DOI 10.51558/2490-3647.2024.9.2.647

UDK 821.163.42'366
81'37

Primljeno: 01. 07. 2024.

Izvorni naučni rad
Original scientific paper

Gabrijela Buljan

TWO FACES OF ONE SUFFIX: SOME THOUGHTS ON USING CORPORA IN USAGE-BASED STUDIES OF WORD FORMATION[1]

This paper compares the semantic profile of a single multifunctional derivational suffix derived from data obtained in two general digital corpora of Croatian. The primary motivation is to explore whether our verdicts about the semantics of affixes may depend on the corpus selected as the source of empirical material. The issue is of vital importance, especially for those studying word formation from a usage-based perspective. If grammar is construed as the cognitive organization of our experience with language (Bybee 2006) and if we turn to large, general digital corpora for evidence of this experience, we must be aware that examining different corpora may lead to different hypotheses about users' internalized grammar. The here-presented semantic analysis of the Croatian nominal suffix *-ar(a)* in the more controlled Croatian National Corpus v3.0 and the liberal web-based corpus hrWaC v2.2 yielded conspicuously different results about its dominant function. This does not mean that similar discrepancies would necessarily be observed with other affixes, and it most certainly does not negate the value of corpora in studying word formation. However, such results do caution us against generalizing corpus-relative findings into some general "truth" about the affixes studied.

Key words: usage-based approach; word-formation; general corpus; semantic structure

1 An early version of this work was presented at the CLARC "Language and Language Data" conference held in Rijeka (Croatia), 28 – 30 September 2023.

1. INTRODUCTION

1.1. Background

The goal of the study of word-formation (WF) is to understand morphologically complex words and how and why we build them. Existing complex words are part of this agenda but are not equally central in different theoretical frameworks. Descriptive studies assemble actual words and describe them within semasiologically (Babić 2002) or onomasiologically (Barić et al. 1997; Silić & Pranjković 2007) organized units like affixes. Since the 1970s, generative morphologists have recognized actual words but, on the whole, their primary interest has been in possible words and modelling morphological competence (Plag 1999: 1, cf. also Aronoff 1976: 17–18; for some generative work involving large morphological databases see Simonović & Kovačević 2022; Milosavljević & Arsenijević 2022).

Authentic language data are central to functionally oriented schools/models like Dikkian Functional Grammar, Greenbergian typology, Givón (1979), Hopper & Thompson (1980) and Cognitive Linguistics. They are all usage-based in that they seek to explain the nature of grammar from how it is used and reject the language structure/knowledge (*langue*, *competence*) vs. language use (*parole*, *performance*) dichotomy. The cognitive linguistic usage-based paradigm (Barlow & Kemmer 2000, Bybee 2010) makes the additional claim that grammar is "the cognitive organization of one's experience with language" (Bybee 2006: 711). In this framework, actual words are the fabric of linguistic experience from which language users build their grammatical representation using domain-general cognitive processes like si-milarity/identity detection, i.e. categorization, chunking, rich memory storage etc. (Bybee 2010). Frequency of exposure to linguistic constructions is a key facilitator of these processes. It explains, among others, our ability to recognize what is or is not a conventional linguistic unit, and our occasional inability to perceive high-frequency derived words as complex due to their frequency-induced cognitive entrenchment (Wheeler & Schumsky 1980, in Taylor 2015: 152).[2]

Linguists who accept that grammar arises from linguistic experience may turn to large digital corpora when building hypotheses about users' internalized grammar. In WF especially, corpora supersede dictionaries as sources of empirical data (Baayen

2 Much evidence has been amassed for frequency effects, but other linguistic, processing-, and speaker-related factors have been found to co-shape grammar (Schmid 2017), like salience of forms/extralinguistic referents, emotional load (Blumenthal-Dramé 2012), cue-outcome contingencies and sociocultural background (Ellis 2006, 2012; Divjak & Caldwell-Harris 2015).

& Lieber 1991). They are not only our windows into vast swathes of linguistic experience; they also allow us to quantify complex words (token frequency) and more schematic constructions (type frequency) when gauging the semantic profile and productivity of affixes. Here, much rests on the assumption that the selected corpus is representative of the language "out there". But what is "the language out there"? For usage-based linguists, it cannot be the language of a fictitious ideal speaker/listener from a fictitious, homogeneous speech community. If, however, it is the language of some real speaker(s)/listener(s) and some real speech community, the question is – which real speech community/-ies exactly (Hohenhaus 2005)? The issue is far from trivial since WF has proven to be sensitive to sociolinguistic factors like age or gender (Clark 1982; Körtvélyessy 2009; Žanić 2010; Säily 2011), other non-structural factors like medium/register (Baayen 1992; Plag et al. 1999; Munat 2007; Miller 2014; Gaeta 2015), including the pragmatics of communication (Dressler & Merlini Barbaresi 1994). For instance, while abstract nominalizing suffixes are common in formal writing (Plag et al. 1999), evaluative morphology is more at home in spoken interactive discourse, in tabloid journalism (Mattiello 2008, Miller 2014, Hummel 2015, Costa 2017), and in "non-serious" speech situations marked by a high degree of familiarity and intimacy (Dressler & Merlini Barbaresi 1994).

