ZW
JW
Zeitschrift für Wortbildung
Journal of Word Formation

2025, 9(1), 24–71
DOI: 10.21248/zwjw.2025.1.132

Ivan Lacić

Competing Strategies in Morphological Approximation: Exploring Prefixoids *kvazi*(-), *nadri*(-), *nazovi*(-), and *pseudo*(-) in Croatian[*]

Abstract: This study explores the phenomenon of affix rivalry within the domain of morphological approximation in Croatian, focusing on the prefixoids *kvazi*(-), *nadri*(-), *nazovi*(-), and *pseudo*(-) as they attach to nominal bases. These prefixoids can be classified as privative, as the derivatives they produce do not fully embody the core characteristics conveyed by their morphological bases. To analyze the rivalry among the prefixoids, the study evaluates their productivity, collocational behavior, and distribution across various textual genres, utilizing data from the *CLASSLA-web.hr* corpus. The findings suggest significant disparities in the productivity and collocational behavior of the prefixoids, with *nazovi*(-) and *kvazi*(-) exhibiting the highest productivity and highly overlapping collocational behavior, whereas *pseudo*(-) and *nadri*(-) reveal more specialized usage patterns. Additionally, a random sample of 500 tokens per prefixoid is annotated for semantic values. Again, *nazovi*(-) and *kvazi*(-) demonstrate substantial overlap, particularly in their mutual application as means for subjective depreciative evaluation, underscoring the insufficiency or pretentiousness of the subject. *Nadri*(-) is more narrowly focused on legal domains, while *pseudo*(-), with its proclivity for scientific contexts, remains distinct but conceptually adjacent to *kvazi*(-) in contexts where imitation is highlighted without necessarily invoking deceit. Overall, the prefixoids present a complex network of interrelationships, yet each prefixoid also establishes a specific niche, balancing between shared semantic roles and distinct, context-dependent uses.

Keywords: affix rivalry, approximation, Croatian, evaluative morphology, prefixoids

1. Introduction

1.1. Approximation in Morphology

Morphological approximation represents a relatively underexplored area within the broader category of evaluative morphology. Evaluative morphology encompasses a diverse range of constructions that serve various semantic functions, all of which pertain to the

[*] I wish to acknowledge Jurica Polančec for his assistance with the preliminary data extraction. I would also like to thank two anonymous reviewers for their valuable comments. Any remaining errors are, of course, my own.

concept of linguistic subjectivity. Subjectivity is central to evaluation, seen as “a mental operation assessing the value of an object or event as more or less desirable and important from the interpreter’s perspective” (Merlini Barbaresi 2015: 38). Here, the speaker makes a judgment about value rather than stating factual information. In evaluative contexts, the standard of comparison is a mental construct, an abstract representation shaped by cultural and/or social factors and commonly shared within a community. As such, it is expected that different speakers may evaluate the same object, action, or person in various ways (Grandi & Körvélyessy 2015; Grandi 2017). While functions such as diminution, augmentation, and intensification have been extensively studied (cf., *inter alia*, Dressler & Merlini Barbaresi 1994; Grandi 2002; Körtvélyessy & Štekauer 2011; Napoli & Ravetto 2017; for Croatian, although primarily on non-morphological means of intensification, see Batinić, Kresić & Pavić Pintarić 2015; Nigoević 2020; Lacić 2024a,b), morphological approximation has only recently begun to receive attention (Amiot & Stosic 2022; Masini, Norde & Van Goethem 2023). The relative paucity of research on approximation within morphological studies is highlighted by the variety of terms and lack of consensus on their definitions. For instance, Grandi & Körtvélyessy (2015) describe this functional domain using the triad *approximation*/*reduction*/*attenuation*, while other terms include *deintensification* (Körtvélyessy 2015), *non-prototypicality* (Cúneo 2015), as well as *non-authenticity*, *fakeness*, or *imitation* (Masini & Micheli 2020). These varying terms underscore the core concept of approximation as a comparison or resemblance to a concept based on specific properties. Crucially, approximative formations convey that category X, despite its similarities, is ultimately not fully present as it is lacking one or more of its prototypical features. As outlined by Masini, Norde & Van Goethem (2023), sources of approximation values are numerous, spanning over negation items, degree and quantity items, diminutives, similative items, taxonomic items, etc. For this study, of a particular interest are markers conveying meanings of fakeness/imitation/pretending, most frequently expressed by *pseudo*(-), derived from Ancient Greek *pseudēs* ‘false’. Such markers are characterized by their “privative” reading, implying that “(a) fake X is not (an) X” (Capelle, Denis & Keller 2018: 9). For instance, a *pseudodoctor* lacks the qualifications of a genuine doctor, even though it may share certain traits with them. Markers of fakeness/imitation/pretending, both cross-linguistically and intra-linguistically, are numerous (e.g., Van Goethem & Norde (2020) explore eight Dutch “fake” morphemes), and the situation is not different for Croatian. In addition to the aforementioned *pseudo*(-), Croatian uses two near-synonymous prefixoids of Slavic origin, namely *nadri*(-) (< IMP.2SG of verb *nadrijeti* ‘to begin’[1]) and *nazovi*(-) (< IMP.2SG of verb *nazvati* ‘to name, to call’). Additionally, it is possible to join to that group also the prefixoid *kvazi*(-) (lat. *quasi* ‘almost’), which primarily conveys a meaning of incompleteness, in line with the prototypical approximation reading. However, we argue (see §3.4) that in Croatian it can also convey the concept of fakeness, just like the aforementioned three prefixoids. With four prefixoids occupying overlapping semantic niches, it becomes apparent that defining specific values of these prefixoids can be challenging due to their capacity to convey multiple, often similar meanings, resulting in affix rivalry. At its most basic, rivalry refers to “the fact that speakers routinely have to make a choice between alternative ways of realizing a certain concept” (Gardani, Rainer & Luschützky 2019: 4). This redundancy prompts linguistic systems to resolve competition through either specialization, akin to Aronoff’s (2019) *habitat niche differentiation*, or the elimination of a form altogether (Bauer, Valera & Díaz-Negrillo 2010). However, as Nagano, Bagasheva & Renner (2024: 3) point out, in word-formation the competition often simply continues since “(i) coexistence rather than disappearance is commonly observed, and (ii) the form of specialization tends to deviate from the elsewhere distribution [cf. Aronoff 2023].”

To the best of our knowledge, no contemporary studies have specifically addressed either the broader concept of morphological approximation in Croatian nor the competitive dynamics among rival forms. This study seeks to address this gap by examining four rival Croatian prefixoids prone to privative use, viz. *kvazi*(-), *nadri*(-), *nazovi*(-), and *pseudo*(-), as a member of <PREF$_{\text{APPRX}}$ + noun> construction as exemplified by (examples taken from CLASSLA-web.hr):

(1) *Mi vodimo svoje male i bezvrjedne živote*, [...] *a cijela armija umišljenih hrvatskih* ***kvaziintelektualaca*** *dobro i predobro živi na naš račun*.

‘We lead our small and worthless lives, [...] while a whole army of pretentious Croatian pseudo-intellectuals (lit. KVAZIintellectuals) lives well and too well at our expense.’

[1] It is noteworthy that in contemporary Croatian the verb *nadrijeti* is mostly used with a meaning of *navaliti*, *nahrupiti*, ‘rush in, come/enter suddenly’ (Šonje 2000).

(2) *No, utjecaj ljevičarskih* ***nadri-intelektualaca*** *na svakodnevnu politiku sve je jači i na Zapadu* [...].

‘However, the influence of the left-wing pseudo-intellectuals (lit. NADRIintellectuals) on everyday politics is growing stronger in the West as well [...]’

(3) *Upravo zato počeli smo pisati Studentsku deklaraciju* [...] *kako bi svijet čuo i glas hrvatskih studenata, a ne samo probranih,* ***nazovi intelektualaca****.*

‘This is precisely why we started writing the Student Declaration [...], so that the world could hear the voice of Croatian students, not just that of the selected, pseudo-intellectuals (lit. NAZOVI intellectuals).’

(4) *Dok su ukus moderne umjetnosti sve više usmjeravali* ***pseudo intelektualci****, dotle se naivna umjetnost* [...] *razvijala sama.*

‘While the taste of modern art was increasingly directed by pseudo-intellectuals, naive art [...] developed independently.’

Naturally, approximation in Croatian can also be conveyed by other morphological processes (e.g., suffixation with -*ić* as in *doktorčić* ‘doctor.DIM’, which, among others, can have a reading synonymous to one with *nadri*(-) or *nazovi*(-)) or by syntactic means (e.g., adjective *tobožnji* ‘so-called/supposed’, as in *tobožnji liječnik*, ‘a supposed doctor’). This study, however, focuses on the competition among the four prefixoids discussed, which can be seamlessly integrated into the same syntactic structures without necessitating modifications in sentence constructions.

1.2. Previous Accounts

As previously noted, reference works concerning morphological approximation in Croatian are non-existent and beside basic dictionary descriptions of the four prefixoids, little is known. As far as dictionaries are concerned, two most comprehensive dictionaries were examined, viz. the *Dictionary of the Croatian language* (Šonje 2000) and VRH – *Large Dictionary of Croatian Standard Language* (Jojić 2015). Furthermore, a database *Croatian LanguagePortal*[2] (HJP), encompassing data from several Croatian dictionaries, was consulted. In Šonje (2000), there are no mentions of *kvazi*(-), while for *pseudo*(-) six entries are listed, but only *pseudoklasicizam* ‘pseudoclasicism’ is relevant for this analysis, as the others pertain to scientific names of animal species. The prefixoid *nadri*(-) is represented by

[2] Available at: https://hjp.znanje.hr/ (accessed 19 January 2025).

four lemmas: *nadriliječništvo* ‘quackery’, *nadriliječnik* ‘quack’, *nadriobrtnik* ‘quack craftsman’, and *nadripisar* ‘quack scribe’. All three profession denoting nouns describe individuals who untruthfully perform an activity for which they do not have the necessary education and competence. Along with the three profession denoting nouns, synonymic versions with *nazovi*(-) are reported. Finally, for *nazovi*(-), Šonje (2000) reports only *nazoviliječništvo* ‘quackery’ and *nazovipisarstvo* ‘scribal quackery’, with no *nadri*(-) formations listed as synonyms. In contrast, VRH (Jojić 2015) includes entries for all four prefixoids. For*kvazi*(-), we find prefixoids *pseudo*(-), *nadri*(-), and *nazovi*(-) mentioned as synonyms, but also adverbs *gotovo* ‘almost’ and *skoro* ‘nearly’. The following formations are found: *kvazijunak* ‘pseudo-hero’, *kvaziliteratura* ‘pseudo-literature’, and *kvaziliteraran* ‘pseudo-literary’. For *pseudo*(-), no synonymic affixes are listed but we do find two nouns, viz. *pseudocivilizacija* ‘pseudo-civilization’ and *pseudočinjenica* ‘pseudo-fact’, as well as the adjective *pseudocivilizacijski* ‘regarding a pseudo-civilization’. All three definitions stress the fakeness of the modified head. Moreover, for *nadri*(-), we find *nazovi*(-) as a synonymous formant, along with profession denoting nouns *nadriliječnik* ‘quack’, *nadripisar* ‘quack scribe’, *nadriobrtnik* ‘quack craftsman’, *nadripjesnik* ‘quack poet’ and *nadriumjetnik* ‘quack artist’, with the respective feminine pairs and derived relational adjectives, such as *nadriliječnički* ‘regarding a quack doctor’ and *nadripisarski* ‘regarding a quack scribe’. Furthermore, for *nazovi*(-), VRH reports nouns *nazovijunak* ‘psuedo-hero’, *nazoviliječnik* ‘pseudo-doctor’, *nazoviliteratura* ‘pseudo-literature’, *nazoviobrtnik* ‘pseudo-craftsman’, *nazovipisar* ‘pseudo-scribe’, and, lastly, *nazovipjesnik* ‘pseudo-poet’. The final resource examined, the *Croatian Language Portal* (HJP), includes entries for all four prefixoids, but does not provide additional insight into the distinctions between them. Similar to VRH, *kvazi*(-) is explained with the prefixoids *pseudo*(-), *nadri*(-), and *nazovi*(-), as well as adverbs *gotovo* ‘almost’ and *skoro* ‘nearly’, *tobože* ‘allegedly’ and *navodno* ‘allegedly’. Regarding *pseudo*(-), the concept of fakeness and falsity is emphasized, and the example *pseudodemokracija* ‘pseudo-democracy’ is provided. The prefixoids, *nadri*(-), *nazovi*(-), *lažno*(-) ‘fake’, and *krivo*(-) ‘lit. wrong’ are listed as synonyms. For *nadri*(-), HJP reports *nazovi*(-) as a synonymous formant, along with examples such as *nadriliječnik* ‘quack’ and *nadriknjiga* ‘quack book’. *Nadri*(-) is defined as conveying concepts of allegation and fakeness, equal to those activated by *nazovi*(-), for which *nadri*(-) is listed as a synonymous prefixoid. An example of *nazovi*(-) is given with the noun *nazoviprijatelj* 'so-called friend'. In conclusion, the existing dictionary descriptions of these prefixoids, where available, are often circular and unhelpful for disambiguation. Each prefixoid is frequently defined using another synonymous prefixoid, reinforcing the shared semantics among them rather than clarifying their individual meanings.

Outside of dictionaries, the primary discussions surrounding these prefixoids are limited to debates on the categorical status of the formations they produce. The earliest available study on this topic, concerning *nadri*(-) is a work by Stjepan Ivšić (1906–1907) entitled *Nešto o riječima složenima s* nadri- [Something about words composed with *nadri*-]. Ivšić (1906–1907) observes that formations with *nadri*(-) were created analogously to the term *nadriknjiga*, denoting a person that has only superficially engaged with a book, thus referring to someone with a minimal grasp of a subject. By analogy, states Ivšić, the term *nadriliječnik* 'quack' was expected to indicate a doctor who is just starting their education/practice, but it does not mean that. Instead, the element *nadri*(-) is used in the sense of fake, self-proclaimed, with *nadriliječnik* 'quack' meaning 'a person without the necessary schooling and professional qualification who performs the duties of a doctor' (Šonje 2000). For that motive, Ivšić suggests that formations with *nadri*(-) should be replaced by *nazovi*(-) (e.g., *nazoviliječnik* 'quack'). However, as later pointed out by Barić (1979, 1980), Ivšić has missed to notice that the meaning of the formations with *nadri*(-) and *nazovi*(-) is equivalent and *nazovi*(-) does not provide a more accurate alternative, as it does not preserve the original meaning. Barić (1980) concludes that *nadri*(-) and *nazovi*(-) act as modifiers of the word they attach to, imparting a meaning of fakeness, pretenses, and classifies them as prefixes. Finally, Barić also mentions *laži*(-) (IMP.2SG of verb *lagati* 'to lie'), but noted its rarity compared to the others[3]. Returning to *nadri*(-), Klajn (2002) concurs with Ivšić (1906–1907) that today

[3] A research based on several contemporary Croatian dictionaries revealed only formations *lažitorba* (lit. LAŽIbag) 'one who lies a lot', and *lažidoktor* (lit. LAŽIdoctor) 'a fake doctor, a person pretending to be a doctor' (Jojić 2015). However, a brief search in the CLASSLA.web-hr corpus revealed over ten nominal formations with *laži*-, as well as two adjectival ones. Since to the best of our knowledge there is no previous work devoted to *laži*-, we deemed useful to report the found formations in a hope they will motivate some future research. Found nominal formations, besides the aforementioned *lažitorba* and *lažidoktor*, are: *lažibogomolja* 'fake place of worship', *lažibumbar* 'fake bumblebee', *lažijezik* 'fake language' (here it is curious to notice that *laži* in the formation *politički lažijezik* 'political fake language' keeps its original meaning and, in that given context, the formation refers to the communication full of lies, not to a fake,

nadri(-) is completely opaque and has lost every connection with the verb *nadrijeti* ‘to begin’ from which it is derived. Like Barić (1980), Klajn emphasizes that *nadri*(-) functions as a prefix rather than a component of a compound, paralleling the usage of borrowed prefixes such as *pseudo*(-) and *kvazi*(-). On the other hand, discussing the categorical status of elements with *nazovi*(-), Klajn (2002) states that, since the connection with the verb *nazvati* ‘to name, to call’ is completely transparent, words like *nazoviprijatelj* ‘pseudo-friend’ should be classified as compounds.

