hana.hledikova@ufal.mff.cuni.cz

Abstract

This paper presents a data-based case study of the interaction between word-formation and the syntactic-semantic characteristics of verbs (i.e., *valency*). We carry out an analysis of Czech verbs created by prefixation using a combination of language resources which contain information about the verbs' derivational and morphemic characteristics (Slavíčková, 1975; Vidra et al., 2021) and valency characteristics (Lopatková et al., 2022; Urešová et al., 2023, 2024). The results show that the addition of a prefix causes a change to the verb's valency characteristics in about a third of the analysed items. The most frequent changes involve a particular set of syntactic elements (namely those denoting the patient, addressee, effect, origin, and direction). The addition of the patient is the most frequent change overall, which is in accordance with previous analyses of prefixation in Slavic languages (cf. Romanova, 2006; Ramchand, 2008; Biskup, 2019). However, the data also document a significant number of deletions, which have not previously been paid as much attention as additions in the theoretical literature. Overall, the effects of prefixation are shown to be highly diverse and individual prefixes are shown to be polyfunctional both in terms of the meaning they add to the base verb and the type of change they cause to the base verb's valency characteristics.

1 Introduction

The addition of a prefix, as in examples (1)-(3),[1] is the most frequent way of forming verbs from other verbs in Czech and other Slavic languages (cf. Körtvélyessy, 2016).

(1) *ps-á-t* > *vy-ps-a-t*
write-THEME-INF out-write-THEME-INF
'to write (impf.)' 'to write something down (from somewhere) (pf.)'

(2) *prac-ova-t* > *vy-prac-ova-t*
work-THEME-INF out-work-THEME-INF
'to work (impf.)' 'to work something out, develop, create (pf.)'

(3) *ps-á-t* > *na-ps-a-t*
write-THEME-INF on-write-THEME-INF
'to write (impf.)' 'to write (pf.)'

The prefixed verbs are related to the base verb both formally and semantically. The prefix typically adds some lexical meaning to the base verb and also makes the verb telic and perfective in grammatical aspect (cf. Biskup, 2019, p. 18). In example (1), the prefix *vy-* adds the concrete spatial meaning 'out' to the meaning of the input verb *psát* 'to write'. However, as is typical of prefixes in Slavic languages in general, Czech prefixes are polyfunctional and some of their meanings are only loosely linked to the concrete spatial meaning (cf. Romanova, 2006, p. 95; Sussex and Cubberley, 2006, pp. 444–447, Biskup,

[1] In all examples in this paper, prefixes are glossed with their primary spatial meaning, although they are polyfunctional and have shifted to other meanings in some of the examples.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 63–72, Fribourg, Switzerland, 4-5 September 2025.

2019, pp. 53–76). In example (2), the same prefix *vy-* has an abstract completive meaning. In some cases, the prefix can even be argued to have no lexical meaning at all and to only make a change to the grammatical aspect of the base verb (cf. e.g., Komárek, 1984), as in example (3).

Both the input verbs and the output prefixed verbs also have a potential to require certain types of syntactic elements, which we will call *valency* (cf. the following Section 2 for a description of the valency theory we use). For example, as is manifested in the sentence in example (4), the verb *psát* 'to write' has the potential to appear with syntactic elements expressing the agent (*Petr* 'Peter') and the piece of writing / information written by the agent (*zpráva* 'a report'). The process of adding a prefix to a verb can be accompanied by changes to this potential. When we derive the verb *vypsat* 'to write something down (from somewhere)' from the verb *psát* 'to write', the prefixed verb requires one more element, namely the place where the information was originally found (e.g., *z encyklopedie* 'from the encyclopedia' in example (5)).

(4) *Petr* ***píš-e*** *zpráv-u.*
Petr write-3SG.PRS report-ACC.SG
'Peter is writing a report.'

(5) *Petr* ***vy-ps-a-l*** *poznámk-y z encyklopedi-e.*
Petr out-write-THEME-3SG.PST note-ACC.PL from encyclopedia-GEN.SG
'Peter wrote down notes from the encyclopedia.'[2]

The availability of language data resources makes it possible to take a data-based, quantitative approach to this interaction between prefixation and the verbs' valency, which the previous literature describes mostly from a formal perspective using a number of selected examples. Derivational resources, such as DeriNet (Vidra et al., 2021) for Czech, capture links between words related by different word-formation processes, including prefixation. However, derivational resources do not include information about the words' semantics and valency characteristics. In this contribution, we combine the derivational data resources with valency lexicons (Lopatková et al., 2022; Urešová et al., 2023, 2024) in order to analyze the patterns of changes in the valency behaviour of verbs that happen in the process of prefixation and to evaluate the expectations from previous theoretical literature against a larger data sample. We investigate this phenomenon on verbs in Czech; however, the methodology is applicable for analyzing other languages for which derivational and valency resources are available.

2 Theoretical background

The syntactic elements required by the verb, which we have described in pre-theoretical terms so far, can be categorized differently depending on the theoretical framework as well as the level of linguistic description. On the surface-syntactic level, the verb in example 5 requires a subject (in this sentence, it is expressed by a noun in the nominative case), an object (here expressed by a noun in the accusative case) and an adverbial (here expressed by a prepositional phrase). On the semantic level, the verb typically denotes a kind of event and therefore implies the involvement of specific participants – in other words, the verb acts as a predicate that requires certain arguments. Therefore, the required syntactic elements can together be called the verb's *argument structure* (cf. e.g., Williams, 2015). When describing a verb's argument structure, the arguments can either be simply numbered in order of their prominence, as in e.g., PropBank (Pradhan et al., 2022), or their function can be further described on different levels of generalization – using a set of abstract thematic roles, such as *Agent/Actor*, *Theme*, *Source* (cf. e.g., Jackendoff, 1990), or more concretely using the frame elements that are part of the situation denoted by the specific verb, such as *Author*, *Text*, *Source text*, as in the FrameNet lexicon (Baker and Sato, 2003) created based on the theory of frame semantics (Fillmore and Baker, 2001).

For Czech, there is a tradition of using the term *valency* when talking about the verb's potential to bind certain elements, which can be referred to as *valency slots* that make up the verb's *valency frame*, and a description of both their form and function is part of the Functional Generative Description framework

[2] An analogical example with this verb in Russian is discussed in Romanova (2006, p. 72).

(FGD, cf. Panevová, 1974, 1975). In this description, the function of syntactic elements is defined using *functor* labels, which capture the role that the elements have in relation to the verb. To distinguish which elements in the sentence are bound by the specific verb, FGD makes the distinction between arguments and adjuncts (similarly to other theories of valency/argument structure). There is a total of five types of arguments: Actor (ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF), which are always part of the verb's valency frame (no matter if they are obligatorily expressed when the verb is used in a sentence or can be left out). The five arguments undergo shifting, which means that if a verb has one argument, it is the Actor, if it has two, it is the Actor and Patient, and if it has three, one of the three other arguments is added based on the meaning. However, some adjuncts, such as Manner (MANN), Location (LOC) or Direction (DIR1 'from', DIR2 'through', DIR3 'to'), can also be part of the verb's valency frame in cases when they are obligatory. This is the case of the DIR1 adjunct *z encyklopedie* 'from an encyclopedia' in example 5 – therefore, the verb *vypsat* 'to write down' (in the sense that it is used in the example) has a valency frame consisting of three valency slots: ACT, PAT, DIR1.

The effects of prefixation on valency have not been described using the FGD framework so far, but they have long been a topic of research in Slavic (Jirsová, 1979; Uher, 1987; Svenonius, 2004; Romanova, 2006; Gehrke, 2008; Ramchand, 2008; Biskup, 2019) as well as in other languages (Wunderlich, 1987; Lieber and Baayen, 1993; Stiebels, 1996; McIntyre, 2003). Studies on Slavic argue that although the specific semantic contribution of individual prefixes is of various types and levels of concreteness, in general the meaning often corresponds to the causation of some kind of resultative state.[3] In cases where the prefix has a spatial meaning, the resultative state is a location, and because a location implies a reference object, this can lead to an addition of a valency slot, as in example 5 where the prefix *vy-* means 'out' and requires a reference object (out of *what*) to be expressed. In cases where the meaning of the prefix is not spatial, the resulting state is more abstract and can involve an additional affected entity, which is not part of the input verb's valency frame – cf. example (6):

(6) *ží-t* > *pro-ží-t* *šťastn-é* *dětstv-í*
live-INF > through-live-INF happy-ACC.SG childhood-ACC.SG
'to live' > 'to experience a happy childhood'

Based on the previous theoretical literature, the expectation is therefore that prefixation will most often lead to the addition of a valency slot expressing a directional modification (DIR1, DIR2, DIR3) in cases where the prefix has a concrete spatial meaning, and to the addition of the PAT valency slot in cases where the meaning is more abstract and the resulting state is not a location.

3 Data and method

First, we compiled a sample of Czech verbs formed by prefixation from another verb. We used a list of verbal lemmas extracted from a Czech corpus of the size of 100 million tokens (Čermák et al., 2000) and semi-automatically annotated their morphemic structure and word-family membership using derivational and morphematic resources (Slavíčková, 1975; Vidra et al., 2021).[4] From the list, we extracted pairs of prefixed verbs and their base verbs (such as those in examples (1)-(3)).

Next, to get information about the valency characteristics of the extracted verbs, we use Vallex (version 4.5, Lopatková et al., 2022) and PDT-Vallex (version 4.5, Urešová et al., 2024), which are valency lexicons of Czech verbs using the FGD framework (cf. Section 2). The verbs' valency frames are defined on the level of individual senses, which means that one verbal lexeme often has multiple valency frames. This makes it possible to carry out the analysis on the level of individual senses of the verbs, which is advantageous because not all senses of the prefixed verb are necessarily connected to all senses of the base verb.

[3] Some authors argue that this is only true for a certain group of prefixes which can affect the valency of the input verb, the so called *lexical prefixes* as opposed to *superlexical prefixes*. However, this distinction has been questioned and superlexical prefixes have been shown to also affect the valency of the input verbs (Biskup, 2019); therefore, we do not work with this distinction.

[4] The annotated list is publicly available under this link: http://hdl.handle.net/11234/1-5824 (Hledíková, 2024).

In Vallex, the individual verbs' valency frames are not connected to each other in any way. In order to link the valency frames of the base verb to those of the prefixed verb on the level of their senses, we use SynSemClass (Urešová et al., 2023), a lexicon which groups individual senses of verbs into classes based on the similarity of their meaning, which is defined as the general type of situation which they denote. The situation types are delimited using a set of semantic roles, and they roughly correspond to FrameNet (Baker and Sato, 2003) frames and rolesets. Because SynSemClass works on the level of individual senses, a polysemous verbal lexeme can be included in several different classes. For instance, the verb *hrát* 'to play' is assigned three different classes for three of its individual meanings, as shown in example (7):

(7) a. 'to play music' - `vec00823`, roleset: *Performer*, *Music*; valency frame: ACT, PAT
b. 'to perform a role' - `vec00611`, roleset: *Performer*, *Role*; valency frame: ACT, PAT
c. 'to play in a competition' - `vec00415`, roleset: *Competitor 1 and 2*, *Competition*, *Targeted*; valency frame: ACT, PAT, ADDR, EFF

Each verbs' entry in SynSemClass is linked to a particular entry in Vallex or PDT-Vallex, so that it is possible to automatically extract the valency frame(s) of the verb in its specific sense (as defined by SynSemClass). Example 7 shows the valency frames from PDT-Vallex (Urešová et al., 2024) linked to each sense of the verb *hrát* 'to play'.

Out of all the pairs of 'prefixed verb – base verb' we have extracted, 1,076 have both the prefixed and base verb available in SynSemClass. For each of these pairs, we manually linked the class(es) available for the prefixed verb to the class(es) available for the unprefixed verb so that only those meanings of the verbs that are related are linked. For example, the verb *prohrát* 'to lose' in class `vec00275` (roleset: *Winner*, *Loser*, *Competition*) is linked to to the verb *hrát* 'to play' in class `vec00415` (roleset: *Competitor 1 and 2*, *Competition*, *Targeted*), but no link is made between *prohrát* 'to lose' and other senses of *hrát* 'to play', because they are not related. When both the prefixed verb and its base verb are in the same class, their senses and corresponding valency frames are linked automatically and do not have to go through the process of manual annotation, which makes the annotation procedure significantly easier. There are 630 verb pairs for which there is at least one such link between the two verbs' senses.[5] The sample contains 869 individual links between a class of the prefixed verb and a class of its corresponding unprefixed verb (e.g., *hrát* 'to play' `vec00415` – *prohrát* 'to lose' `vec00275`).

The valency frames are retrieved and then compared for the unprefixed and prefixed verb in each pair.[6] In case there is an addition or a deletion of one or more valency slots, it is recorded as the valency change that happens in the process of prefixation for the particular item. For example, the verb *hrát* 'to play' `vec00415` has four valency slots (ACT, PAT, ADDR, EFF), while the verb *prohrát* 'to lose' `vec00275` only has three (ACT, PAT, ADDR), and so the change is the deletion of the EFF argument (which we will mark as –EFF).

4 Results and discussion

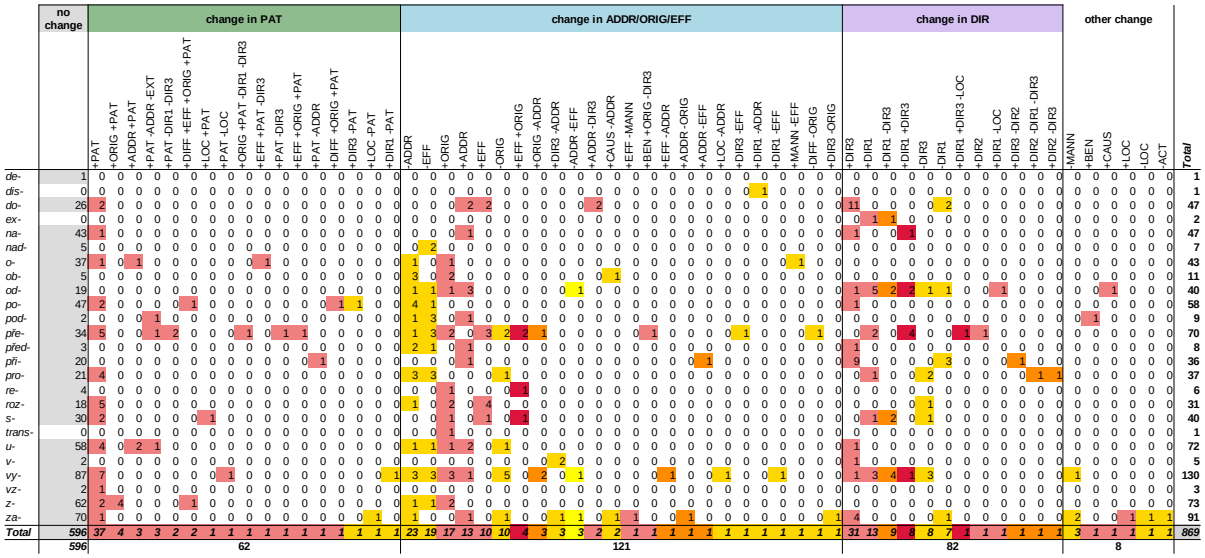
Figure 1 shows the number of items (i.e., relationships between an individual sense of a prefixed verb and an individual sense of its base verb) documented for each type of valency change (addition and/or deletion of a valency slot) and each prefix. The valency changes are presented in groups according to the hierarchy of valency slots (cf. Section 2): 1) no change, 2) change in PAT, 3) change in ADDR/ORIG/EFF, 4) change in DIR,[7] 5) change in other types of adjuncts.

Out of the total 869 items, 597 items (68,7%) have the same valency frame for the unprefixed and prefixed verb. This means that in the majority of cases, prefixation does not lead to any change in the valency slots. In 417 (48%) out of these items, the unprefixed and prefixed verb also have the same class. 105 of them are listed by Uher (1987) as purely aspectual uses of prefixes, i.e., those prefixes that

[5] It is important to note that SynSemClass does not exhaustively capture all possible senses of each verb that it contains.

[6] The level of generalization used in Vallex is advantageous when comparing valency slots across different verbs, because they are less verb-specific than the FrameNet-style rolesets. This is why the analysis is not carried out directly using the rolesets.

[7] Because we expect a large quantity of changes in DIR valency slots in cases where the prefixes have a spatial meaning, we defined this as a special group of changes in adjuncts.



.

Figure 1: Changes in valency frames between individual senses of prefixed verbs and their base verbs. Values are counts of items with each type of change in valency frame (columns) and each prefix (rows). Additions of valency slots are marked with +, deletions with –. Changes with more additions than deletions are marked in shades of red, changes with more deletions than additions in shades of yellow, changes with an equal number of additions and deletions in orange. The changes are classified into 5 groups based on the hierarchy of valency slots. For the interpretation of valency slot labels that have not been mentioned in the text, see the Appendix or the following documentation: https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch07.html

supposedly function as markers of perfective grammatical aspect without adding any lexical meaning. In the rest of the items which exhibit no change either in the class or the valency frame, the prefix can add a modification that is not considered as relevant in distinguishing the classes in SynSemClass, such as phasal or quantifying characteristics (e.g., *řešit* 'to solve' – *dořešit* 'to finish solving', *pršet* 'to rain' – *zapršet* 'to rain a little').

A change in the valency frame is documented for 272 items (31,3%). The data show a total of 59 different individual types of changes, 18 of which involve the addition of one or more valency slots, 10 involve a deletion, and 31 involve both an addition and a deletion. The most frequent changes concern the arguments PAT, ADDR, ORIG and EFF and the directional adjuncts DIR1 and DIR3.

It is clear from Figure 1 that the attestation of the different changes in valency is generally sparse. Overall, individual prefixes are highly polyfunctional in terms of the type of changes that they cause; however, they differ in the distribution of the valency changes they are attested with. For example, the *od-* and *do-* prefixes appear mostly with changes in directional modifications. Some prefixes are also more likely to cause a change than others (cf. *na-*, which does not cause any valency change in all but four items).

The most frequent change, overall, is the addition of the patient (PAT) valency slot. The changes in PAT are additions in the absolute majority (95%) of cases. This confirms the expectation that the resultative nature of the prefixes can bring about an affected entity that was not a necessary part of the input verb's valency. A closer look at the semantics of the items with the addition of PAT shows that they include:

- verbs of creation (*myslet* 'to think' – *vymyslet něco* 'to think up, invent something');
- verbs of destruction (*jet* 'to drive' – *přejet někoho* 'to run somebody over');
- verbs where the prefix has an attenuative meaning (*jíst* 'to eat' – *ujíst něco* 'to eat a bit of sth');

- verbs where the PAT argument denotes an amount of time that is used up (*spát* 'to sleep' – *prospat něco* 'to sleep through something');
- verbs where the prefixed verb is causative while the base verb is inchoative (*sílit* 'to become stronger' – *posílit něco* 'to strengthen, make something stronger').

Table 1 shows an example of a verb pair with an addition of the PAT valency slot for each prefix.

prefix	**# items**	**example**
do-	2	*končit* 'to end' – *dokončit něco* 'to finish something'
na-	1	*být* 'to be' – *nabýt něco* 'to gain something '
o-	1	*zářit* 'to shine' – *ozářit něco* 'to shine onto something'
po-	2	*sílit* 'to become stronger' – *posílit něco* 'to strengthen something'
pře-	5	*jet* 'to drive' – *přejet někoho* 'to run somebody over'
pro-	4	*spát* 'to sleep' – *prospat něco* 'to sleep through something'
roz-	5	*plakat* 'to cry' – *rozplakat někoho* 'to make somebody cry'
s-	2	*trpět* 'to suffer' – *strpět něco* 'to tolerate something'
u-	4	*končit* 'to end' – *ukončit něco* 'to put an end to something'
vy-	7	*pracovat* 'to work' – *vypracovat něco* 'to work out, develop something'
vz-	1	*planout* 'to blaze' – *vzplanout (z něčeho)* 'to flame up (out of something)'
z-	2	*pracovat* 'to work' – *zpracovat něco* 'to process something'
za-	1	*žít* 'to live' – *zažít něco* 'to experience something'
Total	*37*	

Table 1: Prefixes which add the PAT valency slot and the number of items with this change for each prefix, along with an example of a prefixed verb and its base verb.

As for the changes in the ADDR/ORIG/EFF arguments, there is a significant proportion of deletions, especially of ADDR and EFF. As previous literature mostly talks about prefixation leading to valency slot additions, this deserves a further comment. The deletion of ADDR mostly concerns cases where the input verb is a verb of transfer or communication. In the prefixed verb, the resulting state can become more important than the addressee (cf. example (8)) or the addressee more important than the theme (cf. examples (9) and (10)), leading to changing the valency frame 'ACT PAT ADDR' in the base verb to 'ACT PAT' in the prefixed verb.

(8) *hlás-i-t* *něk-omu* *něc-o* > *vy-hlás-i-t* *něc-o*
call-THEME-INF somebody-DAT.SG something-ACC.SG out-call-THEME-INF something-ACC.SG
'to report something to somebody' 'to announce something'

(9) *dar-ova-t* *něk-omu* *něc-o* > *ob-dar-ova-t* *něk-oho*
gift-THEME-INF somebody-DAT.SG something-ACC.SG around-gift-THEME-INF somebody-ACC.SG
'to gift something to somebody' 'to give somebody a gift'

(10) *plat-i-t* *něk-omu* *něc-o* > *pod-plat-i-t* *něk-oho*
pay-THEME-INF somebody-DAT.SG something-ACC.SG under-pay-THEME-INF somebody-ACC.SG
'to pay somebody something' 'to bribe somebody'

The deletions of EFF often concern verbs of cognition and evaluation. As demonstrated in examples (11) and (12), while in the input verb the EFF valency slot specifies the way something is evaluated, in the prefixed verb this is already clear from the prefix.

(11) *soud-i-t* *o* *něč-em* *že* *je* *to špatn-é* > *od-soud-i-t* *něc-o*
judge-THEME-INF about something-LOC.SG that be-3SG.PRS it bad-NOM.SG away-judge-THEME-INF something-ACC.SG
'to judge something to be bad' 'to condemn something'

(12) *hodnot-i-t* *něc-o* *jako důležit-é* > *nad-hodnot-i-t* *něc-o*
value-THEME-INF something-ACC.SG as important-ACC.SG over-value-THEME-INF something-ACC.SG
'to evaluate something as important' 'to overvalue, overestimate something'

The changes in DIR concern items in which the prefix has a concrete spatial meaning. The most frequent change is the addition of DIR3 (*téci* 'to flow' – *přitéci do jezera* 'to flow into the lake'), followed by the addition of DIR1 (*letět* 'to fly' – *odletět z Čech* 'to fly away from Czechia'). This result is in accordance with the previous analyses of prefixes with a spatial meaning, which describe that the prefix denotes a spatial relation requiring the expression of a reference object (*do jezera* 'into the lake', *z Čech* 'from Czechia'). However, the data show that the spatial relation expressed in the valency slot is not necessarily the one corresponding to the cognate preposition of the prefix – for instance, *odletět* with the prefix *od-* 'from' can appear not only with DIR1 (expressing the source location), but also with a DIR3 valency slot expressing the goal location (*odlětet do Čech* 'to fly away to Czechia').

5 Conclusion

The analysis has shown that there is a great variety of valency changes that can accompany the process of prefixation in Czech; however, in about a half of all cases, there was no change either in the verb's semantic class or its valency frame. This result documents a high frequency of cases in which the prefix either adds a meaning that is not relevant to the definition of the semantic classes (e.g., certain temporal or quantifying characteristics) and cases in which (according to some analyses) the prefix adds no lexical meaning at all and only changes the grammatical aspect.

However, the analysis also provides support for the importance of the interaction between prefixation and valency. Cases where the prefix causes a change in the valency frame of the base verb make up about a third of our data sample. The analysis has shown that the most frequent changes are additions (most often of the patient or the direction 'to, towards'), which is in accordance with previous theoretical literature that argues that the addition of a prefix typically adds some kind of resultative meaning and may lead to the verb requiring an additional valency slot specifying the spatial reference object or the affected entity. However, the data also document that the addition of a prefix can lead to deletions, especially in the valency slots denoting the addressee and effect. These are changes that have not been paid much attention in previous literature, which has mostly focused on the additions of the patient and/or spatial modifications.

The individual prefixes were shown to be highly polyfunctional in terms of the types of valency changes they are connected with. The same is true the other way around: One type of valency change can be realized by multiple prefixes. For instance, the addition of a patient is connected with several prefixes, some of which have a meaning that is still closely connected to the spatial meaning of the cognate preposition (e.g., *jet* 'to drive' – *přejet* 'to run somebody over'), while others have an abstract causative/resultative meaning (e.g., *plakat* 'to cry' – *rozplakat někoho* 'to make somebody cry', *pracovat* 'to work' – *zpracovat něco* 'to process something').

The conclusions drawn from the analysis are limited by the level of generalization on which the valency slots are defined and the particular theoretical framework, as well as the coverage of the data resources used. Despite these limitations, this contribution demonstrates that it is possible to combine derivational resources and valency lexicons to investigate the interaction between word-formation and the syntactic-semantic characteristics of verbs. The methodology can be applied to examine this phenomenon in other languages than Czech in the future. A comparison with other Slavic languages would help clarify whether Czech follows Slavic-wide trends or exhibits unique properties. Also, since Slavic prefixes have been argued to have an effect on the syntactic-semantic characteristics of verbs that is similar to that of Germanic prefixes and verb particles (Svenonius, 2004; Ramchand, 2008), a data-based comparison of Slavic and Germanic languages would be beneficial in providing support for these claims.

Appendix

Abbreviations

impf. ... imperfective aspect
pf. ... perfective aspect

Glosses

ACC ... accusative
DAT ... dative
GEN ... genitive
INF ... infinitive
LOC ... locative
NOM ... nominative
PL ... plural
PRS ... present
PST .. past
SG ... singular
THEME ... thematic suffix
3 ... third person

Valency slot labels

ACT ... actor (*Father is working)*
PAT ... patient (*He is cooking lunch*)
ADDR ... addressee (*He sent a present to a friend*)
EFF ... effect (*They appointed him as a chairman*)
ORIG ... origo (*He makes furniture out of wood*)
EXT ... extent (*The child weighs five kilograms*)
DIR1 ... direction 1 (from) (*He came from Prague*)
DIR2 ... direction 2 (through) (*They are walking through the woods*)
DIR3 ... direction 3 (to) (*He came home*)
DIFF ... difference (*The temperature dropped ten degrees*)
LOC ... location (*He kept the children in bed*)
CAUS ... cause (*She suffered from illness*)
MANN ... manner (*He took it seriously*)

Acknowledgments

The study was supported by the Charles University, project GA UK No. 246723, by the SVV project No. 260 821 and by the Charles University Research Centre program No. 24/SSH/009.

References

Collin F. Baker and Hiroaki Sato. 2003. The FrameNet data and software. In Yuji Matsumoto, editor, *The Companion Volume to the Proceedings of 41st Annual Meeting of the ACL*. pages 161–164.

Petr Biskup. 2019. *Prepositions, Case and Verbal Prefixes: The Case of Slavic*. John Benjamins.

František Čermák, Renata Blatná, Jaroslava Hlaváčová, Jana Klímová, Jan Kocek, Marie Kopřivová, Michal Křen, Vladimír Petkevič, Věra Schmiedtová, and Michal Šulc. 2000. *SYN2000: Žánrově Vyvážený Korpus Psané Češtiny*. Ústav Českého Národního Korpusu FF UK, Praha.

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.

Berit Gehrke. 2008. *Ps in Motion: On the Semantics and Syntax of P Elements and Motion Events*. LOT, Utrecht.

Hana Hledíková. 2024. Verbs annotated for morphemic structure in Czech, English, German, Spanish. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5824.

Ray Jackendoff. 1990. *Semantic Structures*. MIT Press.

Anna Jirsová. 1979. Prefixace sloves a slovesná vazba. *Naše Řeč* 62(1):1–7.

Miroslav Komárek. 1984. Prefixace a slovesný vid (k prefixům prostě vidovým a subsumpci). *Slovo a Slovesnost* 45(4):257–267.

Lívia Körtvélyessy. 2016. Word-formation in Slavic languages. *Poznan Studies in Contemporary Linguistics* 52(3):455–501. https://doi.org/10.1515/psicl-2016-0002.

Rochelle Lieber and Herald Baayen. 1993. Verbal prefixes in Dutch: A study in lexical conceptual structure. In Geert E. Booij and Jaap van Marle, editors, *Yearbook of Morphology 1993*, Springer, Dordrecht, pages 51–78.

Markéta Lopatková, Václava Kettnerová, Jiří Mírovský, Anna Vernerová, Eduard Bejček, and Zdeněk Žabokrtský. 2022. VALLEX 4.5. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-4756.

Andew McIntyre. 2003. Preverbs, argument linking and verb semantics: Germanic prefixes and particles. In Geert Booij and Jaap Van Marle, editors, *Yearbook of Morphology 2003*, Springer, Dordrecht, pages 119–144.

Jarmila Panevová. 1974. On verbal frames in functional generative description: Part I. *The Prague Bulletin of Mathematical Linguistics* 22:3–40.

Jarmila Panevová. 1975. On verbal frames in functional generative description: Part II. *The Prague Bulletin of Mathematical Linguistics* 23:17–52.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner, and Martha Palmer. 2022. PropBank comes of age: Larger, smarter, and more diverse. In Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, Jose Camacho-Collados, and Alessandro Raganato, editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, Seattle, Washington, pages 278–288. https://doi.org/10.18653/v1/2022.starsem-1.24.

Gillian Ramchand. 2008. *Verb Meaning and the Lexicon: A First-Phase Syntax*. Cambridge University Press.

Eugenia Romanova. 2006. *Constructing Perfectivity in Russian*. Doctoral thesis, University of Tromsø. https://www.researchgate.net/publication/33417460_Constructing_Perfectivity_in_Russian.

Eleanora Slavíčková. 1975. *Retrográdní Morfematický Slovník Češtiny*. Academia, Praha.

Barbara Stiebels. 1996. *Lexikalische Argumente und Adjunkte: Zum Semantischen Beitrag von verbalen Präfixen und Partikeln*. De Gruyter. https://doi.org/10.1515/9783050072319.

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge University Press.

Peter Svenonius. 2004. Slavic prefixes inside and outside VP. *Nordlyd* 32(2):205–253. https://doi.org/10.7557/12.68.

František Uher. 1987. *Slovesné Předpony*. Univerzita J. E. Purkyně, Brno.

Zdeňka Urešová, Cristina Fernández Alcaina, Peter Bourgonje, Eva Fučíková, Jan Hajič, Eva Hajičová, Georg Rehm, Kateřina Rysová, and Karolina Zaczynska. 2023. SynSemClass 5.0. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5230.

Zdeňka Urešová, Alevtina Bémová, Eva Fučíková, Jan Hajič, Veronika Kolářová, Marie Mikulová, Petr Pajas, Jarmila Panevová, and Jan Štěpánek. 2024. PDT-Vallex: Czech valency lexicon linked to treebanks 4.5. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5814.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. 2021. DeriNet 2.1. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3765.

Alexander Williams. 2015. *Arguments in Syntax and Semantics*. Cambridge University Press.

Dieter Wunderlich. 1987. An investigation of lexical composition: The case of German *be-* verbs. *Linguistics* 25(2):283–331. https://doi.org/10.1515/ling.1987.25.2.283.

InTens – A dataset of Italian intensified derivatives. Description and application in a productivity study

Ivan Lacić
University of Bologna
ivan.lacic2@unibo.it

Abstract

The paper introduces *InTens*, a dataset of Italian intensified adjectival derivatives formed with six evaluative prefixes, namely *arci*, *extra*, *iper*, *stra*, *super*, and *ultra*. Initially, we delineate the process of data extraction and filtration. Subsequently, we address the polyfunctional characteristics of the evaluative prefixes, with semantic annotation of derivatives into two macro-categories: INTENSIFICATION and NON-INTENSIFICATION. After discussing the annotation results, the application of *InTens* is demonstrated through an investigation of the morphological productivity of the prefixes. The analysis underscores the variability in productivity contingent upon the semantic function of the prefixes, an aspect most often overlooked in productivity research.

1 Motivation

While large-scale resources providing a comprehensive view of the word-formation system are available for languages such as French (e.g., *Démonette-2* (Namer et al., 2023)), the situation for Italian, particularly concerning derivation, remains comparatively underdeveloped (for instance, Morph-It! (Zanchetta and Baroni, 2005) does not encompass derivational structure and derivationally related lemmas). This study seeks to make a contribution toward addressing this gap by developing a small-scale resource providing a coherent set of Italian adjectival derivatives expressing INTENSIFICATION, formed with six evaluative prefixes: *arci*, *extra*, *iper*, *stra*, *super*, and *ultra*.[1] This dataset is believed to be novel not only for Italian, but conceptually also for more extensively studied languages, as the majority of intensified derivatives do not form entrenched words with stable references in the lexicon and are therefore frequently omitted from existing resources.

The utility of *InTens* becomes particularly apparent when analyzing the six prefixes as formally suppletive morphemes that produce near-synonyms. In the context of Italian, the limited existing literature on evaluative prefixes has focused exclusively on cataloging these elements (Montermini, 2008), examining them as isolated phenomena (Napoli, 2012), or through a contrastive lens (Calpestrati, 2017), without addressing their competitive interactions. At its most basic, competition "refers to the fact that speakers routinely have to make a choice between alternative ways of realizing a certain concept" (Gardani et al., 2019, p. 4). A particular case of morphological competition is known as affix rivalry. Affix rivalry refers to the interaction between two or more affixes that, in at least some of their uses, can generate words of identical or similar semantic types (Guzmán Naranjo and Bonami, 2023; Huyghe and Varvara, 2023; Nagano et al., 2024). Evaluative morphology (EM) emerges as a particularly apt domain for the investigation of affix rivalry, as a single evaluative function is often expressed by multiple formal exponents,

[1] Our primary focus lies on evaluative prefixes that convey qualitative evaluation (+GOOD). Nevertheless, due to the potential difficulty in discerning between qualitative and quantitative evaluation (Iacobini, 2004; Napoli, 2012), we also considered prefixes capable of expressing both +GOOD and +BIG. To identify such prefixes, a thorough investigation of the existing literature was undertaken. Although space constraints preclude detailing the complete rationale behind selecting six specific prefixes, it should be noted that after the application of numerous exclusion criteria (e.g., the exclusion of predominantly quantitative prefixes such as *maxi-* and *mega-*, and the exclusion of prefixes primarily conveying spatial meaning such as *sopra-* and *sur-*) and restricting our selection to prefixes that productively combine with adjectives ţtepiacobini2004, we were left with the six prefixes under examination.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 73–85, Fribourg, Switzerland, 4-5 September 2025.

thus violating the blocking principle (Grandi, 2015). Many morphological systems, especially those of Standard Average European languages, abound with synonymous evaluative affixes that can be appended to the same base, resulting in derivatives that are semantically largely comparable (Körtvélyessy, 2015). While the examined six prefixes exhibit extensive polysemy, deriving words of various semantic functions – SPATIAL LOCATIVE (e.g., *extraurbano* 'extraurban'), NON-SPATIAL LOCATIVE (e.g., *stragiudiziale* 'extrajudicial'), HIERARCHY (e.g., *arcivescovile* 'archbishop's'), EXCESS (e.g., *strapagato* 'overpaid'), INTENSIFICATION (e.g., *ultramoderno* 'ultra-modern'), etc. – INTENSIFICATION seems to be the only domain in which all six prefixes compete. This is not surprising, as predominantly pragmatic communicative objectives that are met through INTENSIFICATION (differently from prototypical derivational morphology) promote the emergence of morphological competition as an acceptable pleonastic feature of the system (Dressler et al., 2019; Merlini Barbaresi and Dressler, 2020).