In summary, the review of existing research on Croatian prefixoids reveals a predominant focus on the two Slavic prefixoids, with earlier studies primarily concerned with their morphological status. These studies have largely neglected the semantic properties and competitive dynamics involving these prefixoids, both amongst themselves and in relation to non-native forms like *kvazi*(-) and *pseudo*(-). In contrast, English-language studies provide some comparative insights into *quasi*(-) and *pseudo*(-). Bauer, Lieber & Plag (2013) argue that while both elements share the characteristic of forming derivatives that do not represent genuine exemplars of their categories, their distinction lies in the element of falseness, which is inherent in *pseudo*(-) but absent in *quasi*(-). Similarly, Dixon (2014) contends that although the two prefixes appear superficially similar, they exhibit significant semantic differences: *quasi*-X denotes something that possesses some characteristics of X but is not a full X, while *pseudo*-X refers to something pretending to be like X or resembling X without being X. Consequently, while *quasi*(-) is associated with the lack of a key feature, aligning it with approximation, *pseudo*(-) conveys falseness or imitation (a value defined as *disproximation* by Cappelle, Daugs & Hartmann 2023).

1.3. Adopted Approach and Scope of the Paper

The study adopts an onomasiological approach to affix rivalry, focusing on how four different forms vie to express a broader, approximative meaning (many-to-one relationship). Subsequently, the semasiological perspective is also considered, examining the multiple meanings associated with individual prefixoids (one-to-many relationship) (cf. Nagano,

so-called language), *lažipauk* (name for the species Opiliones), *lažipčela* ‘fake bee’, *lažisvetac* ‘fake saint’, *lažištipavac* (name for the species *Pseudoscorpiones*), *lažiučenje* ‘fake teaching’, *laži-iluzija* ‘fake illusion’, *laži-mit* ‘fake myth’, and *laži-španjolac* ‘fake Spanish person’. The two adjectival formations are *lažidesni* ‘fake right’ and *lažisvjetski* ‘fake worldwide’.

Bagasheva & Renner 2024). This paper has several objectives. First, grounded in the understanding that base selection has traditionally been viewed as closely tied to the productivity of each affix (Plag 1999; Bauer 2001), it examines the productivity of the four prefixoids. Second, it seeks to examine the differences in collocational preferences of the prefixoids to evaluate the extent of their shared/diverging conceptual content. Third, based on the corpus examples, it aims to provide a description of the semantic values conveyed by each prefixoid. In doing so, the study aims to contribute to the better understanding of the understudied category of morphological approximation in Croatian. Given the lack of prior studies focused on identifying the factors influencing the choice between the examined prefixoids, whether categorically or tendentially, the factors included in this work are based on their presumed relevance and the feasibility of their extraction from the corpus.

The paper is organized as follows. The next subsection (§1.2) briefly reviews previous account on morphological approximation in Croatian. §2 outlines the methodology. In §3, results of the analyses are presented. Finally, §4 provides a discussion of the key findings, draws conclusions and suggest possible directions of future research on this topic.

2. Data and Method

The analysis draws on data from the recently released CLASSLA-web.hr corpus (Ljubešić & Kuzman 2024; Ljubešić, Rupnik & Kuzman 2024), a contemporary Croatian web corpus consisting of nearly 2.2 billion words. This corpus, available for download, is particularly valuable due to its recency, as it primarily comprises data from the *.hr* internet top-level domain collected in 2021 and 2022, ensuring relevant and up-to-date data for synchronic linguistic research. For each of the four approximative prefixoids under investigation, CLASSLA corpus has been queried for <PREF$_{\text{APPRX}}$> nominal[4] ngrams as well for nominal formations that start with <PREF$_{\text{APPRX}}$-> and <PREF$_{\text{APPRX}}$>, in order to account for constructions in three orthographic variants, i.e. block use (univerbation) (e.g., *nazoviprijatelj* 'pseudo-friend'), hyphenated use (e.g., *nazovi-prijatelj*), and separated use (juxtaposition) (e.g., *nazovi prijatelj*). While this procedure yielded high recall for most of the prefixoids,

[4] A brief corpus search revealed that nominal formations are by far the most frequent. Nonetheless, the analysis of non-nominal formations may be useful in order to assess whether the prefixoids exhibit any categorial constraints or preferences.

manual verification was still required, particularly for *nazovi*(-). As previously noted, Slavic prefixoids *nadri*(-) and *nazovi*(-) originate from the 2nd person imperative forms of the verbs *nadrijeti* 'to begin' and *nazvati* 'to name, to call'. While *nadrijeti* is extremely rare (with only two attestations in the reference corpus), *nazvati* is quite frequent, with 265,980 occurrences. This presented a significant challenge in extracting relevant instances, as many of the examples of debonded forms of *nazovi*(-) were actually imperative constructions, resulting in numerous false positives, as illustrated in (5):

(5) *Mislim da ti je jedini način podmirit dug ili* ***nazovi banku*** *i pitaj dali to mogu prolongirat.*

'I think the only way is to settle the debt or call the bank and ask if they can extend it.'

Additionally, only formations with a base that appears at least 10 times in the corpus were included. This threshold was set after observing that bases with lower frequencies were often either non-existent or orthographically compromised.

The dimensions of the final dataset are illustrated in Table 1.

Tab. 1: Size of the dataset

CLASSLA corpus	Tokens	Types
	21,353	4,270

The analysis is structured in two main steps and comprises a quantitative and qualitative analysis of the data. The data was processed using *R* (version 4.4.1) programming language (R Team 2022). In recent years, quantitative approaches to rivalry have emerged, primarily aimed at investigating the discriminative properties of competing affixes. Consequentially, a range of statistical methods has been employed to assess the impact of (non-)structural factors in resolving rivalry (Huyghe & Varvara 2023; Salvadori, Varvara & Huyghe 2024).

In this study, first, an in-depth analysis of the prefixoids' productivity was conducted. Various approaches have been proposed for quantitatively evaluating productivity, all operating under the premise that a process is more productive when it generates a greater number of lexemes (Fernández-Domínguez 2013). To assess the vocabulary size of the four prefixoids, i.e. the number of types they form in relation to the increasing number of tokens generated by the process, the study first examines the prefixoids' vocabulary growth curves (VGC). Furthermore, recognizing the importance of considering multiple variables, rather than relying solely on a single measure of productivity (Hartmann 2018), three measures of productivity were calculated – Moving-Average Type-Token Ratio (MATTR) (Covington & McFall 2010), Shannon entropy (see, e.g., Hein & Brunner 2020; Evert & Baroni 2022), and Potential Productivity *P* (Baayen 1992, 2009) – each capturing different dimensions of the phenomenon.

MATTR (Covington & McFall 2010) is a sample-size independent variant of the well-known Type-Token Ratio (TTR). MATTR employs a moving window approach to estimate TTR values for each successive window of fixed length. Initially, a window length is chosen, such as 100 words, and the TTR for words 1–100 is computed. The TTR is then calculated for words 2–101, followed by 3–102, and so on, until the end of the dataset. The final score is obtained by averaging all the estimated TTR values. Unlike TTR, MATTR final score is considered unaffected by text length or any statistical assumptions.

In recent years, an information-theoretic approach has been considered and the connection between entropy (Shannon 1948) and productivity in linguistic samples has been observed[5] (Sundquist 2020). Entropy can be understood as a measure of the uniformity of a probability distribution, or, in other words, a measure of uncertainty. It is calculated from a sample of tokens all derived with a specific morpheme. Initially, we tally the occurrences of each type in the sample, resulting in a type frequency distribution. Next, we convert the type frequency distribution to a probability distribution using maximum likelihood estimation. Finally, we compute the entropy of this probability distribution, measured in bits. Higher entropy indicates that tokens are more evenly distributed among types, which suggests greater productivity. Conversely, low entropy means that most attestations correspond to a few highly frequent types, indicating lower productivity (Barðdal et al. in prep.).

Finally, Potential Productivity *P* (Baayen 1992, 2009) refers to the relation between the number of hapax legomena formed with an analyzed affix in a sufficiently large corpus and the total number of tokens of that affix in the corpus (Gaeta & Ricca 2015). While *P* is a time-honored and widely-applied measure, it suffers from a well-known methodological

[5] As noted by Gries (2015), entropy is also related to contextual distinctiveness/diversity and learning. For instance, Goldberg et al. (2004) in their work on argument structure generalizations demonstrate that a more skewed distribution (with higher entropy) is learned more effectively than a more balanced one (with lower entropy).

problem – its high sensitivity to sample size. When sample size differs (and it is to be expected that when comparing more morphemes one will be more frequent than another), that difference will affect the productivity measures calculated for those samples, rendering their comparison not interpretable[6]. To obtain a methodologically sound comparison of Potential Productivity values, the token count for individual patterns must be standardized (Gaeta & Ricca 2006). To address this issue, a method known as bootstrapping was applied. From each of the four samples, the same fixed number of tokens was picked randomly. It was determined that a sample size of 1200 would be used, ensuring repeated sampling without significant overlap for the least frequent prefixoid, namely *nadri*(-), with 1714 tokens. With all groups now standardized to the same size, MATTR, entropy, and Potential Productivity *P* were calculated for each group, and this process was iterated 500 times, each time selecting new random samples with the *with replacement* function. This repetition was crucial for generating more reliable results. By executing this process 500 times, we can ensure a more robust representation of the data, enhancing the validity of the analyses.

After analyzing the morphological productivity of the prefixoids, we shift our focus to a sociolinguistic variable, namely text genre in which the formation appears in, by checking the *genre* tag available in CLASSLA corpus.

Finally, we examine the prefixoids' semantic properties. To examine their distributional/ collocational preferences, Multiple Distinctive Collexeme Analysis (MDCA), a method within the broader framework of collostructional analysis (Stefanowitsch & Gries 2003; Gries & Stefanowitsch 2004) is applied. The results from MDCA serve as input for a multifactorial analysis technique known as Correspondence Analysis (CA) (Benzécri 1973; Greenacre 2017). The examinations draw on the distributional hypothesis and build upon the assumption that "the degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear"

[6] That is because *P* is conceptually related to the aforementioned VGCs, as it represents the slope at its endpoint, when the maximum number of tokens has been observed. Since the position on the curve is often unknown (for instance, whether we have observed the entire curve or only, for example, the first 20% due to a small sample size), comparing *P* values across divergent sample sizes is problematic (Evert & Lüdeling 2001).

(Lenci 2008: 3)[7]. The principle asserts that a correlation between distributional similarity, i.e. linguistic contexts in which an element is observed, and meaning similarity enables us to infer the latter from the former. The working hypothesis posits that an overlap in collocational preferences among *kvazi*(-), *nadri*(-), *nazovi*(-), and *pseudo*(-) would indicate a shared conceptual content, thus classifying them as near-synonyms. On the other hand, differing collocational patterns would indicate that these prefixoids impose distinct construals. To examine the hypothesized near-synonymy among the prefixoids, MDCA is employed to contrast the four near-synonymous constructions (<*kvazi* + noun>, <*nadri* + noun>, <*nazovi* + noun>, <*pseudo* + noun>). In this analysis, orthographic variations are not deemed relevant, and the input data for each construction is given by the sum of the frequencies of its three orthographic variants. MDCA filters out overlapping collocates and focuses solely on the nouns that are idiosyncratic to each prefixoid, enabling the classification of distinctive nouns based on their function and meaning, thereby providing a deeper understanding of the individual specificities of the four prefixoids. The results from MDCA are crucial for the subsequent Correspondence Analysis (CA). CA, a distance-based clustering technique, visually represents cross-tabulations on a two-dimensional plot, facilitating the mapping of correlations between lexical items. In this study, CA demonstrates how prefixoids imply specific construals based on the nouns they are associated with. The input for CA is derived from the cross-tabulation of frequencies of the 50 most distinctive collexemes of each of the four prefixoids, as identified by MDCA. These analytical methods collectively offer a comprehensive exploration of the relationship between the examined prefixoids, enriching our understanding of their semantic properties.

[7] When discussing distributional hypothesis, another approach to capturing the semantics of prefixes has to be acknowledged – vector space models (for previous accounts on approximative prefixes, cf. Van Goethem & Norde 2020; Cappelle, Daugs & Hartmann 2023). Pre-trained models for Croatian include word embeddings derived from the skip-gram model of fastText (Terčon & Ljubešić 2023), and a transformer model *BERTić* (Ljubešić & Lauc 2021). However, these embeddings are based on multiple corpora: the skip-gram model embeddings originate from *hrWaC* corpus (Ljubešić & Klubička 2016) and the *MaCoCu-hr* corpus (Bañón 2022), while BERTić is trained on a combination of ten datasets totaling over 8 billion tokens of written text in Bosnian, Croatian, Montenegrin, and Serbian (closely related languages classified under the same *hbs* identifier (Serbo-Croatian macrolanguage) by the ISO-693-3 standard). Due to time constraints, it was not feasible to train a new model exclusively on the CLASSLA dataset. To maintain methodological consistency with other analyses presented in this study, the exploration of Croatian approximative prefixes using vector space models is deferred to a future study.

As the last step of the semantic analysis, semantic values conveyed by the prefixoids are examined. The initial approach involved analyzing the 50 most distinctive collexemes of each prefixoid, as used for the Correspondence Analysis, under the assumption that these would be semantically representative of each prefixoid. However, as the annotation process progressed, two concerns arose. Firstly, it became apparent that a single lexical item may exhibit diverging meanings depending on its context and the register of use. Although the noun that the prefixoid combines with typically influences its reading, this influence is not always straightforward. For example, in (6), *pseudodokomentarac* 'pseudo-documentary' refers to a recognized film genre, devoid of any subjective connotation, whereas in (7) it carries a pejorating sense (further emphasized by the sentence context and the sarcastic portrayal of invited guests), suggesting that the speaker is ridiculing the entire situation.