Given that rival affixes often exhibit differences in productivity (Gaeta and Ricca, 2015), productivity may serve as one of the distinguishing factors in rivalry. Therefore, it is crucial to quantify it using a valid dataset, namely one that comprises derivatives exclusively conveying INTENSIFICATION. Although it has long been recognized that the focus of productivity studies should be directed towards the productivity of semantic functions within a morphological pattern, rather than the pattern as a whole (Kastovsky, 1986), this aspect is frequently overlooked. Consequently, investigations into productivity that incorporate considerations of affix polysemy are more the exception than the rule (for a polysemy-aware approach within the field of EM, see, e.g., Efthymiou et al., 2015).

2 Dataset creation

To understand the properties of intensified derivatives, extensive data annotated for instances of INTENSIFICATION are essential. Section 2.1 outlines the methodology employed in the construction of this dataset, whereas the annotation procedure is presented in Section 2.2.

2.1 Data extraction

The presented dataset is based on the `itWaC` corpus (Baroni et al., 2009), comprising approximately 2 billion words. While not the largest or most recent web corpus, `itWaC` remains the largest copyright-free, freely downloadable corpus of Italian.[2]

As a first step in the dataset creation, adjectives formed with one of the six specific strings – *arci*, *extra*, *iper*, *super*, *stra*, and *ultra* – were extracted. For each of the six strings under exam, constructions in three orthographic variants, namely solid spelling (e.g., *ipercostoso* 'superexpensive'), hyphenation (e.g., *iper-costoso*), and open spelling (e.g., *iper costoso*) were taken into consideration. Instances in which an evaluative prefix follows the base, assuming an attributive use, were not considered. Although such a construction is possible with nominal bases (e.g., *il finalone iper* 'hyper grand finale'), it is typically not anticipated with adjectival bases, as in Italian adjectives generally do not modify other adjectives (e.g., **una casa costosa super* 'a super expensive house') (Montermini, 2008).

Due to the extraction regime not considering the nature of the extracted word – whether a real derivative or just a unit starting with one of the strings formally identical with affixes – cleaning was required. As an initial refinement step, we excluded adjectives that begin with one of the six target strings, where such a string serves as a component of the base rather than functioning as a prefix (e.g., *stra* in the adjective *stradale* 'street$_{\text{REL}}$'). To do this, we cross-referenced extracted occurrences against a list of Italian adjectives obtained from *lo Zingarelli 2024* dictionary (Zingarelli et al., 2023), then removed any matches. Moreover, many of the extracted adjectival "bases" were partial constituents of real words, such as the string *tegico*, a constituent of the adjective *strategico* 'strategic'. To address this, we utilized

[2]While INTENSIFICATION is often associated with spoken language due to its informal characteristics (Ito and Tagliamonte, 2003), a spoken corpus proved unsuitable for our analysis given the limited size of available corpora and the resulting low frequency of relevant derivatives. For instance, a preliminary analysis of the `KIParla` corpus (Mauri et al., 2019) identified only about 60 intensifying prefix occurrences, rendering quantitative analysis unfeasible. That said, a critical distinction between the diamesic and diaphasic dimensions of language use is necessary. We believe that the productivity of our prefixes depends more on the language formality than on the medium itself. The `itWaC` corpus, covering a wide range of text types, is largely independent of this diaphasic variable, making it an appropriate resource.

Wiktionary data and cross-referenced the extracted forms with a list of Italian adjectives from the machine-readable dictionary `kaikki` (Ylonen, 2022). Adjectives absent in the `kaikki` dictionary were excluded, resulting in a more refined list of adjectives. Furthermore, to minimize noise stemming from spurious forms, bases with a frequency less than 5 in the corpus were excluded from the analysis.[3]

Another challenge encountered was the presence of derivatives in which prefixation is anterior to denominal suffixation (e.g., *arcidiocesano* 'archdiocesan' < *arcidiocesi* 'archdiocese'). The standard choice in these situations has been to select only tokens in which the affix is appended as the final element, thereby excluding inner derivations (Fradin et al., 2008; Baayen, 2009). Consequently, it was decided to omit these instances from the analysis.[4] This decision is based on two considerations: first, it is posited that the presence of an already prefixed base in the formation of a derivative impacts the new derivative by analogy; second, this study seeks to examine the productivity of prefixes with exclusively adjectival bases.

Lastly, since a single morphological family may encompass closely related nouns and adjectives, it is often ambiguous which lexeme should be considered the derivative's base (e.g., *supertechnologico* 'supertechnological' could be derived either via suffixation from the noun *supertechnologia* 'supertechnology' or via prefixation from the adjective *technologico* 'technological'). As noted by Bonami and Thuilier (2019), there is frequently no operational method for determining the base of the derivative when the derivational family presents multiple potential "solutions". Therefore, for practical purposes, derivatives that could theoretically be derived through prefixation (as one among the possible mechanisms of word-formation) were not excluded from the analysis. After these refinement steps, and once the data set had been reduced to a manageable size, a final manual verification of derivatives in context, as attested in the corpus, was conducted.

Following the described refinement processes, a final list of prefixed derivatives was obtained. The complete dataset consists of 4,599 derivative types formed with 2,683 adjectival base types. The distribution of the derivatives across prefixes is illustrated in Table 1.

	Tokens	Types	Hapax legomena
arci	1,318	117	81
extra	75,109	722	235
iper	9,695	988	430
stra	20,924	342	163
super	12,888	1,327	528
ultra	19,279	1,103	492

Table 1: Distribution of the derivatives across prefixes.

A notable disparity in the frequency distribution of both tokens and types among prefixes can be observed. Notably, the prefix *extra* is the most frequent prefix by token count at 75,109 occurrences, followed by *stra* with 20,924 occurrences and *ultra* with 19,279 occurrences. Conversely, the prefixes *super* and *iper* register moderate frequencies of 12,888 and 9,695 tokens respectively, whereas the prefix *arci* demonstrates the lowest frequency with 1,318 tokens. Regarding unique types, the prefix *super* possesses the highest number of distinct types, totaling 1,327, while the prefix *arci* contains the fewest with only 117 unique types.

Whilst this dataset can serve as a basis for the examination of derivatives across all semantic values, the specific focus of this study, that is the analysis of competition within the INTENSIFICATION domain, necessitates a focused dataset comprising derivatives that specifically express INTENSIFICATION.

[3]Note that this procedure did not exclude any hapax legomena that might appear with the prefixes as long as the general frequency of the adjectival base in the corpus is ≥ 5.

[4]While the inclusion of inner cycle derivatives may hold psycholinguistic relevance (Plag, 1999), Gaeta and Ricca (2006) observe that when productivity calculations are performed on a fixed token count, as applied in the present study, nearly identical results are obtained irrespective of the inclusion of inner derivations.

2.2 Annotation for semantic function

Given the high specificity of the required annotations, the manual annotation of the derivatives was selected as the preferred methodological approach. While a detailed semantic annotation of the derivatives could potentially reveal interesting insights regarding the competition between prefixes, it was presumed to be extremely labor intensive and, most importantly, not central to this study, which primarily focuses on the phenomenon of rivalry within the context of INTENSIFICATION. For this reason, it was decided to categorize the derivatives into two distinct semantic macro-categories: (i) INTENSIFICATION and (ii) NON-INTENSIFICATION. Annotators had three labels at their disposal: (i) INTENSIFICATION, (ii) NON-INTENSIFICATION, and (iii) TERMINOLOGY. The label TERMINOLOGY refers to the derivatives that are considered part of the scientific lexicon, for which the annotators were uncertain of a semantic value to assign. Furthermore, in instances of ambiguous interpretations (e.g., *iperattivo* 'hyperactive' can be understood both as conveying INTENSIFICATION and EXCESS), annotators were advised to classify the derivative as an instance of INTENSIFICATION, provided it permits such a potential interpretation. This decision stemmed from the understanding that distinguishing between these two interpretations with precision is both challenging and potentially counterproductive given the relatively low frequency of the derivatives being analyzed.

Considering the size of the dataset ($139,213$ tokens), it was split into two parts. The annotations were carried out by the author in conjunction with four co-annotators. The co-annotators are native speakers of Italian, all of whom are either in the process of obtaining or have already obtained a doctoral degree in linguistics. The relatively homogeneous background of the annotators, coupled with their prior linguistic training, is considered advantageous (Artstein, 2017). Each part of the dataset was annotated by three annotators: the author, along with annotators 1 and 2 for the first part, and the author, along with annotators 3 and 4 for the second part. While explicit annotation guidelines in the form of a distinct document were not formulated, comprehensive procedural instructions were provided to each annotator, with consistent communication maintained throughout the annotation process. All issues and inquiries were addressed and resolved collaboratively.

Due to the large size of the initial dataset, annotating individual tokens was not feasible. As a result, the annotations were performed on a type-based level, a methodology we recognize introduces a certain level of approximation.[5] To assess the trustworthiness of the annotations, a qualitative analysis of the annotations was initially conducted, followed by the computation of inter-annotator agreement metrics. Two distinct measures of agreement were initially selected: raw agreement (RA) (Goodman and Kruskal, 1959) and Fleiss' κ (κ_F) (Fleiss, 1981). However, observing the results obtained, a great discrepancy between RA and κ_F values was seen, and it was noted that κ_F encountered a challenge associated with the degree of homogeneity in the annotations, known as the Kappa paradox. In highly homogeneous annotations, where the majority of annotation points fall into a single category (a very common occurrence in our dataset), the marginal probabilities are expected to be very high. Consequently, the expected agreement by chance alone (as measured by κ_F) is also likely to be considerably elevated, and this will, in turn, make κ_F rather low.

To gain a more accurate understanding of the level of agreement among the annotators, an additional metric was introduced, namely Gwet's AC1 (Gwet, 2008). Gwet's AC1 addresses certain limitations

[5] We recognize that type-based annotations should not preclude polysemy, given that certain derivatives may present both intensive and non-intensive interpretations. To (at least partially) address the assumption of monosemy adopted in this study, a random sample of 20 derivative types with a frequency ≥ 10 per prefix was extracted and an analysis on a randomly selected sample of 300 tokens was conducted. Although no instances of genuine polysemous interpretations, encompassing both intensified and non-intensified readings, were identified, two examples classified as TERMINOLOGY allowing also for a non-terminological intensified interpretation were discovered. Specifically, the derivative *iperelastico* appeared both as an instance of INTENSIFICATION meaning 'very elastic' and in a technical context as a term in continuum mechanics. Additionally, the derivative *ultraperiferico* 'ultra-peripheral', was observed in both the terminological reading, referring to a territory of the EU located outside the European continent, and a genuinely intensified reading describing a property for sale in a remote location. As previously noted, both derivatives were adjudicated as TERMINOLOGY by the annotators and subsequently excluded from the analysis, thus not considering the intensifying reading of the two lexemes. It is evident, however, that for more robust conclusions, a significantly larger token sample should be assessed, and we intend to revisit this matter in future research. In the meantime, considering the almost perfect monosemy detected in our $1,800$ token sample, we believe that it is safe to proceed with the analysis.

observed for κ_F, particularly in contexts of high agreement, where κ_F value is typically diminished disproportionately.[6]

	Raw agreement (%)	Fleiss' κ	Gwet's AC1 (95% CI)
arci	98.4	−0.01	0.99 (0.97–1.00)
extra	85.3	0.60	0.86 (0.83–0.89)
iper	87.5	0.32	0.86 (0.83–0.88)
stra	99.4	0.42	0.98 (0.97–1.00)
super	91.3	0.33	0.90 (0.88–0.92)
ultra	89.8	0.60	0.90 (0.88–0.93)

Table 2: Inter-annotator agreement – first part of the dataset.

	Raw agreement (%)	Fleiss' κ	Gwet's AC1 (95% CI)
arci	98.1	0.49	0.99 (0.94–1.00)
extra	80.7	0.60	0.86 (0.82–0.89)
iper	83.9	0.32	0.88 (0.86–0.91)
stra	94.9	0.54	0.96 (0.93–0.99)
super	86.4	0.35	0.90 (0.88–0.92)
ultra	82.5	0.52	0.87 (0.85–0.90)

Table 3: Inter-annotator agreement – second part of the dataset.

Tables 2 and 3 present the values of the three inter-annotator agreement measures applied to the two parts of the dataset. The highest RA and AC1 values were observed for the prefixes *arci* (RA: 98.4/98.1; AC1: 0.99) and *stra* (RA: 99.4/94.9; AC1: 0.98/0.96), while the remaining prefixes also showed consistently high inter-annotator agreement, with all values exceeding the commonly accepted 0.80 threshold for reliable annotations[7] (Brezina, 2018, p. 89). The particularly high levels of agreement for *arci* and *stra* suggest that derivatives formed with these prefixes seem comparatively less polyfunctional, which likely facilitated annotation and contributed to stronger agreement.

The negative κ_F value for *arci* in the first part of the dataset, along with relatively low κ_F values for *iper* and *super*, stand in stark contrast to their high RA and AC1 values. This discrepancy underscores the significant impact that the selection of agreement metrics can have on the interpretation of inter-annotator reliability, emphasizing the necessity of employing multiple agreement measures for acquiring a comprehensive understanding of reliability. The robustness of the findings is substantiated by a correlation analysis between the three agreement measures. This revealed a perfect positive correlation ($\rho = 1$) between RA and AC1 in the second part of the dataset, and a very strong correlation ($\rho = 0.83$) in the first. In contrast, Fleiss' κ displayed a negative correlation with both RA and AC1 across both datasets. This is consistent with the earlier observation, where it was indicated that κ_F tends to underestimate agreement due to the homogeneous nature of the annotations. In conclusion, the consistently high RA ($> 80\%$) and Gwet's AC1 values (≥ 0.86) provide strong evidence of substantial agreement among annotators. It is therefore reasonable to conclude that the annotations are reliable.

The subsequent phase in the development of the dataset involved the identification of intensified derivatives. This process adhered to specific criteria: if two of the three annotators assigned the same tag to a derivative, it was classified accordingly. Conversely, if each annotator assigned a distinct tag to the same derivative, it was categorized as UNCLEAR. Table 4 presents the annotation results for both parts of the dataset.

Out of $4,599$ derivative types, $3,686$ (80.1%) were classified as instances of INTENSIFICATION (e.g., *arcicostoso* 'super-expensive', *strafigo* 'super-cool'), while 809 (17.6%) derivatives were labeled as NON-

[6] While Silveira and Siqueira (2023) champion Gwet's AC1 as the preferred measure for inter-rater agreement in contingency tables, Vach and Gerke (2023) argue that it shouldn't be considered a comprehensive κ-based metric. To sidestep potential methodological debates arising from favoring just one approach, we do not propose AC1 as a replacement for κ-based measures here. Instead, it acts as a complementary metric that offers increased stability in the face of category prevalence effects.

[7] In this study, RA was determined by a strict criterion: only instances where all three annotators assigned the identical label counted as agreement. While this conservative approach captured full unanimity, it inherently excluded cases of partial agreement (e.g., two out of three annotators agreeing). In contrast, AC1 estimates agreement by calculating pairwise consistency across all annotator combinations, then adjusting for chance. Because AC1 does not demand full unanimity, it is sensitive to high levels of partial agreement. Consequently, the observed agreement contributing to AC1 can sometimes surpass the strictly calculated RA, especially when the consensus is strong but not absolute. Furthermore, AC1's chance agreement correction is less affected by category imbalance than traditional κ-based methods. In datasets dominated by one label, this yields more stable, and often higher, adjusted agreement values. Therefore, instances where AC1 slightly exceeds RA should not be interpreted as inconsistencies. Instead, they demonstrate AC1's ability to offer a more inclusive and robust estimate of annotator consensus, particularly in scenarios characterized by high RA and skewed category distributions.

	Types	INTENSIFICATION	NON-INTENSIFICATION	TERMINOLOGY	UNCLEAR
arci	117	116	1	0	0
extra	722	122	592	1	7
iper	988	904	37	31	16
stra	342	336	6	0	0
super	1,327	1,244	51	10	22
ultra	1,103	964	122	3	14
SUM	4,599	3,686	809	45	59

Table 4: Distribution of derivative types across annotation categories.

INTENSIFICATION (e.g., *extraregionale* 'extra-regional', *ultraindividuale* 'extra-individual). Additionally, 45 instances were labeled as TERMINOLOGY (e.g., *ipercontratto* 'over-contracted', *ultralineare* 'ultra-linear'), and 59 instances as UNCLEAR. Given that the TERMINOLOGY and UNCLEAR classes collectively represented only 2.3% of the dataset, and in order to avoid any reliance on subjective authorial judgment concerning their status, they were excluded from further analysis. Finally, a consolidated table was created, containing 3,686 intensified derivative tokens. Table 5 presents the distribution of intensified derivatives per prefix, along with the percentage of intensified tokens and types within the total sample, as illustrated in Table 1.

	Intensified tokens	% of total tokens	Intensified types	% of total types
arci	1,297	98.4	116	99.2
extra	837	1.1	122	16.9
iper	7,930	81.8	904	91.5
stra	18,581	88.8	336	98.3
super	11,167	86.7	1,244	93.8
ultra	8,257	42.8	964	87.4

Table 5: Token and type counts of intensified derivatives.

For certain prefixes, token and type counts change minimally. Specifically, *arci*'s tokens decrease by just 1.6% and its types by 0.8%. Similarly, *stra*, *super*, and *iper* undergo token reductions ranging from 11.2% to 18.2%, and type reductions spanning 1.7% to 8.5%, highlighting their primary role as intensifying prefixes. In contrast, *ultra* and *extra* demonstrate more pronounced reductions when confined to intensifying forms, as evidenced by a 57.2% decrease in *ultra*'s token count, and a sharp 98.9% reduction in *extra*'s, thereby underscoring *extra*'s predominant non-intensifying use. While *extra*'s token and type count reductions are largely proportional, *ultra* displays a unique pattern characterized by a significantly smaller decrease in the type count, measured at 12.6%. This indicates that limiting *ultra* to the INTENSIFICATION domain removes a limited quantity of high-frequency (mainly terminological) types, while largely preserving the diverse, less frequent intensified types, resulting in a comparatively minor change to its overall type count.

The results highlight the importance of annotating for semantic function, as it provides a more nuanced understanding of prefix usage, particularly within the framework of prefix polyfunctionality and rivalry. Without proper semantic annotation, these distinctions would be overlooked, potentially skewing the interpretation of other analyses. Moreover, existing descriptions regarding the usage of intensifying prefixes are challenged. For instance, Calpestrati (2017) suggests that *extra* in Italian primarily functions as an intensifier (regardless of its low degree of combinability with adjectives), but our findings show

that *extra* plays a marginal role in INTENSIFICATION, as almost all of its occurrences convey a non-evaluative meaning. The observations related to the restricted use of *extra* in intensifying contexts are likewise apparent in French, where it is posited that, analogously to the situation in Italian, the broader proliferation of *extra* is constrained by the existence of other synonymous prefixes, namely *super* and *hyper* (Izert, 2012).

3 Morphological productivity as a function of semantics: a case study

In addition to quantifying the morphological productivity of intensifying prefixes, *InTens* can serve to illustrate the enhanced insights gained by analyzing productivity for a specific semantic function via an annotated dataset. To that end, we find it useful to contrast the prefixes' productivity within the intensifying domain with their broader productivity in generating derivatives of "all senses".

The popularity of studies on morphological productivity, as highlighted by Dal and Namer (2016), comes from the idea that morphology essentially *equals* productivity. Indeed, should morphological theory be concerned solely with processes of word formation that are productive, then identifying which processes are productive and which are not becomes a central concern in morphological research (Baayen and Lieber, 1991). In the quantitative approach, morphological productivity is conceptualized as fluctuating in the frequency of new coinages. In essence, a process is considered more or less productive based on the number of lexemes it generates (Fernández-Domínguez, 2013).

To quantify the productivity of prefixes, a comprehensive study based on type-token ratio (*TTR*), potential productivity ($\mathscr{P}$) (Baayen, 2009), entropy (H) (Shannon, 1948), and population vocabulary size (S) from the finite Zipf-Mandelbrot model for LNRE (Evert and Baroni, 2007; Baroni and Evert, 2014) was conducted. Each measure captures distinct aspects of word formation and usage. The *TTR* measures balance in usage, $\mathscr{P}$ estimates the likelihood of encountering a new type, H can be seen as a measure of unpredictability in the type-frequency distribution, while the population vocabulary size S extrapolates from the observed sample to a sample of arbitrary size, estimating the total number of types that would be observed if the entire population were sampled (Zeldes, 2012). Since "the main upshot of these debates [on morphological productivity] has been the insight that rather than relying on one measure of productivity exclusively, multiple variables should be taken into account and compared" (Hartmann, 2018, p. 84), we believe that using four measures of productivity is a well-grounded decision.

To enable an unbiased comparison of productivity metrics across different prefixes, while avoiding confounding effects introduced by varying token sizes, it is imperative to establish a uniform sample size. Therefore, for both the broader "all senses" dataset, encompassing derivatives of all semantic functions, and the intensified dataset, a fixed number of tokens were randomly chosen for each prefix. In the case of the "all senses" dataset, a sample size of $1,000$ tokens was used, ensuring sufficient overlap for the smallest group, namely the prefix *arci*. For the intensified dataset, the sample size was set at 635 tokens, representing approximately 70% of the dataset for *extra*, the least frequent prefix. We recognize that varying the sample size might influence productivity correlations, but due to practical constraints, testing different sizes was impractical. Larger samples could not be analyzed since the least frequent prefix has only $1,318$ tokens, and much smaller samples would risk introducing too much variability. With all six prefix groups standardized to the same token size, we computed the four productivity measures. To ensure the robustness of our results, this process was iterated 100 times, drawing new random samples (taken `with replacement`) in each iteration.

Figure 1 represents the median values of the four productivity measures for both "all senses" and intensified datasets.

Visual inspection of the plot shows a general trend of increase in the type-token ratio (*TTR*) and potential productivity ($\mathscr{P}$) when prefixes are used to create intensified derivatives. The only exception to this pattern is *extra*, which shows a slight decrease in $\mathscr{P}$. The most significant positive variation in the two measures is apparent with *super*. On the other hand, *stra* and *arci* exhibit minimal changes across all measures. As previously noted, this consistency suggests that the two are predominantly used within the intensifying domain and, therefore, their limitation to intensifying contexts does not notably affect their productivity, as this is their principal area of application. Although variations in *TTR* and $\mathscr{P}$

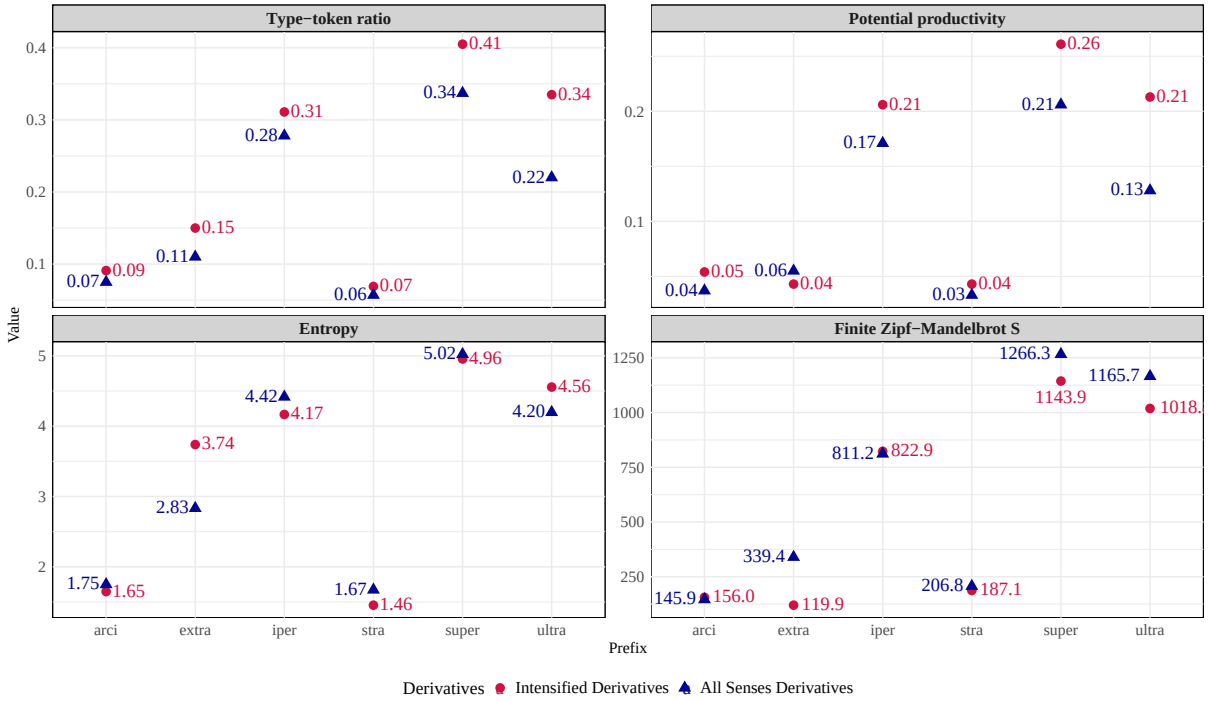


Figure 1: Median productivity values for six prefixes forming derivatives of all semantic functions (represented as triangles) and solely intensified derivatives (represented as dots).

are consistent across all prefixes except for *extra*, larger discrepancies can be seen for entropy (H) and population vocabulary size (S). The observed non-uniformity in H changes indicates alterations in the predictability of the derivative distributions associated with each prefix. For example, the noteworthy increase in H observed for *extra* can be ascribed to the prevalence of frequent, non-intensified derivatives within its general usage. When these highly predictable, frequent forms are excluded to concentrate exclusively on intensified derivatives, the remaining, less frequent items result in a distribution where the occurrence of any particular form becomes less predictable, thereby increasing H. In contrast, the most significant reduction in H is observed for *iper*. This prefix is inclined to produce numerous less frequent, often scientific terms. When these infrequent derivatives are omitted by focusing solely on intensified forms, the ensuing distribution is governed by these more frequent forms, resulting in increased predictability and a reduction in H.

In summary, all measures demonstrate that *super*, *ultra*, and *iper* are the most productive prefixes in the formation of both derivatives of "all senses" and purely intensified derivatives. Notably, these findings align with those of Cartier and Huyghe (2021) concerning French, where *hyper*, *ultra*, and *super* were identified as the most productive high-degree prefixes. Additionally, the same findings can be associated with the perception of *super* as the least intense intensifying prefix in Italian (Calpestrati, 2017). Indeed, should the hypothesis that intensifiers lose their efficacy due to overuse (Mutz, 2015) be accurate, then *super*'s high productivity and low perceived intensity have a strong correlation. On the other hand, *arci* and *stra* exhibit notably low productivity, likely because a single type dominates a large portion of their samples (65% for *arci* with *arcinoto* 'very well-known', and 58% for *stra* with *stragrande* 'vast'). This finding supports the common assertion that low-productivity categories often contain a multitude of high-frequency forms (Plag, 2003).

To assess whether the six prefixes' productivity differs when forming intensified derivatives versus those of all semantic functions, while acknowledging that the intensified dataset constitutes a subset of the comprehensive "all senses" superset, we opted for a permutation test. For each prefix, we compared the observed productivity of its intensified derivatives (calculated from 635 tokens, as previously introduced) against a null distribution generated by repeatedly drawing 1,000 random samples (also 635 tokens,

`with replacement`) from the "all senses" for that same prefix. For each random sample, we calculated the aforecited productivity measures and then derived an empirical two-tailed p-value, indicating the probability of observing a median productivity as extreme as our actual intensified median purely by chance. The permutation results, completed with Holm-Bonferroni adjusted p-values and Cohen's d effect sizes, revealed statistically significant differences in productivity profiles for certain prefixes.

Prefixes *arci*, *iper*, and *super* showed no statistically significant differences across any measure (all $p_{\text{adj}} = 1$). This indicates that for these prefixes, the observed productivity of their intensified derivatives is not statistically distinct from what would be expected by random chance from their overall usage, implying that intensification does not uniquely impact their productivity in a robust way. In contrast, *extra* and *ultra* exhibited multiple highly significant differences. For *extra*, H ($p_{\text{adj}} = 0$) was significant, with a very large effect size ($d = 11.60$). On the other hand, *ultra*'s intensified derivatives demonstrated highly significant differences in TTR ($p_{\text{adj}} = 0$, $d = 5.16$), $\mathscr{P}$ ($p_{\text{adj}} = 0$, $d = 3.64$), and H ($p_{\text{adj}} = 0$, $d = 6.15$). The large magnitudes of these effect sizes underscore the practical significance of these findings, indicating that the intensified uses of *extra* and *ultra* are significantly more productive in these specific aspects than would be expected from their overall derivative pools. Finally, *stra* showed a large negative effect for H ($d = -2.94$), indicating a lower diversity in its intensified forms. However, this difference is not significant after correction ($p_{\text{adj}} = 0.08$).

Overall, our findings suggest that productivity differences of the prefixes involved in the formation of intensified derivatives versus those of all semantic functions do deviate in a statistically robust way, though not uniformly across all prefixes. Methodologically, these results corroborate the need for extensive semantic annotation in affix rivalry studies, as it allows for a clearer distinction between prefixes' various uses across semantic functions, providing a more precise approach to this complex, gradient phenomenon.

4 Concluding remarks

In this study, we outlined a methodology for building a small-scale dataset of intensified Italian derivatives. This work highlighted the critical role of semantic annotation for polyfunctional derivatives, particularly in the context of affix rivalry. Our subsequent illustrative analysis then demonstrated how the productivity of certain prefixes notably changed when the domain of interest was narrowed from "all senses" to the specific semantic function of INTENSIFICATION.

Beyond the presented (mostly methodological) insights and building upon findings of Section 3, our findings compel a reflection on the intricate relationship between morphological competition and productivity. As noted by Fernández-Domínguez (2017), literature presents divergent views on whether a morphological process gains productivity as a result of previously prevailing in competitive contexts, or if it prevails in competition owing to a prior enhancement in its productivity. For instance, Scherer (2015) sees the competition as a language-internal factor whose changes cause variations in the productivity of processes, while van Marle (1988) describes competition as a gradual process where a decrease in productivity of one process leads to the rise of a competing process that will eventually become so productive that it will replace the original one. While our brief productivity analysis provided a synchronic snapshot of this relationship, a comprehensive understanding of the underlying mechanisms necessitates a fine-grained evaluation of diachronic productivity. Nonetheless, what seems clear from the analysis is that the six prefixes persist in coexistence, with no single process definitively supplanting another. The three most productive prefixes – *super*, *ultra*, and *iper* – exhibit comparably elevated levels of productivity. In contrast, the comparatively lower productivity of *arci* and *stra*, combined with their ability to generate highly lexicalized derivatives, suggests that these prefixes have started developing "niche productivity" (in the sense of Lindsay and Aronoff (2013)) and they might persist precisely because of these established forms, as affixes depend on a certain inventory of frequent types to remain cognitively present. This phenomenon could influence morphological competition, given that lexicalized lexemes often act as blocking agents. Notably, when examining the bases *noto* 'known' and *grande* 'big', two of the most idiosyncratic bases that *arci* and *stra* combine with, it is observed that although the frequency of *arcinoto* 'very well-known' and *stragrande* 'vast' is exceptionally high, the four other examined prefixes seldom combine with these two bases.

Furthermore, it should be noted that variations in productivity may arise from factors other than competition, such as sociolinguistic and pragmatic variables (Körtvélyessy, 2010; Merlini Barbaresi and Dressler, 2020), as well as language fashion, and, on the other hand, the emergence of new processes that enter in competition with an existing one can be attributed to influences beyond productivity, such as language change and language contact (Nagano et al., 2024). We believe that is the case for our most productive prefix — *super*. According to Migliorini (1963), *super* started its diffusion after the appearance of the leader word *super-uomo* 'superman' (after Nietzsche's *Übermensch*) and has spread quickly through contact with mass media. However, existing theories regarding the nexus between competition and productivity are based on traditional morphological competition, which may not be applicable in our context, since within EM rivalry is controlled differently compared to typical derivational morphology (Grandi, 2023). Indeed, it is necessary to examine the pragmatics of EM as a catalyst for competition and to investigate the interplay between competition and productivity within that framework. Space constraints, however, limit this inquiry.

Ultimately, when interpreting these productivity results from the perspective of actual language use, it is crucial to acknowledge their contingency on the corpus employed. The itWaC corpus, developed between 2005 and 2007, is somewhat outdated, and possible shifts in the productivity of the prefixes over the past two decades cannot be ruled out. Diachronic studies typically describe a continuous trend of increasing use of multiple evaluative prefixes throughout the 19th and 20th centuries (Cartier and Huyghe, 2021), with certain prefixes expanding their use diachronically, which may affect their productivity. Such changes may also manifest in micro-diachrony, namely within the temporal range from the creation of itWaC to the present day, given that intensifiers constitute a rather volatile area of language that is particularly susceptible to linguistic innovations (Tagliamonte, 2016).

All things considered, we believe that the creation of this detailed dataset represents a valuable resource for advancing the analysis of prefixal intensifiers in Italian. It provides a foundation for a wide range of qualitative and quantitative investigations, all of which are essential to gain a more nuanced understanding of the complex and multifactorial phenomenon of affix rivalry.

Acknowledgments

I am grateful to Olivier Bonami and three anonymous reviewers for their insightful comments on earlier drafts of this manuscript.

References

Ron Artstein. 2017. Inter-annotator agreement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, Springer, pages 297–313. https://doi.org/10.1007/978-94-024-0881-2_11.

Harald Baayen. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling and Merja Kyto, editors, *Corpus Linguistics: An International Handbook*, De Gruyter, pages 899–919. https://doi.org/10.1515/9783110213881.2.899.

Harald Baayen and Rochelle Lieber. 1991. Productivity and English word-formations: A corpus-based study. *Linguistics* 29(5):801–844. https://doi.org/10.1515/ling.1991.29.5.801.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43:209–226. https://doi.org/10.1007/s10579-009-9081-4.

Marco Baroni and Stefan Evert. 2014. The zipfR package for lexical statistics: A tutorial introduction. https://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf.

Olivier Bonami and Juliette Thuilier. 2019. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure* 12(1):4–41. https://doi.org/10.3366/word.2018.0130.

Vaclav Brezina. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899.

Nicolò Calpestrati. 2017. Intensification strategies in German and Italian written language. In Maria Napoli and Miriam Ravetto, editors, *Exploring Intensification: Synchronic, Diachronic & Cross-Linguistic Perspectives*, John Benjamins, pages 305–326. https://doi.org/10.1075/slcs.189.16cal.

Emmanuel Cartier and Richard Huyghe. 2021. La concurrence affixale en diachronie: Le cas des préfixes de haut degré en français. *Linx* 82. https://doi.org/10.4000/linx.8078.

Georgette Dal and Fiammetta Namer. 2016. Productivity. In Andrew Hippisley and Gregory Stump, editors, *The Cambridge Handbook of Morphology*, Cambridge University Press, pages 70–90. https://doi.org/10.1017/9781139814720.004.

Wolfgang U. Dressler, Lavinia Merlini Barbaresi, Sonja Schwaiger, Jutta Ransmayr, Sabine Sommer-Lolei, and Katharina Korecky-Kröll. 2019. Rivalry and lack of blocking among Italian and German diminutives in adult and child language. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*, Springer, pages 123–143. https://doi.org/10.1007/978-3-030-02550-2_5.