(6) *Odlučivši se na znatno kraćenje radnje u odnosu na Shakespeareov predložak, Welles je projekt realizirao pune tri godine* [...] *o čemu je 1978. napravio* ***pseudodokumentarac*** *"Snimajući Otela".*

'Having decided on significantly shortening the plot compared to Shakespeare's template, Welles spent a full three years realizing the project [...], about which, in 1978, he made a pseudo-documentary "Filming Othello"'

(7) *Složit će i* ***pseudodokumentarac*** *o Novom valu unutar kojeg će probrani sa šalovima oko vrata* [...] *reći tri prigodne floskule.*

'He will also put together a pseudo-documentary about the New Wave, in which a selected few with scarves around their necks [...] will say three appropriate clichés.'

For this motive, annotating only the types was deemed insufficient. Additionally, it became clear that the 50 most distinctive collexemes did not adequately capture the full semantic spectrum of each prefixoid, as the range of meanings identified was more limited than the one observed in a random sample of derivatives. To provide a more comprehensive overview, a different approach was adopted. Given the large number of relevant tokens (21,353), annotating all tokens according to their specific semantic value was not feasible. Therefore, we opted to semantically annotate a random sample of 500 tokens per prefixoid, with the expectation that this sample size would be sufficient to represent the range of meanings each prefixoid can convey (cf., a.o., Masini & Micheli (2020) who annotated a sample of 219 tokens, and Micheli (2023) who worked with samples of 100 tokens).

3. Results

3.1. Morphological Productivity

Table 2 summarizes key characteristics of the analyzed data, including the number of tokens, types, and hapax legomena. The frequencies of nominal formations exhibit variability, with token counts ranging from 1,714 to 8,647 and type counts spanning from 277 to 2,533.

Type count represents one of the most straightforward measures of morphological productivity – the greater the number of types a morpheme generates, the more productive it is considered (Bauer 2001; Zeldes 2012) – and is seen as an estimation of the 'extent of use' or the 'profitability' of a morphological category (Corbin 1987). An examination of the results in Table 2 reveals that *kvazi*(-) forms the most types (2,533), followed by *nazovi*(-) (1,498), and *pseudo*(-) (1,446), while *nadri*(-) generates the fewest types, viz. 277, 10.94 times fewer than *kvazi*(-). Consequentially, *prima facie*, *kvazi*(-) appears to be the most productive prefixoid, whereas *nadri*(-) seems the least productive one. However, due to differences in sample sizes across prefixoids, comparing raw type counts is not meaningful.

Tab. 2: Statistical overview of prefix usage

PREFIXOID	Tokens	Types	Hapax legomena
kvazi(-)	8,647	2,533	1,557
nadri(-)	1,714	277	189
nazovi(-)	3,230	1,498	980
pseudo(-)	7,942	1,446	816

In addition to analyzing the number of types formed by each prefixoid, an initial understanding of their productivity can also be gained through the analysis of their vocabulary growth curves (VGCs). VGCs show the vocabulary size, i.e. the number of types, in relation to the increasing number of tokens generated by the examined four processes. Typically, a VGC displays a characteristic shape because the number of types observed for a given number of tokens associated with a particular prefixoid is a monotonically increasing function of the number of tokens: as more tokens are sampled, more types are observed. Initially, the curve rises fairly steeply, but, as more tokens are encountered, the rate at which new types are observed decreases, causing the growth of the curve to slow down. For morphological categories with a finite number of types, the curve eventually plateaus, indicating that no new types are observed beyond a certain token count. A relatively unproductive process typically displays a shallow or asymptotic VGC, where vocabulary growth stabilizes early, reflecting fewer instances of novel types (Baayen 2001; Evert & Lüdeling 2001).

Figure 1 illustrates the VGCs for the four examined prefixoids. Alongside the observed vocabulary curve, depicted as a dashed line, the plot also shows an interpolated growth curve. An empirical growth curve often appears irregular due to variations arising from the non-random distribution of words in a corpus (Baroni & Evert 2014). To achieve a smoother curve, binomial interpolation is used (Baayen 2001). Binomial interpolation utilizes a frequency spectrum to generate expected values of vocabulary size for a given sample size. The interpolated curve obtained using the *vgc.interp* function from *zipfR* (Evert & Baroni 2022) is presented as a solid line. The interpolated curve was evaluated using Root Mean Square Error (*RMSE*) and R^2. The R^2 values are high across all prefixoids, ranging from 0.993 to 0.999, while the *RMSE* values span from 6.070 to 14.684. Both values confirm that the interpolated curves are performing as desired.

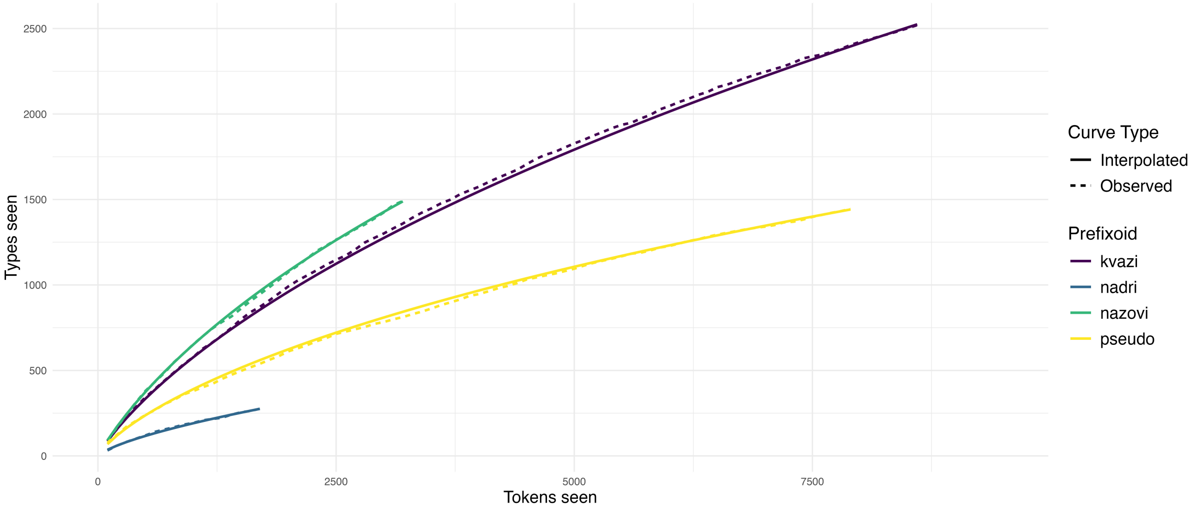

Fig. 1: Vocabulary growth curves for the four prefixoids

The VGCs for the four prefixoids reveal distinct patterns. First, the plot clearly adheres to the previously described general trend: most prefixoids exhibit a rapid initial growth in the number of types, which gradually slows as the number of tokens increases. The prefixoid *nazovi*(-) displays the steepest initial rise, indicating a quicker discovery of new types. On the other hand, *kvazi*(-) and *pseudo*(-) demonstrate a more moderate growth trajectory, while *nadri*(-) exhibits the flattest curve, indicative of its limited capacity to generate novel types.

Following these initial observations of productivity, a more detailed analysis was conducted using the three aforementioned productivity measures: Moving-Average Type-Token Ratio (MATTR), entropy, and Potential Productivity *P*. These measures were selected for their ability to capture different aspects of morphological productivity. While MATTR measures balance in usage, Potential Productivity estimates the likelihood of encountering a new type, and entropy can be seen as a measure of uncertainty (unpredictability) in the type-frequency distribution. To mitigate the impact of varying sample sizes, a standardized random sample of 1,200 tokens (70% of the least frequent prefixoid's token count, in order to allow repeated sampling with minimal overlap) was selected for each prefixoid. For each sample, MATTR (100-word window), entropy and Potential Productivity *P* were computed. Random sampling was repeated 500 times, with all three measures calculated for each iteration. The fluctuations in the values of these three productivity measures across the 500 randomly sampled 1,200-token sets are depicted in Figure 2.

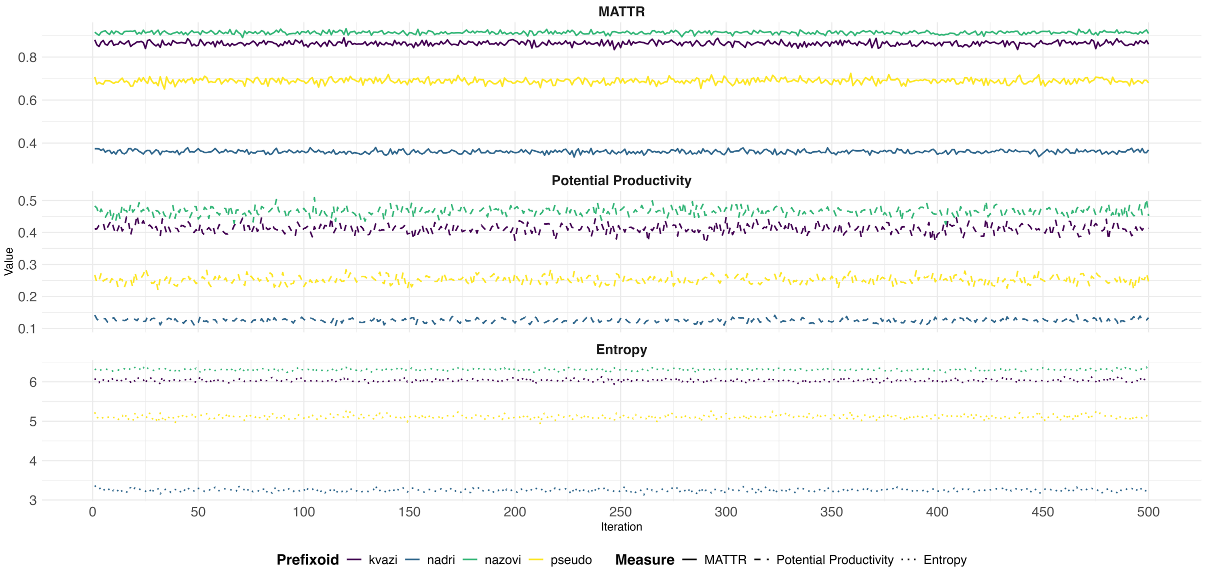

Fig. 2: Fluctuations of the productivity measures values over 500 fixed-size samples

The plot reveals that fluctuations are most pronounced for Potential Productivity values, moderately pronounced for MATTR, and least pronounced for entropy. In fact, min-max normalized Coefficient of Variation across the three measures for all prefixoids amounts to 0 for entropy, 0.152–0.277 for MATTR, and 1 for Potential Productivity, indicating it as the most volatile measure, very sensitive to changes in data composition. These findings validate the chosen method, i.e. bootstrapping with 500 fixed-size samples, while also highlighting the risk of relying on a single random sample, which could result in an "extreme" value rather than an average (representative) one. To facilitate a meaningful comparison of productivity across the prefixoids, Figure 3 presents the MATTR, Potential Productivity, and entropy results, with the median values displayed for each measure and each prefixoid[8]. This approach offers a balanced perspective, mitigating the influence of sample variability and providing a clearer insight into the relative productivity dynamics of each prefixoid.

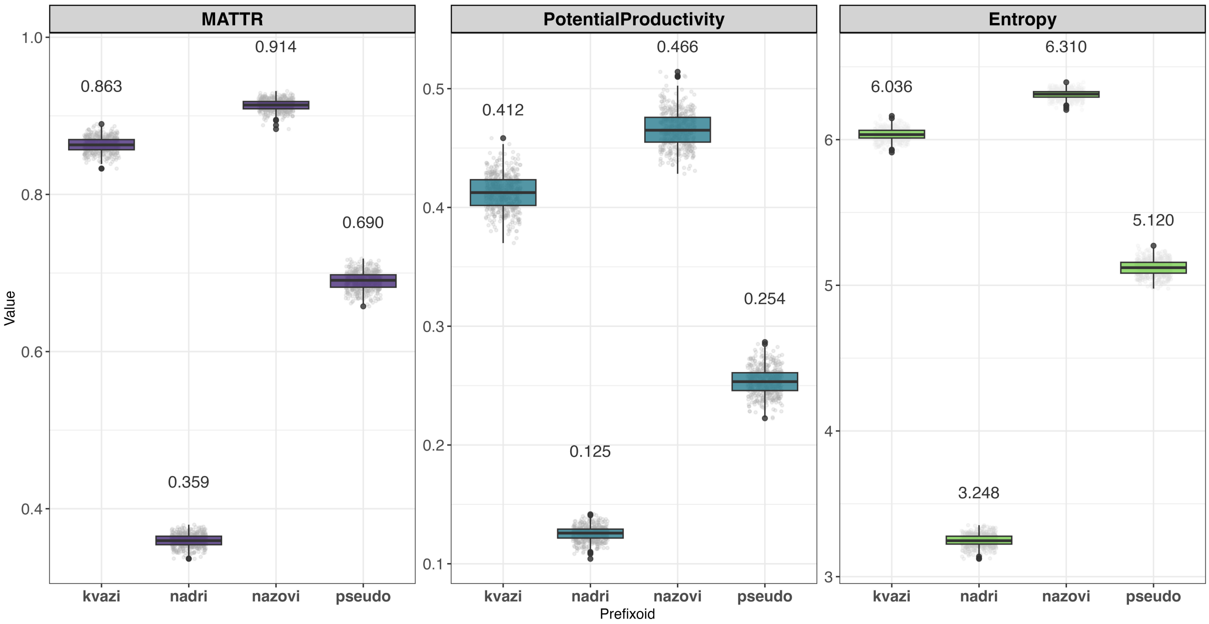

Fig. 3: MATTR, Potential Productivity, and entropy box plots including median values

The results of the fixed-size sample analysis are consistent and the rankings according to three productivity measures align perfectly (Pearson correlation coefficient spans from 0.967

[8] In order to statistically confirm that the observed differences between productivity measures were significant, since the assumption of homogeneity of variances was violated, the Kruskal-Wallis test was performed. The test revealed significant differences in entropy (χ^2 = 1874.1, p < 2.2e–16), MATTR (χ^2 = 1874.1, p < 2.2e–16), and Potential Productivity (χ^2 = 1869.3, p < 2.2e–16) across the four prefixes. Dunn's post-hoc tests with Bonferroni correction further confirmed that all pairwise comparisons between the prefixes were statistically significant ($p < 0.05$). This step ensured that the trends seen in the boxplots are statistically robust, thereby providing a more rigorous basis for interpreting the productivity differences among the prefixes.

for Potential Productivity–entropy correlation to 0.999 for MATTR–entropy correlation). This suggests that, while the adopted measures capture different aspects of productivity, when the fixed-sample size rule is applied, they seem, in fact, highly correlated and can be used interchangeably for comparative purposes. According to all three measures, *nazovi*(-) is the most productive prefixoid, closely followed by *kvazi*(-). *Pseudo*(-) ranks third in productivity, while *nadri*(-) is deemed the least productive, with a notable gap separating it from the other prefixoids. The findings corroborate the well-established postulate that rival affixes typically exhibit differences in productivity levels (cf. Bybee 1985; Baayen & Lieber 1991; Plag 1999; Bauer 2001; Gaeta & Ricca 2015).

3.2. Prefixoids Across Text Genres

Apart from various linguistic factors, competition between rival forms can also be influenced by the register (genre) in which they appear, with specific affixes being favored in certain contexts. To examine the relationship between text genre and prefixoid selection, this study makes use of the genre annotations present in the CLASSLA corpus. The corpus was automatically annotated for genres using the Transformer-based X-GENRE classifier (Kuzman, Mozetič & Ljubešić 2023) and the following genre categories are used: News, Information/Explanation, Promotion, Opinion/Argumentation, Instruction, Legal, Prose/Lyrical, Forum, Other and Mix[9].

To evaluate the dependency between the choice of a prefixoid and the genre it appears in, a Pearson's χ^2 test was conducted. The test was performed on a contingency table that cross-tabulates the frequencies of four Croatian prefixoids across various genres. The test results ($\chi^2 = 5239.2$, df = 30, p-value < 2.2e–16) indicate a statistically significant relationship between the prefixoid used and the genre of the text. This significant p-value highlights a strong association between genre and prefixoid choice, suggesting that different genres may prefer distinct prefixoids.

The bar chart in Figure 4 illustrates the distribution of four Croatian prefixoids across the ten aforementioned genres. The values are standardized within each genre to show the

[9] Description and examples of each genre category are available at https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier (accessed 24 August 2024).

relative frequency of each prefixoid as a percentage of the total number of prefixoids within that specific genre.

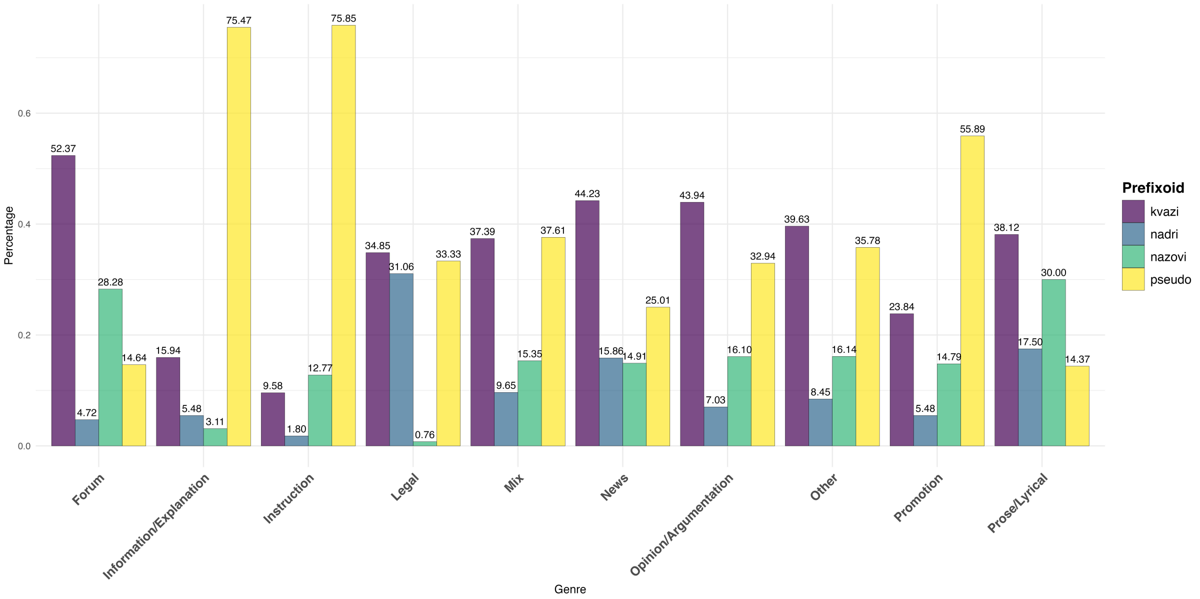

Fig. 4: The standardized frequency of the prefixoids across genres

Observing the bar chart, it is visible that:

kvazi(-) has prominent appearances in genres like Forum (52.4%) and News (44.2%), while it is least prevalent in Instruction (9.58%) and Information/Explanation (15.9%);

nadri(-) presents a notable peak in Legal (31.1%), while it is the least frequent in Instruction (1.80%) and Forum (4.72%);

nazovi(-) exhibits a relatively even presence across all genres with the highest frequencies in Prose/Lyrical (30%) and Forum (28.3%), while only one token (0.76%) formed with *nazovi* is found in Legal;

pseudo(-) is the prominent in Instruction (75.8%) and Information/Explanation (75.5%). It is the least dominant in Prose/Lyrical (14.4%) and Forum (14.6%).

Several compelling observations can be drawn from the genre distribution of the prefixoids. First, a near-perfect complementarity is evident between *nazovi*(-) and *pseudo*(-). The genres in which *nazovi*(-) is the most frequent (Prose/Lyrical and Forum), are in fact genres in which *pseudo*(-) is the least dominant. The complementarity observed between the two elements suggests a form of genre-based niche differentiation that reduces direct competition. This genre-based divergence could likely parallel their distinct collocational preferences, as will be showed in the Correspondence Analysis (Figures 5 & 6), where *nazovi*(-) and *pseudo*(-) are located on the very opposite sides of the biplot, reinforcing their functional and semantic dissimilarity. On the other hand, *kvazi*(-) and *nazovi*(-) display a more aligned genre distribution, particularly in genres like Forum and Prose/Lyrical, in which they are the most frequent prefixoids. This suggests that *kvazi*(-) and *nazovi*(-) may have a closer functional or semantic relationship, which also justifies their overlap in the Correspondence Analysis (Figures 5 & 6). Furthermore, it is curious to observe how the two prefixoids of Slavic-origin exhibit distinct genre-specific usage patterns, especially in Legal. While *nadri*(-) observes its peak in this genre (primarily due to the strong presence of modified profession denoting nouns such as *nadriliječnik* and *nadripisar*), at the same time just a single occurrence formed with *nazovi*(-) is found in Legal, representing the lowest share (0.76%) of any prefixoid across all ten genre categories. These findings highlight genre as a significant factor influencing the selection between rival prefixoids, demonstrating that different communicative contexts, i.e. less-related semantic fields, distinctly shape morphological choices. This underscores the importance of genre-specific factors in affix rivalry, revealing how certain genres favor specific prefixoids while disfavoring others, thereby influencing the functional landscape of morphological approximation.

3.3. Collocational Behavior

Having examined the productivity and genre distribution of the four prefixoids, our focus shifts to their semantics. Rival approximative affixes, even when used in analogous contexts, often exhibit distinct distributional tendencies. To analyze the extent of overlap in collocational preferences between the four prefixoids, first we identify the number of nominal bases they share. Subsequently, we employ Multiple Distinctive Collexeme Analysis (MDCA) to extract the most distinct collexemes of each prefixoids. This approach isolates the bases that are particularly characteristic of each prefixoid, highlighting their unique collocational patterns. We conclude the semantic analysis by visualizing the most distinct collexemes by means of Correspondence analysis (CA). The combined use of MDCA and CA offers a comprehensive understanding of how these rival prefixoids interact within the semantic landscape, providing a clearer picture of their distinct yet overlapping roles.

3.3.1. Shared Bases

Table 3 presents the number of nominal bases shared by each pair of prefixoids.

Tab. 3: Number of shared nominal bases by pairs of prefixoids

PREFIXOID	*kvazi*(-)	*nadri*(-)	*nazovi*(-)	*pseudo*(-)
kvazi(-)		163	658	586
nadri(-)			124	97
nazovi(-)				341
pseudo(-)				

As observed by comparing the values in the Table 3, *kvazi*(-) and *nazovi*(-) share the highest number of bases (658), suggesting they might be more closely related to each other than with other prefixes. In contrast, *nadri*(-) shares the fewest bases with the other prefixoids, indicating a more distinct usage pattern[10]. While the number of shared bases offers some insight into the semantic and functional relatedness of the four prefixoids, it provides only a partial view, and ulterior analyses are required in order to obtain a more complete picture of the overall relationship and the extent of the semantic overlap.

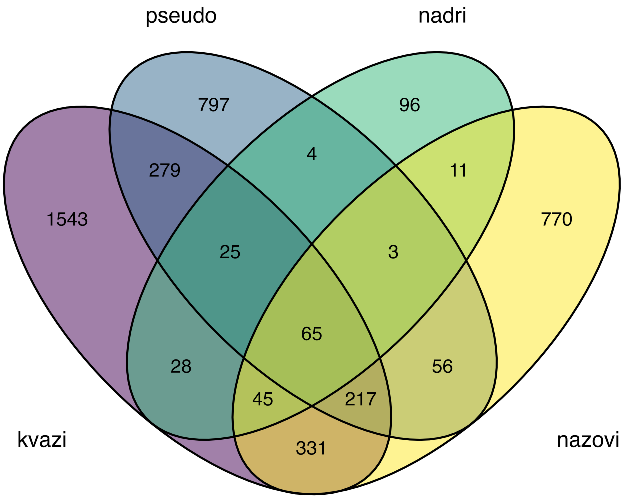

Fig. 5: The overlap of base nouns between the prefixoids

[10] To quantify the degree of overlap between pairs of prefixoids more precisely, Jaccard Similarity Indices were calculated. The Jaccard Similarity Index (JSI) measures similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. The Index ranges from 0 (no overlap) to 1 (complete overlap). As anticipated, the highest JSI was observed between *kvazi*(-) and *nazovi*(-) (0.195), followed by *kvazi*(-) and *pseudo*(-) (0.173), and *nazovi*(-) and *pseudo*(-) (0.131). The lowest JSI was noted between *nadri*(-) and *pseudo*(-) (0.059), suggesting minimal shared bases between these two prefixoids.

Figure 5 presents a four-way Venn diagram that illustrates the shared and unique bases among the four examined prefixoids. Unlike the simple pairwise comparisons shown in Table 3, this diagram provides a detailed visualization of the specific overlap counts between different sets without aggregating them. Each ellipse represents the set of bases associated with a specific prefixoid; for instance, the light blue ellipse represents *kvazi*'s bases. Numbers within the non-overlapping sections indicate bases unique to each prefixoid, such as the 1,543 bases that *kvazi*(-) does not share with the other prefixoids. Conversely, numbers in overlapping sections indicate shared bases, such as the 331 bases shared exclusively between *kvazi*(-) and *nazovi*(-). The central number (65) represents the bases common to all four prefixoids. To determine the total shared bases between two prefixoids, such as *kvazi*(-) and *nazovi*(-), one must sum all overlapping regions between their respective ellipses, including those overlapping with other ellipses. Specifically, *kvazi*(-) and *nazovi*(-) share 658 bases in total: 331 bases shared exclusively between the two, 217 bases shared with *pseudo*(-), 45 bases shared with *nadri(-)*, and, finally, 65 bases shared by all four prefixoids. This visualization highlights the degree of overlap between the prefixoids, providing insight into their relational dynamics based on shared bases. For instance, *kvazi*'s greater base overlap with *nazovi*(-) and *pseudo*(-) than with *nadri*(-) could suggest a closer functional or semantic connection between these prefixes.

Having established the proportion of shared bases, it is relevant to further examine the 65 nominal bases common to all four prefixoids. Notably, 38 of these shared bases are directly related to individuals, either as profession-denoting nouns (e.g., *filozof* 'philosopher', *novinar* 'journalist') or as cultural/political terms (*roker* 'rocker', ljevičar 'leftist'). Additionally, terms such as *demokracija* 'democracy', *komunizam* 'communism', and *institucija* 'institution' are found, reflecting the frequent use of approximation prefixoids in contexts of societal, and particularly political, discourse. What can be observed in these groups, just like in examples (1)–(4), is that the prefixoids exhibit rival behavior and semantic equivalence (at least at a coarse-grained level), with all four signaling a person, object, or concept that is not a typical exemplar of the base noun, often with reference to inadequate performance. While these (minimal) pairs might be useful for highlighting differences between rival prefixoids by neutralizing base differences, they may not capture all distinctions between rival processes, as certain discriminative properties of the base may prevent the formation of competing lexemes (cf. Huyghe & Varvara 2023). For this motive, we leave the study of formations with identical bases for future research and instead focus on the identification of collexemes unique to each approximative construction, believing this approach offers deeper insight into the semantics of the prefixoids.

3.3.2. Multiple Distinctive Collexeme Analysis

The data concerning the quantity of shared bases presented in the previous section made us hypothesize about the nature of the semantic relationship between the four prefixoids. While some prefixoids share more bases (hence, could be semantically more similar), others do not seem to share many bases with other prefixoids, indicating a potentially more peculiar collocational behavior. A possibility of contrasting more related, near-synonymous constructions in their respective synchronic collocational preferences is made possible by Multiple Distinctive Collexeme Analysis (MDCA) (Gries & Stefanowitsch 2004; Stefanowitsch 2013). MDCA enables the identification of collexemes that are unique to specific constructions, moving beyond mere raw frequency by abstracting common elements and focusing on distinct usage patterns. By systematically examining usage-based, pattern-specific properties through statistical analysis, MDCA assesses asymmetries in the relative frequencies of co-occurring lexical items (Stefanowitsch & Flach 2020), highlighting collexemes that occur significantly more frequently with one construction over another. The input required for MDCA is a data frame, which can be formatted as either a raw frequency list (one observation per line) or an aggregated frequency list that includes a third column for the construction's frequency. The MDCA script used in this study is based on Flach's (2021) *collex.covar* function from the package *constructions*. Furthermore, following the approach outlined by Proisl (2022), units of analysis (corpus size) consist of all (722,422,618) nominal tokens in the CLASSLA corpus. Finally, association – or to be more precise, a combination of frequency, association, and dispersion (cf. Gries 2019, 2022) – is measured using the log-likelihood ratio (G^2), "the most frequently used [association] measure"[11] (Gries 2019:

[11] Recent discussions have focused on using measures like residuals of chi-squared to further reduce computational costs. Additionally, there is an ongoing debate on calculating more than one association measure value to address the issue of conflating frequency, mutual/unidirectional association, and dispersion into one measure (as G^2 does) (Gries 2019, 2023; Liao, Gries & Wulff 2024). However, given that the primary objective of this analysis is exploratory/descriptive, it is essential to recognize that the "conflation of frequency and association makes for a good exploratory tool" (Liao, Gries & Wulff 2024: 13), and "an

150). The degree of attraction between the construction and the collexeme (and vice versa, since G^2 is a bidirectional measure) is very significant at (at least) $p < 0.0001$.

Tables 4–7 present the ten most distinctive collexems for each of the four analyzed prefixoids. OBS stands for *observed frequency*, indicating how often each prefixoid-noun pairing appears in the corpus. EXP represents the *expected frequency*, i.e. anticipated by chance under the null hypothesis. COLL.STR.LOGL stands for *collostructional strength log-likelihood*, a measure of the association between the two slots of the construction.