Angeliki Efthymiou, Georgia Fragaki, and Angelos Markos. 2015. Exploring the meaning and productivity of a polysemous prefix: The case of the Modern Greek prepositional prefix *para*. *Acta Linguistica Hungarica* 62(4):447–476. https://doi.org/10.1556/064.2015.62.4.4.

Stefan Evert and Marco Baroni. 2007. zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the ACL, Demo and Poster Sessions*. pages 29–32. https://aclanthology.org/P07-2008/.

Jesús Fernández-Domínguez. 2013. Morphological productivity measurement: Exploring qualitative versus quantitative approaches. *English Studies* 94(4):422–447. https://doi.org/10.1080/0013838X.2013.780823.

Jesús Fernández-Domínguez. 2017. Methodological and procedural issues in the quantification of morphological competition. In Juan Santana-Lario and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, Peter Lang, pages 67–117.

Joseph L. Fleiss. 1981. *Statistical Methods for Rates and Proportions.* Second edition. John Wiley & Sons. https://doi.org/10.1002/0471445428.

Bernard Fradin, Georgette Dal, Natalia Grabar, Stephanie Lignon, Fiammetta Namer, Delphine Tribout, and Pierre Zweigenbaum. 2008. Remarques sur l'usage des corpus en morphologie. *Langages* 171(3):34–59. https://doi.org/10.3917/lang.171.0034.

Livio Gaeta and Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 41:57–89. https://doi.org/10.1515/LING.2006.003.

Livio Gaeta and Davide Ricca. 2015. Productivity. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe–Vol. 2*, De Gruyter, pages 842–858. https://doi.org/10.1515/9783110246278-003.

Francesco Gardani, Franz Rainer, and Hans Christian Luschützky. 2019. Competition in morphology: A historical outline. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*, Springer, pages 3–36. https://doi.org/10.1007/978-3-030-02550-2_1.

Leo A. Goodman and William H. Kruskal. 1959. Measures of association for cross classifications. II: Further discussion and references. *Journal of the American Statistical Association* 54(285):123–163. https://doi.org/10.1080/01621459.1959.10501503.

Nicola Grandi. 2015. The place of evaluation within morphology. In Nicola Grandi and Lívia Körtvélyessy, editors, *Edinburgh Handbook of Evaluative Morphology*, Edinburgh University Press, pages 74–90. https://doi.org/10.1515/9780748681754-010.

Nicola Grandi. 2023. Evaluative morphology in the Romance languages. In Michele Loporcaro, Francesco Gardani, Patricia Cabredo Hofherr, Jeroen Claes, Andreas Dufter, Martin Maiden, and Franz Rainer, editors, *Oxford Research Encyclopedia of Linguistics*, Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.684.

Matías Guzmán Naranjo and Olivier Bonami. 2023. A distributional assessment of rivalry in word formation. *Word Structure* 16(1):87–114. https://doi.org/10.3366/word.2023.0222.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* 61(1):29–48. https://doi.org/10.1348/000711006X126600.

Stefan Hartmann. 2018. Derivational morphology in flux: A case study of word-formation change in German. *Cognitive Linguistics* 29(1):77–119. https://doi.org/10.1515/cog-2016-0146.

Richard Huyghe and Rossella Varvara. 2023. Affix rivalry: Theoretical and methodological challenges. *Word Structure* 16(1):1–23. https://doi.org/10.3366/word.2023.0218.

Claudio Iacobini. 2004. Prefissazione. In Maria Gorssman and Franz Rainer, editors, *La Formazione delle Parole in Italiano*, De Gruyter, pages 97–163.

Rika Ito and Sali Tagliamonte. 2003. *Well* weird, *right* dodgy, *very* strange, *really* cool: Layering and recycling in English intensifiers. *Language in Society* 32(2):257–279. https://doi.org/10.1017/S0047404503322055.

Małgorzata Izert. 2012. Préfixes *extra-* et *supra-* comme intensificateurs de la propriété en français contemporain. *Kwartalnik Neofilologiczny* 59:437–446. https://journals.pan.pl/publication/102550/edition/88565/kwartalnik-neofilologiczny-2012-no-4.

Dieter Kastovsky. 1986. The problem of productivity in word formation. *Linguistics* 24(3):585–600. https://doi.org/10.1515/ling.1986.24.3.585.

Lívia Körtvélyessy. 2010. *Vplyv Sociolingvistických Faktorov na Produktivitu v Slovotvorbe [On the Influence of Sociolinguistic Factors upon Productivity in Word-Formation]*. SLOVACONTACT.

Lívia Körtvélyessy. 2015. Evaluative morphology and language universals. In Nicola Grandi and Lívia Körtvélyessy, editors, *Edinburgh Handbook of Evaluative Morphology*, Edinburgh University Press, pages 61–73. https://doi.org/10.1515/9780748681754-009.

Mark Lindsay and Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. In Nabil Hathout, Fabio Montermini, and Jesse Tseng, editors, *Morphology in Toulouse: Selected prodeedings of Décembrettes 7*. pages 133–153.

Caterina Mauri, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, and Francesco Suriano. 2019. KIParla corpus: A new resource for spoken Italian. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *CLiC-it 2019 – Proceedings of the Sixth Italian Conference on Computational Linguistics*. CEUR.

Lavinia Merlini Barbaresi and Wolfgang U. Dressler. 2020. Pragmatic explanations in morphology. In Vito Pirrelli, Ingo Plag, and Wolfgang U. Dressler, editors, *Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon*, De Gruyter, volume 405–452, pages 405–451. https://doi.org/10.1515/9783110440577-011.

Bruno Migliorini. 1963. Fortuna del prefisso *super-*. In Bruno Migliorini, editor, *Saggi sulla Lingua del Novecento*, Sansoni, pages 61–69.

Fabio Montermini. 2008. *Il Lato Sinistro della Morfologia: La Prefissazione in Italiano e nelle Lingue del Mondo*. FrancoAngeli.

Katrin Mutz. 2015. Evaluative morphology in a diachronic perspective. In Nicola Grandi and Lívia Körtvélyessy, editors, *Edinburgh Handbook of Evaluative Morphology*, Edinburgh University Press, pages 142–154. https://doi.org/10.1515/9780748681754-015.

Akiko Nagano, Alexandra Bagasheva, and Vincent Renner. 2024. Towards a competition-based word-formation theory. In Alexandra Bagasheva, Akiko Nagano, and Vincent Renner, editors, *Competition in Word-Formation*, John Benjamins, pages 1–31. https://doi.org/10.1075/la.284.01nag.

Fiammetta Namer, Nabil Hathout, Dany Amiot, Lucie Barque, Olivier Bonami, Gilles Boyé, Basilio Calderone, Julie Cattini, Georgette Dal, et al. 2023. Démonette-2: A derivational database for French with broad lexical coverage and fine-grained morphological descriptions. *Lexique* 33:6–40. https://doi.org/ff10.54563/lexique.1242.

Maria Napoli. 2012. Uno stra-prefisso: L'evoluzione di *stra-* nella storia dell'italiano. *Rivista Italiana di Linguistica e Dialettologia* 14:89–112.

Ingo Plag. 1999. *Morphological productivity: Structural constraints in English derivation*. De Gruyter. https://doi.org/10.1515/9783110802863.

Ingo Plag. 2003. *Word-Formation in English*. Cambridge University Press. https://doi.org/10.1017/9781316771402.

Carmen Scherer. 2015. Change in productivity. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe–Vol. 3*, De Gruyter, pages 1781–1793. https://doi.org/10.1515/9783110375732-014.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423.

Paulo Sergio Panse Silveira and Jose Oliveira Siqueira. 2023. Better to be in agreement than in bad company: A critical analysis of many kappa-like tests. *Behavior Research Methods* 55(7):3326–3347. https://doi.org/10.3758/s13428-022-01950-0.

Sali Tagliamonte. 2016. *Teen Talk: The Language of Adolescents*. Cambridge University Press. https://doi.org/10.1017/CBO9781139583800.

Werner Vach and Oke Gerke. 2023. Gwet's AC1 is not a substitute for Cohen's kappa: A comparison of basic properties. *MethodsX* 10:102212. https://doi.org/10.1016/j.mex.2023.102212.

Jaap van Marle. 1988. On the role of semantics in productivity change. In Geerd Booij and Jaap van Marle, editors, *Yearbook of Morphology 1988*, Foris, pages 139–154. https://doi.org/10.1007/978-94-017-3710-4_7.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In Nicoletta Calzolari et al., editors, *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*. pages 1317–1325. https://aclanthology.org/2022.lrec-1.140.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. In *Proceedings from the Corpus Linguistics Conference Series 2005*. University of Birmingham. https://hdl.handle.net/11585/15321.

Amir Zeldes. 2012. *Productivity in Argument Selection: From Morphology to Syntax*. De Gruyter. https://doi.org/10.1515/9783110303919.

Nicola Zingarelli, Mario Cannella, Beata Lazzarini, and Andrea Zaninello. 2023. *Lo Zingarelli 2024: Vocabolario della Lingua Italiana*. Zanichelli.

Paying the inheritance tax: Novel and preserved overabundance in Latin prefixed verbs

Matteo Pellegrini, Eleonora Litta and Federica Iurescia
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli 1, 20123, Milan
{matteo.pellegrini,eleonoramaria.litta,federica.iurescia}@unicatt.it

Abstract

This paper analyses how derivational history influences inflectional behaviour by focusing on the phenomenon of overabundance – the presence of multiple forms within the same paradigm cell – in Latin prefixed verbs. By integrating data from multiple sources – namely, the PrinParLat lexicon and the LASLA corpus – this work investigates two key aspects: the emergence of novel overabundance patterns in prefixed derivatives, and the preservation (or lack thereof) of overabundance patterns from base lexemes to their derivatives. Results show that even for derivational processes where inflectional behaviour is expected to be inherited from the base to the derivative, this is not exceptionless, but an "inheritance tax" needs to be paid: such processes can both introduce new overabundance patterns and fail to preserve existing ones.

1 Introduction

While the attestation of different form variants to express the same grammatical meaning has long been recognised in linguistics, it is only in recent times that overabundance – defined as the availability of multiple forms ("cell-mates") in the same paradigm cell of a given lexeme – has been identified as an object of inquiry relevant to discussion on theoretical morphology. This happened mostly thanks to works by Thornton, who first emphasised the relevance of the phenomenon in the framework of a canonical approach to inflection (Corbett, 2005), as a departure from the expectation of inflectional paradigms to canonically display "uniqueness of realisation" (i.e, one form per morphosyntactic property set), providing a thorough investigation of several cases of overabundance in Italian verb inflection (Thornton, 2011). More recently (Thornton, 2019), she sketched a typology of overabundance itself, whose canonicity has been related to factors such as the relative frequency of the available cell-mates, the presence of conditions influencing the choice between them, and the degree to which the phenomenon is widespread across lexemes or cells. Following her lead, several aspects of this issue have been investigated in more detail, such as theoretical consequences on the architecture of morphology (Stump, 2015) and on the structure of the lexicon (Pellegrini, 2023a), or tendencies regarding the distribution of cell-mates in corpora (Guzmán Naranjo and Bonami, 2021). Furthermore, many empirical assessments of overabundance phenomena have appeared for several languages – see, among others, Bermel and Knittl (2012) on Czech and Lečić (2015) on Croatian.

Overabundance pertains to inflection, but it is by now a well-established fact that the derivational history of lexemes can have an influence on their inflectional behaviour. Bonami and Boyé (2006) highlighted the case of French adjectives in *-eur* whose irregular inflection (feminine forms in *-euse* and *-rice*) is determined by the morphological process by which they are created, while Stump (2001, Chapter 4) focused on cases where irregular inflection of complex lexemes is inherited from their base. More recent work quantified the impact of such facts on inflectional predictions systematically, either by taking implicative entropy as a quantitative measure (Pellegrini, 2023b, Chapter 6) or by applying other computational techniques, such as boosting trees classifiers (Bonami and Pellegrini, 2022).

As a consequence of these and other phenomena of interaction between derivational history and inflectional behaviour, derived lexemes appear to be a fertile ground for making existing overabundance

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 87–98, Fribourg, Switzerland, 4-5 September 2025.

phenomena more widespread in the lexicon on the one hand, and for generating novel overabundance phenomena on the other hand. Our aim is to delve deep into these issues, investigating a case study on Latin prefixed verbs, that promise to provide interesting material regarding both possibilities. As for the former, if a base lexeme is overabundant, we expect complex verbs that inherit their inflectional behaviour from their base to be overabundant too – thus perpetuating the overabundance phenomenon at hand across lexemes and reinforcing its entrenchment in the lexicon. However, this is not exceptionless, and even when this happens, it can be interesting to look at the frequency of variants in the base and in derivatives to see if there are relevant differences (see Section 4). By the same reasoning, if a base lexeme is not overabundant, we should expect derived verbs to behave similarly. However, in Latin there are other forces at play that counter this tendency and potentially give rise to novel overabundance phenomena in a systematic (although not exceptionless) fashion (see Section 3). In sum, in both cases inflectional behaviour is inherited from the base to the derivative, but not entirely: an "inheritance tax" is paid in the process.

In this work, we provide an investigation strongly grounded in data of these different possibilities. We adopt an approach that integrates data taken from lexical resources – namely, PrinParLat (Pellegrini et al., 2025) – with data extracted from corpora – namely, the LASLA corpus (Fantoli et al., 2022). Both data sources, as well as the WFL derivational lexicon (Litta and Passarotti, 2019) that we use to identify prefixed verbs and their bases, are available as part of the LiLa Knowledge Base of interoperable resources for Latin (Passarotti et al., 2020), which allows for seamless integration of the pieces of information provided in different places.

The remainder of the work is structured as follows. In Section 2, we detail the procedure that we followed to obtain the data on which our analysis is based. Section 3 discusses cases where novel overabundance phenomena arise in derivatives. In Section 4, we look at the preservation in derivatives of overabundance phenomena observed in the bases. Section 5 concludes and highlights possibilities for future research.

2 Data

The initial source of inflectional data for this work is PrinParLat,[1] a recently developed lexicon where the principal parts of 8,017 Latin verbs are recorded. Principal parts are defined as a set of inflected forms of a lexeme from which the content of all the other cells of that same lexeme can be inferred, thus summarising its overall inflectional behaviour (Stump and Finkel, 2013). In PrinParLat, the cells that are used are FUT.ACT.IND.3.SG and PRS.ACT.INF (for forms of the so-called "present system"), PRF.ACT.IND.1SG (for forms of the so-called "perfect system"), PRF.PTCP.NOM.N.SG and FUT.PTCP.NOM.N.SG (for other participial forms). PrinParLat builds its forms from the database of a morphological analyser for Latin, Lemlat 3.0 (Passarotti et al., 2017), that on its turn ultimately relies on reference Latin dictionaries (Georges, 1998; Glare, 2012; Gradenwitz, 1904). Of course, a morphological analyser aims at being capable to recognise as many forms as possible: as a consequence, whenever multiple forms are available as principal parts according to its data sources, those are all recorded in PrinParLat, making it a good starting point for an analysis of overabundance.

For our purposes, we extract all the forms available as principal parts in PrinParLat for all verbs derived by prefixation and their corresponding bases. To identify the relevant verbs, we use WFL,[2] a word-formation lexicon of Latin that provides information on bases, derivatives and affixes involved in the derivational history of each lexeme. The outcome of this procedure is a sample of 5,900 lexemes. This allows us to investigate what Aigro and Vihman (2023) call "potential overabundance", i.e., to identify those cases where there is the possibility of displaying more than one form.

However, relying solely on data from PrinParLat carries the risk of overestimating the impact of overabundance. To maximise Lemlat's ability to analyse forms, even marginal variants that occur very rarely in texts are recorded, and consequently, they are also included in PrinParLat. As a result, there is no clear way to distinguish these marginal variants from cases where there is real competition between

[1] http://lila-erc.eu/data/lexicalResources/prinparlat/Lexicon.
[2] https://lila-erc.eu/data/lexicalResources/WFL/Lexicon.

cell-mates in texts – what Aigro and Vihman (2023) call "realised overabundance". To be able to evaluate this difference, we take advantage of the LiLa Knowledge Base (Passarotti et al., 2020) which provides access to 1.7 million tokens from a corpus of Classical Latin with finely detailed annotations of the morpho-syntactic features of word forms: the LASLA corpus.[3] First, we use the principal parts of PrinParLat to generate full paradigms of lexemes in our sample, by applying a set of rules that generate all the word forms that can possibly fill the other paradigm cells. Then, we look for these possible forms in the LASLA corpus, and keep only forms that are attested therein, together with their frequencies.[4] Lastly, we aggregate frequencies of different forms that are based on the same stem (or inflection) variant: this makes data less sparse and allows for better focus on slabs of the paradigm comparable to the ones represented by the principal parts of PrinParLat – namely, the present system, perfect system, and perfect and future participial word forms. The outcome of this procedure results in a smaller, but more ecological sample of 3,145 lexemes. On the other hand, one shortcoming of this corpus-based approach is that some potentially relevant overabundance phenomena might be overlooked only because the lexemes and/or cells they appear in are themselves rare. This motivates our choice to adopt an integrated approach and consider both lexical and textual data.

An important but problematic issue concerns the distinction between purely orthographic variants and phonological alternations. For instance, variation is found between forms spelled with <nm> or <mm> (e.g., *inminēre* vs. *imminēre* 'project over'). However, this is very likely to be a purely orthographic fact that does not reflect any linguistic reality, with both spellings corresponding to phonetic [immineːre] (Cser, 2020, pp. 161 ff.). While some degree of uncertainty on the actual pronunciation of Latin is inevitable given the nature of historical language data, we tried as far as possible to only consider variation that is likely to be reflected in the phonetic reality of surface forms. To do that, we performed an automatic transcription of the forms we extracted into IPA, using a custom script based on the recent description of Latin phonology provided by Cser (2020).

Lastly, we performed a manual annotation of our sample, specifying what processes cause each case of overabundance, among the ones mentioned in Section 3.

3 Novel overabundance phenomena in derivatives

In Latin, various linguistic factors contribute to the emergence of novel overabundance phenomena in prefixed verbs that are not observed in their corresponding base verbs. On the one hand, this might be due to the non-systematic application of morpho-phonological rules whose context is only met in the derivative, and not in the base. This is trivially the case of processes that occur at the boundary between the derivational prefix and the lexical base. Another factor contributing to overabundance is the synchronic consequence of a process of weakening of short vowels in non-initial syllables that was active in Old Latin. Lastly, the emergence of overabundance can be caused by analogical processes: when the inflectional behaviour of the base is irregular, in many cases it is preserved as such in the derivative, but it is also possible for some kind of analogy-driven regularisation to happen.

After discussing each of these possibilities in detail in Subsections 3.1, 3.2 and 3.3, respectively, we show their quantitative impact on the data of our sample in Subsection 3.4.

3.1 Processes occurring at the prefix-base boundary

In this subsection, we focus on cases where overabundance is due to phenomena related to the addition of prefixes and their interaction with the bases they combine with. We start with an overview of the different kinds of processes that take place, based on the comprehensive analysis provided by Cser (2020, Chapter 7), to which the reader is referred for further details.

The most common scenario involves assimilation processes of various kinds between consonant-final prefixes and consonant-initial bases. In some cases, assimilation is partial: for instance, regressive voice assimilation is found in *ad-serere* [atserere] 'join to', and regressive place assimilation is found in

[3] http://lila-erc.eu/data/corpora/Lasla/id/corpus.

[4] We do that instead of directly inferring forms from corpora because that would have introduced a lot of noise due to typos or occasional mistakes in the annotation of features.

compōnere ‘unite’ (from *con-* + *pōnere* ‘put’). In other cases, there is total assimilation, like in *alligare* ‘bind to’ (from *ad-* + *legare* ‘bind’) and *asserere* (from *ad-* + *serere*). Some assimilation processes are exceptionless and thus always predictably yield a single output form: this is the case, for instance, of the regressive place assimilation of *componere* just mentioned. Other ones, however, do not apply across the board: for instance, looking at the examples given at the beginning of this paragraph, it can be observed that the outcome of the application of *ad-* prefixation to *serere* can be either *adserere* ([atserere], with partial assimilation) or *asserere* ([asserere], with total assimilation).

Another relevant process is the loss of prefix-final [s] when followed by a base beginning with a voiced consonant. With the prefix *dis-*, the application of the process is systematic, as in examples like *dīgerere* ‘force apart’ from *gerere* ‘carry’ (in contrast to transparent forms such as *dis-currere* ‘run to and from’). With the prefix *trans-*, however, variation is found – for example, between the transparent form *trans-mittere* and *trāmittere* ‘send across’, where [s] is lost between voiced consonants, accompanied by regular nasal loss and vowel lengthening.

With prefixes ending in [b], there is the possibility of finding the insertion of [s] before voiceless stops – e.g., *suscitāre* [suskitaːre] ‘lift up’ from *sub-* + *citāre* ‘set in motion’, *abscēdere* [apskeːdere] ‘depart’ from *ab-* + *cedere* ‘move’, *ostendere* [ostendere] ‘stretch out’ from *ob-* + *tendere* ‘stretch’. However, the application of this process is very limited, and cases where the prefix only appears without [s] are much more common – e.g., *sub-timēre* ‘fear a little’. Most relevantly for our purposes, cases of overabundance are also sometimes attested – e.g., *succipere/suscipere* ‘take up’ from *sub* + *capere* ‘take’.

Other processes with highly restricted applicability are found with the prefixes *ex-* and *ab-*. Regarding *ex-*, it surfaces as [eː] before voiced consonants — e.g., *ēmanēre* ‘stay without’ from *ex-* + *manēre* ‘stay’. Furthermore, when it is followed by a base beginning with [f], a variant [ek] can be found: for instance, with the base *fodere* ‘dig’, besides the transparent *exfodere* and *effodere* with total assimilation, *ecfodere* is also attested (meaning ‘dig out’). Such forms can be interpreted as arising either from regressive assimilation of [s] to [f] followed by degemination ([eksfodere] → *[ekffodere] → [ekfodere]), or from a reanalysis of forms like *exsilīre* ‘spring out’ from *salīre* ‘spring’ as [ek+siliːre] – whereas in such forms only one [s] is found because of degemination, with [eks+siliːre] → *[eksiliːre] (Cser, 2020, p. 167). Regarding *ab-*, it always surfaces as [aː] before bases beginning with [m] and [w] (e.g., *āmovēre* ‘remove from’ from *movēre* ‘move’), but it can surface as either [a] or [au] before the few bases beginning with [f], as shown in the PRF.ACT.IND.1.SG of the verb ‘be away from’ *āfuī* and in *aufugere* ‘run away’, respectively.[5]

Lastly, for the prefix meaning ‘before’, the forms *ante-* and *anti-* are found, with no underlying phonological conditioning – as shown by the contrast between *anti-cipare* ‘take before’ and *ante-cēdere* ‘go before’ – but with little room left for overabundance due to quite systematic lexical selection: the only base with which the two variants are considered to be potentially applicable according to lexical data is {*ante/anti-*}*stare* ‘stand before’, but only the former variant is found in our corpus data.

Moving to vowel-final prefixes, most of the action takes place when they combine with vowel-initial stems. In such cases, we can find either contraction to a single long vowel – as happens with the prefix *dē-*, yielding *dēgere* ‘carry on’ from *agere* ‘lead’ – or insertion of [d] to avoid a hiatus – as happens with the prefix *re-*, yielding *redintegrare* ‘make whole again’ from *integrare* ‘make whole’. Those processes are applied quite systematically, so cases of overabundance are seldom attested – see, e.g., *prohibēre* vs. *prōbēre* ‘hold in front’ from *pro* + *habēre* ‘have’, where, however, the former is much more frequent than the latter (329 occurrences vs. 2 in our corpus). For the prefix *re-*, we also find a few examples of overabundance due to the possibility of lengthening the initial consonant of the base, like in *reddūcere* ‘lead back’ from *re-* + *dūcere* ‘lead’.

To sum up, a wealth of phenomena can be identified with different levels of systematicity, ranging from fully automatic phonological processes (e.g., place assimilation in *compōnere*), to ones that are still phonological in nature but are not exceptionless (e.g., total assimilation in *asserere* vs. *adserere*), to phonologically conditioned allomorphy (selection of *ā-* allomorph in *āmovere*), up to lexically conditioned ones (form variant selection in *anticipare* vs. *antecēdere*). In this context, we are more interested in cases

[5] Concerning the last variant *au-*, some scholars even consider it as not etymologically related to the prefix *ab-*, but we follow De Vaan (2003–2004) who convincingly argues for the two to be allomorphic variants of a same prefix.

that are not systematic, and especially in those where different outcomes are found for the same lexeme, thus giving rise to overabundance.

Table 1 lists all the prefixes that give rise to overabundance phenomena in our dataset. For each of them, it gives the number and percentage of cases where the interaction of the prefix with the base gives rise to overabundance out of the total number of verbs displaying the prefix, and the most frequent segments that are found in initial position in the bases for which overabundance is found, both in lexical data from PrinParLat and in corpus data from LASLA.[6] Prefixes are sorted according to the number of verbs that display them according to lexical data from PrinParLat. Since the kinds of phenomena that take place are mostly the same in different sub-paradigms, we will only discuss data and results on the present system: the picture would be fully comparable, although not identical, if we also considered data pertaining to the perfect system and participial forms.[7]

Prefix	**Overabundant/total verbs (%)**		**Most frequent base initials**		**Examples**
	PrinParLat	**LASLA**	**PrinParLat**	**LASLA**	
con-	59/544 (10.85 %)	14/297 (4.71 %)	[l]: 27 [r]: 13	[l]: 8 [r]: 3	*conlaudāre/collaudāre* *conruere/corruere*
ex-	39/476 (8.19 %)	7/282 (2.48 %)	[f]: 15 [w],[s]: 4	[f]: 6 [s]: 1	*effringere/ecfringere* *essurīre/ēsurīre*
in-	46/438 (10.5 %)	13/247 (5.26 %)	[r]: 21 [l]: 19	[r]: 8 [l]: 5	*inrīdere/irrīdēre* *inlīdere/illīdere*
de-	11/411 (2.68 %)	2/230 (0.87 %)	[s]: 4 [r]: 3	[e],[r]: 1	*deerrare/dērrāre* *dērigere/dīrigere*
re-	14/407 (3.44 %)	3/222 (1.35 %)	[a]: 3 [k],[p]: 2	[d],[k],[p]: 1	*redūcere/reddūcere*
ad-	184/363 (50.69 %)	65/213 (30.52 %)	[s]: 49 [p]: 30 [l]: 21	[s]: 21 [p]: 11 [f]: 8	*adserere/asserere* *adpetere/appetere* *adfīgere/affīgere*
per-	5/305 (1.64 %)	3/160 (1.88 %)	[l]: 4 [j]: 1	[l]: 2 [j]: 1	*perlegere/pellegere* *periero/pēiero*
prae-	2/257 (0.78 %)	-	[k],[m]: 1	-	*praeēminēre/praeminēre*
sub-	45/246 (18.29 %)	8/111 (7.21 %)	[k]: 12 [m],[f]: 9	[m]: 3 [k],[f]: 2	*submittere/summittere* *suscensēre/succensēre*
ob-	24/205 (11.71 %)	2/120 (1.67 %)	[k]: 7 [m],[f]: 5	[f],[m]: 1	*obfundere/offundere*
dis-	12/173 (6.94 %)	4/108 (3.7 %)	[m]: 3 [j],[r],[s]: 2	[j]: 2 [r],[n]: 1	*disiungere/dīiungere* *disrumpere/dirrumpere/dīrumpere*
circum-	3/154 (1.95 %)	-	[a],[i],[d]: 1	-	*circumīre/circuīre*
prō-	3/154 (1.95 %)	1/88 (1.14 %)	[h],[j],[d]: 1	[i]: 1	*prohibēre/prōbēre*
ab-	11/97 (11.34 %)	-	[f]: 2 [p],[b]: 1	-	*abfugere/aufugere/āfugere*
trans-	23/74 (31.08 %)	5/36 (13.89 %)	[s]: 8 [m],[u]: 3	[d],[j],[m], [n],[w]: 1	*transmittere/trāmittere* *transvolāre/trāvolāre*
ante-/anti-	2/21 (9.52 %)	-	[e],[s]: 1	-	*antestāre/antistāre*
ambi-/an-	1/9 (11.11 %)	-	[t]: 1	-	*am(p)truāre/antruāre*
por-	1/2 (50.0 %)	1/2 (50.0 %)	[r]: 1	[r]: 1	*porrigere/porgere*
sēmi-	1/2 (50.0 %)	-	[u]: 1	-	*sēmiustulare/sēmustulare*

Table 1: Overabundance due to processes occurring at the prefix-base boundary (present system)

As is to be expected, the presence of potential overabundance as observed in lexical data is remarkable,

[6]Note that, in some cases, there is not a full overlap between what emerges from lexical data on the one hand and corpus data on the other hand. When the most frequent base initials are not the same according to the two data sources, we provide examples for those attested in corpus data; we only provide examples from lexical data when no corpus attestations are found.

[7]The slight differences would be due to cases of verbs for which forms of other sub-paradigms are not attested in corpora nor recorded in dictionaries, or to cases of strong suppletion that cause the initial segment of the base to be different in different stems, like in the verb meaning 'bring', with present stem *fer-* and perfect stem *tul-*.

but only some of those cases are also found to be realised in corpus data: the numbers and percentages are always lower if data from the LASLA corpus are considered rather than data from the PrinParLat lexicon. If we exclude some very marginal cases like *por-* and *semi-*, the prefix that most frequently gives rise to overabundance phenomena is *ad-*, due to several bases with which either partial voice assimilation (not reflected in writing, e.g., *ad-petere* [atpetere] 'strive after' and *ad-fīgere* [atfiːgere] 'fix upon') or total assimilation (*appetere*, *affigere*) can take place. It is followed by *trans-*, where overabundance is due to the optional application of [s] deletion before voiced consonants, and *sub-*, where it is due to optional application of assimilation. Interestingly, similar rates of overabundance are found with the nasal-ending prefixes *con-* and *in-*, despite the fact that the former is analysed by Cser (2020, p. 162) as ending with a placeless nasal that surfaces in various ways depending on the context, rather than with the coronal nasal. The most common source of overabundance with those prefixes is variation between presence and lack of assimilation before [r] and [l]. On the other hand, prefixes ending with [b] are not so homogeneous, with *sub-* giving rise to overabundance more frequently than *ob-* and *ab-*, thus suggesting a more relevant role of lexical stipulation, which might differ from one prefix to another (or even from one lexeme to another), differently from what we observed with *con-* and *in-*. A broader generalisation that emerges from our data is that the liquid consonants [l] and [r] appear to be particularly oscillating in the capability to trigger regressive assimilation. As a consequence, bases beginning with those segments – especially [l] – are the ones that most frequently give rise to overabundance when combining not only with nasal-final prefixes, but also with *per-* and *ad-*.

3.2 The Old Latin vowel weakening

In Old Latin, short vowels in non-initial syllables underwent a process of weakening, mostly yielding [i] from all vowels in open syllables, [e] from [a] and [u] from [o] in closed syllables – e.g., *[ku.pi .do.taːs] → [ku.pi.di.taːs] *cupiditās* 'desire', *[ske.les.tos] → [ske.les.tus] *scelestus* 'wicked'.[8] Differently than many of the processes described in the previous section, this one is not synchronically active in Classical Latin anymore (Cser, 2020, p. 104). Nevertheless, it has morphological consequences that are still visible regarding the shape of prefixed verbs, as vowels in the initial syllable of the base end up being in non-initial position in the derivative. For instance, the verb FACIO 'make' has PRS.ACT.INF *facere*, PRF.ACT.INF *fēcisse*, and PRF.PASS.PTCP *factus*. When prefixed (e.g., with *con-*), the [a] is not in the initial syllable of the derivative, and it is thus raised to [i] when it is in open syllable (PRS.ACT.INF *conficere*) and to [e] in closed syllable (PRF.PASS.PTCP *confectus*), while the long vowel of the PRF.ACT.IND *confēcisse* is left unchanged. Again, what is more relevant for our purposes is the presence of cases in which forms with and without vowel weakening are attested for the same lexeme, thus giving rise to overabundance, as happens with *ēligere* and *ēlegere* 'pick out' (from *ex-* + *legere* 'gather').

Table 2 summarises our findings on this aspect. As the example just given shows, the effects of this process potentially vary in different sub-paradigms, motivating our choice to present data also concerning the perfect system and the perfect participle in this section.[9] For each sub-paradigm, we give information on the shape of the stem of bases, and the number of derivatives formed by prefixation that display overabundance due to the presence of both morphotactically transparent variants and variants that are affected by vowel weakening, both according to lexical data from PrinParLat and according to corpus data from LASLA, with examples.

Data are very sparse, especially for the perfect system and participial forms. This is by itself an interesting finding, that shows that even if the process of vowel weakening is not active in the synchronic phonology of Classical Latin anymore, this does not generate much overabundance: whether a given derivative will or will not display the consequences of that process appears to be mostly specified lexically.

However, there are some bases for which it seems to be more common to find overabundance due to the presence of both transparent and weakened forms. This is the case of SPARGO 'strew' and TRACTO 'draw violently' – where the effects are visible in all sub-paradigms, cf. present stem *dētract-/dētrect-*, perfect stem *dētractāv-/dētrectāv-*, and perfect participle stem *dētractāt-/dētrectāt* – as well as LEGO 'gather' –

[8] Further details and other kinds of behaviour can be found in Weiss (2009, Chapter 13).

[9] For future participle forms, the picture is entirely comparable to the one of perfect participles, but data are sparser, hence our decision to omit those results in what follows.

	PrinParLat Bases	No.	LASLA Bases	No.	Examples
Present stem	*sparg-*	7	*tract-*	3	*consparg-/consperg-, dētract-/dētrect-, ēleg-/ēlig-, perem-/perim-*
	leg-,tract-	5	*em-,leg-,sparg-*	2	
	cant-,sed-,iac-, em-,frang-,farc-	3	*iac-,prem-,sid-, sec-,quaer-,carp-, mand-,len-,nec-, perg-*	1	
	carp-,part-,pang-, calc-,sacr-,rep-	2			
	(26 bases)	1			
Perfect stem	*spars-,tractāv-*	5	*tractāv-*	2	*contractāv-/contrectāv-, circumstet-/circumstit-, enecu-/enicāv-*
	stet-,cantāv-,licu-,sacrāv-, carps-,partīv-,fars-	2	*necu-,stet-*	1	
	(13 bases)	1			
Prf. ptcp. stem	*spars-,tractāt-*	5	*iact-*	2	*superiact-/superiect-, dēlenīt-/dēlinīt-, āmandāt-/āmendāt-, dispans-/dispess-*
	pass-,fart-,nōt-,cantāt-,partīt-pact-,iact-,sacrāt-,carpt-	2	*lenīt-,mandāt,pass*	1	
	(11 bases)	1			

Table 2: Overabundance due to the consequences of Old Latin vowel weakening

where the effects are only visible in the present stem *ēleg-/ēlig-*, since vowel weakening cannot give rise to different variants in the perfect stem *ēlēg-* and perfect participle stem *ēlect-*.

3.3 The role of analogy

Another potential source of overabundance phenomena in derivatives that are not attested in their bases is due to the role of analogy. When the inflectional behaviour of the base is irregular, in many cases it is preserved as such in the derivative: for instance, the 1st conjugation verb EXCUBO 'sleep outside' has PRF.ACT.IND.1SG *excubuī* like its base CUBO 'lie down' (PRF.ACT.IND.1SG *cubuī*). However, it is also possible for some kind of analogy-driven regularisation to happen: for instance, RECUBO 'lie back' has PRF.ACT.IND.1SG *recubāvī*, like most other 1st conjugation verbs. When both possibilities are available for the same verb, overabundance emerges – e.g., *incubāvī* vs. *incubuī* 'lie upon'.