Tab. 4: The 10 most distinctive nominal collexemes of *nadri*(-)

***nadri*(-)**	**OBS**	**EXP**	**COLL.STR.LOGL**	**ΔP1**	**ΔP2**
liječništvo ‘medical profession’	453	35.2	2435.84	0.2625	0.9415
liječnik ‘doctor’	393	32.2	1934.40	0.2267	0.8881
pisarstvo ‘scribal profession’	141	11.0	731.27	0.0817	0.9282
pisar ‘scribe’	72	72	370.64	0.0417	0.9253
ljekarstvo ‘medicine’	13	1.0	66.50	0.0075	0.9228
obrtnik ‘craftsman’	12	0.9	61.38	0.0070	0.9228
veterinarstvo ‘veterinary medicine’	9	0.7	46.02	0.0052	0.9226
liječenje ‘treatment’	10	0.9	40.65	0.0057	0.7560
obrt ‘craft/trade’	6	0.5	30.67	0.0035	0.9225
majstor ‘handyman’	17	4.7	21.88	0.0077	0.2015

Tab. 5: The 10 most distinctive nominal collexemes of *kvazi*(-)

***kvazi*(-)**	**OBS**	**EXP**	**COLL.STR.LOGL**	**ΔP1**	**ΔP2**
parcijala ‘repetition of some university courses’	84	33.2	156.60	0.0096	0.60745
menadžer ‘manager	127	56.5	152.82	0.0133	0.4965
kristal ‘crystal’	78	30.8	145.39	0.0089	0.6073
renta ‘rent’	54	21.30	100.56	0.0062	0.6066
čestica ‘particle’	50	19.70	93.09	0.0057	0.6065
novinar ‘journalist’	198	118.5	87.04	0.0150	0.2688
navijač ‘fan’	67	30.4	75.35	0.0069	0.4769
političar ‘politician’	147	84.9	73.91	0.0117	0.2917
grupa ‘group’	47	19.3	72.79	0.0052	0.5656
intelektualac ‘intellectual’	158	94.4	69.72	0.0119	0.2691

association measure combining (a lot of) frequency and (a little bit of) association is still a good option” as it provides a “heuristically useful amalgam of two kinds of information” (Gries 2023: 372).

Tab. 6: The 10 most distinctive nominal collexemes of *nazovi*(-)

***nazovi*(-)**	**OBS**	**EXP**	**COLL.STR.LOGL**	**ΔP1**	**ΔP2**
prijatelj 'friend'	48	11.6	92.14	0.0121	0.5066
album 'album'	20	3.4	65.27	0.0055	0.7914
sud 'court'	24	5.3	52.06	0.0062	0.5664
bog 'god'	18	3.7	43.33	0.0047	0.6215
banka 'bank'	12	2.1	37.06	0.0033	0.7618
hrvat 'Croat'	27	8.6	34.21	0.0061	0.3485
grad 'city'	14	3.01	32.89	0.0036	0.5756
pjesma 'song'	10	1.8	27.80	0.0030	0.6837
sloboda 'freedom'	17	4.7	26.42	0.0045	0.3989
komentar 'comment'	10	2.3	20.49	0.0028	0.5170

Tab. 7: The 10 most distinctive nominal collexemes of *pseudo*(-)

***pseudo*(-)**	**OBS**	**EXP**	**COLL.STR.LOGL**	**ΔP1**	**ΔP2**
znanost 'science'	1014	397.5	1649.99	0.1197	0.5964
žitarica 'cereal'	365	134.2	721.96	0.0448	0.6395
kod 'code'	236	86.3	479.27	0.0291	0.6412
cista 'cyst'	152	55.6	307.66	0.0187	0.6387
stvarnost 'reality'	167	63.3	291.47	0.0201	0.6044
jezik 'language'	91	33.6	173.62	0.0111	0.6261
gen 'gene'	66	24.1	133.14	0.0081	0.6363
gravidnost 'pregnancy'	66	24.1	133.14	0.0081	0.6362
stablo 'tree'	65	23.8	131.12	0.0080	0.6362
događaj 'event'	70	26.06	118.92	0.0084	0.5952

Tables 4–7 provide additional information, specifically the ΔP value (Ellis 2007; Ellis & Ferreira-Junior 2009), which addresses the limitations of G^2 related to its bidirectionality. Unlike G^2, ΔP is unidirectional (asymmetric), meaning it does not conflate p(word2|word1) and p(word1|word2) into a single value. This distinction allows ΔP to identify cases where collexeme 1 strongly attracts collexeme 2, but not vice versa. Additionally, ΔP reflects association independently of frequency, meaning changes in corpus size do not affect the association value. ΔP is divided into two values: ΔP1 measures the predictiveness of the collexeme (slot 2) for the construction (slot 1), whereas ΔP2 quantifies the predictive capacity of the construction (slot 1) for the collexeme (slot 2) (Gries & Ellis 2015; Gries 2019). As anticipated, an analysis of <PREF$_{\text{APPRX}}$ + noun> constructions reveals that the constructions are more predictive of the noun than the other way around, as indicated by significantly higher ΔP2 values compared to ΔP1. Among the four prefixoids, *nadri*(-) stands out as the

most predictive when considering its ten most distinctive collexemes, with a very high mean ΔP2 of 0.8331. The highest predictiveness is observed with *nadriliječništvo* 'quackery', where the construction almost impeccably anticipates its collexeme (ΔP2 = 0.94).

When overlapping collexemes (nouns correlated with two or more analyzed prefixoids) are filtered out by MDCA, the distinct collocational patterns of the prefixoids become more pronounced. In fact, Aronoff's (2019) *habitat niche differentiation* clearly emerges for *nadri*(-): nine out of the ten most preferred collexemes denote a profession or a person exercising a profession, while the remaining non-profession noun, viz. *liječenje* 'treatment', denotes a job done by the aforementioned medical professional. Moreover, differences in the level of formality of the preferred collexemes are observed. Whereas *nadri*(-) and, especially, *pseudo*(-) combine with nouns that belong to a more formal and sometimes scientific register, *nazovi*(-) and *kvazi*(-) seem to combine more felicitously with nouns from general lexicon.

While MDCA has offered valuable insights, it is important to acknowledge its limitations, especially due to the small number of displayed collexemes for each prefixoid. On the other hand, as noted by Desagulier (2014), increasing the number of collexemes can make it harder to draw meaningful generalizations from the data. Therefore, rather than relying solely on comparing MDCA output tables, it is advisable to use a technique that allows for visualizing the relationships between (a) prefixoids, (b) nouns, as well as (c) prefixoids *and* nouns, by converting the initial matrix into a low-dimensional space. In line with Desagulier's (2014, 2015) approach, and as already applied in Lacić (2024a), this study will use the output of MDCA as input for Correspondence Analysis (CA).

3.3.3. Correspondence Analysis

Correspondence Analysis (CA) is a multifactorial exploratory statistical technique used to explore relationships and patterns within categorical data (Benzécri 1973; Greenacre 2017). In CA, rows and columns of a contingency table are represented as points in Euclidean space, with their proximity indicating the strength of association. The χ^2 distance, multivariate statistical distance measure akin to Euclidean distance but weighted by the inverse of the average row profile, measures differences between profiles, positioning rows and columns with similar counts closer together[12].

CA was conducted using the 50 most distinctive collexemes of each of the four prefixoids and the raw frequency of each construction as input. The threshold of 50 most distinctive collexemes was decided in order to maintain the degree of attraction between the construction and the collexeme statistically significant – all included collexemes have log-likelihood significant at (at least) $p < 0.001$. Furthermore, the hypothesis of independence regarding the input data can be rejected, with $\chi^2 = 18093.73$; p-value = 0. In addition, Cramér's V of 0.790 indicates a significant association between the rows and the columns, supporting the notion of a meaningful relationship between the prefixoids and nouns they combine with. CA uses the input frequencies to juxtapose (a) line profiles, i.e. distinctive collexemes (nouns); (b) column profiles, i.e. prefixoids; (c) line profiles *and* column profiles, i.e. nouns *and* prefixoids. The *CA* function from the *FactoMineR* package was employed to run CA. Figures 6 and 7 (with collexemes translated in English) display the output of CA.

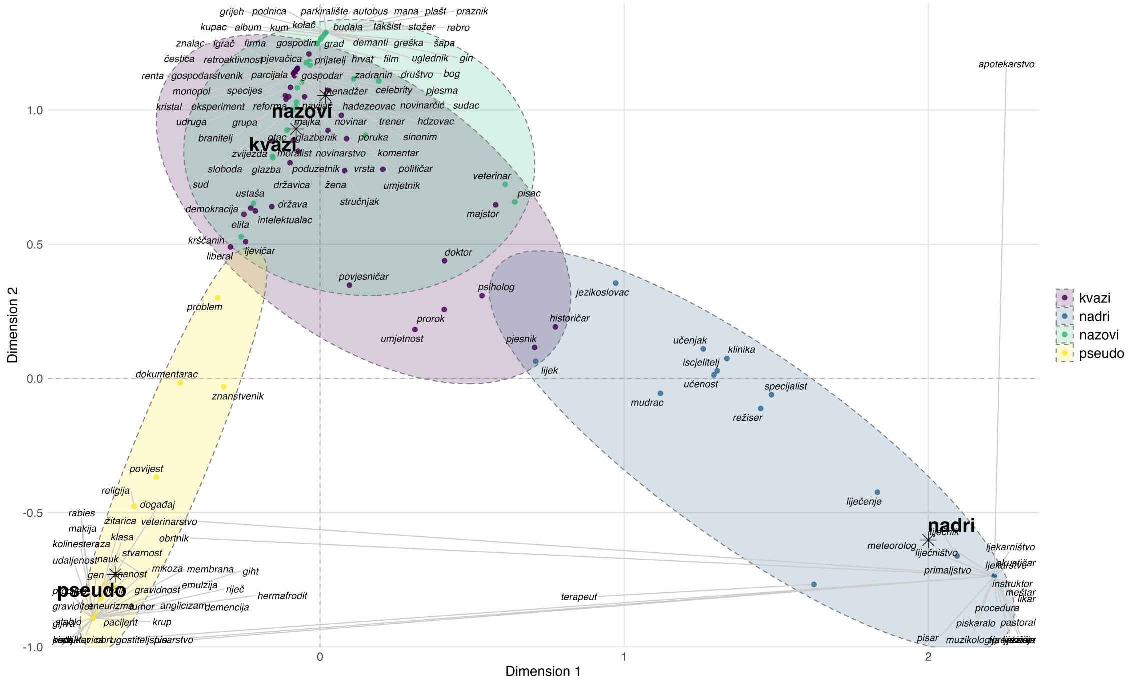

Fig. 6: CA biplot of the <PREF$_{\mathrm{APPRX}}$ + noun> construction in CLASSLA corpus

[12] However, interpreting the proximity between rows and columns should be done cautiously, as there is no direct interpretation of row-to-column or column-to-row distances (Levshina 2015).

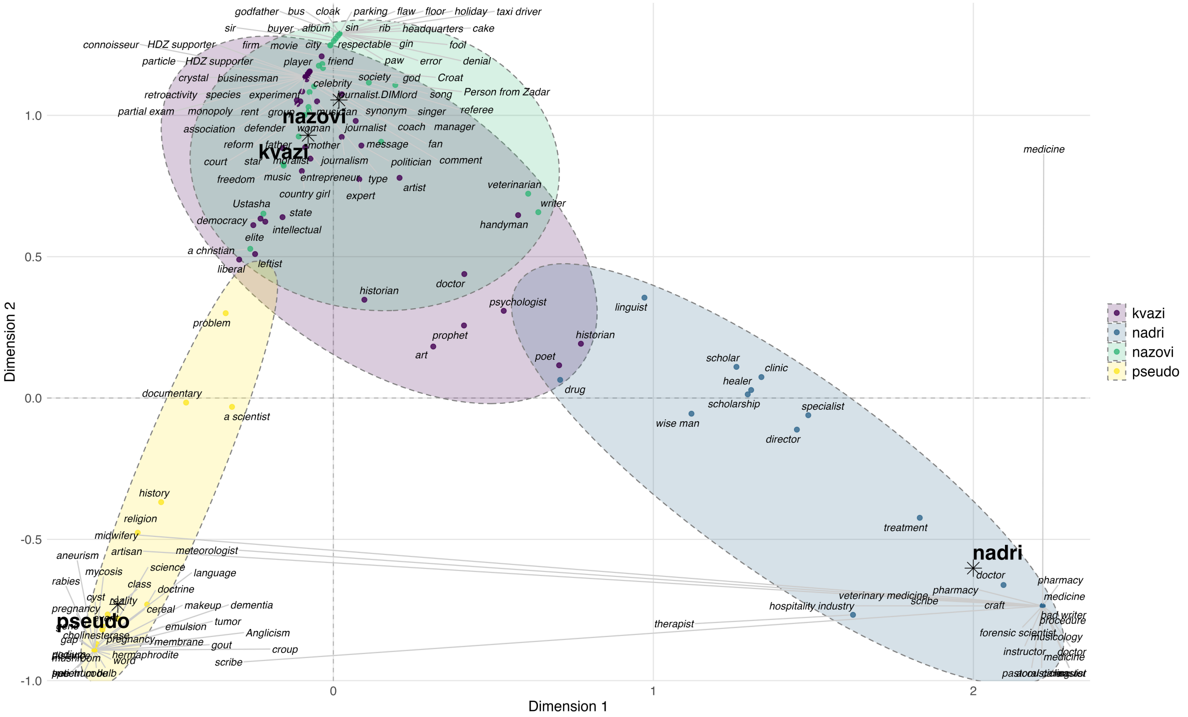

Fig. 7: CA biplot of the <$\text{PREF}_{\text{APPRX}}$ + noun> construction in CLASSLA corpus

The CA biplot is constructed using two principal axes of inertia, which intersect to define the average profile of all points in the data cloud. The technique decomposes the overall inertia (Φ^2) – weighted (co)variance obtained dividing the χ^2 statistic by the total sample size – by identifying representative dimensions that condense as much information as possible within each axis corresponding to a dimension. Typically, for reasons of practicality, a plot displays only two dimensions, selected based on their eigenvalues, which measure the amount of information (variation) present along each axis (Levshina 2015; Greenacre 2017). In this analysis, the first axis (dimension 1) represents 45.98% of Φ^2, while the second axis (dimension 2) represents 37.81% of Φ^2. There is also a third dimension with an eigenvalue of 16.21%, which is not included in the biplot. Whilst including the third dimension with means of an interactive, tri-dimensional plot would provide additional information regarding the relationships between the analyzed variables, the first two dimensions already account for 83.79% of the variation contained in the input table, allowing for a sufficiently accurate interpretation of the results.

We can start examining the plot and how it juxtaposes four prefixoids by contrasting the two main dimensions. Dimension 1 (horizontal axis), primarily contrasts *pseudo*(-) and *nadri*(-). Positioned distinctly in the lower left quadrant, *pseudo*(-) occupies a unique space linked with academic, scientific, or technical terminology that often involves imitation aspects (e.g., *psuedoznanost* 'pseudoscience', *pseudodemencija* 'pseudodementia', *pseduocista* 'pseudocyst'). This prefixoid diverges from the societal roles more characteristic of *kvazi*(-) and *nazovi*(-) and instead aligns with terminology that carries an inherent critique of validity or authenticity, particularly in specialized/technical domains. Conversely, on the lower right side, where *nadri*(-) dominates, the focus shifts towards contexts related to medicine, expertise, and professional roles that involve deception or unauthorized practice (e.g., *nadriliječnik* 'quack', *nadri-iscjelitelj* 'quack healer', *nadriliječenje* 'quack treatment'). Its unique placement underscores its more specific and legally loaded use, setting it apart from the broader, more socially nuanced applications of the other prefixoids. This isolation reflects its primary association with professional deceit rather than broader conceptualizations of inauthenticity or approximation. In contrast, observing Dimension 2 (vertical axis), a separation of *pseudo*(-) and *nadri*(-) from *kvazi*(-) and *nazovi*(-) is evident. The separation could potentially be interpreted through the lens of the semantic roles these prefixoids play. In fact, as seen from observing the preferred collexemes, *pseudo*(-) attaches to terms in academic, scientific, and technical domains, while *kvazi*(-) and *nazovi*(-) are more common in colloquial or socially-oriented language (e.g., *kvaziprijatelj* 'pseudo-friend', *nazovipolitičar* 'pseudo-politician') that foster more subjective interpretation. The vertical axis, therefore, might reflect a contrast between formal, domain-specific language (right side) and more general, socially embedded language (left side), which would align with the general theory that formal registers and specialized vocabulary often cluster together due to their shared contexts.