In our manual annotation, we relied on a broad definition of the notion of analogy, using it as a catch-all term for every phenomenon that was not generated by the phonological processes described in the previous subsections, and that can thus be considered as morphological in nature. This choice is motivated by the difficulty in discriminating different types with neatly identifiable boundaries, due to the presence of many disparate and partly overlapping phenomena. As a consequence, the data are quite heterogeneous, and it is difficult to directly extract quantitative generalisations. In what follows, we try to offer a qualitative discussion instead, highlighting some remarkable processes observed in different sub-paradigms.

The present system is the place where conjugation distinctions are relevant in Latin (Clackson, 2011, p. 113). As a consequence, we focus on cases of analogical conjugation shifts. According to lexical data,[10] the 1st and 4th conjugations are the strongest attractors. For instance, the verb *vādere* 'go' belongs to the 3rd conjugation, but when it is prefixed with *in-*, also 1st conjugation forms like PRS.ACT.INF *invādāre* are considered to be possible, besides the inherited *invādere*. The same happens to the 3rd conjugation verb *glūbere* 'peel' yielding *deglūbāre/deglūbere* and to the 4th conjugation *puvīre* 'strike' yielding *oppuviāre/oppuvīre*. Shifts to the 1st conjugation are consistent with its status as a highly productive class in Classical Latin (Dressler, 2002), preserved in its descendants in Romance languages. Shifts to the 4th conjugation are instead considered to be possible with derivatives of the mixed conjugation verb *linere* 'spread' – e.g., *circumlinere/circumlinīre* – and with the second conjugation verb *sedēre* 'sit' yielding derivative *insidīre/insidēre*. Since the mixed conjugation owes its name to displaying forms inflected

[10]Corpus data are unfortunately too sparse to offer insight on these phenomena.

according to the 3rd conjugation in some cells and forms inflected according to the 4th conjugation in other cells (Dressler, 2002), there is a remarkable sharing of forms between the mixed conjugation and the 4th, which, coupled with the lower type frequency of the former (Pellegrini, 2023b, p. 35), makes analogical shifts to the latter more likely.

Outside of the present system, conjugation distinctions are only partly relevant, so the analogical processes that we will mention concern patterns of stem formation instead. In the perfect system, a very common scenario[11] is the extension of the pattern of perfect stem formation including a long theme vowel followed by *-v-*, as found in the frequent and productive 1st conjugation and in the 4th conjugation. An example of this process has already been provided above for *incub-āv-ī* alongside inherited *incubu-ī*. In addition, we can mention that the perfect of STO 'stay' is *stetī*, but derivatives with prefixes *re-* and *prae-* also admit analogical perfects like *rest-āv-ī* in addition to *restit-ī* (with inherited inflectional behavior and regular weakening of [e] to [i]). Another pattern of stem formation that is frequently[12] found in cases where it is not attested in the base is the one where the perfect stem is identical to the present stem. This pattern is commonly found in derivatives of bases that display reduplicated perfects, like TONDEO 'shave' with PRS.ACT.INF *tond-ēre* and PRF.ACT.IND.1SG *totond-ī*, whose derivative DĒTONDEO 'shear off' displays PRS.ACT.INF *dētond-ere* and PRF.ACT.IND.1SG *dētond-ī* alongside expected *detotond-ī*. However, we also find sporadic examples of extension of perfects in *-u-* – e.g., *concrēd-u-ī* 'I intrusted' besides inherited *concrēdidī* from CRĒDO 'believe' – and sigmatic perfects – e.g., *compul-s-ī* 'I collected' besides inherited *compul-ī* from PELLO 'push'.

A roughly comparable situation is found with the stem of the perfect participle. The pattern that is most commonly[13] extended is the one with the long theme vowel + *-t-*, typical of 1st and 4th conjugation verbs – e.g., *incub-āt-um* besides inherited *incubitum* from CUBO 'lie'. However, other common stem formation processes, such as the addition of plain *-t-* or *-s-* regardless of the base's theme vowel, are sometimes found in cases where they are not applied to bases – e.g., respectively, *depul-t-um* 'expelled' besides inherited *depul-s-um* from PELLO 'push', and *ador-s-um* 'risen up' besides inherited *adortum* from ORIOR 'rise'. As these examples show, shifts from and towards these stem formation patterns happen in both directions, indicating a remarkable competition between the two in similar contexts.

3.4 A quantitative assessment of the impact of novel overabundance in derivatives

Table 3 shows the overall impact of the facts described in the previous subsections on our sample, considering both lexical and corpus data. From this analysis it emerges that among prefixed derivatives there is a remarkable number of novel overabundance phenomena that are not attested in bases. This is more evident when observing lexical data than when examining our corpus data, where some overabundance phenomena are not actually attested. However, the overall proportion of overabundant lexemes is not that different in bases and derivatives: this is shown in the right columns of Table 3, where the difference between the percentage of overabundant lexemes in derivatives and in bases is given for each slab. This is likely to be due to the fact that there are overabundance phenomena in the base that are not preserved in derivatives (see Section 4 for quantitative data). We can speculate that this is related to the frequency of lexemes: derivatives are often less frequently used than their bases, making different cell-mates less likely to appear in corpora and, consequently, less likely to be recorded in lexicons. Indeed, if we perform the same counts considering only lexemes with a token frequency of 100 or more – where overabundance is much less likely to go unattested by chance – the overall impact of overabundance is significantly more pronounced in derivatives than in bases.

Table 4 highlights the remarkable quantitative differences in the individual contribution of each of the possible causes of novel overabundance phenomena identified above: morpho-phonological processes occurring at the boundary between prefix and base have the lion's share, while vowel weakening and

[11] 18 cases according to lexical data, 5 of them also attested in corpus data.

[12] 18 cases according to lexical data.

[13] 11 cases according to lexical data.

[14] In the perfect, a case of systematic overabundance – optionally contracted perfects like PRF.ACT.INF *amāvīsse* vs. *amāsse* 'have loved' – is documented in corpus data and not in lexical data. Since this phenomenon concerns all verbs regardless of derivation, the lower frequency of derivatives compared to bases is likely to motivate this outlier.

	Novel overabundance		Diff. between % overabundant lexemes in derivatives and bases	
	Full dataset	**Freq. >100**	**Full dataset**	**Freq. >100**
Present system (PrinParLat)	639/4,676 (13.67 %)	72/329 (21.88 %)	+1.39	+10.14
Perfect system (PrinParLat)	471/4,295 (10.97 %)	72/310 (23.23 %)	–3.32	+11.4
Perfect participle (PrinParLat)	431/3,953 (10.9 %)	63/312 (20.19 %)	–3.92	+5.01
Present system (LASLA)	151/2,258 (6.69 %)	34/293 (11.6 %)	+3.08	+8.69
Perfect system (LASLA)	73/1,847 (3.95 %)	24/294 (8.16 %)	–23.93[14]	–9.58
Perfect participle (LASLA)	82/1,658 (4.95 %)	24/262 (9.16 %)	+1.7	+7.74

Table 3: Novel overabundance in derivatives

other analogical processes are responsible for a much lower number of cases.

	Prefix-base boundary	**Vowel reduction**	**Analogy**
Present system (PrinParLat)	497	73	75
Perfect system (PrinParLat)	362	37	116
Perfect participle (PrinParLat)	354	39	75
Present system (LASLA)	128	19	13
Perfect system (LASLA)	67	4	12
Perfect participle (LASLA)	80	5	8

Table 4: Distribution of sources for novel overabundance in derivatives

4 Preservation of overabundance in derivatives

If a verb is overabundant, we expect complex verbs that inherit their inflectional behaviour from it to be overabundant too. For example, according to lexical data, overabundance in PRS.ACT.INF *vertere-vortere* 'turn' is mirrored by systematic preservation of the overabundance phenomenon in all derivatives (e.g., *advertere-advortere* 'turn toward'). However, this is not always the case. According to lexical data, the lexeme TENEO 'keep' shows different cell-mates in PRF.ACT.IND.1SG (*tenuī/tetinī/tenīvī*), but no prefixed derivative preserves this overabundance pattern, inheriting only the perfect stem formation in *-u-*, with regular application of vowel reduction (e.g., *con-tinuī* 'keep together'). To investigate this phenomenon, integrating information from corpus data proves particularly useful, as details on the actual attestation of cell-mates in both the base and the derivative might be crucial for a better interpretation of results. In fact, if we look at corpus data, we can see that all 212 attestations of perfective forms of TENEO display the perfect stem formation pattern in *-u-*. Overabundance is thus very marginal (if at all present), and it is no surprise that it is not attested in derivatives.

As for the forms *vertere/vortere*, they are both attested in the base verb, but the former is more frequent, and this competition is also found in some derivatives, but in other ones (almost) only the majoritary variant is attested in texts, as shown in Table 5.

Meaning	**Present stem**		**Perfect stem**		**Perfect participle stem**	
	Variants	**Freq.**	**Variants**	**Freq.**	**Variants**	**Freq.**
'turn'	*vert-/vort-*	314/29	*vert-/vort-*	127/3	*vers-/vors-*	171/3
'turn to'	*advert-/advort-*	61/23	*advert-/advort-*	29/4	*advers-/advors-*	5/1
'precede'	*antevert-/antevort-*	1/2	*antevert-/antevort-*	1/1	-	
'turn away'	*dēvert-*	12	*dēvert-*	13	-	
'turn upside down'	*ēvert-*	79	*ēvert-*	31	*ēvers-/ēvors-*	81/1

Table 5: Corpus data on stem variants of VERTO/VORTO and derivatives

Table 6 presents the overall impact of these facts on our data, providing information on the number and percentage of cases where overabundance phenomena observed in bases are not transmitted to their

derivatives. It is interesting to notice how, despite the bias that different sources of data potentially introduce, the picture remains remarkably similar when we look at lexical and corpus data. This scenario turns out to be more frequent than preservation of overabundance, which we might expect based on the principle that these lexemes should inherit the inflectional patterns of the base. Again, the fact that frequency tends to be lower in derivatives than in bases might play a role, and indeed if only lexemes with token frequency higher than 100 are considered, the proportion of cases where overabundance is not preserved decreases, but still remains remarkable.

	Non-preserved overabundance	
	Full dataset	**Freq. >100**
Present system (PrinParLat)	610/765 (79.74 %)	55/70 (78.57 %)
Perfect system (PrinParLat)	732/1,013 (72.26 %)	37/96 (38.54 %)
Perfect participle (PrinParLat)	572/720 (79.44 %)	25/48 (52.08 %)
Present system (LASLA)	101/123 (82.11 %)	18/25 (72.0 %)
Perfect system (LASLA)	462/644 (71.74 %)	41/95 (43.16 %)
Perfect participle (LASLA)	87/114 (76.32 %)	12/27 (44.44 %)

Table 6: Preserved overabundance in derivatives

5 Conclusions and future work

Our analysis confirms that overabundance phenomena frequently arise in prefixed derivatives even when there is no overabundance in their base. This is most frequently due to morpho-phonological adjustments at the boundary between prefix and base, but can also be due to analogy-driven processes. Despite these newly introduced overabundance patterns, the proportion of overabundant lexemes remains relatively stable between bases and derivatives. Indeed, while derivation can introduce novel overabundance, it can also eliminate pre-existing variation: while some base lexemes transmit their overabundance patterns to derivatives, many do not. Although this discrepancy is partly explained by corpus frequency – derivatives tend to be less frequent than their bases, reducing the likelihood of accounting for multiple inflectional variants – the ratio of preservation of bases' overabundance in derivatives remains quite low.

These findings suggest that even in cases where we expect inflection of derivatives to be inherited from their base, not only are there other forces at play that counter this tendency, but those have a remarkable impact on the lexicon: inheritance taxes are high in the taxation system of morphology. This confirms the value and significance of comprehensive investigations of the impact of derivational history on inflectional behaviour that take a quantitative, rather than qualitative approach.

We close the paper mentioning some aspects that were not dealt with in this paper for reasons of space, but would be interesting to explore in future work. In our analysis of corpus data, we mostly focused on identifying the mere presence of cell-mates, only offering sporadic mentions of their relative frequency. However, it would be useful to be able to distinguish between cases in which two variants are attested with roughly the same frequency and those in which one of them is overwhelmingly preferred. Elaborating on the observations made in Section 4 on TENEO and VERTO/VORTO, this can help to explain why some overabundance phenomena attested in bases are not preserved in derivatives: if one of the cell-mates is much more frequently used than the other one in the base, of course the lack of attestation of the latter in derivatives is much less unexpected. In future work, it could be interesting to try to find ways to take this aspect into account more systematically.

Another aspect that would be worth a more in-depth exploration concerns the factors that might motivate whether a specific morpho-phonological or analogical process is applied or not also in non-overabundant cases. In our analysis, we focused on cases of overabundance, but the same processes that give rise to variation within lexemes in such cases have been shown to also give rise to variation across lexemes that systematically display only one form. For instance, regarding vowel weakening, we focused on cases where both the morphotactically transparent and the weakened form are attested for the same lexeme, as in *eligere* and *elegere* from *legere*. However, we have seen that there are many more cases in which

either only the weakened form – e.g., *diligere* 'esteem' – or only the morphotactically transparent one – e.g., *interlegere* 'cull' – is attested. It would be interesting to explore whether specific factors can be found to motivate the selection of a transparent or weakened form (e.g., the time of formation, or the transparency of the relation between base and derivative), and whether instances of overabundance can be explained as intermediate cases according to these factors. A similar line of reasoning is also potentially helpful for cases of analogical extension and especially for morpho-phonological processes occurring at the prefix-base boundary, where it would be interesting to compare our generalisations with the broader ones drawn by Cser (2020) on the phonology of prefixes (also) in non-overabundant lexemes.

References

Mari Aigro and Virve-Anneli Vihman. 2023. Realised overabundance in Estonian noun paradigms: A corpus study. *Word Structure* 16(2-3):154–175.

Neil Bermel and Luděk Knittl. 2012. Morphosyntactic variation and syntactic constructions in Czech nominal declension: Corpus frequency and native-speaker judgments. *Russian Linguistics* 36(1):91–119.

Olivier Bonami and Gilles Boyé. 2006. Deriving inflectional irregularity. In Stefan Müller, editor, *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar, Varna*. CSLI Publications, Stanford, CA, pages 361–380. http://cslipublications.stanford.edu/HPSG/2006/bonami-boye.pdf.

Olivier Bonami and Matteo Pellegrini. 2022. Derivation predicting inflection: A quantitative study of the relation between derivational history and inflectional behavior in Latin. *Studies in Language* 46(4):753–792.

James Clackson. 2011. The forms of Latin: Inflectional morphology. In James Clackson, editor, *A Companion to the Latin Language*, Wiley-Blackwell, Oxford, pages 105–117.

Greville G. Corbett. 2005. The canonical approach in typology. In Zygmunt Frajzyngier, Adam Hodges, and David S. Rood, editors, *Linguistic diversity and language theories*, John Benjamins, Amsterdam, pages 25–49.

András Cser. 2020. The phonology of Classical Latin. *Transactions of the Philological Society* 118(S1):1–218.

Michiel De Vaan. 2003–2004. Latin *au-* 'away', an allomorph of *ab*. *Anuari de Filologia* 25(26):141–147.

Wolfgang U Dressler. 2002. Latin inflection classes. In Machtelt A Bolkestein, Caroline Kroon, H. Wim Pinkster Harm, Remmelink, and Rodie Risselada, editors, *Theory and Description in Latin Linguistics: Selected Papers from the XIth International Colloquium on Latin Linguistics*, Brill, Leiden, pages 91–110. https://doi.org/10.1163/9789004409057_008.

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 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, pages 26–34.

Karl Ernst Georges. 1998. *Ausführliches lateinisch-deutsches Handwörterbuch*. Wissenschaftliche Buchgesellschaft, Darmstadt. Reprint of first edition of 1913–1918, Hannover, Hahnsche Buchhandlung.

Peter Geoffrey William Glare. 2012. *Oxford Latin Dictionary*. Oxford University Press, Oxford, 2nd edition.

Otto Gradenwitz. 1904. *Laterculi Vocum Latinarum: Voces Latinas Et a Fronte Et a Tergo Ordinandas*. Hirzel, Leipzig.

Matías Guzmán Naranjo and Olivier Bonami. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa* 6. https://doi.org/10.5334/gjgl.1626.

Dario Lečić. 2015. Morphological doublets in Croatian: the case of the instrumental singular. *Russian Linguistics* pages 375–393.

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina. Volume I: Words and Sounds*, De Gruyter, Berlin, pages 224–239. https://doi.org/http://doi.org/10.1515/9783110647587-015.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 package for morphological analysis of Latin. In Gerlof Bouma and Yvonne Adesam, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Linköping University Electronic Press, Gothenburg, pages 24–31. https://aclanthology.org/W17-0506.pdf.

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* 48(1):177–212. https://doi.org/http://doi.org/10.4454/ssl.v58i1.277.

Matteo Pellegrini. 2023a. Flexemes in theory and in practice: Modelling overabundance in Latin verb paradigms. *Morphology* 33(3):361–395. https://doi.org/10.1007/s11525-023-09414-7.

Matteo Pellegrini. 2023b. *Paradigm Structure and Predictability in Latin Inflection: An Entropy-based Approach*. Springer, Cham. https://doi.org/10.1007/978-3-031-24844-3.

Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2025. PrinParLat: A lexicon of principal parts of Latin verbs linked to the LiLa knowledge Base. *Language Resources and Evaluation* https://doi.org/10.1007/s10579-025-09847-y.

Gregory Stump. 2015. *Inflectional Paradigms: Content and Form at the Syntax-Morphology Interface*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781316105290.

Gregory Stump and Raphael A. Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139248860.

Gregory T. Stump. 2001. *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press, Cambridge.

Anna M. Thornton. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*, Oxford University Press, Oxford, pages 357–381.

Anna M Thornton. 2019. Overabundance: A canonical typology. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*. Springer, Cham, pages 223–258. https://doi.org/10.1007/978-3-030-02550-2_9.

Michael Weiss. 2009. *Outline of the Historical and Comparative Grammar of Latin*. Beech Stave Press, Ann Arbor/New York.

Improving the quality of morphological segmentation using self-training methods

Michal Olbrich
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
olbrich@ufal.mff.cuni.cz

Zdeněk Žabokrtský
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
zabokrtsky@fal.mff.cuni.cz

Abstract

This paper explores the use of self-training methods to improve the quality of morphologically segmented datasets. Such datasets often suffer from inconsistencies—most notably, missing annotations of some segmentation boundaries. We present two experiments. In our first experiment, we leveraged the generalization capabilities of neural networks to detect such unannotated segmentation points. Then, these suggested corrections were manually reviewed by linguistically trained annotators, and their accuracy was assessed. In the second experiment, we evaluated how well this correction procedure performs in a fully automatic setting. First, we apply it to the original, unaltered data. Second, we investigate how effectively the model can predict gold-standard labels when trained on asymmetrically corrupted data, where a portion of the segmentation boundaries has been deliberately removed. We then measure how accurately the model is able to recover these missing boundaries compared to the original dataset.

1 Introduction

Morphological analysis is the decomposition of words into morphemes, the smallest linguistic units that carry meaning (Haspelmath, 2020). For example, the English word *copyists* can be segmented into *copy-ist-s*. Segmentation can be performed either manually, based on grammatical rules and the linguistic intuition of annotators, or automatically. Automatic segmentation is typically performed using either an unsupervised approach (Creutz and Lagus, 2007), a semi-supervised approach (Bodnár et al., 2020), or a supervised approach that uses training data, nowadays often using deep neural networks (Bolshakova and Sapin, 2020; Peters and Martins, 2022; Morozov et al., 2024). This study focuses on the use of neural networks for manually annotated segmentation data.

The creation of manually annotated morphemic data sets is time-consuming due to the vastness of a language's vocabulary. For illustration: the Czech dictionary of inflected forms MorfFlex 2.1 includes 126 million word forms (Hajič et al., 2024). In contrast, manually compiled morphological segmentation dictionaries typically contain only tens of thousands of words or fewer (Slavíčková, 1975; Ološtiak et al., 2015; Sánchez-Gutiérrez et al., 2018).

Our initial goal was to develop an improved neural network-based model for automatic morphological segmentation of Czech. However, similar to the observations of Garipov et al. (2023) and Morozov et al. (2024), we encountered a range of inconsistencies in the annotated resources available. These inconsistencies hinder both the training and evaluation of neural models. As a result, we shifted our focus toward detecting these inconsistencies and, where feasible, correcting them automatically.

2 Dataset

2.1 Czech dataset

For the purposes of this study, a morphological segmentation dataset created for the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022) with some additional sources was used. The total number of segmented words in the compiled collection is 56,459. The SIGMORPHON dataset

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 99–108, Fribourg, Switzerland, 4-5 September 2025.

consists of 34,000 segmented word forms and is itself a compilation of multiple sources. Approximately 11,000 words were taken from the MorfoCzech dictionary (Pelegrinová et al., 2021). Approximately 13,000 entries are verbs extracted from the digitized portion of Slavíčková's dictionary (Slavíčková, 1975), while the rest is a compilation of several other sources, such as 19,000 verb lemmas segmented by Hledíková (2024).

This multi-source nature increases the dataset's size but also introduces challenges related to the inconsistency of the annotation rules. Such inconsistencies can be observed, for instance, as conflicting segmentations of the same word across various sources.

2.2 English and Slovak datasets

Our primary goal was to improve the quality of a morphological segmentation dataset for Czech. For comparison purposes, however, we included two additional languages in our experiments: Slovak and English. Slovak was chosen because of its close linguistic similarity to Czech and the availability of a high-quality, linguistically annotated dataset. As a source of Slovak data, we used a digitized version of the Retrograde Morphemic Dictionary of Slovak (Ološtiak et al., 2015). English was included as a generally accessible reference language. For English, we used data from the Universal Segmentations project (Žabokrtský et al., 2022), specifically the converted version of the MorphoLex dictionary (Sánchez-Gutiérrez et al., 2018).

Language	Train size	Test size	Avg. boundaries/word	Avg. word length
Czech	52,458	4,000	2.6	8.1
English	64,623	4,000	1.2	8.3
Slovak	65,430	4,000	2.9	8.6

Table 1: Basic quantitative properties of the datasets used in this study: training and test set sizes, average number of segmentation boundaries per word, and average number of characters per word.

2.3 Identification of annotator inconsistencies in the Czech dataset

Based on an analysis of available sources for Czech, several challenges related to problematic segmentation have been identified. Czech is an inflectionally rich language, with individual morphemes that can serve multiple functions. For example, a single morpheme may simultaneously mark the plural and genitive case. Another challenge for Czech, as well as other languages, that complicates machine segmentation is allomorphy—the phenomenon where a morpheme appears in different forms. For instance, the root morpheme ***př*** can take various shapes, such as *o-**př**-ít* ('to support'), *pod-**pěr**-a* ('a support'), or *pod-**půr**-n-ý* ('supportive'). Ideally, training data should include as many variants as possible to improve the model's ability to predict segmentation for unseen words.

Another theoretical issue concerns the segmentation of loanwords. One approach is to leave borrowed words unsegmented, while another extreme is to segment them according to the morphological rules of the source language. A middle-ground approach involves performing partial segmentation based on specific criteria. One such criterion is the frequency of words that share the same root. If multiple words with the same root exist, it makes sense to treat that root as an independent morpheme. For example, compounds with Latin roots, such as *terapie*, can be further segmented into *canis-terapie* and *chemo-terapie*. The treatment of loanwords also depends on how long ago they entered the language. Older borrowings may have undergone phonological changes that obscure or eliminate the original morpheme. A striking example is the Czech word *žák* ('pupil'), which originates from the Greek *dia-kon-os* (Rejzek, 2001), where the original morphological structure has been completely lost. A case of partial change is *abatyše* ('abbess'), derived from Latin *abbat-issa* (Rejzek, 2001). Here, the Czech adaptation has altered the structure to the point that segmentation into *abat-yše* would be meaningless, as these morphemes do not occur elsewhere in Czech. In contrast, from the same Latin root *abbat*, the Czech words *opat* ('abbot') and *opat-ství* ('abbey') retain a recognizable and segmentable morpheme.

Another complex case in morphological segmentation involves words with the suffix *-árn(a)*, which denotes places associated with crafts or production, such as *pekárna* ('bakery'), *kovárna* ('blacksmith's workshop'), *továrna* ('factory'), and *elektrárna* ('power plant'). The segmentation of such words depends on whether they derive from the person performing the activity or from the activity itself (Šimandl et al., 2016). For instance, *pekárna* could be derived from *pekař* ('baker'), leading to the segmentation *pek-ár-n-a*, or from *pečení* ('baking'), in which case the segmentation would be *pek-árn-a*. An interesting case is *továrna* ('factory'), which comes from *tovar* ('goods'), a borrowing from Turkic-Tatar languages (Rejzek, 2001). Here, the segmentation *továr-n-a* is in place, since *ár* is part of the root rather than an independent morpheme. Further comparison can be seen in Table 2. Ambiguities such as these introduce inconsistencies in the dataset, making it more difficult for automatic segmentation algorithms to generalize effectively.

Original word	**Segmented form**	**Gloss**
pek**árna**	pek-**ár-n-a**	bakery
tisk**árna**	tisk-**árn-a**	printing house
ocel**árna**	ocel-**árn-a**	steel mill
kas**árna**	kas**árn-a**	barracks
kav**árna**	kav-**árn-a**	coffee shop
tov**árna**	tov**ár-n-a**	forge

Table 2: Examples of different segmentations of words ending with **árna** from the dataset.

3 Method

In our experiment, words were converted to one-hot encoded sequences of characters without diacritics as input to the neural network. All sequences were padded to a length of 32, as none of the words in our dataset were longer. A binary feature indicating the presence of diacritics was added for each character. The labels consisted of binary sequences that match the length of the word, where 0 indicated the absence of a segmentation boundary between adjacent characters, and 1 marked the presence of a boundary as shown in Table 3. During inference, the model predicted the probability of a segmentation boundary for each position using a sigmoid activation function, with a threshold set at 0.5.

Training example	**copy-ist-s**							
Features	c	o	p	y	i	s	t	s
Labels	0	0	0	1	0	0	1	0

Table 3: A simplified illustration of input features and corresponding segmentation labels for a neural network model.

3.1 Neural model

Based on our previous experiments with various neural network architectures for morphological segmentation (Olbrich and Žabokrtský, 2025), we selected a 1D ResNet-based architecture (He et al., 2016) utilizing convolutional neural networks. Other candidate models included architectures combining convolutional layers with Bi-directional LSTM (Bi-LSTM) (Hochreiter and Schmidhuber, 1997), a multilayer perceptron, gradient boosted decision trees (GBDT), and for comparison also common baseline methods: Morfessor (Virpioja et al., 2013) and ULM (Kudo, 2018). Among these, the 1D ResNet architecture performed best. This finding is consistent with the results of Morozov et al. (2024) and Bolshakova and

Sapin (2020), where convolutional network-based models also proved to be the most effective. We also experimented with the incorporation of word embeddings but found that this did not improve performance.

Specifically, we employed a network with fifteen layers of ResNet blocks, a convolutional kernel width of 240, cosine learning rate decay, label smoothing of 0.1 and a dropout rate of 0.3. The model was trained for 35 epochs, as training for more epochs led to overfitting. This setup led to a word accuracy of 87.6% and 94.3 morpheme F1 on the test set. The test set consists of 4,000 words, randomly chosen from the dataset.

3.2 Self-training and pseudo-labeling

Self-training is a semi-supervised machine learning technique in which a model is initially trained on a labeled dataset and then used to generate predictions for unlabeled data; the most confident of these predictions are then treated as pseudo-labels and incorporated back into the training process. This iterative approach can improve performance in scenarios where labeled data is limited, under the assumption that the model's confident predictions are generally correct (Ouali et al., 2020).

In our case, we adapt this concept to a slightly different setting. Morphological segmentation is framed as a binary sequence labeling task, where each training example (i.e., a word) is associated with a sequence of binary labels indicating segmentation boundaries. As a result, one example corresponds to multiple output labels. Due to the nature of the task, different examples may share identical or very similar character sequences—in particular, the same morphs—but exhibit inconsistent segmentation annotations.

To address this, we sought to standardize the training data by identifying and correcting annotation inconsistencies. This included both adding missing segmentation boundaries and removing redundant ones. To that end, we leveraged principles from semi-supervised learning in two complementary ways: (i) to flag potentially inconsistent segmentations for manual inspection; and (ii) to identify candidate boundaries suitable for automatic correction.

Our approach is based on the assumption that when a model predicts a segmentation boundary—that is, when the output of the sigmoid activation exceeds 0.5—at a position where the gold label indicates a non-boundary ('0'), it may be because the model has generalized this pattern from similar examples. In such cases, the prediction may reflect a more consistent or linguistically plausible segmentation than the original annotation. If the data contain noise or omissions, this model-derived "signal" serves as a useful heuristic for identifying likely annotation errors, particularly missing boundaries.

3.3 Metrics

Similar to previous experiments (Batsuren et al., 2022; Olbrich and Žabokrtský, 2025), we report three evaluation metrics:

First, *word accuracy*, defined as the proportion of correctly segmented words:

$$\text{Word Accuracy} = \frac{\#\text{Correctly segmented words}}{\#\text{Total words}}$$

Second, *morpheme-level precision*, *recall*, and *F1 score*, using true positives (TP), false positives (FP), and false negatives (FN):

$$\text{Precision}_{\text{morph}} = \frac{TP_{\text{morph}}}{TP_{\text{morph}} + FP_{\text{morph}}}, \quad \text{Recall}_{\text{morph}} = \frac{TP_{\text{morph}}}{TP_{\text{morph}} + FN_{\text{morph}}},$$
$$\text{F1}_{\text{morph}} = 2 \cdot \frac{\text{Precision}_{\text{morph}} \cdot \text{Recall}_{\text{morph}}}{\text{Precision}_{\text{morph}} + \text{Recall}_{\text{morph}}}$$

Third, *segmentation boundary precision*, *recall*, and *F1 score*, computed analogously at the level of individual boundary positions (rather than morphemes).

Fourth, for the asymmetric noise injection setup (see Section 3.4), we additionally report *Boundary Recovery Accuracy (BRA)*. This metric measures how well the model recovers segmentation boundaries that were deliberately removed from the training data. It is defined as:

$$\text{BRA} = \frac{\#\text{Recovered boundaries}}{\#\text{Removed boundaries}}$$

This metric focuses only on the artificially missing boundaries and evaluates whether the model is able to generalize the segmentation behavior from the remaining (potentially noisy) data. It serves as a proxy for the model's ability to correct or recover annotation inconsistencies under controlled label-noise conditions.

3.4 Asymmetrical noise injection

Since it is difficult to evaluate whether the predicted boundaries missing from the training data are truly correct—unless a higher-quality reference is available—we opted to simulate this scenario using asymmetric noise injection.

In this setup, noise is introduced by randomly removing a certain proportion of the segmentation boundaries from the training labels, while keeping the original test set unchanged. The model is then trained on such partially corrupted data and evaluated on its ability to recover the omitted boundaries.

This controlled experiment allows us to compute standard evaluation metrics such as Boundary Recovery Accuracy (BRA) and precision. Although we acknowledge that this kind of uniform noising does not fully reflect the complexity and biases of real-world annotation errors, we consider it a useful proxy for assessing the model's ability to recover missing information.

3.5 Manual detection of incorrectly annotated segmentations

Manually reviewing all 56,459 words in our dataset would have been a very time-consuming task. Therefore, we decided to narrow down the selection of problematic segmentations for manual inspection. To predict suspicious cases, we utilized our custom ResNet model. We trained this model on the entire dataset and then used it to generate morphological segmentation predictions.

Since the model does not achieve perfect accuracy, it sometimes predicts segmentation boundaries that are not present in the manually annotated training data, and in other cases, it fails to mark boundaries where they are annotated. However, this does not necessarily mean that the model's predictions are incorrect; rather, it may indicate an error in the manual annotations.

The set of potentially incorrect segmentations identified by the model was then subjected to manual review by linguists. They either confirmed or rejected the model's predictions or made further modifications based on their linguistic judgment. Additionally, the linguists received guidelines on how to approach ambiguous cases, as outlined in Section 2.3.

Finally, we replaced the original segmentations with the manually corrected ones and repeated the entire experiment using this revised dataset.

This process was repeated iteratively, first on the original uncorrected dataset and then on the dataset with corrections. For a detailed analysis, see the Results section.

4 Results

4.1 Qualitative analysis of predictions

In this section, we present a selection of segmentation boundaries predicted by the model that were not present in the original gold annotations. The examples cover a range of phenomena, including missing prefixes, suffixes, and undersegmented roots. In many cases, the model correctly inferred linguistically plausible boundaries, suggesting an ability to generalize from observed patterns. However, we also highlight examples where the model's predictions appear questionable or over-segmented, reflecting the limitations of generalization in the presence of ambiguous or noisy training data. These cases provide insight into both the strengths and weaknesses of the model when used for automatic correction of inconsistencies.

4.1.1 Czech

#	Word	Gold segmentation	Predicted segmentation	Gloss
1	divadlo	div-a-dlo	div-a-dl-o	theater
2	vojenského	vojen-sk-ého	voj-en-sk-ého	military (gen. sg.)
3	polotovar	polo-tovar	pol-o-tovar	semi-finished product
4	prosmyknout	pro-smyk-nou-t	pro-s-myk-nou-t	to slip through
5	vzájemné	vzájem-né	v-zá-jem-né	mutual (nom. pl. n.)
6	půlmiliardový	půl-miliard-ov-ý	půl-mili-ard-ový	worth half a billion
7	záplata	záplat-a	zá-plat-a	patch
8	akcionář	akcion-ář	akci-on-ář	shareholder

Table 4: Examples of segmentation boundary inconsistencies between gold annotations and model predictions. Hyphens (-) indicate segmentation boundaries. Commentary on each case is provided below.

Commentary on selected examples from Table 4:

1. Missing segmentation of the ending "o".
2. The true root is "voj"—compare to other examples: *voj-sk-o*, *voj-ák*.
3. Missing segmentation of the interfix "o".
4. Undersegmented "smyk"—the true root is "myk", similarly found in *za-myk-a-t*, *vy-myk-a-t*; this connection may be semantically unclear for some speakers.
5. Semantically not fully transparent, but "v" and "zá" are prefixes: "v" → *zájem*, "za" → *jmout*.
6. Segmentation of the borrowing *miliarda*; both "mili" and "arda" are productive in Czech, e.g., *mili-metr*, *bili-arda*.
7. Missing segmentation of the prefix "zá".
8. Example of a borrowing—the true root is "akci", as in *akci-e*, *akci-ov-ý*.

4.1.2 Slovak

#	Word	Gold segmentation	Predicted segmentation	Gloss
1	rozhodne	rozhod-ne	roz-hod-ne	decides
2	frajerkárstvo	frajer-k-ár-stv-o	fraj-er-k-ár-stv-o	philandering
3	vyparatiť	vyparat-i-ť	vy-parat-i-ť	to make mischief
4	prosperovať	prosper-ov-a-ť	pro-sper-ov-a-ť	to prosper
5	dvíhačka	dvíh-a-čk-a	dvíh-a-č-k-a	jack (lifting device)

Table 5: Examples of segmentation boundary inconsistencies between gold annotations and model predictions in Slovak. Hyphens (-) indicate segmentation boundaries.