When it comes to clusterings, a distinct cluster can be observed formed with *kvazi*(-) and *nazovi*(-) and their collexemes. Clusters are visually enhanced by confidence ellipses around the groups of data points that are associated with each prefixoid. When ellipses overlap, like in the case of *kvazi*(-) and *nazovi*(-), it indicates that the base words associated with those prefixoids are similar or used in similar contexts, suggesting a potential functional or semantic similarity between the prefixoids. Furthermore, the size of each ellipse provides insight into the variability or spread of the base words for each prefixoid. Larger ellipses indicate greater variability in how the prefixoid is used with different base words. Returning to the clusters, the distinct position of *pseudo*(-) with scientific and medical terms corroborates the prefixoid's specialized role in denoting false or imitative phenomena within technical and academic discourses. Conversely, the clear separation of *nadri*(-) suggests another specialized usage, namely denoting fraudulent or unqualified individuals in professional roles.

Following this analysis, the study will proceed with a semantic annotation of selected examples to further elucidate the nuanced meanings of each prefixoid.

3.4. Semantic Annotation

As outlined in §2, the semantic annotation process is based on a 500-token random sample for each prefixoid. The annotation was done manually by one annotator. Notably, the identified semantic values align substantially with those documented by Masini & Micheli (2020) for the Italian *simil-* and Vassiliadou et al. (2023) for *pseudo-* in French, indicating cross-linguistic parallels in how approximation is manifested. To avoid contributing to the already present terminological imbroglio, the labels adopted here partially adhere to the established terminology from the aforementioned studies. In what follows, we first detail the seven identified semantic values, each of which represents a specific manifestation within the broader category of approximation[13]. Subsequently, for each prefixoid, we present the semantic values identified within the 500-token sample. With Y being the output of the prefixoid modification process, and X the nominal base to which the prefixoids apply, the identified semantic values can be defined as follows:

"CLOSE-TO" EVALUATION (CTE): This category captures instances where Y is coming close the state of X, embodying an almost-but-not-quite identity with X. This formation can often be paraphrased using adverbs such as *gotovo* 'almost' and *skoro* 'nearly', emphasizing a degree-based proximity.

FAKENESS (F): In this value, Y deliberately deceives by pretending to be X without authentically being X. The element of intentional deceit is central, aligning with scenarios where imitation serves to mislead observers into believing Y is genuinely X.

[13] It should be noted that, from a functional perspective, not all values are approximating at the same level. While concepts such as SUBJECTIVE DEPRECIATIVE EVALUATION or VAGUENESS can be viewed as more closely aligned with approximation, FAKENESS or KIN-CATEGORIZATION are more related to category creation (Masini & Micheli 2020). Given space constraints, we leave a deeper exploration of this complex issue for future research.

ILLEGALITY (ILL): This value is particularly relevant in contexts involving unauthorized or unlawful actions. Here, Y represents an individual or activity that operates outside the boundaries of legal norms.

IMITATION (IM): Involves Y replicating or reproducing aspects of X without the intent of being mistaken for X. Unlike FAKENESS, the absence of intentional deceit distinguishes this category, allowing for a less pejorative reading where imitation is acknowledged but not condemned.

KIN-CATEGORIZATION (K-C): This value identifies Y as an entity closely related to X but distinct in its defining characteristics. Y's identity is rooted in X but deviates enough to form a separate category, drawing on shared traits without fully merging.

SUBJECTIVE DEPRECIATIVE EVALUATION (SDE): Y is perceived as an inferior or flawed version of X, subject to the speaker's negative evaluation. It captures a speaker's critical stance, attributing substandard or undesirable qualities to Y in comparison to the prototypical X.

VAGUENESS (V): This value arises when Y is an entity whose nature is uncertain and has a tenuous relationship to X. It may be seen as a borderline or ambiguous member of the category X, suggesting fuzzy category distinctions.

3.4.1. *Kvazi*(-)

The analyzed sample of *kvazi*(-) formations reveals that this prefixoid is predominantly used in Croatian to negatively evaluate the referent, suggesting it fails to represent its category appropriately. This evaluative function, termed SUBJECTIVE DEPRECIATIVE EVALUATION, reflects instances where *kvazi*(-) highlights perceived deficiencies in the subject, as demonstrated in the example (8), where athletes are disparaged, possibly due to their lifestyle choices. Examples provided in this subsection, as well as in the following ones, derive from the aforementioned CLASSLA corpus.

(8) *Najbolji naš igrač prošle sezone, igra većim srcem nego bilo tko drugi,* [...] *ne opija se po vikendima za razliku od 95% hrvatskih* ***kvazisportaša***.

'Our best player last season, he plays with a bigger heart than anyone else, [...] he doesn't get drunk on weekends, unlike 95% of Croatian pseudo-athletes (lit. KVAZIathletes).'

Furthermore, a prototypical function of *kvazi*(-) is noted, viz. degree modifier-like behavior, in which *kvazi*(-) indicates the sense of 'coming close' to the concept expressed by the head noun. In such cases, *kvazi*(-) can be paraphrased with adverbs like *gotovo* 'almost' and *skoro* 'nearly', exemplifying a genuine approximation value, as seen in (9). We call this value "CLOSE-TO" EVALUATION.

(9) *Europski sindikalni pokret zabrinut je zbog prijedloga o ekonomskom upravljanju* [...] *koji bi zemlje članice sveo na status* ***kvazikolonije****.*

'The European trade union movement is concerned about the proposals on economic governance [...] which would reduce member countries to the status of a quasi-colony.'

Moreover, *kvazi*(-) occasionally conveys the value of FAKENESS, where it implies intentional deception, as in the case of fake sales tactics described in (10). Here, *kvazi*(-) indicates that X pretends to be genuine to mislead:

(10) [...] *stvarno ne vidim smisla ovdje dodatno reklamirati* ***kvazi akcije*** *trgovina koje prvo dignu cijene 50% pa spuštaju 30%.*

'I really don't see the point of additionally advertising fake-sales (lit. KVAZIsales) of stores that first increase the prices by 50% and then lower them by 30%.'

Lastly, *kvazi*(-) is also found in expressions classified as IMITATION, where *kvazi*-X denotes an imitation or reproduction of X without the intention of being perceived as authentic, as illustrated by *kvazibolonjez* 'pseudo-bolognese' in (11), a plant-based version of the original Italian dish.

(11) *I htjedoh reći – onaj* ***kvazibolonjez*** *od crvene leće je odličan, pogotovo kad se ohladi.*

'And I wanted to say – that red lentil pseudo-bolognese (lit. KVAZIbolognese) is great, especially when it's cold.'

Table 8 presents the semantic classification of the analyzed *kvazi*(-) formations.

Tab. 8: Semantic classification of the 500-token *kvazi*(-) sample

PREFIXOID	VALUE	TOKENS
kvazi(-)	SUBJECTIVE DEPRECIATIVE EVALUATION	492
	"CLOSE-TO" EVALUATION	4
	FAKENESS	2
	IMITATION	2

3.4.2. *Nadri*(-)

The analysis of 500 tokens containing *nadri*(-) reveals that 446 tokens (89.20% of the total) correspond to four specific nouns: *nadriliječnik* ('quack'), *nadripisar* ('quack scribe'), and their related nouns denoting the activities performed by such individuals, *nadriliječništvo* ('quackery') and *nadripisarstvo* ('scribal quackery'). In these contexts, *nadri*(-) refers to individuals who deceitfully engage in a profession for which they lack the requisite education and competence, or to the illicit activities carried out by such individuals, as illustrated in (12):

(12) *U Statutu Hrvatske liječničke komore [...] stoji da Komora "promiče znanstvene postupke dijagnostike i liječenja a suzbija* ***nadriliječništvo***" (čl. 33. t. 7).

'The Statute of the Croatian Medical Chamber [...] states that the Chamber "promotes scientific diagnostic and treatment procedures and suppresses quackery (lit. NADRItreatment)" (art. 33, cl. 7).'

As discussed earlier, in these predominantly legal contexts, *nadri*(-) conveys a specific, well-defined sense, which we term ILLEGALITY. Out of the 446 tokens of the aforementioned four nouns, 423 were annotated as ILLEGALITY due to their direct reference to the legal terminology. Although ILLEGALITY might overlap with FAKENESS or IMITATION (as a quack could be perceived as a fake doctor), there are nuanced differences. The criterion of intentionality, essential for FAKENESS (Masini & Micheli 2020), is not always clear-cut. For example, a quack doctor may deliberately pose as a real doctor for profit, as seen in (13), where Dulcamara from Donizetti's famous opera buffa *L'elisir d'amore* exploits the villagers' gullibility by selling wine as a supposed magical love potion:

(13) Ljubavni napitak *slavna je komična opera, u kojoj je sjajno iznio komične karaktere likova, posebno* ***nadriliječnika*** *Dulcamaru, koji iskorištava seosku lakovjernost i dobrohotnost.*

'*L'elisir d'amore* is a famous comic opera, in which he brilliantly brought out the comic characters of the characters, especially the quack (lit. NADRIdoctor) Dulcamara, who exploits the gullibility and benevolence of the village.'

In this instance, it might be possible to classify the occurrence as an instance of FAKENESS, given the assumed intentionality of deception. However, in other instances, such as in (14), there is no clue on whether subjects in question are intentionally posing as licensed professionals, or they are categorized as quacks by contemporary standards while in their era they were recognized in fact as “medical” figures.

(14) *U srednjem vijeku dolazi do stagnacije u području liječenja zubi jer se tim poslom počinju baviti brijači i drugi* ***nadriliječnici****.*

‘In the Middle Ages, there was stagnation in the field of dental treatment, as barbers and other quacks (lit. NADRIdoctors) began to deal with this business.’

To avoid ambiguous interpretations and given the strong association of *nadri* with legal contexts, ILLEGALITY is deemed the most accurate classification for these occurrences. Finally, in certain contexts, *nadri*(-) is used to negatively assess a subject, devoid of any reference to illegality or clues about the subject’s formal education (15) or even confirming that the subject, in fact, holds a relevant degree (16). In these cases, *nadri*(-) reflects SUBJECTIVE DEPRECIATIVE EVALUATION.

(15) *Nije ovo nikakva kritika tebi već više debilnom jeziku i još debilnijim našim* ***nadri jezikoslovcima****.*

‘This is not a criticism of you, but of a stupid language and our even more stupid quack linguists (lit. NADRIlinguists).’

In (15), the linguists in question may have formal degrees, yet their ideas are considered irrelevant or nonsensical. Conversely, in (16), it is explicitly stated that “quacks” possess a degree, but, for instance, their treatment methods are viewed negatively.

(16) [...] *najviše* ***nadriliječnika*** *se skriva iza diplome liječnika.*
‘[...] most quacks (lit. NADRIdoctors) are hiding behind a doctor’s degree.’

As a result, the 23 tokens from the group of 446 (*nadriliječnik*, *nadriliječništvo*, *nadripisar*, *nadripisarstvo*) and the remaining 54 tokens were annotated as SUBJECTIVE DEPRECIATIVE EVALUATION.

Table 9 presents the semantic classification of the 500 analyzed *nadri*(-) expressions.

Tab. 9: Semantic classification of the 500-token *nadri*(-) sample

PREFIXOID	VALUE	TOKENS
nadri(-)	ILLEGALITY	423
	SUBJECTIVE DEPRECIATIVE EVALUATION	77

3.4.3. *Nazovi*(-)

Nazovi(-), just like *nadri*(-), makes part of a legal terminology and formation such as *nazoviiliječništvo* 'quackery' is synonymous to *nadriliječništvo* 'quackery', as it indicates the illegal activity of providing medical assistance by a non-professional person. However, in the analyzed sample, no tokens of *nazoviliječništvo* 'quackery' or *nazoviliječnik* 'quack' were identified, and only a single occurrence of *nazoviliječništvo* was found in the entire corpus, indicating the rarity of such formation. This scarcity could be attributed to a phenomenon known as *statistical preemption* (Boyd & Goldberg 2011), where the frequent occurrence of *nadriliječnik* in contexts in which also *nazoviliječnik* would have been adequate has led to its entrenchment (due to its frequency), thereby preempting the use of *nazoviliječnik*. Regarding *nazovi*(-), as illustated in (17), we observe how the prefix in most cases carries a value of SUBJECTIVE DEPRECIATIVE EVALUATION.

(17) *Granica dobrog ukusa debelo je prekoračena u ovom* ***nazovi filmu****.*

'The limit of good taste is grossly overstepped in this pseudo-film (lit. NAZOVI film).'

Additionally, *nazovi*(-), due to its origin as an imperative form, can express a function no other analyzed prefix carries, namely one of a hedge (Lakoff 1973). More precisely, *nazovi*(-) can be used as a (non-morphological) mechanism of coming close or matching with what is intended to be signified, as seen in (18) and (19).

(18) *Današnjim klincima rat je pomrsio račune,* [...] *i nitko ih ne bi trebao kriviti za tu* ***nazovi hladnoću****.*

'For today's kids, the war messed up their plans [...], and nobody should blame them for that so-called coldness (lit. NAZOVI coldness).'

(19) *Išla sam na Hitnu, dobila injekciju, kroz neku* ***nazovi bocu*** *sam disala. Dr. mi je rekla da sam pobrala neku bakteriju.*

'I went to the emergency room, got an injection, I breathed through some type of a bottle (lit. NAZOVI bottle). Doctor told me that I've cached some bacteria.'

In these cases, we argue that *nazovi*(-) does not serve carry subjective depreciative value towards the concept indicated by the noun but serves as a hedge, collocating the designated noun as an atypical member of its category or even questioning its very categorization as the head noun. It is usually referred to such a value as VAGUENESS.

Finally, instances of FAKENESS were identified, as in (20), in which it is clear that the modified noun indicates individuals who are intentionally trying to imitate being a patriot without actually being a genuine patriot.

(20) *Mene je beskrajno stid zbog njih, ali sam istovremeno ponosan sto ne pripadam toj bestidnoj sorti* ***nazovi rodoljuba****.*

‘I am beyond of ashamed because of them, but at the same time I am proud that I do not belong to that shameless breed of so-called patriots (lit. NAZOVI patriots).’

Lastly, one example of KIN-CATEGORIZATION was found. The formation *nazovi brak* ‘pseudo-marriage’, illustrated in (21), makes a reference to civil unions between same-sex partners, which, in this context, carry a clearly pejorative connotation.

(21) *Sve druge kombinacije što ih nameće moderni svijet* [...] *neprirodne su prema biblijskim načelima: poligamija,* ***nazovibrakovi*** *između pripadnika/ca istog spola.*

‘All other combinations imposed by the modern world [...] are unnatural according to biblical principles: polygamy, pseudo-marriages (lit. NAZOVImarriages) between same-sex partners.’