Commentary on selected examples from Table 5:

1. Missing segmentation of the prefix *roz-*.
2. Missing boundary between *fraj* and *er*; compare with another entry from the dictionary with the same root: *frajerôčka* → *fraj-er-ôč-k-a* (Ološtiak et al., 2015).
3. Missing segmentation of the prefix *vy-*.
4. An example where the model correctly predicts the Latin structure; however, in this context, the root should be *prosper*.
5. Missing boundary within *a-č-k-a*; compare with other entries such as *trh-a-č-k-a*, *stíh-a-č-k-a* (Ološtiak et al., 2015).

4.1.3 English

#	Word	Gold segmentation	Predicted segmentation
1	incurability	incur-abil-ity	in-cur-abil-ity
2	extravagantly	extravagant-ly	extravag-ant-ly
3	disconsolate	disconsolate	dis-consolate
4	atomically	atom-ic-ally	atom-ic-al-ly
5	unless	unless	un-less

Table 6: Examples of gold vs. predicted morphological segmentations in English. Hyphens (-) indicate segmentation boundaries.

Commentary on selected examples from Table 6:

1. Missing segmentation of the prefix *in-*, which is a common negative prefix in English.
2. Missing segmentation of the suffix *-ant*, which in this context acts as an adjectival or nominal suffix separate from the root *extravag-*.
3. The model segments *disconsolate* as *dis-consolate*, recognizing the negative prefix *dis-*, while the gold treats it as unsegmented, possibly reflecting its lexicalized status.
4. The predicted segmentation adds an extra boundary within the suffix: *-al-ly* instead of the gold's *-ally*, reflecting possible morphological decomposition into adjective + adverbial suffix.
5. While the model segments *unless* into *un-less*, this word is generally considered not transparently segmentable into smaller productive morphemes in contemporary English usage. Historically, however, it can be analyzed diachronically (etymologically) as a combination of *un* and *less*.

4.2 Detection of incorrectly annotated segmentations

In the first iteration of this experiment, the model produced different segmentations for 1,168 words compared to manual annotations, resulting in a Word Accuracy of 97.9%. Among these predictions, annotators corrected 328 words (27.5%), from which 278 were exactly predicted by the model. From the perspective of segmentation borders, the total number of added borders by annotators was 433, from which 377 were detected by the model and on top of that 56 were added by the annotators. In contrast to that, only 6 segmentation borders were removed by the annotators. After correcting those 439 erroneous segmentation borders and updating the data set while maintaining the same train-test split, the Word Accuracy on the test set improved to 88.1%, marking an increase of 0.8%. Repeating the same experiment on the corrected dataset led to a further improvement in word accuracy to 88.8% and the morpheme F1 score reached 94.9%.

These results indicate that the identification and correction of annotation inconsistencies contribute to better model performance, supporting the importance of dataset quality in neural network-based morphological segmentation.

4.3 Recovery of missing boundaries under noisy supervision

To simulate the model's ability to correct erroneous or inconsistent annotations, we conducted controlled experiments with asymmetric label noise, where 5% or 10% of segmentation boundaries were randomly removed from the training data. The model was then evaluated on its ability to recover these boundaries using the Boundary Recovery Accuracy (BRA) metric, alongside boundary-level precision (see Tables 7 and 8).

The results show a clear trade-off between recall and precision as training progresses. In the early training phase (15 epochs), the model achieves the highest BRA scores across all three languages, suggesting a stronger capacity to recover missing boundaries. However, this comes at the cost of lower precision—i.e., a higher number of false positives.

As training continues to 25 and 35 epochs, the BRA consistently declines, while precision improves. This suggests that the model becomes increasingly conservative in placing boundaries, likely due to overfitting to the partially corrupted training data. Interestingly, the 35-epoch models, despite their lower recovery rates, produce highly precise segmentations. This makes them suitable for identifying outliers or suspiciously annotated instances with minimal false positives.

We interpret these results as evidence that early-stage models are more effective for detecting annotation omissions, while later-stage models are better suited for minimizing over-segmentation noise in downstream corrections.

Language	**15 Epochs**		**25 Epochs**		**35 Epochs**	
	BRA	**Precision**	**BRA**	**Precision**	**BRA**	**Precision**
Czech	0.91	0.81	0.83	0.91	0.59	0.96
Slovak	0.93	0.89	0.86	0.96	0.61	0.99
English	0.92	0.64	0.76	0.88	0.44	0.94

Table 7: Boundary recovery results after injecting asymmetric label noise (5% of segmentation boundaries removed). BRA denotes Boundary Recovery Accuracy—i.e., the proportion of correctly predicted removed boundaries. Precision denotes the proportion of all predicted segmentation boundaries that were correct. The model used was a 1D ResNet trained for 15, 25, and 35 epochs, respectively.

Language	**15 Epochs**		**25 Epochs**		**35 Epochs**	
	BRA	**Precision**	**BRA**	**Precision**	**BRA**	**Precision**
Czech	0.90	0.90	0.87	0.92	0.77	0.96
Slovak	0.94	0.94	0.91	0.97	0.83	0.99
English	0.87	0.86	0.75	0.93	0.66	0.94

Table 8: Boundary recovery results after injecting asymmetric label noise (10% of segmentation boundaries removed). BRA denotes Boundary Recovery Accuracy—i.e., the proportion of removed boundaries correctly predicted. Precision denotes the proportion of all predicted segmentation boundaries that were correct. The model used was a 1D ResNet trained for 15, 25, and 35 epochs, respectively.

5 Conclusion

This study investigated the impact of annotation inconsistencies on neural morphological segmentation and explored strategies for detecting and correcting such errors. We demonstrated that standard segmentation models, when trained on noisy data, often generalize in a linguistically meaningful way, allowing them to highlight likely annotation mistakes. Through a combination of qualitative analysis and controlled experiments with asymmetric noise injection, we showed that model predictions can serve as a useful signal for identifying undersegmented or incorrectly labeled examples.

Our experiments with boundary recovery revealed that while models are capable of recovering missing boundaries early in training, they tend to overfit to noisy labels with prolonged training, reducing their ability to generalize. Nevertheless, models trained longer exhibited higher boundary precision, suggesting their utility for low-recall, high-precision correction strategies. Using predicted boundaries as suggestions for manual review led to a successful correction of over 400 erroneous segmentation points in the Czech dataset.

These corrections translated into measurable improvements in both word accuracy and morpheme-level performance. After one round of semi-automatic correction, word accuracy improved by 0.8%, and a subsequent iteration yielded further gains. This highlights the importance of high-quality annotations in morphologically rich languages, where boundary decisions are often ambiguous.

Overall, our findings suggest that semi-supervised learning techniques—particularly boundary-based pseudo-labeling—are promising tools for improving annotation consistency in low-resource or morphologically complex settings.

Acknowledgements

This work has been supported by the Charles University Research Center program No. 24/SSH/009. The authors also thank Dávid Držík for digitizing the Retrograde Morphhemic Dictionary of Slovak, as well as the authors of the dictionary for making the data accessible.

References

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. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In Garrett Nicolai and Eleanor Chodroff, editors, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Seattle, Washington, pages 103–116. https://doi.org/10.18653/v1/2022.sigmorphon-1.11.

Jan Bodnár, Zdeněk Žabokrtský, and Magda Ševčíková. 2020. Semi-supervised induction of morpheme boundaries in Czech using a word-formation network. In Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors, *Text, Speech, and Dialogue. 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*. Springer International Publishing, Cham, pages 189–196. https://doi.org/10.1007/978-3-030-58323-1_20.

Elena I. Bolshakova and Alexander S. Sapin. 2020. An experimental study of neural morpheme segmentation models for Russian word forms. In Alexander Elizarov and Natalia Loukachevitch, editors, *Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020)*. Kazan, Russian Federation, pages 79–89. https://ceur-ws.org/Vol-2780/paper7.pdf.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)* 4(1). https://doi.org/10.1145/1187415.1187418.

Timur Garipov, Dmitry Morozov, and Anna Glazkova. 2023. Generalization ability of CNN-based morpheme segmentation. In Arutyun Avetisyan, editor, *2023 Ivannikov Ispras Open Conference (ISPRAS)*. IEEE, Moscow, Russian Federation, pages 58–62. https://ieeexplore.ieee.org/document/10508171.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. 2024. MorfFlex CZ 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. https://hdl.handle.net/11234/1-5833.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology* 30(2):117–134. https://doi.org/10.1007/s11525-020-09355-5.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Lisa O'Conner, editor, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pages 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hana Hledíková. 2024. Verbs annotated for morphemic structure in Czech, English, German, Spanish. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5824.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 66–75. https://doi.org/10.18653/v1/P18-1007.

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, and Anna Glazkova. 2024. Automatic morpheme segmentation for Russian: Can an algorithm re-place experts? *Journal of Language and Education* 10(4):71–84. https://doi.org/10.17323/jle.2024.22237.

Michal Olbrich and Zdeněk Žabokrtský. 2025. Morphological segmentation with neural networks: Performance effects of architecture, data size, and cross-lingual transfer in seven languages. In *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*. Springer. To appear.

Martin Ološtiak, Ján Genči, and Soňa Rešovská. 2015. *Retrográdny morfematický slovník slovenčiny [Retrograde Morphemic Dictionary of Slovak]*. Filozofická fakulta Prešovskej univerzity v Prešove.

Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. An overview of deep semi-supervised learning. https://arxiv.org/abs/2006.05278.

Kateřina Pelegrinová, Viktor Elšík, Radek Čech, and Ján Mačutek. 2021. MorfoCzech 1.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5202.

Ben Peters and Andre F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In Garrett Nicolai and Eleanor Chodroff, editors, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Seattle, Washington, pages 131–138. https://doi.org/10.18653/v1/2022.sigmorphon-1.14.

Jiří Rejzek. 2001. *Český etymologický slovník [Czech Etymological Lexicon]*. LEDA.

Claudia H. Sánchez-Gutiérrez, Hugo Mailhot, S. Hélène Deacon, and Maximiliano A. Wilson. 2018. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods* 50:1568–1580. https://doi.org/10.3758/s13428-017-0981-8.

Josef Šimandl, Zdenka Rusínová, and Vladimír Petkevič. 2016. *Slovník afixů užívaných v češtině*. Charles University, Karolinum.

Eleonora Slavíčková. 1975. *Retrográdní morfematický slovník češtiny [Retrograde Morphemic Dictionary of Czech]*. Academia.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. https://aaltodoc.aalto.fi/server/api/core/bitstreams/78f1f8d4-c7a4-49e5-992e-85bd70f06ed4/content.

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards universal segmentations: UniSegments 1.0. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 1137–1149. https://aclanthology.org/2022.lrec-1.122/.

Derivational morphemes as markers of borrowed words in Czech

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

Vojtěch John
Charles University,
Faculty of Mathematics and Physics,
john@ufal.mff.cuni.cz

Abstract

In Czech, borrowed words are often marked by the presence of morphemes of foreign origin. Non-native derivational affixes such as *un-*, *dys-*, and *anti-* frequently appear in these words, typically alongside foreign stems. However, it remains unclear to what extent these morphemes independently influence the classification of a word as borrowed or act as a marker for borrowed words. This study investigates the role of derivational morphemes in marking borrowed words in four parts of speech (POS): nouns, verbs, adjectives, and adverbs. We extract native and borrowed words from DeriNet, which categorizes words based on their POS tags and loanword status. Using a multinomial naive Bayes classifier, we perform binary classification to distinguish native and borrowed words, extracting classification probabilities for the morphemes that make up these words. We compare these results with an attention-based binary LSTM classifier and RobeCzech, a pre-trained RoBERTa model for Czech that we fine-tune for our classification task. Our findings show that derivational morphemes receive high classification probabilities and attention scores in borrowed words similar to lexical roots, suggesting that they are salient in signaling borrowings. These results highlight the diagnostic role of derivational morphemes in marking borrowed vocabulary in Czech.

1 Introduction

When a foreign root or stem is borrowed from a language, the morphological integration within the linguistic subsystem of the recipient language is completed by derivational and inflectional morphemes (Poplack et al., 1988). However, sometimes there are specific strategies to mark these foreign elements; for example, Blaha (2022) shows that verbal conjugation with the affix *-ova* is highly productive for borrowed roots (nominal), which can be seen as a strategy for *marking* loanwords in Czech (cf. Stephen and Žabokrtský, 2023). This is also the case for Russian, where the borrowed stems are productively attached to the verbalizer *-ova* (Wohlgemuth, 2009). The additional piece of evidence comes from the languages spoken in the Balkans, where the formative affix *-s-* is productively used as a loanverb marker (Gardani et al., 2015).

The path through which an affix enters a language can also influence the loanword marking strategies. Affixes can be transferred as direct (borrowing of the affix directly from a donor language) or indirect borrowings (through multimorphemic words) (Seifart, 2015). In Czech, we find instances that mimic these two pathways. For example, the Czech noun *dysfunkční* ('dysfunctional') is composed of a Latin origin derivational morpheme *dys-* and a root morpheme *funk(c)*, and *houslista* ('violinist') is composed of a native stem *housl* and a borrowed affix *-ist*. However, it remains unclear which component of a word—its root, the derivational affix, or the inflectional affix—serves as the primary indicator of its status as a loanword. On the other hand, heavily borrowed concepts, especially in the domains of science and technology, may give the impression that foreign roots are the primary markers of loanwords. The borrowed affixoids can sometimes function as roots themselves—for example, *supr-* in *suprový* ('super')—highlighting that morpheme lexicality exists on a continuum rather than as a binary distinction.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 109–118, Fribourg, Switzerland, 4-5 September 2025.

Therefore, a nuanced approach to extracting borrowing signals should involve analyzing the loan status of individual morphemes rather than entire words, focusing on various types of morphemes or affixes. However, to our knowledge, no linguistic resource contains such an annotation, and there is no established method to determine which of the morphemes act as markers of loanwords. This could also provide us with valuable information on adaptation strategies for borrowed morphemes. In our experiments, we use the task of labeling loanwords to extract *borrowedness* scores for individual morphemes that tell us to what extent the morphemes encompass the loan status.

We train a multinomial naive Bayes classifier and extract the feature probabilities of the morphemes. The features are weighted on likelihoods, normalizing the impact of morpheme frequencies, which provides a ranking of morphemes or the borrowedness score. We compare the results using the attention mechanism (Vaswani et al., 2017), using an LSTM-based binary classifier. The intuition is that the attention weights would correspond to the morphemes that define the class labels. Additionally, we rope in a pre-trained RobeCzech (Straka et al., 2021) model for visualising the attentions on the subword level to present a comparative perspective. Training data is curated from DeriNet (Olbrich et al., 2025). All our training data and codes are made publicly available at github.com/abishekjs/MorphAttention.

2 Related work

Considerable research has been conducted in the area of language contact to address the challenges of morphological integration (Poplack and Dion, 2012; Poplack, 2018). Given the fact that morphemes can be borrowed directly or indirectly, Coghill (2015) presents a detailed overview of the morphological integration of Arabic verbal lexemes into Neo-Aramaic dialects where, due to phonotactic constraints, the verbal marker is also often borrowed along with the verbal lexeme. Although we do not deal with phonotactics in our study, it is worth mentioning these potential underlying reasons for borrowing decisions. A more recent study by Camacho (2024) for new coinages or borrowings in Quechua shows that borrowing words that are phonotactically similar to Quechua are morphologically easier to integrate. Introducing a different approach, Boano et al. (2024) show that the probabilistic distribution of different verb conjugation classes is influenced by the etymological origin of the borrowed verbal stems concerning Latin, providing a pivot to identify foreign affixes.

We also perform the task of morpheme classification, which has been frequently used for morpheme segmentation. Ruokolainen et al. (2014) implement a sequential classifying and labeling letters of a given word using conditional random fields (CRFs). John and Žabokrtskỳ (2023) classify root morphs using a bidirectional LSTM-CRF tagger for Czech. With respect to extending word-level labels to the morpheme level, Stephen et al. (2024) perform an unsupervised alignment of morphological categories and morphemes. The results show that the characteristics of the word level stem from the morpheme level.

3 Data

For our experiments, we use DeriNet v2.3 (Olbrich et al., 2025). DeriNet (ufal.mff.cuni.cz/derinet) is a lexical network that models the relations between words in the Czech lexicon. DeriNet contains tags for native words and loanwords that were extracted in a supervised manner from multiple corpora based on language-specific rewrite rules. For example, the suffixes like *-ace*, *-bus*, *-fob*, etc. are hand-coded as foreign suffixes and the words containing them are labeled as loanwords. Most of the words labeled as loanwords are recent borrowings into Czech. Morphological segmentations are also provided in the dataset along with etymologies for a few thousand entries. We use the morph classifier by John (2024) to classify the morphemes into three categories: roots, derivational morphemes, and inflectional morphemes (see Table 1). The architecture of the classifier consists of three parts: character embedding using convolutional layers of varied sizes, bidirectional LSTM, and a final CRF.

4 Methodology

Our methodology consists of two different experiments, namely using the naive Bayes classifier and the attention-based LSTM classifier. We also discuss the results along with the experiments. For a

POS	Borrowed	Native	Roots	Derivational affixes	Inflectional affixes
Noun	100079	197132	39850	4460	56
Verb	13378	42930	6714	719	2
Adjective	85320	199802	26927	3150	14
Adverb	45874	109465	20454	2318	21
Total	244651	549329	46246	5533	66

Table 1: Data overview with the frequency counts of native and borrowed words, along with the counts of affixes across those words conditioned by the POS categories.

comparative account on the subword level, we visualize the attentions extracted from the RobeCzech based binary classifier.

4.1 Experiment 1: Naive Bayes classifier

In the first experiment, we used the multinomial naive Bayes classifier from Scikit-Learn (Pedregosa et al., 2011). All morphemes in a native word are labeled as native, all morphemes in a borrowed word are labeled as borrowed, and as features, we take the unique ID of the morpheme and its role in a given word (root vs. derivational vs. inflectional). Thus, we take into account that the same morpheme might be derivational on some occasions and root on others (e.g., *před-* in *předpokládaný* 'assumed' vs. in *přednosta* 'principal'). We extract predicted probabilities or the *borrowedness* score, which measures how strongly a given morpheme, in a specific role (e.g., *-ova* as a derivational suffix), is a marker of the borrowed status. For example, borrowed roots are expected to appear mostly or exclusively in words marked as borrowed in DeriNet. As a result, their borrowedness score should be close to 1, since they are the only borrowed elements in the word. In contrast, inflectional affixes tend to attach to both native and borrowed stems without strong preference. Therefore, their borrowedness score should be roughly proportional to the overall ratio of borrowed to native words in the dataset. If, for instance, 30% of the words are borrowed, we expect the inflectional affixes to have a borrowedness score around 0.3. We set a frequency threshold of ≥ 20 so that the classifier relies on only such morphemes that appear frequently enough to provide robust predictions.

The mean (Table 2) of the feature likelihoods or predicted probabilities of the classifier correspond to the distribution of borrowed and native words in the data. Strangely, the derivational affixes show a higher score than roots and inflectional suffixes, which might be because of affixoids, which appear quite frequently in the data. The variances (Table 2) are interesting, showing that the derivational affixes are almost as salient as the roots with regard to the origin of the words.

Statistics	Roots	Derivational affixes	Inflectional affixes
Mean	0.38	0.44	0.38
Variance	0.18	0.15	0.08

Table 2: The mean and variance of the feature likelihoods of the data.

Furthermore, there is an interesting point in the frequencies of borrowed and native morphemes in the lexicon: the more a given morpheme predicts that a word is borrowed, the less frequent it is. Even though the linear correlation coefficient is very small (–0.10 for roots, –0.07 for derivational affixes, and –0.2 for inflectional affixes), the t-test for distance correlation, which also captures non-linear dependencies, shows really strong association of borrowedness score and frequency, with a p-value of 0.0 for both derivational affixes and roots – while for inflectional affixes with a p-value of 0.17, the outcome is less

fuzzy.

We present the relative borrowedness scores for the derivational morphemes in Table 3. The scores indicate how much more or less predictive a morpheme is compared to the mean borrowedness score of derivational affixes across all the POS categories.

Derivational affixes	**Frequency**	**RB Scores**
fon	29	0.557
kilo	635	0.557
trans	718	0.557
ova	118329	0.003
sk	18101	0.005
iv	7515	0.007
haz	50	-0.440
řík	32	-0.440
přah	27	-0.440

Table 3: The relative borrowedness (RB) scores for the top, middle and the least ranking derivational affixes.

It should be noted, however, that quite often, roots got misidentified as derivational affixes, and this method brought these errors to light. Generally, the more lexical a morpheme is, the more salient it tends to be for borrowedness identification. This occurs for two reasons: firstly, because it might cause the word to be classified as borrowed; secondly, because borrowed derivational affixes, even if reanalyzed, do not tend to be much productive. If they are productive, it is only in very specific contexts, and it is the more lexical ones (*-fikace*, *-ismus*). These affixes then cause the whole word to be more plausibly regarded as borrowed.

In Table 4, we present the derivational affixes with the highest RB scores. Affixes such as *bio* and *auto* have a higher RB score in nouns, adjectives, and adverbs, while the affixes *para* and *de* mostly confer loanword status to verbs. The important observation is that most of these affixes are affixoids. Usually borrowed as part of neoclassical compounds like *antibiotika* ('antibiotic') or *automat* ('dispenser'), these affixes are productive enough to participate in word-formation processes, while also saliently preserving their foreign origin.

POS	**Derivational affixes**											
Noun	bio	ion	ex	ment	ism	multi	auto	anti	kilo	ing	inter	kom
Verb	kom	ment	ion	de	ex	di	iz	is	ur	para	isova	syn
Adjective	ion	ex	ment	multi	bio	inter	kom	kilo	trans	auto	ent	mikro
Adverb	ion	ment	ex	multi	kom	di	para	bio	anti	inter	trans	auto

Table 4: The derivational affixes with highest (left to right) relative borrowedness (RB) score segregated by POS tags.

The classifier struggles with compounds, noisy classification, and segmentation, and lacks contextual knowledge about the given morpheme. Additionally, it does not address homomorphy. To overcome these issues, we incorporate the attention mechanism into the classifiers.

4.2 Experiment 2: Neural classifiers

The motivation to use neural classifiers is to exploit the attention mechanism. In case of a binary classification task as we have here, the attention mechanism dictates how much importance input tokens get to predict the target labels.

4.2.1 Attention-based LSTM architecture

We want a mechanism that can distribute the borrowedness labels to individual morphemes. To do that, we train a lightweight classifier using a custom architecture. Firstly, we embed individual morphemes using bidirectional LSTM cell; consequently, we pass these embeddings to four attention heads. Then, we sum up the attention-weighted embeddings and pass the resulting vector to one dense layer of size 512 and a sigmoid classification layer. In such a setting, attention scores should correspond to the saliency of the morphemes in context. We train the classifier for 8 epochs in batches of size 256 on a subset of training data so that we do not have too many negative examples (we randomly select at most twice as many negative examples as there are positive examples). As the architecture of our classifier is designed mainly to correctly distribute the attentions, we do not expect we will achieve very good results; however, reasonably good results would confirm that the attention does correspond to borrowedness. For comparison, we used naive Bayes and LinearSVM classifiers from Scikit-learn, using character trigrams or morphemes.

The results are presented in Table 5. For attention extraction, we used the mean values of the attention heads.

Method	**ADJ**			**ADV**			**NOUN**			**VERB**		
	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.
C-LSVC	96.5	95.6	97.7	96.1	94.3	97.2	95.6	93.6	96.3	97.5	97.0	98.7
C-NB	87.0	92.6	93.8	86.0	92.8	93.4	87.3	92.2	92.8	86.4	94.6	95.1
M-LSVC	95.3	93.8	96.9	95.5	89.6	95.7	92.9	93.4	95.3	98.6	95.9	98.7
M-NB	92.3	93.8	95.9	90.6	91.8	94.8	91.6	90.6	94.0	95.9	93.6	97.5
Neural	93.3	89.8	87.0	92.0	90.0	82.0	92.1	88.4	88.5	88.5	88.0	60.5

Table 5: Results for the baselines (NB = Naive Bayes; LSVC = linear SVC; M- = morpheme-based version; C- = character-based version) and the neural classifier (in %).

As for the results (Table 6), the attention really seems to pick at least one of the borrowed morphemes in borrowed words and generally tends to pick roots or derivational affixes (surprisingly, in verbs, it is more often derivational affixes). In Figure 1, we observe that the attention scores for the adverb *dynamicky* ('dynamic') are strongly attached to the root morpheme *dynam-* and the derivational morpheme *-ic*. In case of the adjective *komplexní* ('complex'), the attention picks the root, while in the noun *motocyklista* ('motorcyclist'), attention is high for the derivational morphemes *moto-* and *-ist*. The verb *telefonovat* ('to call') attracts attention on the morpheme *fon*.

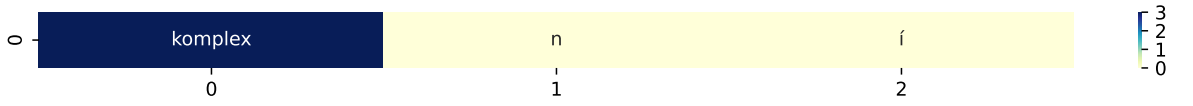
4.2.2 Finetuning RobeCzech

To test how the subword embeddings can be leveraged for the classification task, we fine-tune RobeCzech (Straka et al., 2021). The underlying architecture is inspired by the RoBERTa (Liu et al., 2019) model, which is trained in a monolingual setup for Czech data. We use RobeCzech as the architecture backbone and build a binary classification model. The embedding for each subword is accessed from the last hidden state of the `[CLS]` token. We apply a dropout layer followed by a single linear layer that acts as the classification head. The attention weights extracted from the `[CLS]` indicate how much each part of the subword contributes to the target classification of native and borrowed tags. This provides insight into which subwords, rather than roots or derivational morphemes, the model considers most

Classifier	R	D	I
Adjective	159487	119590	6045
Adverb	135210	20101	28
Noun	245616	50622	973
Verb	21984	34222	102

Table 6: The number of words for all the POS categories in which the roots (R), derivational affixes (D), and inflectional affixes (I) get the highest custom classifier attention scores.

(a) dynamicky



(b) komplexní

(c) motocyklista

(d) telefonovat

Figure 1: The attention weights across morphemes extracted using our custom classifier.

relevant for the classification. This can reveal patterns such as whether certain subwords consistently receive more attention in loanwords, offering a form of interpretability grounded in the model's internal decision-making process. Although attention does not provide a complete explanation, it offers a useful and intuitive lens into what the model *focused on* when making its predictions. Furthermore, we present the word embedding (Figure 2) and subword embedding (Figure 3) space extracted from RobeCzech.

Model	**Binary accuracy**	**Precision**	**Recall**
Noun	98.0	96.5	97.6
Verb	98.3	95.0	98.0
Adjective	98.2	95.2	98.8
Adverb	97.7	94.6	97.8

Table 7: Results for the finetuned RobeCzech classifiers on the test sets after training for epochs = 6, dropout = 0.1 and batch size = 32.

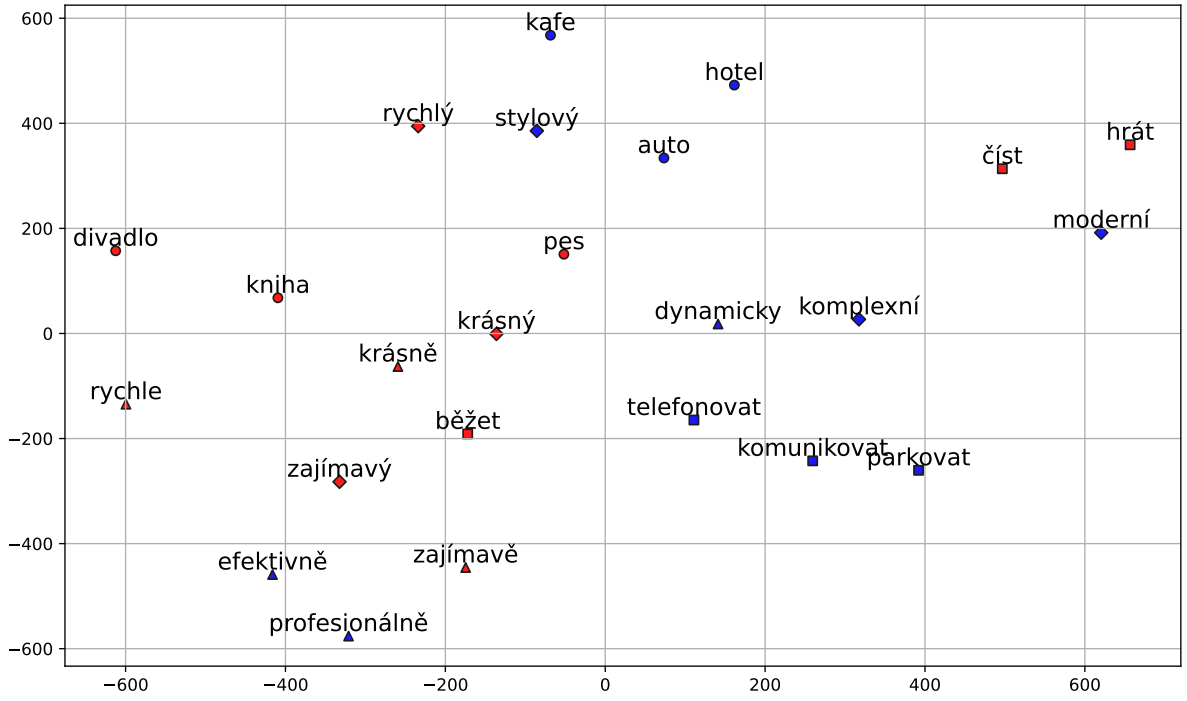


Figure 2: The word embeddings of words from our training set extracted from RobeCzech. The native words and loanwords are projected for nouns ● verbs ■ adjectives ◆ and adverbs ▲. The native and loanwords are well distributed. This t-SNE plot is created using Scikit-learn with perplexity of 10 and 10000 iterations.

The attention maps (Figure 4) show that the attention spans more than one subword to perform the binary class predictions. This suggests that a pre-trained language model like RobeCzech leverages the broader lexical context of the subwords to inform its predictions.

5 Conclusion

We used DeriNet to train a multinomial naive Bayes classifier and an attention-based binary LSTM classifier to classify Czech words as native or borrowed. We extracted feature probabilities and attention scores from the classifiers for individual morphemes. Our results indicated that the derivational morphemes serve as reliable markers of the borrowed status of words, along with root morphemes in Czech. The

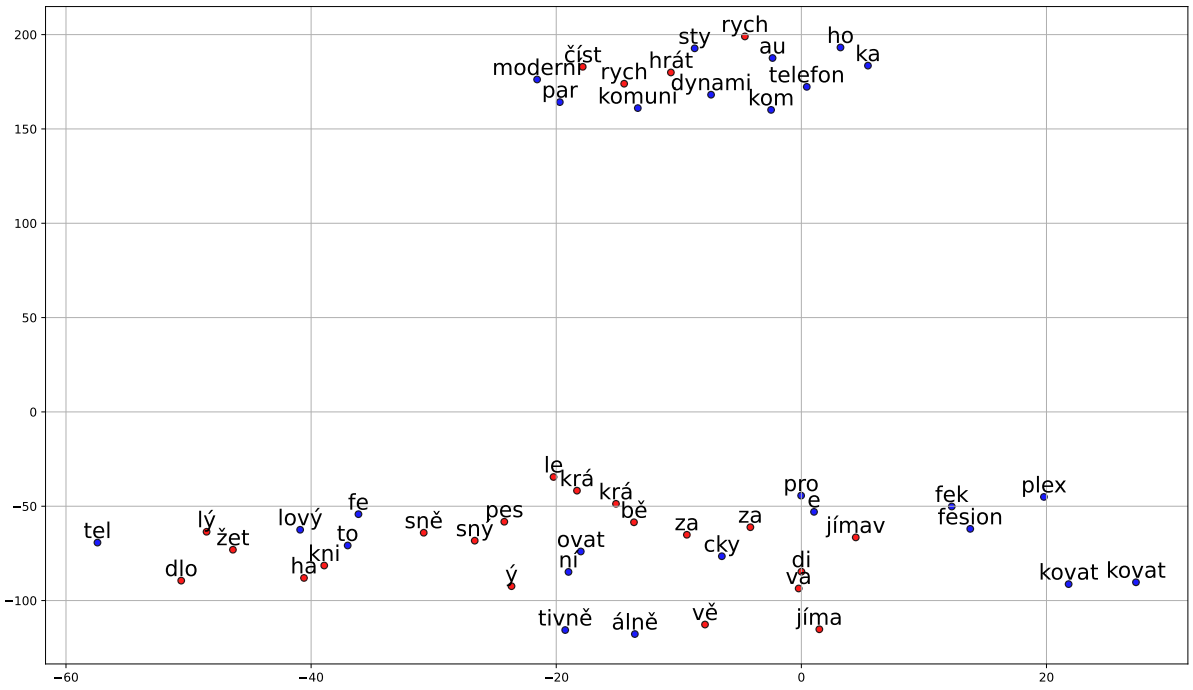
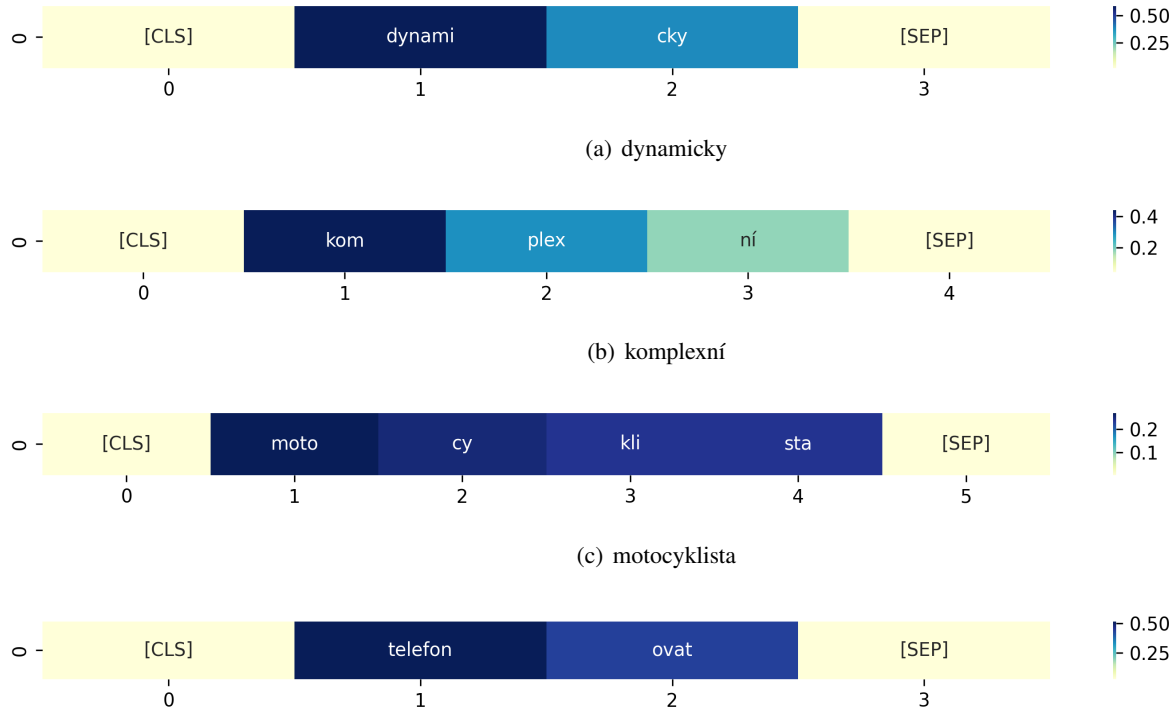


Figure 3: The subword embeddings of words plotted in Figure 2 extracted from RobeCzech. The native words and borrowed words are projected for nouns ● verbs ■ adjectives ◆ and adverbs ▲. Here, we find a separation between the lexical and grammatical morphemes, which indicates that the model differentiates contextually between these two classes. This t-SNE plot is created using Scikit-learn with perplexity of 10 and 10000 iterations.



(a) dynamicky

(b) komplexní

(c) motocyklista

(d) telefonovat

Figure 4: The attention weights across the subword tokens extracted from the RobeCzech based binary classifier.

comparison of results was further analyzed on the basis of the part-of-speech (POS) categories of the words. We also provided a glimpse of attention visualization based on the attention weights of RobeCzech based binary classifier.

6 Limitations

The morphological segmentations were not checked for inconsistencies, as it could lead to deviating from the underlying annotation guidelines of DeriNet v2.3. We did not account for any language internal reanalysis of borrowed stems or the extent of nativization of borrowed stems, due to the lack of additional annotations in the dataset. We were more interested in the interpretability of the classifiers, hence we did not focus on providing qualitative information regarding the etymology of the borrowed affixes or affixoids.

Acknowledgments

We thank two anonymous reviewers for their valuable feedback. This study is supported by the Charles University project GA UK No. 101924; Charles University Research Centre program No. 24/SSH/009 and partially supported by SVV project number 260 698.

References

Ondrej Blaha. 2022. Dynamics of conjugation pattern "kupuje" in contemporary Czech (on material of journalistic texts, 1990-2019). *Bohemica Olomucensia* 14(2):40–54. https://bohemica.upol.cz/pdfs/boh/2022/02/03.pdf.

Valeria Irene Boano, Marco Passarotti, and Riccardo Ginevra. 2024. Querying the lexicon der indogermanischen verben in the LiLa knowledge base: Two use cases. In Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, editors, *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia, pages 22–31. https://aclanthology.org/2024.ldl-1.3/.

Luis Camacho. 2024. Automating the proposition of neologisms for the Quechua language. *Journal of the International Phonetic Association* 54(3):922–938. https://doi.org/10.1017/S0025100324000227.

Eleanor Coghill. 2015. Borrowing of verbal derivational morphology between Semitic languages: The case of Arabic verb derivations in Neo-Aramaic. In Francesco Gardani, Peter Arkadiev, and Nino Amiridze, editors, *Borrowed Morphology*, De Gruyter Mouton, Berlin, München, Boston, pages 83–108. https://doi.org/doi:10.1515/9781614513209.83.

Francesco Gardani, Peter Arkadiev, and Nino Amiridze. 2015. Borrowed morphology: An overview. In Francesco Gardani, Peter Arkadiev, and Nino Amiridze, editors, *Borrowed Morphology*, De Gruyter Mouton, Berlin, München, Boston, pages 1–24. https://doi.org/doi:10.1515/9781614513209.1.

Vojtěch John. 2024. *Morph classifier*. Univerzita Karlova, Matematicko-fyzikální fakulta.

Vojtěch John and Zdeněk Žabokrtskỳ. 2023. The unbearable lightness of morph classification. In Kamil Ekštein, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue. 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings*. Springer, pages 105–115. https://doi.org/10.1007/978-3-031-40498-6_10.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692.

Michal Olbrich, Viktória Brezinová, Šárka Dohnalová, Vojtěch John, Lukáš Kyjánek, Aleš Papáček, Emil Svoboda, Magda Ševčíková, Jonáš Vidra, and Zdeněk Žabokrtský. 2025. DeriNet 2.3. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. https://ufal.mff.cuni.cz/derinet.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830. https://dl.acm.org/doi/pdf/10.5555/1953048.2078195.

Shana Poplack. 2018. *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford University Press. https://doi.org/10.1093/oso/9780190256388.001.0001.

Shana Poplack and Nathalie Dion. 2012. Myths and facts about loanword development. *Language Variation and Change* 24(3):279–315. https://doi.org/10.1017/S095439451200018X.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26:47–104. https://doi.org/10.1515/ling.1988.26.1.47.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In Shuly Wintner, Stefan Riezler, and Sharon Goldwater, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Association for Computational Linguistics, Gothenburg, Sweden, pages 84–89. https://doi.org/10.3115/v1/E14-4017.

Frank Seifart. 2015. Direct and indirect affix borrowing. *Language* 91(3):511–532. https://www.jstor.org/stable/24672164.

Abishek Stephen, Vojtěch John, and Zdeněk Žabokrtský. 2024. Unsupervised extraction of morphological categories for morphemes. In Elmar Nöth, Aleš Horák, and Petr Sojka, editors, *Text, Speech, and Dialogue: 27th International Conference, TSD 2024, Brno, Czech Republic, September 9–13, 2024, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, page 239–251. https://doi.org/10.1007/978-3-031-70563-2_19.

Abishek Stephen and Zdeněk Žabokrtský. 2023. Understanding borrowing through derivational morphology: A case study of Czech verbs. In Matea Filko and Krešimir Šojat, editors, *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology*. Croatian Language Technology Society, Zagreb, Croatia, pages 49–59. https://derimo.ffzg.unizg.hr/media/uploads/papers/derimo2023-06.pdf.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. In Kamil Ekštein, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue. 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*. Springer International Publishing, pages 197–209. https://doi.org/10.1007/978-3-030-83527-9_17.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus, editors, *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, page 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349.

Jan Wohlgemuth. 2009. Backmatter. In *A Typology of Verbal Borrowings*, De Gruyter Mouton, Berlin, New York, pages 303–459. https://doi.org/doi:10.1515/9783110219340.bm.

Offset vectors and affix meaning in English nominalizations

Martin Schäfer
Universität Leipzig
martin.schaefer@uni-leipzig.de

Abstract

This paper uses distributional semantics to investigate the offset vectors of English *-ity* and *-ness* derivatives. Overall, the two sets of offset vectors are clearly distinct, but not in any categorical sense. To further explore the offset vectors and possible sub-patterns within them, 100 bases of four different classes are combined with different average vectors in an exploratory analogy task. This shows differentiated effects in line with the idea that morphological structure in the bases encodes similar meaning. The results in the analogy task also again confirm a clear difference between *-ity* and *-ness*.

1 Introduction

In distributional semantics, one very simple approach to represent affix meaning is to use offset vectors, that is, vectors arrived at by simply subtracting the base vector from the complex form vector in words that are either inflectionally or derivationally related. The starting points for this study are two recent distributionally semantics-backed findings on English inflectional and derivational morphology: Shafaei-Bajestan et al. (2024, 381) conclude for English plural inflection that "the semantics of shift vectors is changing in close association with the semantics of the singular and plural words." Schäfer (2025), on the English *-ity*/*-ness* affix rivalry, reports that the distributional vectors of adjectival bases successfully predict the affix choice. The aim of this paper is to explore to what extent the formers' inflection-based offset vector findings also obtain for the derivational data of the latter. More specifically, this study is interested in (1) whether the offset vectors of *-ity* base-derivative pairs are distinct from the *-ness* pairs and (2) whether there are further patterns associated with specific subsets of bases within the offset vectors.

2 Background

Within distributional semantics, there has been a focus on word formation, either via derivation or via compounding (for an overview of work on derivation, see Boleda, 2020). Inflection itself has not been focused on so much, even though Mikolov et al. (2013)—the study introducing the word2vec algorithm—used a test set that (among others) contains five different inflectional relations across the whole spectrum available in English (positive forms of adjectives compared to the corresponding comparative and superlative forms, plain forms of verbs compared to their *-ing*, past tense, and third person singular forms, and nouns in the singular to their plural forms). Their study also pioneers the idea of using offset vectors in the implementation of an analogy task, starting with the question in (1a), illustrated with the example in (1b).

(1) a. Which word `d` is similar to word `c` in the same sense as word `b` is to word `a`?
b. Which word is similar to *tough* in the same sense as *luckier* is to *lucky*?

To answer this question via word embeddings, they propose two simple steps:

1. Calculate a synthetic vector by subtracting the vector for `a` from the vector for `b` and adding the vector for `c`, e.g., $\text{vector}_{\text{luckier}}$ - $\text{vector}_{\text{lucky}}$ + $\text{vector}_{\text{tough}}$.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 119–127, Fribourg, Switzerland, 4-5 September 2025.

2. The word vector that is most similar to the synthetic vector, that is, the vector that is its nearest neighbor in terms of cosine similarity, is the answer to the question.

While Mikolov et al. (2013) were not interested in morphology proper, Bonami and Paperno (2018) provide a very careful operationalization of the differences between inflection and derivation in terms of quantitative aspects accessible through offset vectors, focusing not on the absolute difference in terms of mean cosine similarities across a given pairing of forms but on the difference in variance for such a pairing.

Guzmán Naranjo and Bonami (2023) use offset vectors to assess rivalry in word formation in French by looking at 35 derivational processes (including conversion) across 21,990 base-derivative pairs. In their first experiment, they used a cosine distance matrix created from the pairwise similarities of each average offset vector with the others and clustered the average vectors to explore whether the resulting tree is similar to 7 trees produced by human expert judges intended to capture semantic similarities and differences between the processes. The resulting comparison reveals that the vector-based tree is qualitatively and quantitatively very comparable to the human-produced trees. In their second experiment, they train classifiers on their data in order to see whether the classifier can correctly predict to which process a given offset vector belongs. This is done for each subset of base-derivative part of speech configurations. For example, their dataset contains four types of adjective to noun derivations (among them *-ité*, the original source of English *-ity* derivatives). For this subclass, their classifier is at chance level, that is, the classifier is not able to successfully distinguish the different offset vectors stemming from pairs across these four processes. For other processes, the classifiers are much better, and overall there is always a cline in the similarity between the offset vectors, across all processes.

Shafaei-Bajestan et al. (2024) use offset vectors (their shift vectors) to explore pluralization in English. They find that these show a non-random structure, and cluster with regard to the semantic classes of the nouns. This is important because it is in clear contrast to what one would expect on the traditional view that there is one plural affix in English and that this inflectional affix simply always encodes the same functional meaning, which would lead one to expect that the offset vectors across singular-plural forms show ideally no or very little variation and are not linked to the semantics of the nouns. In addition, they show that semantic class-based average offset vectors clearly outperform the simple average offset vector in predicting the correct plural form, that is, using the analogy task introduced above, the nearest neighbor of the synthetic vector combining base and semantic-class based average plural vector more often has the actual plural form among its top semantic neighbors.

Schäfer (2025) uses distributional semantics to investigate the affix rivalry between *-ity* and *-ness*. Distinguishing between doublet-bases (i.e., bases that come with both *-ity* and *-ness* forms, such as *aggressive* → *aggressivity/aggressiveness*) and non-doublets, I show that for non-doublets the distributional vectors of the bases already predict which affix they take. Considering the question of whether the two affixes themselves are synonymous, I conclude that the results are "in line with the idea that they induce similar meaning shifts" (Schäfer, 2025, p. 37). The paper does not consider offset vectors, which allow one to address this question much more directly.

The first research question addressed in this study is whether the offset vectors across the *-ity/-ness* non-doublet bases are distinct from each other or not. The expectation is that they are distinct, given that Schäfer (2025) shows that their bases are distinct and that Shafaei-Bajestan et al. (2024) show that distinct patterns in the bases are associated with distinct patterns even in the offset vectors of inflectionally related forms. The second research question, inspired again by Shafaei-Bajestan et al. (2024), is whether one also finds other sub-regularities within the offset vectors of both affixes associated with the different classes of bases involved.

3 Study 1: *-ity/-ness* offset vectors

3.1 Methods

This study uses the vector-set provided by Schäfer (2024), a curated set of vectors for 1,475 adjectives and their *-ity* derivatives and 1,802 adjectives and their *-ness* derivatives. This set contains all pairs of adjectival base and their *-ity/-ness* derivatives from the tagged ukWaC corpus (Baroni et al., 2009) that are

also contained in the 1 million item fastText vectorsets published with Mikolov et al. (2017). The specific fastText vectors used, `wiki-news-300d-1M.vec.zip` available at https://fasttext.cc/docs/en/english-vectors.html, were trained on 16 billion tokens, using a corpus that itself was concatenated from 5 different web-sources and web-derived corpora (see Mikolov et al., 2017 for the details). This set of fastText vectors does not include subword information, an advantage for this study because subword information would automatically help to differentiate between our target derivatives via their different endings, *-ity* and *-ness*.

As this study is explicitly interested in possible differences in the offset vectors, the 130 doublets—that is, pairs that occur with both affixes (e.g. *dense* with *density*/*denseness*)—were excluded. This leaves a set of 3,014 base-derivative pairs, 1,343 base/*-ity* pairs and 1,671 base/*-ness* pairs.

Offset vectors are calculated by subtracting the base vector from the derived vector for each pair of forms in the dataset. For downstream analysis, t-SNE is used for visualization and Linear Discriminant Analysis (LDA) for statistical corroboration, following Shafaei-Bajestan et al. (2024) and Schäfer (2025) and making the results here thereby directly comparable to these studies.

The software implementations of both t-SNE as well as LDA I use come from Python's `scikit-learn` library (Pedregosa et al., 2011). All preparatory steps and all further analysis are reproducible via the scripts and data available at https://doi.org/10.6084/m9.figshare.29425184.v1.

3.2 Results

Figure 1 shows the t-SNE visualization of the 300-dimensional vectors of the offset vectors on a two-dimensional plane. Blue circles represent the projections of the vectors of the bases of *-ness* derivatives, red crosses represent the vectors of bases of *-ity* derivatives.

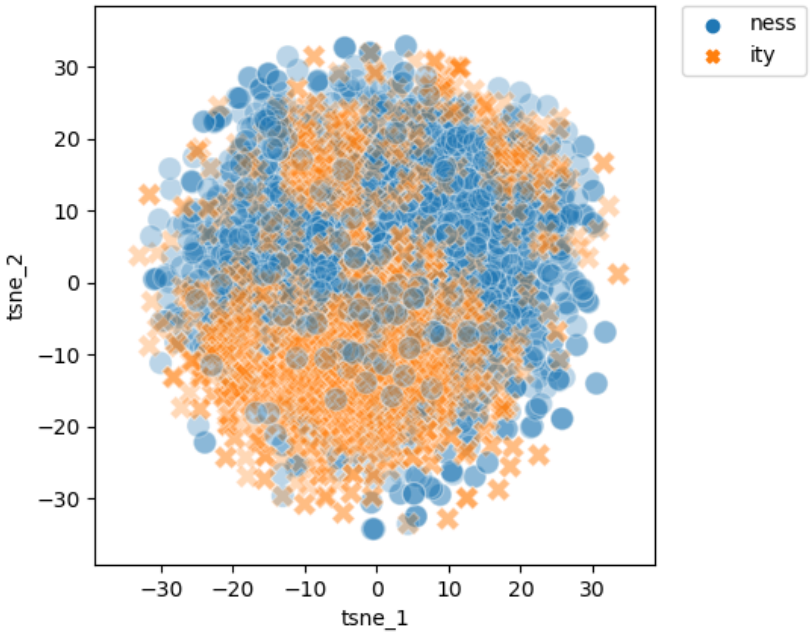


Figure 1: Projection of the offset vectors for all base-derived form pairs into two dimensional space using the t-SNE dimension reduction technique.

We see that the offset vectors form clusters, with *-ity* offset vectors concentrating in the lower half, with two additional clusters in the upper middle and to the upper right of the top half. The *-ness* offset vectors are concentrated in the upper half. While the dominant clustering at the top and in the lower half shows that the *-ity* and *-ness* offset vectors are clearly distinct, the considerable amount of overlap also shows that there is by no means a categorical distinction between the two sets of offset vectors. The clear

clustering is supported by an LDA classifier, trained to predict the vectors as either *-ity* or *-ness* offset vectors: the average weighted F1 score is 0.838 (0.019 std). If we compare this to a baseline classifier that simply assigns the most frequent category to everything, the resulting weighted F1 is 0.395 (with, for the most frequent category, *-ness* derivatives, precision at 0.554 and recall at 1, and for the other category both set to zero). The simple accuracy score for the classifier is 0.839, with the baseline here being the same as the precision of the most frequent category, 0.554 (= 1,671/(1,343 + 1,671)). In other words, on both metrics, the classifier performs well in classifying the vectors into *-ity* and *-ness* offset vectors, while the baseline classifier performs poorly.

3.3 Discussion

As expected, and in contrast to the results for the offset vectors of French deadjectival derivations in Guzmán Naranjo and Bonami (2023), the offset vectors of *-ity* and *-ness* are clearly different. Interestingly, the F1 score for the offset vectors is on the same level as the same evaluation scores reported for the bases in Schäfer (2025)—0.849—with both scores slightly higher than the one reported for the derivatives (0.744). Importantly, and similar to the plural-offset vectors investigated by Shafaei-Bajestan et al. (2024), this difference is not categorical, and the offset vectors associated with each affix show considerable variation. Shafaei-Bajestan et al. (2024) state for the clusters in their vectors that are associated with specific semantic classes that "[t]hese examples [vehicles vs. fruits; cognition vs. state nouns] suggest informally that for concrete nouns, their affordances play a crucial role, while for abstract nouns, different construals influence their plural usages". It is not clear whether a similar point can be made for the difference encoded by *-ity* and *-ness* offset vectors. One aspect that has been prominently discussed for *-ity* and *-ness*, but not so much for plural formation in English, is a possible link to genre or text type, which would also be encoded in the corresponding vectors. However, the current dataset does not allow one to explore this further.

Another aspect that was not investigated in detail here is possible frequency effects. This is difficult because the absolute frequencies for both sets of bases and derivatives are markedly different. The simplest solution, zooming in on the subset that has a frequency ratio around 1 across forms (here: between .5 and 1.5), yields only a small subset of 249 pairs, with almost 4/5 *-ity* pairs. Consequently, the baseline classifier is relatively high—0.671—almost reaching the average F1 score of the LDA for this subset (0.685, 0.093 std), which is, in turn, markedly lower than the classifier for the whole set reported above.

4 Study 2: inside the *-ity*/*-ness* offset vectors

Since both sets of offset vectors display considerable variation, an obvious question is whether this variation is patterned in non-random ways even within the form pairs. The success of cross-class average vectors in Shafaei-Bajestan et al. (2024) invites the question of whether more focused average vectors emerge if they are restricted to specific semantic classes of the adjectival bases. Schäfer (2025) argues that there are no good ways of semantically classifying the adjectival bases involved but partially explores an idea by Riddle (1985): morphological patterns in the bases possibly encode semantic similarity if one affix always adds a consistent semantic core of shared meaning. This idea can also be explored in offset vectors, hypothesizing that average offset vectors of a semantically (i.e., morphological) similar set of adjectives have more in common with each other than those that don't.

4.1 Methods

Similar again to Shafaei-Bajestan et al. (2024), the analogy task of Mikolov et al. (2013) will be used here. Using the same dataset as in Study 1, the performances of the five average vectors in (2) are explored.

(2) a. **all**: average offset vector across all pairs
b. **ity**: average offset vector across all *-ity* pairs
c. **ness**: average offset vector across all *-ness* pairs
d. **ble**: average offset vector across all 547 *-ble* bases that take only *-ity*

e. **ed**: average offset vector across all 173 *-ed* bases

While the first three explore again the differentiation between *-ity* and *-ness*, the last two implement the idea that morphological subclasses of bases can be used as substitutes for meaning-wise similar adjective bases. The two biggest meaningful morphological subgroups for each affix reported in Schäfer (2025) are chosen, with *-ble* bases for *-ity* and *-ed* bases for *-ness*. For the purposes of this study, the ten *-ble* bases that take *-ness* (e.g., *nimbleness*) were excluded. All *-ed* bases occur only with *-ness* in this dataset.

The performance is tested with the analogy task: when adding the average vector to the base vector, is the target vector (that is, the actual *-ness* or *-ity* form associated with the base vector) contained in the nearest neighbors of the synthetic vector? Four groups of bases are used as test sets:

(a) **other *-ity***: 25 bases with no discernable morphological pattern that have only *-ity* derivatives (*sublime, secure*).

(b) **other *-ness***: 25 bases of the same type that have only *-ness* derivatives (*harsh, smart*).

(c) ***-ble* [*-ity*-only]**: 25 *-ble* bases that only have *-ity* derivatives (*lovable, notable*).

(d) ***-ed* [*-ness*-only]**: 25 *-ed* bases that only have *-ness* derivatives (*directed, guarded*)

These four sets were initially randomly selected. This resulted in many low-quality pairs in the first group, exposing problems in the tagging of the underlying dataset used, with pairings like *la/laity* or *discontinue/discontinuity*. Eleven pairs were manually exchanged against valid adjective-derivative pairs. The other groups were not affected much, but three pairs in the group (b) **other *-ness*** and one in group (d) ***-ed* [*-ness* only]** were also replaced manually. The four sets of 25 adjectives are reproduced in the Appendix.

For each adjective, five composed vectors are created by simply adding the five average vectors—**all**, **ity**, **ness**, **ble**, or **ed**—to the vector of each adjective. To determine the nearest neighbors of the composed vectors, only the subset of the 200,000 most frequent vectors in the vectorset was used. This decision was made to keep the task computationally feasible. Since some of the base vectors and many of the target vectors are not contained in this subset, these were manually added to the smaller subset. That is, the cosine similarities of all composed vectors to the most frequent 200,000 other vectors plus all base and derivative vectors not in this set were calculated. These similarities are then ranked in order of closeness to the probes, the five composed vectors for each base, and the position of the derivative in these nearest neighbors for each probe is determined. For example, each vector is added to the vector of the only *-ness* base *smooth* to yield five composed vectors, the five probes. The target word in this case is *smoothness*.

All vector manipulations and the calculations of cosine similarities were done with Python, while R (v4.4.1; R Core Team 2024) was used for the further exploration of the resulting sets of similarities. All preparatory steps and all further analysis are reproducible via the scripts and data available at https://doi.org/10.6084/m9.figshare.29425184.v1.

4.2 Results

The results across the test set are shown in Tables 1–4, which each subset in its own table. No first rank was achieved, with the adjectival base occupying that position in all cases. The variation of interest takes place mostly across ranks four to five, with no higher rank scoring more than two, and the most frequent rank always ranks two.

Within the **all**, **ity**, and **ness** vectors, we see that in every subgroup of probes, **ness** performs best, followed by **all**, followed by **ity**. This ordering is very clear in the **other ity** and **other ness** probes, the differences becoming smaller in the **-ed** probes and even smaller in the **ble** probes. The composed vectors using the average vectors drawn from specific morphological subclasses of bases—the **ble** and the **ed** vectors—show a more differentiated behavior. The **ble** vectors perform best for the set of **-ble** bases, but even for these, they do not perform as well as the **ness** vectors. For the three other groups, they always perform worse than the **ness**, **all**, and **ed** vectors, but better than **ity** for the **other *ity*** group and for the two

remaining groups similar to **ity**. The **ed** vectors perform best for the **-ed [only *-ness*]** bases, third best for the **-ble [only *-ity*]** bases, and second best for the **other *-ity*** and **other *-ness*** bases.

If we consider the four test sets of adjectival bases, the test set ***ble* [*-ity* only]** very clearly contains the best performing probes, with 16 or even 17 second ranks for each type of composed vectors, and the worst ranks all around rank 50. The second best performing test set is **other *-ness***, where the best composed vectors, using the **ness** average offset vector, achieved 10 second ranks, with the other four achieving 8, 7, and 2 times 6 second ranks. The **other *-ity*** test set follows, and the ***-ed* [only *-ness*]** test set performs worst, with the **ble** vectors yielding 7 second ranks, and the **ity** and **ble** vectors each only achieving 4 second ranks.

Rank	all	ity	ness	ble	ed
Rank 2	6	4	9	5	8
Rank 3	5	3	4	3	5
Rank 4	4	4	4	6	2
Rank 5	2	3	0	1	1
Rank 6-10	3	3	3	3	3
Rank 11-50	3	6	2	5	3
Rank >50	2	2	3	2	3

Table 1: Test set (a) **other *-ity***.

Rank	all	ity	ness	ble	ed
Rank 2	7	6	10	6	8
Rank 3	4	2	1	2	3
Rank 4	1	2	2	3	2
Rank 5	3	3	2	2	2
Rank 6-10	3	2	4	3	4
Rank 11-50	2	4	1	3	1
Rank >50	5	6	5	6	5

Table 2: Test set (b) **other *-ness***.

Rank	all	ity	ness	ble	ed
Rank 2	16	16	17	16	16
Rank 3	1	1	1	3	2
Rank 4	2	0	3	0	1
Rank 5	1	1	1	1	3
Rank 6-10	4	4	2	4	2
Rank 11-50	0	2	1	1	0
Rank >50	1	1	0	0	1

Table 3: Test set (c) ***-ble* [*-ity* only]**.

Rank	all	ity	ness	ble	ed
Rank 2	5	4	6	4	7
Rank 3	3	1	4	2	3
Rank 4	2	3	1	2	2
Rank 5	2	2	3	2	2
Rank 6-10	2	3	2	4	2
Rank 11-50	5	5	3	5	4
Rank >50	6	7	6	6	5

Table 4: Test set (d) ***-ed* [-ness only]**.

4.3 Analogy: discussion

Even on a relatively small test set of 25 bases each, the analogy task has been successful in showing differentiated effects for base-structure specific average vectors as well as for the *-ity* and *-ness* average offset vectors. We can observe clear differences between the different types of composed vectors, as well as between the four different test sets. For the four different test sets, the greatest differences emerge between the two groups that were based on a shared morphological structure of the bases: the ***ble* [only *ity*]** test set yields the best results, the ***-ed* [only *-ness*]** test set performs worst. This supports the idea that the considerable variation that emerged in Study 1 is at least partially linked to systematic differences between different types of bases. It also supports the general idea that shared base morphology can be used as a stand-in for shared base semantics, and, in turn, that these semantic differences are tied

to differentiated effects of the affixes on the bases. From the vantage point of the different composed vectors, a similar result emerges: the average vectors based on all *-ble* bases combining only with *-ity* and the *-ed* bases which only occur with *-ness* also show clearly different effects, and that they are optimized for their respective bases emerges from them performing best on their respective bases relative to their performance on the other test sets, and also, for the **ed** vectors, performing best in absolute terms, and with the **ble** performing second best.

That there are clear differences between *-ity* and *-ness* also emerges from the results, most clearly when looking at the performance of the composed vectors across all sets: **ness** vectors always perform best, **ity** vectors always performing worst. The overall better performance of the *-ness* related average vectors can perhaps be linked to its greater productivity and its less distinct lexicalization effects (Bauer et al., 2013). That the ***-ble* [*-ity* only]** test set results in the best rankings overall, and very similar results for all composed vectors, points to this group perhaps constituting the prototypical bases involved in this process, in line with the fact that they also form the largest distinct subgroup of bases in the set of all bases. Another interesting observation with regard to this test set is the relatively good ranking achieved by even the worst-performing probes. All three other test sets contain several outliers, which plausibly are linked to lexicalization effects. Relevant examples are the lowest-ranked examples in the three groups (always across all 5 probes): *minority*, *otherness*, and *signedness*.

In comparison to the results reported in Shafaei-Bajestan et al. (2024) for the plural offset vectors, the average vectors here perform poorly. In their data, the semantic-class specific average vectors produced 86% of top 3 hits, and the bases themselves reached 74% of top 3 hits. In my experiments, the bases always reach rank 1, that is, the base vector is never shifted sufficiently far away through the addition of the average vectors. In terms of the percentages for targets in the top three across the four test sets, test set (a) results range from 28% to 52%, test set (b) results range from 32% to 44%, test set (c) reaches 68% to 76%, and test set (d) ranges from 20% to 40%. A plausible explanation for this huge gap is the less stable nature of derivational vs. inflectional relationships (cf. Bonami and Paperno, 2018).

5 Conclusion

Two research questions were addressed: (1) Are the offset vectors across the *-ity/-ness* non-doublet bases distinct from each other or not? (2) Are there sub-regularities within the offset vectors of both affixes?

For the first question, offset vectors across a large set of *-ity* and *-ness* base-derivative pairs were calculated. The t-SNE base visualization of the resulting vectors showed clear but non-categorical differences, in line with the expectations. For the second question, a small test set of bases in combination with five average vectors was explored with the analogy task. Apart from the difference between the **ity** and **ness** vectors themselves, using morphological properties of the bases instead of true semantic classification turned out to show sub-regularities, with not only the ***-ble* [only *-ity*]** test set maximally different from the ***-ed* [*-ness* only]** test set, but also the corresponding average vectors always performing best in the corresponding test sets.

Many things could not be explored here, ranging from a closer look at the average vectors in terms of their distributional properties to the consideration of completely alternative approaches. In particular, the analogy task could also be approached by following Marelli and Baroni (2015), that is, using a regression analysis to create matrices representing an affix and creating composed vectors not through vector addition but through matrix multiplication.

Appendix

Testset (a) : 25 adjectives that only occur with *-ity* derivatives without any clear morphological pattern:
austere, biosecure, clandestine, complex, controverse, electron-dense, exterior, feminine, infirm, malign, mature, mediocre, minor, modern, obscure, sacrosanct, sane, sanguine, secure, semi-nude, severe, sincere, sovereign, sublime, uncivil

Testset (b) : 25 adjectives that only occur with *-ness* derivatives without any clear morphological pattern:
bleak, brave, close, cool, done, firm, front, gross, harsh, implicit, other, over-eager, shrill, smart, smooth, sober, sombre, sure, true, uncommon, unique, unsure, winsome, wise, wooden

Testset (c) : 25 *-ble* bases that only occur with *-ity*:
alienable, bistable, defeasible, differentiable, driveable, inapplicable, infallible, lovable, monitorable, navigable, non-sustainable, notable, patentable, potable, producible, retrievable, serializable, spreadable, testable, unattainable, unpalatable, unsatisfiable, unsociable, untenable, wettable

Testset (d) : 25 *-ed* bases that only occur with *-ness*:
adapted, boneheaded, closed, devoted, directed, disconnected, embedded, farsighted, guarded, ill-prepared, many-sided, open-minded, pointed, prepared, ragged, rooted, rugged, self-centered, signed, simple-minded, situated, surefooted, unbounded, unrelated, wretched

Acknowledgments

This paper profited from comments, advice, and feedback from the audience of the English linguistics colloquium at the University of Leipzig in the spring of 2024, as well as from comments and feedback from a talk at the 10th International Conference of the German Cognitive Linguistics Association at the University of Osnabrück in the fall of the same year. I also thank the two anonymous reviewers of the paper, who both gave very good, helpful, and friendly feedback and advice.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3):209–226. https://doi.org/10.1007/s10579-009-9081-4.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press, Oxford.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6:213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303.

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2):173–195. https://doi.org/10.1418/91864.

Matías Guzmán Naranjo and Olivier Bonami. 2023. A distributional assessment of rivalry in word formation. *Word Structure* 16(1):87–114. https://doi.org/10.3366/word.2023.0222.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122(3):485–515. https://doi.org/10.1037/a0039267.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. https://arxiv.org/abs/1712.09405.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830. https://dl.acm.org/doi/pdf/10.5555/1953048.2078195.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Elizabeth M. Riddle. 1985. A historical perspective on the productivity of the suffixes *-ness* and *-ity*. In Jacek Fisiak, editor, *Historical Semantics—Historical Word-Formation*, De Gruyter Mouton, pages 435–462. https://doi.org/10.1515/9783110850178.435.

Martin Schäfer. 2024. A distributional semantics analysis of the two English suffixes *-ity* and *-ness*. https://doi.org/10.6084/m9.figshare.23538207.v1.

Martin Schäfer. 2025. The role of meaning in the rivalry of *-ity* and *-ness*: Evidence from distributional semantics. *English Language and Linguistics* pages 1–46. https://doi.org/10.1017/S1360674324000443.

Elnaz Shafaei-Bajestan, Masoumeh Moradipour-Tari, Peter Uhrig, and R. Harald Baayen. 2024. The pluralization palette: Unveiling semantic clusters in English nominal pluralization through distributional semantics. *Morphology* 34(4):369–413. https://doi.org/10.1007/s11525-024-09428-9.

Assessing morphological productivity in a corpus language: A diachronic study of Ancient Greek deverbal nominal suffixes

Silvia Zampetta
University of Pavia
silvia.zampetta01@universitadipavia.it

Abstract

This paper explores the morphological productivity of six Ancient Greek deverbal nominal suffixes: *-eía*, *-mos*/*mós*, *-sia*, *-sis*, *-tis*, and *-tus*. While previous research has primarily examined these suffixes from a morphophonological and comparative Indo-European perspective, a quantitative diachronic analysis has yet to be conducted. By analyzing a corpus of approximately four million tokens spanning from the 8th century BC to the 6th century AD, this study applies different productivity measures to determine how these suffixes evolved across historical periods. It combines Baayen's (1989, 1992, 1993, 2009) corpus-based statistical methods with LNRE models of word frequency distribution (Baayen, 2001; Evert, 2004; Baroni and Evert, 2006; Evert and Baroni, 2006). Crucially, in these corpus-based statistical methods, the notion of morphological productivity is synchronic and linked to the number of *hapax legomena*, i.e., words with a frequency of one in a given corpus. The analysis reveals that *-sis* is the dominant suffix and has a central role in the deverbal nominal derivation processes. The suffix *-mos*/*-mós* is quite productive in all historical phases and shows a peak in the Hellenistic period. The suffixes *-eía* and *-sia* demonstrate a relatively low degree of productivity. In contrast, *-tis* and *-tus* should be considered unproductive.

1 Introduction

Ancient Greek (AG) deverbal nominal suffixes have been widely studied from an Indo-European perspective (e.g., Debrunner, 1917; Chantraine, 1933; Benveniste, 1948; Risch, 1974), primarily focusing on morphophonology and cross-linguistic comparison with other ancient Indo-European languages. More recent works include Napoli (2009) on *-mo- in a diachronic and typological framework, and Civilleri (2010) on synchronic nominalizations.

However, a comprehensive diachronic and quantitative investigation into their morphological productivity is still lacking. At the same time, most empirical research on productivity in derivational morphology has focused on modern languages, mainly due to the availability of large electronic corpora and computational tools. The earliest studies centered on English—where such resources first became available—and were subsequently extended to other languages, such as German (Evert and Lüdeling, 2001), Italian (Gaeta and Ricca, 2002, 2005, 2006; Varvara, 2019, 2020), and, most relevant to the present work, Old Italian (Štichauer, 2009), which introduced a diachronic dimension.

To date, no study has systematically and diachronically analyzed AG deverbal nominal suffixes using the quantitative tools developed in corpus-based morphological analysis. This paper aims to fill this gap by analyzing six AG deverbal suffixes, i.e., *-eía*, *-mos*/*-mós*, *-sia*, *-sis*, *-tis* and *-tus* (and their allomorphs) across a corpus[1] (~4 million tokens) spanning from the 8th century BC to the 6th century AD, divided into four chronological sub-corpora—Archaic, Classical, Hellenistic, and Imperial—that are comparable

[1]The corpus was constructed using the *Thesaurus Linguae Graecae* (TLG). This resource is inherently limited in the range of data that can be extracted from it. However, at the time of dataset construction, the TLG represented the best available option.