Table 10 presents the semantic classification of the 500 analyzed *nazovi*(-) tokens.

Tab. 10: Semantic classification of the 500-token *nazovi*(-) sample.

PREFIXOID	VALUE	TOKENS
nazovi(-)	SUBJECTIVE DEPRECIATIVE EVALUATION	334
	VAGUENESS	92
	FAKENESS	73
	KIN-CATEGORIZATION	1

3.4.4. *Pseudo*(-)

Among the four examined prefixoids, *pseudo*(-) exhibits the highest frequency in scientific and technical language. In these contexts, *pseudo*(-) functions as a left constituent of neoclassical compounds, usually combining with neoclassical Final Combining Forms as *pseudonim* ‘pseudonym’ or a non-classical scientific terms such as *pseudohipokalcemija* ‘pseudohypocalcemia’. Here, *pseudo*(-) conveys a classifying (privative) sense, signaling that the modified noun does not belong to the category of its head noun. Given that these terms are predominantly part of scientific jargon, they do not convey a subjective negative evaluation of the subject, but rather indicate an exclusion from the category, as illustrated in (22).

(22) *Transabdominalnom ultrazvučnom pretragom se postavlja sigurna dijagnoza* ***pseudo-gravidnosti*** *na temelju odsutnosti placentoma, fetusa i fetalnih membrana.*

‘A transabdominal ultrasound examination establishes a safe diagnosis of pseudo-pregnancy based on the absence of the placentome, fetus and fetal membranes.’

These instances are annotated as conveying SCIENTIFIC NON-EVALUATIVE (SNE) meaning, which lies outside the scope of evaluative morphology since no explicit evaluation is expressed. Another identified value is SUBJECTIVE DEPRECIATIVE EVALUATION, where *pseudo*(-) does not indicate the exclusion from the head’s category (thus, *pseudo*X is, in fact, X) but the head’s qualitative depreciation, serving as a derogatory marker, as shown in (23). This value often appears in formations with native elements rather than scientific terms, as in *pseudo poezija* (‘pseudo poetry’), which refers to poorly written poetry devoid of artistic merit.

(23) *Loša erotska* ***pseudo poezija*** *na stranu, budite liberalni sa isprobavanjem novih stvari u postelji.*

‘Bad erotic pseudo poetry aside, be liberal with trying new things in bed.’

Finally, an instance has been identified where it is not possible (at least with certainty) to determine whether the occurrence exemplifies category-inclusion or category-exclusion. In this context, as highlighted by Vassiliadou et al. (2023), “semantic vagueness meets [...] subjective vagueness: as *pseudo*(-) exploits the existence of borderline cases and underlines the negative side of X, it questions the very categorization of X”. In the example provided in (24), it is unclear whether *pseudoljubav* ‘pseudo-love’ denotes a form of love that is not genuine (an imitation), a depreciated form of love, or even calls into question the concept of love itself.

(24) *Nedostatak ljubavi, koji se često očituje i kroz uvjetovanost ljubavi i razne oblike* ***pseudoljubavi****, može pridonijeti razvoju otuđenosti* [...]

‘Lack of love, which is often manifested through the conditionality of love and various forms of pseudo-love, can contribute to the development of alienation [...]’

This example, annotated as VAGUENESS, highlights the complex interplay between semantic and subjective vagueness, underscoring how *pseudo*(-) can destabilize clear-cut categorizations, particularly in contexts where the boundaries of concepts are fluid or disputed.

Table 11 presents the semantic classification of the 500 analyzed *pseudo*(-) expressions.

Tab. 11: Semantic classification of the 500-token *pseudo*(-) sample.

PREFIXOID	VALUE	TOKENS
pseudo(-)	SCIENTIFIC NON-EVALUATIVE	198
	SUBJECTIVE DEPRECIATIVE EVALUATION	301
	VAGUENESS	1

3.4.5. Distribution of Semantic Values

Following an analysis of the distribution of semantic values expressed by each of the four prefixoids within the 500-token sample per prefixoid, it is pertinent to visualize these distributions across all prefixoids. Table 12 illustrates the distribution of semantic values among the derivatives constructed with the four prefixoids.

Tab. 12: Distribution of semantic values among derivatives (500 tokens per prefixoid). For each prefixoid and semantic value, the token count is provided, with the type count indicated in parentheses.

VALUE / PREFIXOID	CTE	F	ILL	IM	K-C	SDE	V	SNE
kvazi(-)	4 (3)	2 (1)	/	2 (2)	/	492 (52)	/	/
nadri(-)	/	/	423 (6)	/	/	77 (10)	/	/
nazovi(-)	/	73 (12)	/	/	1 (1)	334 (41)	92 (9)	/
pseudo(-)	/	/	/	/	/	301 (19)	1 (1)	198 (36)

Observing the table, it can be confirmed how all four prefixoids are polyfunctional, with the distribution of the derivatives in relation to semantic value varying significantly among the prefixoids. Prefixoids *kvazi*(-) and *nazovi*(-) convey the highest number of semantic values (4), while *pseudo*(-) conveys 3, and *nadri*(-) only 2 values. To quantify the prefixoid polyfunctionality, we consider both the proportion of semantic values shared between them and the frequency of lexical realization (number of types) of these values, as per Huyghe et al. (2023). In alignment with the methodology of Salvadori, Varvara & Huyghe

(2024), we calculate the Hill-Shannon diversity index *D*[14] (cf. Roswell, Dushoff & Winfree 2021) utilizing the *MeanRarity*[15] package. According to *D*, *nazovi*(-) (*D* = 2.56) displays the highest diversity, followed by *pseudo*(-) (*D*=2.06) and *nadri*(-) (*D*=1.94), while *kvazi*(-) (*D*=1.55) is determined to be the least diverse.

While the necessity for a scalar evaluation of rivalry, that is, a metric[16] to delineate the degrees of rivalry for accurately assessing the intensity of a rivalry relationship, is apparent in occurrences of rivalry among multiple affixes (Huyghe et al. 2023; Salvadori, Varvara & Huyghe 2024), the present study delivers a comparatively clear scenario. It can be noted how five semantic values are associated exclusively with one specific prefixoid: "CLOSE-TO" EVALUATION and IMITATION to *kvazi*(-), ILLEGALITY to *nadri*(-), KIN-CATEGORIZATION to *nazovi*(-), and SCIENTIFIC NON-EVALUATIVE to *pseudo*(-). Conversely, the values of FAKENESS are expressed by *kvazi*(-) and *nazovi*(-), VAGUENESS is expressed by *nazovi*(-) and *pseudo*(-)[17], and, lastly, SUBJECTIVE DEPRECIATIVE EVALUATION is conveyed by all four prefixoids. Consequently, based on the examination of a 500-token sample for each prefixoid, it can be inferred that competition among the four prefixoids predominantly emerges within contexts characterized by the speaker's negative evaluation of a subject. In these instances, prefixoids are utilized to signify undesirable attributes in contrast to the prototypical instance. Conversely, for other, less frequent semantic values, distinct preferences concerning the selection of the prefixoid become evident. This observation does not come as a surprise, as the value referred to as SUBJECTIVE DEPRECIATIVE EVALUATION is, in fact, a very frequent part of language use, whereas functions like KIN-CATEGORIZATION or VAGUENESS appear less frequently in common discourse. The increased prevalence of

[14] Unlike the conventional Hill index, Hill-Shannon diversity emphasizes neither rare nor common species (in this context, semantic values). It is defined with the exponent that determines the rarity scale on which the mean is taken (*l*) of 0 and calculated with the base of the natural logarithm, *e*, raised to the power of the traditional Shannon entropy index (for detailed formalizations, see Roswell, Dushoff & Winfree 2021).

[15] https://mikeroswell.github.io/MeanRarity/articles/Using_MeanRarity.html (accessed 12 November 2024).

[16] Several methods that could (either implicitly or explicitly) account for affix rivalry have been explored. Fernández-Domínguez (2017) introduces a competition index that evaluates the prevalence of a derivative in relation to its rivals. Guzmán Naranjo & Bonami (2023) measure the similarity between different word-formation representations by calculating average vector offsets between the distributional representations of derivatives and their bases. Lastly, Salvadori, Varvara, & Huyghe (2024) utilize semantic annotation of derivatives to estimate the number of (un)shared functions among rival affixes, applying incidence-based measures (Sørensen index) and abundance-based measures (Percentage similarity coefficient).

[17] However, it has to be noted that only one instantiation of VAGUENESS with *pseudo*(-) is observed so it could be argued that VAGUENESS is, in fact, exclusively a property of *nazovi*(-).

negative evaluative contexts, consequently, provides more opportunities for competitive usage of these prefixoids.

4. Conclusions

This usage-based study has introduced and applied several statistical methods to analyze four competing <PREF$_{\text{APPRX}}$ + noun> constructions, revealing that the combinations of the analyzed prefixoids and nouns are not entirely unconstrained.

The analysis reaffirms that affix rivalry is a gradient phenomenon, necessitating the examination of a comprehensive range of factors to elucidate its nature. First, it was shown that prefixoids vary significantly in their productivity, with *nazovi*(-) identified as the most productive and *nadri*(-) the least productive, exhibiting a productivity difference of 3.72 times between the two (based on Potential Productivity). Second, the collocational preferences of the prefixoids were scrutinized. The hypothesis of a significant semantic similarity between *kvazi*(-) and *nazovi*(-), premised on a great number of shared bases between the two, was corroborated by MDCA-based Correspondence analysis. This analysis revealed that the two prefixoids cluster together, with ellipses embracing the collexemes that are associated with two prefixoids overlapping almost perfectly, thereby suggesting a functional and semantic similarity between them. In contrast, distinct semantic behavior of *nadri*(-) and *pseudo*(-) was observed, with no overlap in their collocational preferences. The distinct clustering of *pseudo*(-) with scientific and medical terms corroborates the prefixoid's specialized role in denoting false or imitative phenomena within technical discourse, while the separation of *nadri*(-) implies another specialized usage, viz. denoting fraudulent or unqualified individuals in professional roles. Lastly, a sample of 500 tokens for each prefixoid was analyzed, yielding several semantic readings and proposing a classification based on the typical values conveyed by these prefixoids.

Particularly intriguing are the findings regarding *kvazi*(-) and its possibility to express fakeness. As already noted, the reference literature examines *quasi*(-) in English (almost) exclusively as an approximative prefixoid. Bauer, Lieber & Plag (2013) highlight how *quasi(-)*, unlike *pseudo(-)*, lacks the element of falseness, while Dixon (2014: 171) uses a contrastive pair to illustrate the falseness *pseudo*(-) lacks to convey: whereas *quasi-cripple* represents "someone who has some small thing wrong with them (say, missing two fingers from one hand) but not really so serious to justify the label 'cripple'", a *pseudo-cripple* indicates "someone who has nothing at all wrong with them but pretends to be a cripple (perhaps, so that they can take part in the para-Olympics)". Cappelle, Daugs & Hartmann (2023) build upon this by asserting that *quasi(-)* predominantly has an approximative meaning, with *pseudo(-)* being characterized as disproximative. In the end, however, they acknowledge that *quasi*(-) can also convey disproximation, but the impression remains that that reading according to the authors is rather infrequent. In fact, Masini, Norde & Van Goethem (2023: 11) in their introduction to the Special Issue stress how according to Cappelle, Daugs & Hartmann (2023) the "idea of deficiency (...) is not prominent in *quasi*-". In this study, we assert that *kvazi*(-) in Croatian can convey both what Cappelle, Daugs & Hartmann (2023) define as *disproximation* (understood as a more inclusive, wider term), and *fakeness/falseness* (understood as "intentional act of deception or obfuscation" (Bauer, Lieber & Plag 2013: 416)). On a general note regarding semantic annotation, once again *kvazi*(-) and *nazovi*(-) display the most overlap, particularly in their shared use of SUBJECTIVE DEPRECIATIVE EVALUATION. Both prefixoids often function as tools of criticism, emphasizing the insufficiency or pretentiousness of the subject. There is also a notable, albeit less frequent, overlap between *kvazi*(-) and *pseudo*(-) in expressing IMITATION, highlighting their shared role in marking resemblance without authenticity. *Nadri*(-), while sharing the FAKENESS value with the other prefixoids, is more narrowly focused on legal contexts, limiting its broader evaluative overlap. *Pseudo*(-), with its scientific leanings, remains distinct but conceptually adjacent to *kvazi*(-) in contexts where imitation is highlighted without necessarily invoking deceit. Overall, the prefixoids display a complex web of relationships, yet each prefixoid also carves out a specific niche, balancing between shared semantic functions and unique, context-driven applications.

The adopted approach is expected to offer novel insights into the study of approximative prefixoids in Croatian. However, needless to say, the presented findings shed some light only on selected aspects of this previously overlooked phenomenon, and further research is needed to enhance our understanding of the analyzed prefixoids. An intriguing question remains regarding the original nature of these constructions. For instance, in the case *nadri*(-), while it cannot be confirmed due to limited resources, one could hypothesize that *nadriliječnik* 'quack' and the related term *nadriliječništvo* 'quackery' exemplify what Rainer (2013) refers to as a *leader word*, i.e. a word which serves as a model for the formation of new words and the development of a new construction. If this is indeed the case, it would suggest that the construction has witnessed a broadening of meaning (from what we defined as ILLEGALITY to SUBJECTIVE DEPRECIATIVE EVALUATION), giving rise to a more general approximative construction. However, it is equally, if not more, plausible that the reverse has occurred: a construction initially conveying SUBJECTIVE DEPRECIATIVE EVALUATION has become more specialized, entering legal terminology. For now, this question remains open. Furthermore, from a methodological perspective, a comprehensive understanding of rivalry resolution necessitates a holistic view of the potential differences between affixes (Huyghe & Varvara 2023). Future research should therefore incorporate regression modeling or classifier approaches, using the factors presented here along with additional variables. Moreover, due to prefixoids' different etymologies and presumably diverging periods in which they enter the lexicon, a diachronic study, comparing results from different time periods, may be relevant to account for this case of affix rivalry. However, given the scarcity of resources for Croatian (sufficiently large diachronic corpus or a tool such as Google Books Ngram Viewer), this task poses significant challenges. In addition, prefixoids' preferences regarding the etymology of their bases should be investigated. While some authors suggest that *kvazi*(-) is found with non-native words while *nadri*(-) and *nazovi*(-) combine with Croatian words (Ham 2015), a corpus-based study is needed to verify these claims. Finally, similar to Vassiliadou et al. (2023), instances of prefixoids appearing in quotation marks or parentheses, as seen for all four prefixoids, should be explored to determine whether they reflect merely an orthographic variation or actual uncertainty regarding the modified head's status (e.g., whether a term like (*nadri*)*liječnik* denotes a legitimate doctor or a potential quack).

References

Amiot, Dany & Dejan Stosic. 2022. Evaluative morphology: From evaluation to approximation and semi-categorization. In Héléne Vassiliadou & Marie Lammert (eds.), *A crosslinguistic perspective on clear and approximate categorization*, 53–94. Newcastle upon Tyne: Cambridge Scholars Publishing.