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 129–140, Fribourg, Switzerland, 4-5 September 2025.

in terms of token count and literary genres.[2] In addition to providing the first large-scale quantitative productivity measurement of the deverbal nominal derivation domain in AG, this study also aims to evaluate the extent to which modern productivity measures can be applied to an ancient language, and to identify the limitations and challenges that arise in such applications.

The extraction of lemmas containing the six suffixes under investigation was conducted using the Liddell-Scott-Jones lexicon from the Perseus Library (www.perseus.tufts.edu), followed by manual verification to ensure that the final list included only the deverbal nouns relevant to this study,[3] excluding other parts of speech, non-deverbal derived nouns, compounds, proper nouns, borrowings, baseless formations, and cases of suppletion. This process yielded a final dataset of 1,905 types and 50,637 tokens.

Using this dataset, a quantitative analysis was performed to identify patterns in the formation of deverbal nouns over time. The theoretical framework is the corpus-based quantitative approach to morphological productivity first proposed by Baayen (1989, 1992, 1993, 2001, 2009; Baayen and Lieber 1991; Baayen and Renouf 1996), which crucially links morphological productivity to the number of *hapax legomena*—i.e., words with a frequency of one—occurring in a sufficiently large corpus. For this purpose, three productivity measures were applied: (i) the P measure (Baayen, 1989) or potential productivity (Baayen, 2009); (ii) the P^* measure[4] (Baayen, 1993) or expanding productivity (Baayen, 2009); and (iii) the Large Number of Rare Events (LNRE) models of word frequency distribution (Baayen, 2001; Evert, 2004; Baroni and Evert, 2006; Evert and Baroni, 2006).

In addition to measuring individual suffix productivity, the study also examines potential interactions among the suffixes. By applying Kendall's Tau correlation, it investigates whether the resolution of rivalry between affixes can be inferred from their diachronic productivity trends.

The paper is structured as follows. Section 2 provides a general overview of productivity measures. Section 3 presents the quantitative analysis, detailing suffix frequency (3.1), Baayen's P measure (3.2), P^* measure (3.3), the LNRE models (3.4), and a correlation-based exploration between suffixes (3.5). Section 4 summarizes the key findings and outlines the main issues encountered in the use of these metrics in a corpus language.

2 Measures of productivity

The productivity of a given affix refers to its potential to form new words and the extent to which this potential is actually realized in language use (Plag, 2006, p. 553). This definition considers productivity both as a qualitative and a quantitative property—a distinction long discussed in the literature, starting with Danielle Corbin (Corbin, 1987; Carstairs-McCarthy, 1992; Bauer, 2001; Thornton, 2005), who proposes a differentiation between *availability* and *profitability*. The former refers to the possibility of an affix being used to create new formations, while the latter concerns the actual extent to which an affix is used to produce a large number of new words. Since fully unproductive morphological processes are relatively rare, what is more relevant is the variation in output across different affixes. This is why productivity research has often emphasized the quantitative aspect—that is, profitability—over the qualitative one. In order to evaluate this aspect of morphological productivity, various methods have been developed, traditionally divided into dictionary-based and corpus-based measures.

In dictionary-based methods, productivity is often estimated by counting the number of attested word types with a given affix at a certain point in time. However, this method reflects historical usage rather than current productivity: a large number of established words does not necessarily indicate that the affix is still synchronically productive. Another dictionary-based approach involves identifying neologisms—newly coined words—attested in specific periods, typically using historical dictionaries. This method,

[2] Since we are dealing with a corpus language, we must work with what has been preserved by tradition. For instance, the Archaic period is less well-documented than later periods, a factor that crucially affects the composition of the sub-corpus. Additionally, to ensure philological consistency, only texts with available critical editions, commentaries, and translations were included.

[3] For a discussion of the limitations of fully automatic processing in quantitative morphological studies, see Evert and Lüdeling (2001).

[4] The label P^* is proposed here for practical reasons.

however, depends heavily on the lexicographer's selection criteria and may underrepresent semantically transparent words that do not require dictionary entries to be understood.

A different line of inquiry, introduced by Aronoff (1976), proposes measuring productivity by calculating the ratio between the number of actual words—that is, existing and attested formations with a given affix—and the number of possible words, i.e., words that could, in principle, be formed using the same affix. The higher the ratio, the higher the productivity of a given affix. While conceptually appealing, this approach faces serious challenges. First, the number of potential formations is theoretically unbounded, since new base words can always enter the language. More critically, the measure leads to wrong predictions: highly productive affixes, for which the number of possible combinations is virtually infinite, may yield a very low productivity index. Conversely, completely unproductive affixes may score artificially high. These logical inconsistencies make the measure problematic for empirical application.

To overcome the limitations of dictionary-based and actual/possible-word-based methods, corpus-based methods have been developed, especially following the work of Harald Baayen and collaborators (Baayen, 1989, onward). These approaches rely on large electronic corpora and use the number of *hapax legomena*—words occurring only once in a corpus—as an indicator of productivity. While a *hapax legomenon* is not necessarily a neologism, in large corpora such words are often unfamiliar to most speakers and thus indicative of ongoing word-formation processes. This idea aligns with psycholinguistic models of morphological processing, which suggest that complex, low-frequency words can be interpreted by decomposing them into known morphemes using productive rules stored in the mental lexicon. Productive morphological processes therefore tend to generate many low frequency (and especially one-off) forms, while unproductive ones are typically associated with a few high frequency, well-established words. Based on these principles, Baayen (1992, 1993, 2009) proposed two main measures of morphological productivity: potential productivity (P) and expanding productivity (P^*).

P is the ratio of hapaxes with a given affix (h) to the total number of tokens with that affix (N): $P = \frac{h}{N}$. This measure estimates the probability of encountering a new type after sampling N tokens with that affix, and it reflects the affix's speed and capacity to expand its lexical inventory. However, P is a decreasing function that approaches zero as N increases (Baayen and Lieber, 1991), leading to an overestimation of rare suffixes and thus producing counterintuitive results when suffixes with very different token frequencies are compared.

P^* is the ratio of hapaxes with a given affix (h) to the total number of hapaxes in the corpus (H): $P^* = \frac{h}{H}$. Since H remains constant, P^* allows for fairer comparisons between affixes with different token frequencies. Unlike P, which can be interpreted independently, P^* is inherently comparative, making it suitable for analyzing affixes with varying token frequencies in the corpus and aligning well with the goals of this study. However, a direct consequence of H being constant is that comparing P^* for the six suffixes is equivalent to directly comparing the number of their hapaxes, regardless of their respective total frequency. Following this reasoning, Bauer (2001, p. 155) objects that P^* "asks 'What proportion of new coinages use affix A?' rather than asking 'What proportion of words using affix A are new coinages?'", the latter being more relevant to assess productivity in a strict sense. Gaeta and Ricca (2006) also highlight this conceptual limitation, preferring the interpretation of productivity as the rate of lexical growth of an affix over time: *P(N)*. Nonetheless, they acknowledge that P^* is a practical and efficient proxy, precisely because the results it yields are consistent with those obtained through their variable-corpus approach, which measures productivity by calculating P at equal values of N across different affixes. This method avoids the bias introduced by token frequency differences in the P index, but requires reducing the sample size to the lowest available N, which may underrepresent the productivity of more frequent affixes, and ignores the broader distribution of frequencies across types. These limitations can be addressed using statistical models called LNRE (Baayen, 2001; Evert, 2004; Baroni and Evert, 2006; Evert and Baroni, 2006). Starting from the observed data, these models allow for the estimation of P for any value of N (also greater than the N empirically observed), enabling comparison between affixes with very different token frequencies. The most widely used models currently available are Generalized Inverse Gauss-Poisson (GIGP; Baayen, 2001), finite Zipf-Mandelbrot and Zipf-Mandelbrot (fZM and ZM; Evert, 2004; Baroni and Evert, 2006), implemented in the `zipfR` package for R (Baroni and Evert, 2006).

3 Measuring the productivity of Ancient Greek deverbal nominal suffixes

3.1 Distribution and relative frequency across time

This section presents key quantitative data for each suffix, including type count (V), token count (N), number of *hapax legomena* (h), and relative frequency (R_f) within the overall corpus.

As mentioned above, the full corpus is divided into four diachronic subcorpora. The Archaic corpus (8th–6th c. BC) contains 277,876 tokens; the Classical corpus (5th–4th c. BC), 1,231,944 tokens; the Hellenistic corpus (3rd–1st c. BC), 1,121,023 tokens; and the Imperial corpus (1st–6th c. AD), 1,288,522 tokens.

Archaic period				
Suffix	V	N	h	R_f(‰)
-eía	8	19	4	0.068
-mos	40	358	17	1.288
-sia	11	158	4	0.569
-sis	145	547	73	1.969
-tis	8	24	3	0.086
-tus	12	45	7	0.162

Table 1: Type count (V), token count (N), number of *hapax legomena* (h), and relative frequency (R_f) of AG deverbal nominal suffixes in the Archaic period.

Classical period				
Suffix	V	N	h	R_f(‰)
-eía	46	865	17	0.702
-mos	156	2,005	64	1.628
-sia	33	320	12	0.26
-sis	792	10,238	302	8.31
-tis	8	195	2	0.158
-tus	1	1	1	0.001

Table 2: Type count (V), token count (N), number of *hapax legomena* (h), and relative frequency (R_f) of AG deverbal nominal suffixes in the Classical period.

Hellenistic period				
Suffix	V	N	h	R_f(‰)
-eía	50	1,632	11	1.456
-mos	160	1,267	78	1.130
-sia	39	914	9	0.815
-sis	537	10,065	206	8.978
-tis	11	352	2	0.314
-tus	2	4	1	0.001

Table 3: Type count (V), token count (N), number of *hapax legomena* (h), and relative frequency (R_f) of AG deverbal nominal suffixes in the Hellenistic period.

Imperial period				
Suffix	V	N	h	R_f(‰)
-eía	66	1,222	12	0.948
-mos	217	3,116	87	2.418
-sia	45	1,151	13	0.893
-sis	367	13,391	279	10.39
-tis	8	493	2	0.383
-tus	0	0	0	0

Table 4: Type count (V), token count (N), number of *hapax legomena* (h), and relative frequency (R_f) of AG deverbal nominal suffixes in the Imperial period.

As shown in Tables 1 to 4, the suffixes differ considerably in their distribution across the four historical periods. While some suffixes show relatively stable patterns, others display notable fluctuations. For instance, *-sia* and *-eía* vary in frequency over time, whereas *-sis* consistently emerges as the most frequent suffix, followed by *-mos*/*-mós*. By contrast, *-tis* and *-tus* are the least frequent overall, with *-tus* appearing almost exclusively in the Archaic period.

The diachronic distribution of suffix usage is also visualized in Figure 1, which displays relative frequency trends across periods. To assess whether these changes are statistically significant, a chi-squared test with simulated p-values (10,000 replicates) was conducted on raw frequency counts. The test yielded a significant result ($\chi^2 = 3236.7$, $p < 0.001$), indicating that suffix usage and historical period are not independent. However, the effect size as measured by Cramér's V ($V = 0.149$) indicates that the association, while statistically significant, is relatively weak.

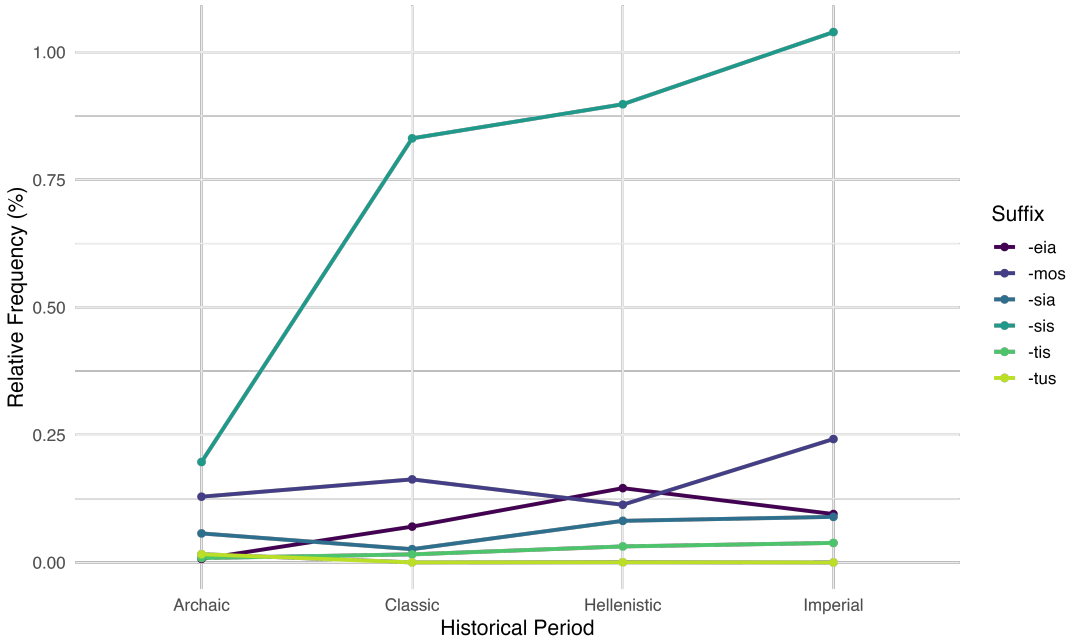


Figure 1: Relative frequency (%) of the six suffixes across the four historical periods.

Suffix	*P*-Archaic	*P*-Classical	*P*-Hellenistic	*P*-Imperial
-eía	0.211	0.019	0.007	0.009
-mos	0.047	0.032	0.062	0.028
-sia	0.025	0.038	0.009	0.011
-sis	0.133	0.029	0.020	0.021
-tis	0.125	0.010	0.006	0.004
-tus	0.156	1	0.25	0

Table 5: *P*-index values for AG deverbal nominal suffixes in the four historical periods.

3.2 *P* measure

Table 5 presents the values of Baayen's *P* measure across the four historical subcorpora[5] for the six suffixes under investigation. As mentioned in Section 2, a key limitation of *P* is its tendency to overestimate productivity for affixes with low token frequency, which can lead to counterintuitive results. As Gaeta and Ricca (2006) emphasize, this distortion becomes particularly significant when the affixes being compared exhibit a large difference in token frequency. This study exemplifies this issue, calling for a cautious interpretation of the results. Notably, *P* values in the Archaic period appear consistently inflated. This can be explained by the relatively small size of the Archaic corpus, due to the limited documentation from this early phase. A direct consequence is that most suffixes have a small number of occurrences, leading to a systematic overestimation of productivity in this period compared to later phases. However, this overestimation is not limited to the Archaic phase alone. For instance, *-tus*—the rarest suffix in the dataset—shows very high values for *P*.[6] Conversely, *-sis*—which is by far the most frequent suffix throughout the four historical phases and contributes the highest number of types and hapaxes—registers very low *P* values, particularly in the Classical and Hellenistic phases. Although the

[5]Productivity results are presented separately for each historical phase, because the operative notion of productivity used in this study is inherently synchronic. Combining data from different periods would erase diachronic variation and potentially lead to misleading conclusions, as it would mix productive behavior from earlier and later stages of the language.

[6]The case of *-tus* is particularly illustrative of *P*'s limitations: despite reaching a productivity index of 100%, this suffix is attested only once in the Classical corpus (one type, one token – Ar. *Pax* v. 1164). This paradox underscores how *P* can be misleading when applied to extremely low-frequency affixes.

relationship between frequency and productivity is far from straightforward,[7] it is clearly problematic that the least frequent suffix (largely confined to Archaic epic poetry) appears as the most productive, whereas the most frequent one scores among the lowest. These results confirm previous findings that *P* tends to overestimate the productivity of low-frequency affixes. In a dataset such as the one of this study, where affix distributions are highly uneven and vary significantly across periods, the distortion introduced by *P* becomes particularly pronounced. This reinforces the broader methodological consensus that *P*, while informative in some contexts, is insufficient on its own for evaluating morphological productivity, especially in diachronic studies or in cases involving strongly unbalanced frequency distributions.

3.3 *P** measure

To further investigate morphological productivity, the number of *hapax legomena* (*h*) for each suffix across the four historical corpora is reported in Table 6. Since the *P** measure is calculated as the ratio of affix-specific hapaxes to the total number of hapaxes in the corpus, these raw counts serve as the basis for assessing expanding productivity and identifying general trends in affix usage over time.

Based on these counts, a chi-squared test was conducted to determine whether the distribution of hapaxes across historical periods is significantly dependent on the suffix. The test yielded a chi-squared value of 71.50 with $p < 0.001$ (simulated, 10,000 replicates), indicating a statistically significant association. To evaluate the strength of this dependence, Cramér's *V* was also calculated, yielding a value of 0.14. This suggests a small effect size, which nevertheless may offer valuable insights into the dynamics of suffix rivalry and the changing productivity of affixes over time.

Overall, *-sis* appears as the dominant suffix across all periods, suggesting its central role in deverbal nominalization. The second most productive suffix is *-mos*/*-mós* that peaks in the Hellenistic period, possibly reflecting stylistic preferences.[8] A notable deviation is *-tus*, which ranks third in the Archaic period but declines sharply thereafter, disappearing by the Imperial era, suggesting early restriction to specific genres. The suffixes *-sia* and *-eía* consistently show low productivity, with *-eía* producing a slightly higher number of hapaxes until the Imperial period, where *-sia* overtakes *-eía*. This variation, however, does not reach statistical significance and may reflect a marginal fluctuation rather than a meaningful trend. Finally, *-tis* does not appear to be productive in any period.

Suffix	***h*-Archaic**	***h*-Classical**	***h*-Hellenistic**	***h*-Imperial**
-eía	4	17	11	12
-mos	17	64	78	87
-sia	4	12	9	13
-sis	73	302	206	279
-tis	3	2	2	2
-tus	7	1	1	0

Table 6: Number of hapaxes for AG deverbal nominal suffixes in the four historical periods.

3.4 LNRE models

As outlined in the previous sections, both *P* and *P** present notable limitations. While *P* is a more robust measure of productivity, it is negatively sensitive to affix frequency distribution. *P**, on the other hand, is a useful metric that enables more balanced comparisons, but provides less detailed insight into affix-specific productivity, being mathematically simpler and thus less informative. To overcome these issues and allow for more robust comparisons across differently distributed suffixes, this section

[7] Most studies on derivational morphology from a historical linguistic perspective consider productivity solely based on frequency or type/token ratio. Furthermore, expressions such as 'X is productive' are often presented as self-evident (Sandell, 2015), reflecting an approach that considers productivity categorically as availability.

[8] The analysis of literary genres will be addressed in a forthcoming study.

introduces statistical LNRE models, which provide a principled method for estimating productivity based on the overall frequency distribution of the suffixes. Among the three main LNRE models introduced in Section 2—ZM, fZM and GIGP—this study employs only the Zipf-Mandelbrot (ZM) model.[9] This choice is motivated by the fact that the ZM model performs more reliably with small sample sizes, whereas fZM and GIGP often fail to produce stable parameter estimates (Evert and Baroni, 2006). Based on the observed frequency distribution, the model infers a set of parameters that generalize the behavior of the suffixes, making it possible to estimate productivity (*P*) for values of *N* larger than those observed in the corpus.

Archaic Period					
Suffix	*h*	*EV1* 1000	*P* 1000	*EV2* 2000	*P* 2000
-eía	4	73.53	7.353	107.77	5.388
-mos	17	62.04	6.204	83.52	4.176
-sia	4	33.29	3.329	48.25	2.413
-sis	73	215.64	21.564	325.97	16.299
-tis	3	54.29	5.429	76.24	3.812
-tus	7	115.99	11.599	190.14	9.507

Table 7: Estimated productivity values using the ZM model for each suffix in the Archaic period.

Classical Period					
Suffix	*h*	*EV1* 1000	*P* 1000	*EV2* 2000	*P* 2000
-eía	17	48.82	4.882	63.69	3.184
-mos	64	114.32	11.432	164.4	8.22
-sia	12	55.41	5.541	74.77	3.738
-sis	302	269.22	26.922	381.3	19.065
-tis	2	19.36	1.936	25.81	1.291
-tus	1	–	–	–	–

Table 8: Estimated productivity values using the ZM model for each suffix in the Classical period.

Hellenistic Period					
Suffix	*h*	*EV1* 1000	*P* 1000	*EV2* 2000	*P* 2000
-eía	11	45	4.5	55.8	2.79
-mos	78	143.02	14.302	204.51	10.226
-sia	9	39.94	3.994	47.23	2.361
-sis	206	202.55	20.255	274.75	13.737
-tis	2	16.45	1.645	20.37	1.019
-tus	1	–	–	–	–

Table 9: Estimated productivity values using the ZM model for each suffix in the Hellenistic period.

Imperial Period					
Suffix	*h*	*EV1* 1000	*P* 1000	*EV2* 2000	*P* 2000
-eía	12	64.84	6.484	80.25	4.013
-mos	87	132.18	13.218	183.73	9.186
-sia	13	47.05	4.705	61.58	3.079
-sis	279	279.22	27.922	379.88	18.994
-tis	2	10.27	1.027	12.71	0.636
-tus	0	–	–	–	–

Table 10: Estimated productivity values using the ZM model for each suffix in the Imperial period.

Tables 7 to 10 report, for each suffix and each historical phase, the following information: the attested number of hapaxes (*h*), the expected number of hapaxes for $N = 1000$ (*EV1*-1000) and $N = 2000$ (*EV2*-2000), and the corresponding productivity values *P*-1000 and *P*-2000. Outside the Archaic period, *-tus* is excluded because its total number of tokens falls below the minimum threshold ($N \geq 5$) required for using the model. This, however, serves as indirect evidence of the suffix's lack of productivity in later stages—an observation also confirmed by the *P** values, which contrast with the misleadingly high *P* scores.

That said, the small size of the Archaic corpus—already identified as a source of distortion in Section 3.2—may also affect the performance of the ZM model. As Evert and Baroni (2006) note, extrapolation quality can degrade rapidly when the estimation size becomes too small. This is particularly relevant

[9] Relying solely on the models *goodness-of-fit* was particularly challenging in this case, as applying the models to six different suffixes across four historical periods—each with markedly different distribution—yielded highly variable results depending on the suffix. For the sake of consistency, I therefore decided to apply the same model to all suffixes.

for suffixes like *-eía*, *-sia*, *-tis*, and *-tus*, which are represented by very few types and tokens, making it difficult for the model to accurately capture the shape of the distribution and potentially leading to biased estimates. While Evert and Baroni (2006) express skepticism regarding the use of LNRE models for extrapolating productivity based solely on hapax counts, the results for the Classical, Hellenistic, and Imperial periods in this study show strong alignment between the ZM model estimates and the values obtained via *P**. This consistency reinforces the reliability of both approaches: on the one hand, it provides a mathematical foundation that strengthens the *P** results; on the other, it helps validate the ZM model's output in contexts where it might otherwise be questionable due to limited data.

Finally, Figures 2-5 present the extrapolated vocabulary growth curves for each period, summarizing the core findings of the ZM model (while keeping in mind the limitations of the Archaic dataset). These curves clearly indicate that *-sis* is by far the most productive suffix for deverbal noun formation in AG, followed by *-mos*/*-mós*. The suffixes *-eía* and *-sia* show minimal productivity, while *-tis* should not be considered productive at all.[10]

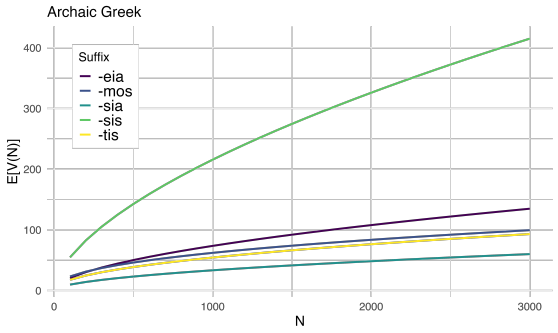


Figure 2: Extrapolated vocabulary growth curves for each suffix in the Archaic period using the ZM model.

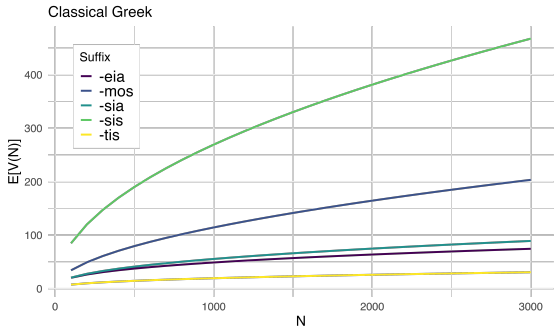


Figure 3: Extrapolated vocabulary growth curves for each suffix in the Classical period using the ZM model.

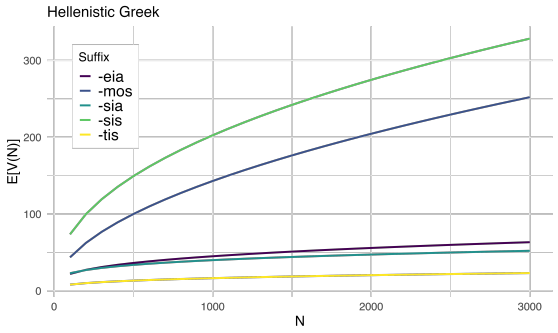


Figure 4: Extrapolated vocabulary growth curves for each suffix in the Hellenistic period using the ZM model.

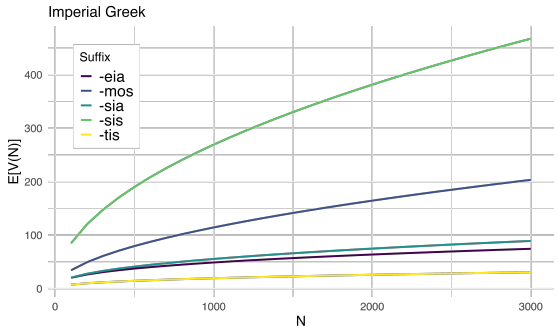


Figure 5: Extrapolated vocabulary growth curves for each suffix in the Imperial period using the ZM model.

[10]The suffix *-tus*, as mentioned in the Section, is excluded due to insufficient data.

3.5 Suffix interaction and resolution of rivalry

The data extracted through the quantitative analysis were also used to explore possible dynamics of rivalry between suffixes. This step was motivated by hypotheses found in the literature regarding the potential competition among deverbal nominal suffixes in AG. More specifically, I aimed to: (i) test Chantraine's (1933) observation that *-sis* might stand in a competitive relationship with *-mos*/*-mós* and *-sia*; and (ii) examine whether additional patterns of competition might emerge within the morphological domain of deverbal nominal derivation.

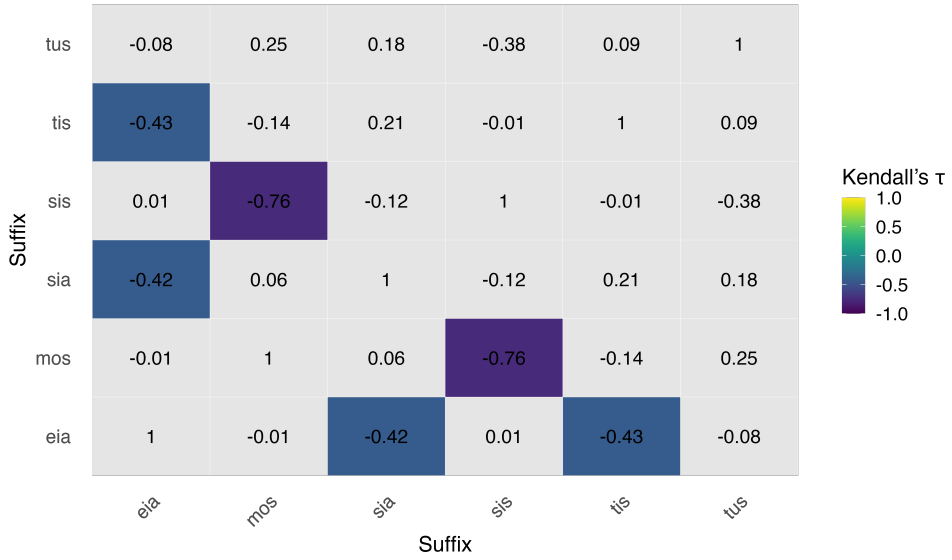


Figure 6: Kendall's Tau correlation heatmap showing pairwise productivity correlations among six AG deverbal nominal suffixes, computed on the combined diachronic–genre series of *P** scores.

Figure 6 presents the results of Kendall's Tau correlation analysis, chosen for its suitability in handling non-normally distributed data. The heatmap shows the pairwise correlations among the six suffixes considered in this study. For each suffix-pair, Kendall's Tau was computed on the combined diachronic–genre series of *P** scores. A strong and statistically significant negative correlation emerges between *-sis* and *-mos*/*-mós* ($p < 0.001$), indicating that an increase in the productivity of one corresponds to a decrease in the other. A similar, though less pronounced, negative correlation is observed between *-sia* and *-eía* ($p = 0.012$). Additionally, *-eía* shows a weak but significant negative correlation with *-tis* ($p = 0.014$). By contrast, no significant correlation was found between *-sis* and *-sia*. However, this does not necessarily imply the absence of competition. Rather, it may suggest that, if a competitive dynamic did exist, it was not resolved through clear functional differentiation or temporal displacement. Any potential functional overlap or genre-specific usage between these suffixes must therefore be assessed through qualitative analysis. Importantly, the significant negative correlations observed in the heatmap do not indicate the presence of ongoing competition, but rather suggest that, where competition may have occurred in the past, it was eventually resolved. Such resolution could take the form of functional specialization or the gradual decline of one suffix in favor of another. In this sense, the correlation data capture the outcome of a competitive process, not its existence or dynamics. Taken together with the low global association measured by Cramér's *V*, the correlation data suggest that observed rivalry is likely limited to specific affix pairs rather than reflecting a broad restructuring of the derivational system.

These findings will be further contextualized through qualitative analysis in order to better understand how such interactions manifest within specific literary contexts, how competing suffixes may have diverged functionally, and whether cases of overabundance—where multiple suffixes coexist—may underlie relationships not captured by correlation alone.

4 Conclusions

This study provides the first quantitative and diachronic analysis of the productivity of six AG deverbal nominal suffixes, applying Baayen's *P* and *P**, and using an LNRE model of word frequency distribution (Baayen, 1992, 1993, 2001, 2009; Baroni and Evert, 2006; Evert and Baroni, 2006). Overall, the results show that the three measures have limitations, largely due to the quantity and distributional characteristics of the available data for AG: *P* tends to overestimate the productivity of low-frequency affixes; *P**, while useful for comparison, is overall not fully informative on its own; and the ZM model can yield biased results when applied to small datasets. However, both *P** and the ZM model largely point in the same direction, suggesting that, for assessing morphological productivity in a corpus language, the integration of multiple quantitative measures is essential—along with a careful awareness of the data's shape and distribution. Moreover, data deficiency and sparsity are not the only challenges in applying quantitative methods to diachronic research. As Štichauer (2009) already noted, the non-random nature of diachronic corpora must also be taken into account. Such corpora are typically affected by inhomogeneity and repetition/clustering effects. Regarding inhomogeneity, the corpus used in this study consists of texts with diverse properties: different genres, different authors, and varying proportions of text types across the four historical subcorpora. This reflects a structural imbalance, as certain types of texts are simply absent in specific periods. In addition, repetition effects are often present, as most texts are authored by single individuals, leading to localized overuse of specific formations. A typical and unlucky case is when an author introduces a new formation (a hapax candidate) and repeats it multiple times, or when a high-frequency item appears almost exclusively in one or two texts rather than being evenly distributed across the corpus. Another key caveat when working with AG corpora is the risk of PoS-tagging errors in automatically annotated texts. In several cases, homographic forms can correspond to different parts of speech. For example, *amúxeis* could be either a plural noun in the direct case derived from the verb *amússō* or the second-person singular future form of the same verb. Only a close reading of the original passages allows for correct disambiguation based on context, since automatic taggers may assign the wrong category. To mitigate the impact of such errors, a sample-based manual review of the data is being conducted to assess the potential margin of error.

Despite the methodological challenges outlined above, I believe that quantitative measures of morphological productivity can be effectively applied to AG—provided that they are used with caution and in combination. The integration of different metrics, along with a critical awareness of corpus structure, frequency distribution, and annotation quality, allows for meaningful analysis even in historical languages with limited and heterogeneous documentation. In the case of this study, the findings highlight clear differences in suffix productivity, reinforcing the dominant role of *-sis* and *-mos*/*-mós* in deverbal nominalization, while revealing the limited productivity of *-tis* and *-tus* over time (*-tus* producing *hapax legomena* only in Archaic epic poetry). The suffix *-sis* consistently emerges as the most productive across all periods, combining high frequency with the ability to generate new types, thus confirming its central role in the deverbal nominal domain. While less dominant, *-mos*/*-mós* shows substantial productivity, peaking during the Hellenistic period, potentially reflecting literary trends. The suffixes *-eía* and *-sia* display moderate but unstable productivity, suggesting a more specialized and constrained role in word formation, possibly influenced by stylistic and literary conventions. In addition to individual suffix productivity, the study also explored potential interactions between suffixes through a correlation-based analysis. Significant negative correlations—particularly between *-sis* and *-mos*/*-mós*, *-sia* and *-eía*, and *-eía* and *-tis*—suggest that, in some cases, morphological competition may have been resolved over time. By contrast, the lack of significant correlation in other pairs, such as *-sis* and *-sia*, points to unresolved or more complex dynamics, which cannot be captured quantitatively. These findings highlight the need for further qualitative investigation to assess the nature and extent of functional overlap, and to determine whether cases of overabundance or genre-specific distribution underlie the observed trends. To further explore these hypotheses, future work will examine suffix productivity across literary genres in order to assess whether the observed shifts reflect genuinely active word-formation processes or are instead shaped by genre-specific usage patterns. In parallel, a qualitative analysis is currently underway to investigate suffix polyfunctionality and possible instances of morphological rivalry.

References

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. MIT Press, Cambridge.

Harald Baayen. 1989. *A Corpus-Based Approach to Morphological Productivity: Statistical Analysis and Psycholinguistic Interpretation*. Ph.D. thesis, Vrije Universiteit, Amsterdam.

Harald Baayen. 1992. Quantitative aspects of morphological productivity. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1991*, Dordrecht, Springer, pages 109–149. https://doi.org/10.1007/978-94-011-2516-1_8.

Harald Baayen. 1993. On frequency, transparency and productivity. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1992*, Dordrecht, Springer, pages 181–208. https://doi.org/10.1007/978-94-017-3710-4_7.

Harald Baayen. 2001. *Word-Frequency Distributions*. Kluwer Academic, Dordrecht.

Harald Baayen. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, New York, volume 2, pages 899–919. https://doi.org/10.1515/9783110213881.2.899.

Harald Baayen and Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* (29):801–844. https://doi.org/10.1515/ling.1991.29.5.801.

Harald Baayen and Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language* (72):69–96. https://doi.org/10.2307/416794.

Marco Baroni and Stephanie Evert. 2006. The zipfr package for lexical statistics: A tutorial introduction https://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf.

Laurie Bauer. 2001. *Morphological Productivity*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511486210.

Émile Benveniste. 1948. *Noms d'agent et noms d'action en indo-européen*. Adrien-Maisonneuve, Paris.

Andrew Carstairs-McCarthy. 1992. *Current Morphology*. Routledge, London.

Pierre Chantraine. 1933. *La formation des noms en grec ancien*. Libraire C. Klincksieck, Paris.

Germana Olga Civilleri. 2010. *Nomi deverbali nel continuum nome-verbo: il caso del greco antico*. Ph.D. thesis, Università di Roma Tre, Roma.

Danielle Corbin. 1987. *Morphologie dérivationnelle et structuration du lexique*, volume 1. Niemeyer, Tübingen.