Aronoff, Mark. 2019. Competitors and alternants in linguistic morphology. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds.), *Competition in inflection and word-formation*, 39–66. Berlin: Springer.

Aronoff, Mark. 2023. Three ways of looking at morphological rivalry. *Word Structure* 16(1). 49–62.

Baayen, Harald & Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29. 801–844.

Baayen, Harald. 1992. A quantitative approach to morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 109–149. Dordrecht: Kluwer.

Baayen, Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 900–919. Berlin & New York: De Gruyter.

Bañón, Marta, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff & Jaume Zaragoza. 2023. *Croatian web corpus MaCoCu-hr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1516 (accessed 27 August 2024).

Barðdal, Jóhanna, Renata Enghels, Quentin Feltgen, Sven Van Hulle & Peter Lauwers. in prep. Productivity in diachrony. In Adam Ledgeway, Edith Aldridge, Anne Breitbarth, Katalin E. Kiss, Joseph Salmons & Alexandra Simonenko (eds.), *The Wiley Blackwell Companion to Diachronic Linguistics*. Oxford: Wiley-Blackwell.

Barić, Eugenija. 1979. Dosadašnje proučavanje složenica u hrvatskom i srpskom jeziku [Previous study of compounds in the Croatian and Serbian language]. *Rasprave* 4–5. 17–29.

Barić, Eugenija. 1980. Imeničke složenice s glagolskim prvim dijelom [Noun compounds with the verbal first part]. *Rasprave* 6–7. 17–30.

Baroni, Marco & Stephan Evert. 2014. *The zipfR package for lexical statistics: A tutorial introduction*. https://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf (accessed 17 July 2024).

Batinić, Mia, Marijana Kresić & Anita Pavić Pintarić. 2015. The intensifying function of modal particles and modal elements in a cross–linguistic perspective. *Rasprave* 41(1). 1–27.

Bauer, Laurie, Rochelle Lieber & Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

Bauer, Laurie, Salvador Valera & Ana Díaz-Negrillo. 2010. Affixation vs. conversion: The resolution of conflicting patterns. In Franz Rainer, Wolfgang U. Dressler, Dieter Kastovsky & Hans Christian Luschützky (eds.), *Variation and change in morphology*, 15–32. Amsterdam: Benjamins.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Benzécri, Jean-Paul. 1973. *L'analyse des données. 2. L'analyse des correspondances*. Paris: Bordas.

Boyd, Jeremy K. & Adele E. Goldberg. 2011. Learning what *not* to say: the role of statistical preemption and categorization in *a*-adjective production. *Language* 87(1). 55–83.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: Benjamins.

Cappelle, Bert, Pascal Denis & Mikaela Keller. 2018. Facing the facts of fake: A distributional semantics and corpus annotation approach. In Beate Hampe & Susanne Flach (eds.), *Yearbook of the German Cognitive Linguistics Association* 6, 9–42. Berlin & New York: De Gruyter.

Cappelle, Bert, Robert Daugs & Stefan Hartmann. 2023. The English privative prefixes *near-*, *pseudo-* and *quasi-*: Approximation and 'disproximation'. *Zeitschrift für Wortbildung/Journal of Word Formation* 7(1). 52–75.

Corbin, Danielle. 1987. *Morphologie dérivationelle et structuration du lexique*. Tübingen: Max Niemeyer.

Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/ (accessed 10 June 2024).

Covington, Michael A. & Joe D. McFall. 2010. Cutting the Gordian knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94–100.

Cúneo, Paola. 2015. Toba. In Nicola Grandi & Lívia Körtvélyessy (eds.), *The Edinburgh handbook of evaluative morphology*, 625–633. Edinburgh: Edinburgh University Press.

Desagulier, Guillaume. 2014. Visualizing distances in a set of near-synonyms: *rather*, *quite*, *fairly*, and *pretty*. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 145–178. Amsterdam: Benjamins.

Desagulier, Guillaume. 2015. Forms and meanings of intensification: a multifactorial comparison of *quite* and *rather*. *Anglophonia – French Journal of English Linguistics* 20. https://journals.openedition.org/anglophonia/558 (accessed 8 July 2024).

Dixon, Robert M. W. 2014. *Making new words: Morphological derivation in English*. Oxford: Oxford University Press.

Dressler, Wolfgang U. & Lavinia M. Barbaresi. 1994. *Morphopragmatics. Diminutives and intensifiers in Italian, German, and other languages*. Berlin & New York: De Gruyter.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Island and the distinctiveness of their occupancy. Francisco José Ruiz de Mendoza Ibáñez (ed.), *Annual review of cognitive linguistics: Volume 7*, 187–220. Amsterdam: Benjamins.

Ellis, Nick C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

Evert, Stephan & Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie & Shereen Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 conference*, 167–175. Lancaster: University Centre for Computer Research on Language.

Evert, Stephan & Marco Baroni. 2022. *Package 'zipfR'. Statistical Models for Word Frequency Distributions*. https://cran.r-project.org/web/packages/zipfR/zipfR.pdf (accessed 9 July 2024).

Fernández-Domínguez, Jesús. 2013. Morphological productivity measurement: Exploring qualitative versus quantitative approaches. *English Studies* 94(4). 422–447.

Fernández-Domínguez, Jesús. 2017. Methodological and procedural issues in the quantification of morphological competition. In Juan Santana-Lario & Salvador Valera (eds.), *Competing Patterns in English Affixation*, 67–117. Lausanne: Peter Lang.

Flach, Susanne. 2021. *Collostructions: An R implementation for the family of collostructional methods* (Version 0.2.0) [R Script].

Gaeta, Livio & Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.

Gaeta, Livio & Davide Ricca. 2015. Productivity. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation: An international handbook of the languages of Europe*, 842–858. Berlin & New York: De Gruyter.

Gardani, Francesco, Franz Rainer & Hans Christian Luschützky. 2019. Competition in morphology: A historical outline. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds.), *Competition in inflection and word-formation*, 3–36. Berlin: Springer.

Goldberg, Adele E., Devin M. Casenhiser & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15(3). 289–316.

Grandi, Nicola. 2017. Evaluatives in Morphology. In *Oxford Research Encyclopedia of Linguistics* (accessed 9 August 2024)

Grandi, Nicola & Livia Körvélyessy. 2015. Introduction: why evaluative morphology. In Nicola Grandi & Livia Körvélyessy (eds.), *Edinburgh handbook of evaluative morphology*, 3–20. Edinburgh: Edinburgh University Press.

Grandi, Nicola. 2002. *Morfologie in contatto: le costruzioni valutative nelle lingue del Mediterraneo*. Milan: FrancoAngeli.

Greenacre, Michael. 2017. *Correspondence analysis in practice. Third edition.* Boca Raton: Chapman & Hall/CRC Press.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics* 9(1). 97-129.

Gries, Stefan Th. 2015. More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.

Gries, Stefan Th. 2019. *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Leiden: Brill.

Gries, Stefan Th. 2022. Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics* 1(1).

Gries, Stefan Th. 2023. Overhauling collostructional analysis: Towards more descriptive simplicity and more explanatory adequacy. *Cognitive Semantics* 9(3). 351–386.

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

Ham, Sanda. 2015. Makro, mikro [Macro, micro]. *Jezik: časopis za kulturu hrvatskoga književnog jezika* 62. 144–145.

Hartmann, Stefan. 2018. Derivational morphology in flux: A case study of word-formation change in German. *Cognitive Linguistics* 29(1). 77–119.

Hein, Katrin & Annelen Brunner. 2020. Why do some lexemes combine more frequently than others? An empirical approach to productivity in German compound formation. In Jenny Audring, Nikos Koutsoukos & Christina Manouilidou (eds.), *Rules, patterns, schemas and analogy. Proceedings of Mediterranean Morphology Meetings 12*, 28–41. Patras: University of Patras.

Huyghe, Richard, Alizée Lombard, Justine Salvadori & Sandra Schwab. 2023. Semantic rivalry between French deverbal neologisms in *-age*, *-ion* and *-ment*. In Sven Kotowski & Ingo Plag (eds.), *The Semantics of Derivational Morphology. Theory, Methods, Evidence*, 125–158. Berlin: De Gruyter.

Huyghe, Richard & Rossella Varvara. 2023. Affix rivalry: Theoretical and methodological challenges. *Word Structure* 16(1). 1–23.

Ivšić, Stjepan. 1906–1907. Nešto o riječima složenima s *nadri-* [Something about words composed with *nadri-*]. *Nastavni vjesnik* 15. 525– 527.

Jojić, Ljiljana (ed.). 2015. *VRH – Veliki rječnik hrvatskoga standardnog jezika* [VRH – Large Dictionary of Croatian Standard Language]. Zagreb: Školska knjiga.

Klajn, Ivan. 2002. *Tvorba reči u savremenom srpskom jeziku. Prvi deo: slaganje i prefiksacija* [Word formation in the contemporary Serbian language. Part first: composition and prefixation]. Beograd – Novi Sad: Zavod za udžbenike i nastavna sredstva – Institut za srpski jezik SANU – Matica srpska.

Körtvélyessy, Lívia & Pavol Štekauer (eds.). 2011. Diminutives and augmentatives in the languages of the world. *Lexis: e-journal in English lexicology* 6. 5–25.

Körtvélyessy, Lívia. 2015. *Evaluative morphology from a cross-linguistic perspective*. Cambridge: Cambridge Scholars Publishing.

Kuna, Branko. 2006. Nazivlje u tvorbi riječi [Terminology in Word-Formation]. *Filologija* 46. 165–182.

Kuzman, Taja, Igor Mozetič & Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction* 5(3). 1149–1175.

Lacić, Ivan. 2024a. A corpus–based study of maximizer–adjective patterns in Croatian. *Language Sciences* 102.

Lacić, Ivan. 2024b. An insight into the Croatian degree modifier paradigm and its clustering profiles. *Suvremena lingvistika* 50(97). 85–112.

Lakoff, George. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* 2. 458–508.

Levshina, Natalia. 2015. *How to do linguistics with R. Data exploration and statistical analysis.* Amsterdam: Benjamins.

Liao, Shengyu, Stefan Th. Gries, & Stefanie Wulff. 2024. Transfer five ways: applications of multiple distinctive collexeme analysis to the dative alternation in Mandarin Chinese. *Corpus Linguistics and Linguistic Theory*, aop.

Ljubešić, Nikola & Davor Lauc 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In Bogdan Babych (ed.), *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, 37–42. Association for Computational Linguistics.

Ljubešić, Nikola & Filip Klubička. 2014. {bs,hr,sr}WaC –Web corpora of Bosnian, Croatian and Serbian. In Felix Bildhauer & Roland Schäfer (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, 29–35. Association for Computational Linguistics.

Ljubešić, Nikola & Peter Rupnik, Taja Kuzman. 2024. *Croatian web corpus CLASSLA-web.hr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1929 (accessed 10 May 2024).

Ljubešić, Nikola & Taja Kuzman. 2024. CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. *arXiv:2403.12721*.

Masini, Francesca & M. Silvia Micheli. 2020. The morphological expression of approximation: The emerging *simil-* construction in Italian. *Word Structure* 13(3). 371–402.

Masini, Francesca, Muriel Norde & Kristel Van Goethem. 2023. Approximation in morphology: A state of the art. *Zeitschrift für Wortbildung/Journal of Word Formation* 7(1). 1–26.

Merlini Barbaresi, Lavinia. 2015. Evaluative morphology and pragmatics. In Nicola Grandi & Livia Körvélyessy (eds.), *Edinburgh handbook of evaluative morphology*, 32–42. Edinburgh: Edinburgh University Press.

Nagano, Akiko, Alexandra Bagasheva & Vincent Renner. 2024. Towards a competition-based word-formation theory. Core research questions and major hypotheses. In Alexandra Bagasheva, Akiko Nagano & Vincent Renner (eds.), *Competition in word-formation*, 1–31. Amsterdam: Benjamins.

Napoli, Maria & Miriam Ravetto. 2017. Exploring intensification: synchronic, diachronic and cross-linguistic perspectives. Amsterdam: Benjamins.

Nigoević, Magdalena. 2020. *Intenzifikacija u jeziku: s primjerima iz hrvatskog i talijanskog jezika* [Intensification in language: With examples from the Croatian and Italian Language]. Split: Filozofski fakultet Sveučilišta u Splitu.

Norde, Muriel. 2009. *Degrammaticalization*. Oxford: Oxford University Press.

Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin & New York: De Gruyter.

Proisl, Thomas. 2022. *Use words, not constructions!* A new perspective on the unit of analysis in collostructional analysis. *International Journal of Corpus Linguistics* 27(3). 349–379.

Rainer, Franz. 2013. Formación de palabras y analogía. Aspectos diacrónicos. In Isabel Pujol Payet (ed.), *Formación de palabras y diacronía*, 141–172. A Coruña: Universidade da Coruña.

Roswell, Michael, Jonathan Dushoff & Rachael Winfree. 2021. A conceptual guide to measuring species diversity. *Oikos* 130(3). 321–338.

Salvadori, Justine, Rossella Varvara & Richard Huyghe (2024). Measuring affix rivalry as a gradient relationship. In Alexandra Bagasheva, Nagano, Akiko & Vincent Renner (eds.), *Competition in word-formation*, 104–138. Amsterdam: Benjamins.

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

Šonje, Jure (ed.). 2000. *Rječnik hrvatskoga jezika* [Dictionary of the Croatian language]. Zagreb: Leksikografski zavod Miroslav Krleža – Školska knjiga.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Stefanowitsch, Anatol & Susanne Flach. 2020. 'Too big to fail but big enough to pay for their mistakes': A collostructional analysis of the patterns [*too ADJ to V*] and [*ADJ enough to V*]. In Gloria Pastor Corpas & Jean-Pierre Colson (eds.), *Computational phraseology*, 247–272. Amsterdam: Benjamins.

Stefanowitsch, Anatol. 2013. Collostructional analysis. In Thomas Hoffman & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*, 209–306. Oxford: Oxford University Press.

Sundquist, John D. 2020. Productivity, richness, and diversity of light verb constructions in the history of American English. *Journal of Historical Linguistics* 10(3). 349–388.

Tafra, Branka & Petra Košutar. 2009. Rječotvorni modeli u hrvatskom jeziku [Word-Formation models in the Croatian language]. *Suvremena lingvistika* 35(67). 87–107.

Terčon, Luka & Nikola Ljubešić. 2023. *Word embeddings CLARIN.SI-embed.hr 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1790 (accessed 21 June 2024).

Van Goethem, Kristel & Muriel Norde. 2020. Extravagant "fake" morphemes in Dutch. Morphological productivity, semantic profiles and categorical flexibility. *Corpus Linguistics and Linguistic Theory*, 16(3). 425–458.

Vassiliadou, Hélène, Francine Gerhard-Krait, Georgia Fotiadou & Lammert, Marie. 2023. *Pseudo*(-) in French and Greek: Categorization and approximation. *Zeitschrift für Wortbildung/Journal of Word Formation* 7(1). 234–262.

Zeldes, Amir. 2012. *Productivity in argument selection: From morphology to syntax*. Berlin & New York: De Gruyter.

Ivan Lacić
Alma Mater Studiorum – University of Bologna
Department of Classical Philology and Italian Studies
via Zamboni, 32
40126 Bologna, Italy
ivan.lacic2@unibo.it