Albert Debrunner. 1917. *Griechische Wortbildungslehre*. Carl Winters Universitätsbuchhandlung, Heidelberg.

Stephanie Evert. 2004. A simple LNRE model for random character sequences. In Gérald Purnelle, Cédrick Fairon, and Anne Dister, editors, *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004)*. pages 411–422. http://www.corpus.bham.ac.uk/PCLC.

Stephanie Evert and Marco Baroni. 2006. Testing the extrapolation quality of word frequency models. In Pernilla Danielsson and Martijn Wagenmakers, editors, *Proceedings of Corpus Linguistics*. https://marcobaroni.org/publications/cl2005/cl-052-pap-final.pdf.

Stephanie Evert and Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In Paul Rayson, Andrew Wilson, Tong McEnery, Andrew Hardie, and Shereen Khoja, editors, *Proceedings of the Corpus Linguistics 2001 Conference*. UCREL, Lancaster, pages 167–175. https://doi.org/10.18452/13417.

Livio Gaeta and Davide Ricca. 2002. Corpora testuali e produttività morfologica: i nomi d'azione italiani in due annate della stampa (1996-1997). In Roland Bauer and Hans Goebl, editors, *Parallela IX. Testo-variazione-informatica/Text-Variation-Informatik. Atti del IX Incontro italoaustriaco dei linguisti, Salzburg, 1-4 novembre 2000*. Egert, Wilhelmsfeld, pages 223–249.

Livio Gaeta and Davide Ricca. 2005. Aspetti quantitativi della produttività morfologica. In Tullio De Mauro and Isabella Chiari, editors, *Parole e numeri. Analisi quantitative dei fatti di lingua*. Aracne, pages 107–124.

Livio Gaeta and Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* (44):57–89. https://doi.org/10.1515/LING.2006.003.

Maria Napoli. 2009. Il morfo *-mo-: uno studio diacronico e qualche nota tipologica. In Gian Franco Nieddu Ignazio Putzu, Giulio Paulis and Pierluigi Cuzzolin, editors, *La morfologia del greco tra tipologia e diacronia. Atti del VII Incontro Internazionale di Linguistica Greca*. Franco Angeli, Milano, pages 314–335.

Ingo Plag. 2006. Productivity. In Bas Aarts and April McMahon, editors, *Handbook of English Linguistics*, John Wiley & Sons, Hoboken, pages 537–556. https://doi.org/10.1002/9780470753002.ch23.

Ernst Risch. 1974. *Wortbildung der homerischen Sprache*. De Gruyter, Berlin/New York.

Ryan Paul Sandell. 2015. *Productivity in Historical Linguistics: Computational Perspective on Word-Formation in Ancient Greek and Sanskrit*. Ph.D. thesis, University of California, Los Angeles. https://escholarship.org/uc/item/2z1476f8.

Anna Maria Thornton. 2005. *Morfologia*. Carocci, Roma.

Rossella Varvara. 2019. Misurare la produttività morfologica: i nomi d'azione nell'italiano del ventunesimo secolo. In Bruno Moretti, Aline Kunz, Silvia Natale, and Etna Krakenberger, editors, *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana*. Officinavenutono, Milano, pages 187–201. https://www.societadilinguisticaitaliana.net/wp-content/uploads/2019/08/eBookSLI_vol_2.pdf.

Rossella Varvara. 2020. Constraints on nominalizations: Investigating the productivity domain of Italian *-mento* and *-zione*. *Zeitschrift für Wortbildung / Journal of Word Formation* 4(2):78–99. https://doi.org/10.3726/zwjw.2020.02.05.

Pavel Štichauer. 2009. Morphological productivity in diachrony: The case of the deverbal nouns in *-mento*, *-zione* and *-gione* in Old Italian from the 13th to the 16th century. In Fabio Montermini, Gilles Boyé, and Jesse Tseng, editors, *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*. Cascadilla Proceedings Project, pages 138–147. https://www.lingref.com/cpp/decemb/6/paper2241.pdf.

CroComp – Lexicon of Croatian Compounds

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

Abstract

This paper deals with the development of the Lexicon of Croatian compounds – CroComp. We discuss the theoretical aspects of compounding in Croatian and their application in the creation of the Lexicon. Lexical entries in CroComp provide information on the individual elements of each compound, the word-formation pattern used in compounding, and the affixes present in the compound. Lexical entries also include information about the morphological structure of each compound. CroComp currently contains around 1,300 entries.

1 Introduction

Croatian is a South Slavic language characterized by a rich inflectional and word-formational morphology. Besides having a rich morphological system, numerous morpho-phonological processes occur at morpheme boundaries. There are also vowel alternations within the lexical morphemes (*ablaut*). These processes frequently result in allomorphy of affixes and stems.

The main word-formation processes in Croatian are derivation and compounding, along with conversion, blending, acronym formation, and others. The key difference between derivation and compounding lies in the number of bases involved: derivation uses one base, while compounding involves two or more bases. Inflection in Croatian relies exclusively on suffixation, whereas word-formation involves suffixation, prefixation, simultaneous suffixation and prefixation, and ablaut. Comprehensive lists of derivational processes can be found in Filko et al. (2020) and Šojat and Filko (2023). In compounding, an additional element - a linking vowel or interfix - is commonly used to connect the bases forming the compound.

In this paper, we focus on the word-formation of compounds. We analyze their structure in terms of the lexical stems and affixes used in compounding. We also analyze their morphological structure, segmenting the compounds into the morphemes they consist of. The results of the analysis are presented in a computational lexicon specifically designed to store and display morphological data of this kind.

The paper is structured as follows: Section 2 presents previous work on Croatian morphology and language resources. Section 3 addresses compounding in general. Section 4 presents the types of compounds identified in our analysis and describes CroComp, a computational lexicon designed to present the structure of compounds. Section 5 summarizes the research findings and provides an outline of future work.

2 Previous work

So far, computational processing of Croatian morphology has mainly focused on inflection, although there are language resources that include derivational data. The Croatian inflectional morphology is covered by several large lexica with paradigms and inflectional patterns (Tadić and Fulgosi, 2003; Ljubešić et al., 2016) used for lemmatization, MSD (MorphoSyntactic Descriptors) and POS (Part of Speech) tagging, and similar NLP (Natural Language Processing) tasks. There are currently two derivational databases available for the Croatian language: DerivBase.HR (Šnajder, 2014)] and CroDeriv (Filko et al., 2020; Šojat and Filko, 2023).

Proceedings of the 5th International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025),
pages 141–150, Fribourg, Switzerland, 4-5 September 2025.

DerivBase.HR provides a structured inventory of derivational relationships between lexical units. It includes more than 100,000 lemmas grouped into approximately 56,000 derivational families. Each family consists of words that are morphologically related through derivation, primarily via suffixation. The resource focuses on the most productive word classes in Croatian and captures regular patterns of word-formation. The resource emphasizes suffixal derivation, particularly between nouns, verbs, and adjectives.

CroDeriv, on the other hand, offers a broader linguistic resource that includes both derivational and morphological information, making it particularly useful for computational applications and linguistic research. The lexical entries in CroDeriv provide information about the morphological structure of words and their derivational links with other words. Each lexeme is segmented into morphemes, the derivational stem used for the derivation is indicated, and the applied derivational process is specified. The morphological segmentation of the lexemes is based on a two-layered approach: segmentation at the surface and deep layer. Allomorphs are identified and marked for their type at the surface layer of the analysis. Allomorphs are connected to their representative morphs at the deep layer of presentation. The representative morph is the one from which other allomorphs can be established with the least number of morpho-phonological rules. In Figure 1 below, we present an entry from the current version of CroDeriv for the verb *nadopisivati* 'to add by writing $_{\text{IMPF}}$'. In this example, the morphological structure at both the surface and deep layers is identical, meaning that there are no morpho-phonological changes in the surface structure:

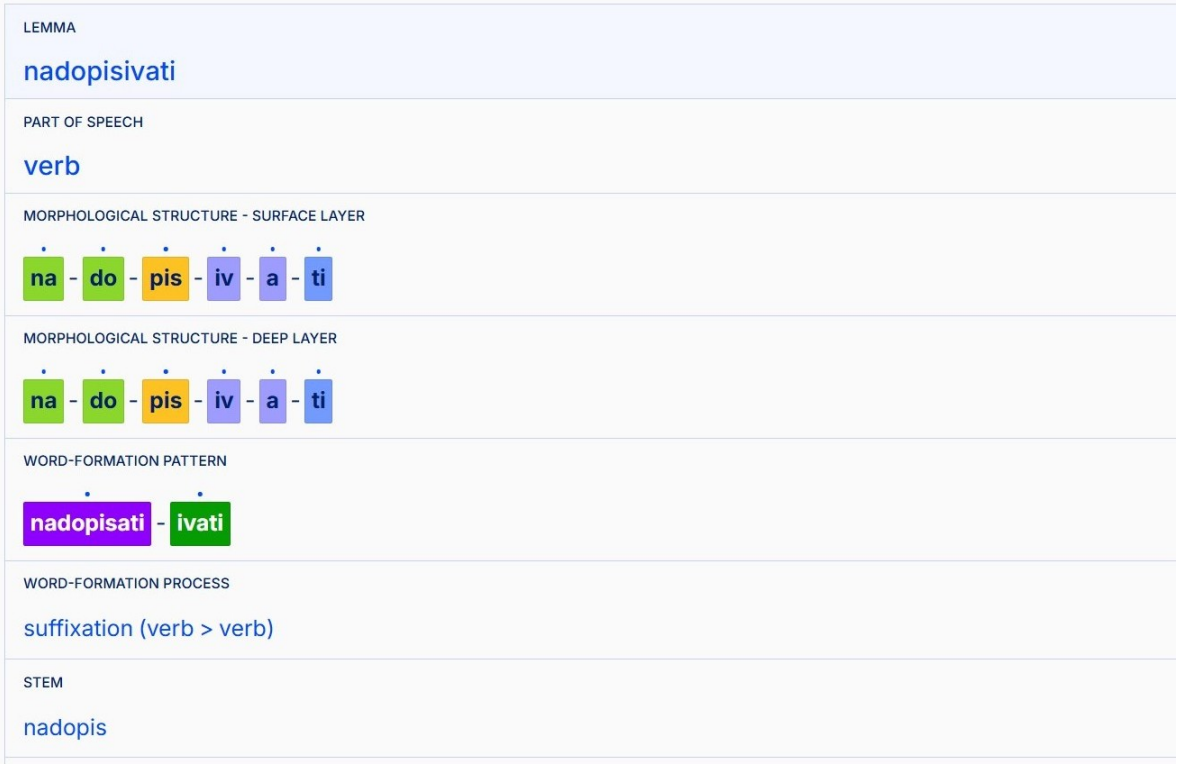


Figure 1: The lexical entry for *nadopisivati* in the CroDeriv search interface

So far, only derivation has been processed in CroDeriv. One of the tasks in its further development is the expansion of the lexicon with compounds. For this purpose, a list of approximately 1,300 compound words was collected. The list was obtained from monolingual dictionaries and publicly available corpora. Before we describe how we process compounds, in the next section we will briefly list the main types of compounding in Croatian.

3 Compounding in Croatian

The most comprehensive overview of word-formation in Croatian is given in Babić (2002). The author discusses the formation of compound nouns, adjectives, verbs, and adverbs. The formation of compound nouns and adjectives dominates in terms of the number of different word-formation patterns. Word-formation patterns refer to combinations of words belonging to various POS and affixes that participate in compounding. A brief overview of word-formation is given by Grčević (2016), who relies entirely on Babić (2002) in his presentation. Word-formation in Croatian is also addressed by Barić et al. (1995), Silić and Pranjković (2005), and others.

Babić (2002) analyzes compounds according to two criteria: (1) according to the POS of the elements in the compounds, and (2) according to whether affixes are added in the compounding process. Affixes used in compounding are suffixes, interfixes, and prefixes. Combinations of elements belonging to different POS and various affixes produce compound nouns, adjectives, verbs, and adverbs. Based on these criteria, the author determines the main types of compounding in Croatian:

1. Proper compounds, also referred to as 'pure' compounds – compounds consisting of a stem and a word usually joined into a single lexeme by an interfix, without suffixation or prefixation: *romanopisac* (roman-o-pisac) 'novelist', *ribolov* (rib-o-lov) 'fishing';
2. Suffixal compounds – combinations of two stems and a suffix: *ženomrzac* (žen-o-mrz-ac-) 'misogynist', *častohlepan* (čast-o-hlep-an) 'ambitious, pushy';
3. Prefixal compounds – combinations of two stems and a prefix (or a prefix and a suffix) – only for verbal compounds: *omalovažiti* (o-malo-važiti) 'to belittle', *odobrovoljiti* (o-dobr-o-volj-iti) 'to appease, to cheer up';
4. Fusions or coalescences (in Croatian 'sraslice')[1] – conjoined words bearing all the inflectional markers which appear in a corresponding syntactic phrase, without affixation. Marković (2012, p. 65) points out that such compounds are created by the fusion of phrases or firm collocations, usually when the meaning of both elements is unified and thus becomes independent. In compounding types 1 to 3 above, in his opinion, there is no syntactic connection between the bases, and the semantic connection is established by interfixes (which can be realized also as ϕ (zero)). He adds that fusion is a separate word-formation procedure, but with the same result as compounding. Klajn (2002, p. 65) emphasizes that the main criterion for distinguishing these two processes is not the presence or absence of interfixes, but the syntactic relationship between words. The transition from a phrase to a compound is marked by a change in meaning and a leveling of accent. In accordance with the above definitions, we treat fusions / coalescences as compounds whose elements can also appear as phrases, i.e., in the same morphological form as they appear in a compound. The elements that appear in a compound word can appear in the same or reverse order in phrases. Finally, fusions are characterized by the absence of interfixes and suffixes. For example:
   - *hvale* 'praise' + *vrijedan* 'worthy' = *hvalevrijedan* 'praiseworthy' – from phrases *hvale vrijedan* or *vrijedan hvale*;
   - *dan* 'day' + *gubiti* 'waste, lose' = *dangubiti* 'waste time' – from phrases *dan gubiti* or *gubiti dan*.[2]
5. Semi-compounds – elements joined by a hyphen in writing, without any inflectional marker on the first element. For example: *bob-staza* 'bobsleigh track', *boks-meč* 'boxing match', *čarter-let* 'charter flight', *remek-djelo* 'masterpiece', *vagon-restoran* 'dining car' and many others, nowadays mostly of English origin. Some authors consider the hyphen to be a type of interfix (cf. Tafra and Košutar 2009, p. 97). Regardless of the spelling norm, semi-compounds are frequently written without a hyphen.[3]

[1] Here we use two terms to explain this type of compounds. In addition to *fusion* and *coalescence*, other terms such as *merging* or *reduced phrases* are used elsewhere.

[2] There are about 15 such compounds in the current version of CroComp.

[3] For now, we do not include this type of compounds in CroComp.

We should mention here that compound words in Croatian can consist of a lexical stem and an affixoid (e.g., *bioznanost* 'bioscience'), or only of affixoids (e.g., *biolog* 'biologist'). Affixoids are prefixoids and suffixoids of Slavic, Greek and Latin or some other origin. Affixoids are also referred to as 'bound stems' or 'bound lexical morphemes' in Croatian grammars and textbooks. For now, we are not dealing with such compounds. Further in this article, we focus on compounding types 1 to 3 listed above. Compounds formed in this way are included in the current version of CroComp – Lexicon of Croatian Compounds.

4 CroComp – Lexicon of Croatian Compounds

In Section 2, we indicated that one of the tasks in the future development of CroDeriv is the expansion with compound words. For this purpose, we compiled an initial list of 1,300 compounds that belong to the main parts of speech – nouns, adjectives, verbs, and adverbs. In the first step of the analysis, the compounds were divided according to their POS. In the second step, we aimed to determine: 1. which words served as the bases for the formation of the compound, and 2. whether affixes were used in the process. If so, which affixes were used. Our primary objective was to determine if any of the compound elements had already been recorded in CroDeriv and whether direct links could thus be established. If the elements could not be straightforwardly linked to an existing entry in CroDeriv due to various alternations of stems and / or ambiguous semantic interpretation, we manually analyzed the compound. Furthermore, we aimed to identify which word-formation patterns are applied in the formation of compounds in Croatian. The word-formation pattern is determined on the basis of which parts of speech are in the first and second positions in the compound words and which affixes are used in the compounding process. During this work, we decided to create a morphological lexicon that would contain only Croatian compound words. Having these data available in one place would make various researches on compounds significantly easier. We named this lexicon CroComp – Lexicon of Croatian Compounds.

So far, we have analyzed 802 nouns, 484 adjectives, 22 verbs, and 4 adverbs, which together make up a total of 1,312 compounds. These data show that compounding is productive for nouns and adjectives and that there are patterns that allow the creation of new compound words. Compound verbs and adverbs are very rare in Croatian, and new compounds in these POS are seldom formed in the contemporary language. Based on the structure of the compounds analyzed, we have identified the following POS that can appear as the first or second element in the compounds:

- First element: 1. noun, 2. adjective, 3. pronoun, 4. numeral, 5. verb, 6. participle, 7. adverb;
- Second element: 1. noun, 2. pronoun, 3. verb, 4. participle, 5. preposition, 6. numeral.

The label *numeral* encompasses cardinal and ordinal numbers, as well as the so-called numeral nouns. In the following, we use the label *num-card* in individual patterns for cardinal numbers, *num-ord* for ordinal numbers, and *num-noun* for numeral nouns.

Next, we present the main word-formation patterns for each part of speech. The patterns are ordered by frequency, which is indicated in parentheses next to each one. The most frequent patterns are listed at the top, and the least frequent patterns are listed at the bottom of the list for each word class.

4.1 Nouns

Based on the combinations of members from the groups of the first and second elements, along with the addition of various interfixes and suffixes, we identified 28 different word-formation patterns for nouns. The noun formation patterns were grouped into six main types based on the first element in the compounds, that is, according to the part of speech to which the first element belongs.

Each pattern is illustrated with two examples, provided it includes more than one compound. 'Interfix' or 'suffix' in parentheses in main types means that in some patterns it is not realized.

Alongside the Croatian words used to illustrate the patterns, we show in parentheses the bases from which the compounds are formed, as well as the interfix that connects them, e.g., *pismonoša* (pism-o-noša) 'postman', literally: *letter-o-carrier*:

1. ***noun*** + *interfix* + *verb / noun / participle (+ suffix)* (524):

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

2. ***adjective*** + *interfix* + *noun / verb (+ suffix)* (108):

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

3. ***numeral*** + *interfix* + *noun / verb (+ suffix)* (76):

- num-card + interfix + noun + suffix (37): *jednoglasje* (jedn-o-glasje) ‘unanimity’; *dvoglasnik* (dv-o-glasnik) ‘diphthong’
- num-card + interfix + noun (11): *dvotočka* (dv-o-točka) ‘colon’; *trokut* (tr-o-kut) ‘triangle’
- num-noun + interfix + noun (8): *četverored* (četver-o-red) ‘four-row formation’; *peterokut* (peter-o-kut) ‘pentagon’
- num-card + interfix + verb + suffix (6): *dvopek* (dv-o-pek) ‘zwieback’; *trosjed* (tr-o-sjed) ‘three-seater sofa’
- num-ord + interfix + noun (5): *prvoborac* (prv-o-borac) ‘veteran fighter’; *prvorodilja* (prv-o-rodilja) ‘first-time mother’
- num-ord + interfix + verb + suffix (4): *prvorotkinja* (prv-o-rotkinja) ‘primipara’; *prvokup* (prv-o-kup) ‘right of first refusal’
- num-ord + interfix + noun + suffix (3): *prvopričesnik* (prv-o-pričesnik) ‘first communicant’; *drugoligaš* (drug-o-ligaš) ‘second division team’
- num-noun + interfix + noun + suffix (1): *četveronožac* (četver-o-nožac) ‘quadruped’
- num-noun + interfix + verb + suffix (1): *četveropreg* (četver-o-preg) ‘four-horse carriage’

4. ***pronoun*** + *interfix* + *noun / verb / pronoun (+ suffix)* (36):

- pronoun + interfix + noun (33): *svojevolja* (svoj-e-volja) ‘self-will’; *samoobrana* (sam-o-obrana) ‘self-defense’
- pronoun + interfix + noun + suffix (1): *samoglasnik* (sam-o-glasnik) ‘vowel’
- pronoun + interfix + verb + suffix (1): *samoljublje* (sam-o-ljublje) ‘narcissism’
- pronoun + interfix + pronoun + suffix (1): *samosvojan* (sam-o-svojan) ‘independent, distinctive’

5. ***verb /*** *participle (+ interfix)* + *noun* + *(suffix)* (29):

- verb + noun (25): *vadičep* (vadi-čep) ‘corkscrew’; *razbibriga* (razbi-briga) ‘pastime’
- verb + interfix + noun (2): *cjepidlaka* (cjep-i-dlaka) ‘nitpicker’; *primopredaja* (prim-o-predaja) ‘handover’
- verb + interfix + noun + suffix (1): *grizodušje* (griz-o-dušje) ‘remorse’
- participle + interfix + noun (1): *rođendan* (rođen-ø-dan) ‘birthday’

6. ***adverb*** + *verb / noun / participle (+ suffix)* (29):

    - adverb + verb + suffix (19): *dalekovod* (daleko-vod) ‘transmission line’; *novotvorina* (novo-tvorina) ‘innovation’
    - adverb + noun + suffix (6): *višeglasje* (više-glasje) ‘polyphony’; *mnogobožac* (mnogo-božac) ‘polytheist’
    - adverb + noun (3): *mnogokut* (mnogo-kut) ‘polygon’; *višeboj* (više-boj) ‘all-around competition’
    - adverb + participle + suffix (1): *novotvorenica* (novo-tvorenica) ‘neologism’

In this list, it can be observed that interfixes are not used in compounds composed of a verb and a noun. These are so-called imperative compounds – a group of compounds characterized by a very specific combination of elements and overall meaning. The meaning of the compound mostly cannot be inferred from the meanings of its individual elements, as almost all of them are exocentric. In most cases, the verbal component corresponds to the imperative form, although this is not always the case. Generally, such compounds are described as lacking an interfix, but even this is not always possible – for example, *cjepidlaka* ‘nitpicker, pedant, hair-splitter’. In that case, the first element does not correspond to any form of the verb *cijepati* ‘to split’.

The affixes used in compound nouns are interfixes and suffixes. The interfixes are: *-o-, -e-, -ø-, and -i-*. The suffixes are: *-(a)c, -(a)š, -a, -ač, -ajnica, -ar, -aš, -ašica, -če, -ica, -ić, -ijan(a)c, -ik, -ina, -ja, -je, -jenje, -ka, -nica, -nik, -nja, -stvo, -t, -telj, -ø*. The brackets around the parts of the suffix indicate how it is represented at the deep layer of morpheme segmentation.

4.2 Adjectives

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

1. ***adjective*** + *interfix* + *noun / adjective / participle (+ suffix)* (219):

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

2. ***noun*** + *interfix* + *verb / noun / adjective (+ suffix)* (93):

    - noun + interfix + verb + suffix (70): *dobrotvoran* (dobr-o-tvoran) ‘charitable’; *osvetoljubiv* (osvet-o-ljubiv) ‘vindictive’
    - noun + interfix + noun + suffix (12): *danonoćni* (dan-o-noćni) ‘round-the-clock, day and night’; *slobodouman* (slobod-o-uman) ‘liberal-minded’
    - noun + interfix + adjective (11): *krvožedan* (krv-o-žedan) ‘bloodthirsty’; *vodoravan* (vod-o-ravan) ‘horizontal’

3. ***numeral*** + *interfix* + *noun / participle / numeral (+ suffix)* (74):

    - num-card + interfix + noun + suffix (57): *dvonog* (dv-o-nog) ‘two-legged’; *jednostranački* (jedn-o-stranački) ‘single-party’
    - num-ord + interfix + noun + suffix (5): *trećerazredan* (treć-e-razredan) ‘third-rate’; *prvoklasan* (prv-o-klasan) ‘first-class’
    - num-noun + interfix + noun + suffix (4): *četverodijelni* (četver-o-dijelni) ‘four-part’; *četveroruk* (četver-o-ruk) ‘four-armed’

- num-card + interfix + noun (3): *tročlan* (tr-o-član) ‘three-member’; *dvočlan* (dv-o-član) ‘binomial, two-member’
- num-ord + interfix + participle + suffix (2): *prvorođeni* (prv-o-rođeni) ‘firstborn’; *drugorođeni* (drug-o-rođeni) ‘second-born’
- num-noun + interfix + noun (1): *četveročlan* (četver-o-član) ‘four-member’
- num-card + interfix + num-card (1): *trojedan* (tr-o-jedan) ‘triune’
- num-card + interfix + participle + suffix (1): *dvosjekli* (dv-o-sjekli) ‘double-edged’

4. ***adverb*** + *verb / noun / participle / adjective (+ suffix)* (61):

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

5. ***pronoun*** + *interfix* + *verb / participle / noun / pronoun (+ suffix)* (33):

   - pronoun + interfix + noun + suffix (14): *ovogodišnji* (ov-o-godišnji) ‘this year’s’; *svojeručan* (svoj-e-ručan) ‘handwritten, handmade’
   - pronoun + interfix + verb + suffix (11): *svojeglav* (svoj-e-glav) ‘headstrong’; *samodopadan* (sam-o-dopadan) ‘complacent’
   - pronoun + interfix + adjective (6): *samodostatan* (sam-o-dostatan) ‘self-sufficient’; *samopouzdan* (sam-o-pouzdan) ‘self-confident’
   - pronoun + interfix + participle (1): *samonikao* (sam-o-nikao) ‘self-sown’
   - pronoun + interfix + participle + suffix (1): *samozvani* (sam-o-zvani) ‘self-proclaimed’

6. ***verb*** + *interfix* + *noun* + *suffix* (4):

   - *buljook* (bulj-o-ok) ‘bug-eyed’; *vrtoglav* (vrt-o-glav) ‘dizzy, vertiginous’

The affixes used in compound adjectives are interfixes and suffixes. The interfixes are: *-o-, -e-, -i-, -u-*. The suffixes are: *-(a)n, -ani, -az(a)n, -en, -iv, -ljiv, -nati, -ni, -nji, -ovni, -ovski, -ski, -ø*.

4.3 Verbs

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

1. ***pure compounding*** (13):

   - adverb + verb (7): *krivotvoriti* (krivo-tvoriti) ‘to forge’; *zlouporabiti* (zlo-uporabiti) ‘to abuse’
   - noun + interfix + verb (5): *rukovoditi* (ruk-o-voditi) ‘to manage’; *hodočastiti* (hod-o-častiti) ‘to go on a pilgrimage’
   - pronoun + interfix + verb (1): *samoupravljati* (sam-o-upravljati) ‘to self-govern’

2. ***prefixation + compounding*** (4):

   - prefix + adverb + verb (3): *odugovlačiti* (o-dugo-vlačiti) ‘to procrastinate’; *omalovažiti* (o-malo-važiti) ‘to belittle’

- prefix + adjective + interfix + verb (1): *oživotvoriti* (o-živ-o-tvoriti) 'to bring to life'

3. ***prefixation + compounding + suffixation*** (4):
   - prefix + adjective + interfix + noun + suffix (3): *udobrovoljiti* (u-dobr-o-voljiti) 'to appease'; *odobrovoljiti* (o-dobr-o-voljiti) 'to appease'
   - prefix + noun + interfix + noun + suffix (1): *ovjekovječiti* (o-vjek-o-vječiti) 'to immortalize'
4. ***suffixation + compounding*** (1):
   - num-card + interfix + noun + suffix + se (1): *dvoumiti se* (dv-o-umiti se) 'to be in doubt'

4.4 Adverbs

So far, we have only 4 compound adverbs on our list. We have noted three word-formation patterns. The interfix is *-o-*, and the suffix is *-ke*.

- numeral + interfix + noun + suffix (2): *četveronoške* (četver-o-noške) 'on all fours'; *dvonoške* (dv-o-noške) 'on two legs'
- adverb + preposition (1): *maloprije* (malo-prije) 'a short while ago'
- adverb + noun (1): *sutradan* (sutra-dan) 'the next day'

From the main types of word-formation patterns shown above, it is evident that in compounds where the adverb is the first element, we do not record an interfix. We will briefly explain that decision in Section 5. Before that, we will present the structure of the lexical entries in CroComp.

4.5 Lexical entries in CroComp

An example of an entry in CroComp that contains all listed elements is shown in Figure 2. It illustrates the analysis of the noun *pismonoša* 'postman'. Lexical entries in CroComp are structured as follows:

Headword The entry begins with the headword. Each headword is accompanied by information about the part of speech (POS). Additional information for headwords includes *gender* (masculine/feminine/neuter) and *number* (singular/plural) for nouns, *definiteness* (definite/indefinite) for adjectives, and *aspect* (perfective/imperfective/biaspectual) for verbs. We also specify whether it is a compound or a fusion/coalescence (cf. Section 3).

Elements In the elements section, we list the words that served as the bases for forming the compound, as well as the affixes used in the process. Here, we also specify the POS of each base and the type of affix.

POS/affix In the POS/affix section, we list the word-formation pattern as described above in Subsections 4.1–4.4. Here we do not state the type of affix, but rather specify the affixes used in the process.

WF (Word-formation In the WF section, we specify the exact bases and affixes involved in the formation of the compound. Brackets are used here to describe the alternations that occur to the bases and affixes in the word-formation process. In some lexical entries, we provide additional information about morpho-phonological alternations or word-formation processes.

SS (surface structure) and DS (deep structure) The entry concludes with the morphological segmentation of the compound. In the SS section, we show the segmentation at the surface level, while DS displays their morphological structure at the deep layer.

The lexicon is compiled in the online dictionary writing system Lexonomy (Měchura, 2017). CroComp is publicly available and can be viewed at: https://lexonomy.zzl.ffzg.unizg.hr/twgi63yh.

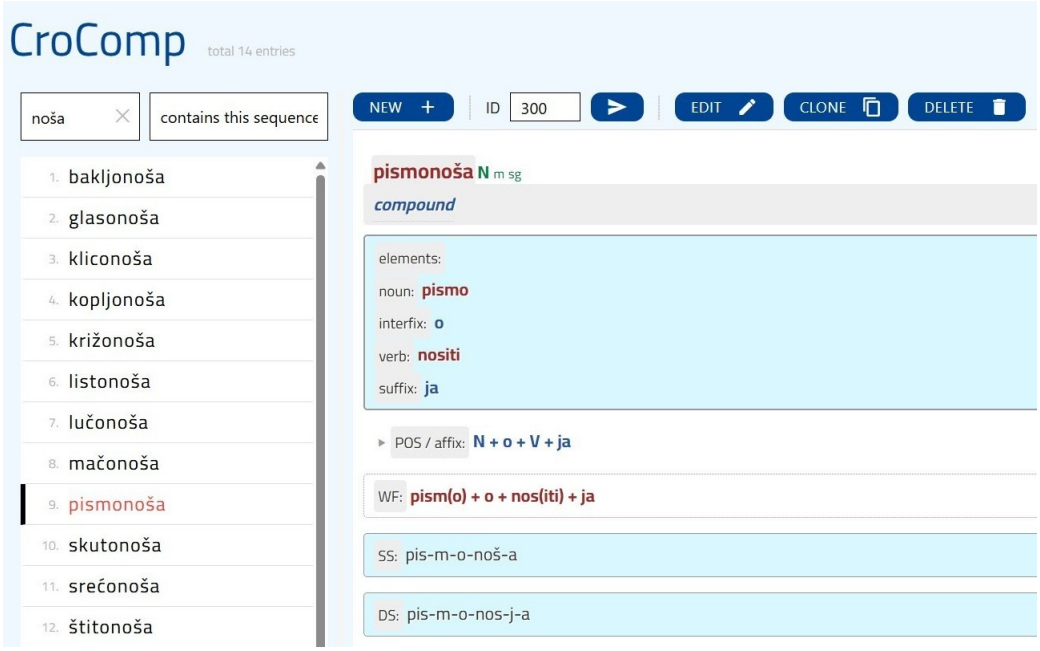


Figure 2: The lexical entry for *pismonoša* in the CroComp search interface

5 Conclusion and future work

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

In some word-formation patterns, we do not assume the presence of an interfix. This applies to the *verb + noun* pattern in the case of noun formation and to all combinations where an adverb appears as the first element. The first group includes compounds often referred to as *imperative compounds*, although the verb form in many cases cannot be equated with the imperative form. This group requires special attention and further elaboration in future work. The second group includes all combinations in which the adverb is the first component. Due to the specific nature of this part of speech and its semantic role in the phrases from which such compounds have developed, we do not assume the use of an interfix in these combinations. In many cases, interpreting the first element as an adverb is challenging, as adverbs are often homographic with adjectives.

As expected, our analysis confirmed that compounding is most productive in the formation of nouns and adjectives, with a wide variety of patterns and affix combinations, whereas verbal and adverbial compounds are significantly less frequent. However, this approach clearly illustrates the word-formation patterns and the corresponding groups of words derived from them.

In future work, we plan to expand CroComp in several directions. First, we aim to increase the number of entries by incorporating additional compounds from corpora and dictionaries. Second, we intend to integrate CroComp more closely with CroDeriv, enabling cross-referencing between derivational and compound structures. Finally, we hope to explore the application of CroComp in natural language processing tasks, such as automatic morphological analysis. CroComp is a valuable resource for both theoretical linguistic research and practical computational applications, offering a detailed and structured insight into the compounding processes of the Croatian language.

Acknowledgments

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

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.

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.

Mario Grčević. 2016. Croatian. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe*, De Gruyter Mouton, Berlin / Boston, volume 4, pages 2998–3016.

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.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Pariz, pages 4264–4270. https://aclanthology.org/L16-1000/.

Ivan Marković. 2012. *Uvod u Jezičnu Morfologiju*. Number 6 in Biblioteka Thesaurus. Disput, Zagreb. OCLC: 815718585.

Michal Měchura. 2017. Introducing Lexonomy: An open-source dictionary writings and publishing system. In Iztok Kosem, Jelena Kallas, Carole Tiberius, Simon Krek, Miloš Jakubíček, and Vít Baisa, editors, *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference*. Lexical Computing, Leiden, pages 662–667. https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf.

Josip Silić and Ivo Pranjković. 2005. *Gramatika Hrvatskoga Jezika: Za Gimnazije i Visoka Učilišta*. Školska Knjiga, Zagreb. OCLC: ocm70847560.

Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian morphological lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*. ACL, Budimpešta, pages 41–46. https://aclanthology.org/W03-2906/.

Branka Tafra and Petra Košutar. 2009. Rječotvorni modeli u hrvatskom jeziku. *Suvremena Lingvistika* 35(67). https://doi.org/811.163.42'37.

Jan Šnajder. 2014. DERIVBASE.HR: A high-coverage derivational morphology resource for Croatian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. ELRA, Reykjavik, pages 3371–3377. https://aclanthology.org/L14-1068/.

Krešimir Šojat and Matea Filko. 2023. Processing Croatian morphology: Roots, segmentation, and derivational families. In Matea Filko and Krešimir Šojat, editors, *Proceedings of the Fourth International Workshop on Resources and Tools for Derivational Morphology*. Hrvatsko Društvo za Jezične Tehnologije, Zagreb, pages 61–70. https://derimo.ffzg.unizg.hr/media/uploads/proceedings/derimo2023.pdf.

University of Fribourg
Avenue de l'Europe 20
CH-1700 Fribourg
Switzerland

https://events.unifr.ch/derimo2025/en/