This paper explores how our usage-based verdicts about the semantics of one Croatian suffix may diverge depending on where we sample users' linguistic experience. I examine the Croatian nominal suffix *-ar(a)* to see whether its recently observed semantic "drift" toward the evaluative function (Buljan 2024) will stand once we switch to a different data source. To that end, two databases will be formed, each featuring suffixations sourced from a different general corpus of Croatian, viz. the Croatian National Corpus (HNK, v3.0; Tadić 2005; 2009) and hrWaC (Ljubešić & Klubička 2014). The two corpora have been the go-to sources for recent empirical work on the Croatian language in general, including its WF (e.g. Brdar 2016; Bogunović et al. 2022; Filko 2020; Buljan 2023a,b, 2024) but are considerably different from each other in their textual content.

In the following, I briefly present *-ar(a)*'s description in authoritative studies (section 1.2). An overview of HNK and hrWaC follows in section 1.3. In section 2, I justify the choice of descriptive semantic categories and my approach to pattern quantification. Results and discussion are presented in section 3. The paper ends with conclusions in section 4.

1.2. On the suffix and its description

In Babić's (2002) monograph on Standard Croatian WF, *-ar(a)* is featured as a multifunctional suffix that builds nouns denoting people (*vračara* 'a fortune-teller' < *vračati* 'to tell fortune'), objects (*pepeljara* 'a receptacle for cigarette ash' < *pepeo* 'ash'), animals (*muzara* 'a milking cow' > *muzna* 'milking'), plants (*jajara* 'an egg-shaped plum' < *jaje* 'an egg'), etc. Its multiple functions notwithstanding, the suffix is regarded as primarily locative, since "almost half the derivatives denote places" (2002: 130; cf. Barić et al. 1997, also Kiršova 1999).

Valuable as Babić's monograph is, its data are quite dated – especially from our contemporary perspective. The bulk that made it into the monograph's 2002 edition had been carried over from its earlier editions, where the earliest year for data inclusion was set at 1860. Babić's study is possibly also biased away from the most spontaneous, informal language in favor of the language of high literature, academic/educational texts, broadsheet journalism and general dictionaries (2002: 5, 601–604). Since the focus was on the standard variety, some data selectivity is natural and expected (cf. Kendall 2011). The downside, however, is that we may know little about how the suffix was used by "ordinary" people at their most spontaneous and whether anything comparable to the following examples (from hrWaC) had existed before:

1. ***Pusara*** *veličine trokrilnog ormara Mayi i Dyni.*
   'A big kiss the size of a triple wardrobe to Maya and Dyna'
   pus(a) 'a kiss' > *pus-ar(a)* 'a kiss$_{aug}$'

2. *Ovi kaj rade u državnom i javnom sektoru su ko zaštićena kasta. Boli ih* ***đonara***.
   'Those working in the public sector are like a privileged caste. They don't give a rat's ass (lit. their shoe-sole$_{eval}$ hurts them)'
   đon 'a shoe-sole' > *đon-ar(a)* 'a shoe-sole$_{eval}$'

3. *E sad trea jedno 3 litre rakije, pa onda 4* ***gajbare*** *piva, naravno karlovachkog ...*
   'Right, now, we'll need about 3 liters of brandy, some 4 crates$_{eval}$ of beer, Karlovačko of course ...'
   gajb(a) 'a crate' > *gajb-ar(a)* 'a crate$_{eval}$'

Examples (1–3) feature *-ar(a)* in a modificational pattern of derivation (Dokulil 1968); it builds augmentative/pejorative or more generally evaluative[3] nouns that mean roughly the same as their bases (a big kiss is a kiss, only a big one). This function of *-ar(a)* has remained unrecognized in authoritative, "standard-oriented" WF studies, unless one counts as variants of *-ar(a)* the extended forms *-eskar(a), -uskar(a)* and *-onjar(a)*, which Babić (2002) exemplifies with several (pejorative) augmentatives like *popeskara* 'a priest$_{aug/pej}$', *braduskara* 'a beard$_{aug/pej}$' and *babonjara* 'an old woman$_{pej}$'.[4, 5]

A recent study using fresh, hrWaC-sourced data (Buljan 2024) showed, however, that *-ar(a)* sits comfortably in the "evaluative" territory, and that its evaluative function has surpassed or threatens to surpass its erstwhile dominant locative function in extent of use (type frequency) and productivity (as per Baayen's measure *productivity in the narrow sense* 1992). I will now explore the generalizability of this observation, but only for the type frequency measure.[6] Suffixations by extended forms like *-endar(a)* in *rup-endar(a)* 'a big hole' < *rup(a)* 'a hole' will be excluded, but not those where suffixation by *-ar(a)* is interpretable as a case of recursive suffixation of an already suffixed base, as in *ruk(a)* 'a hand' > *ruč-erd(a)* 'a big hand (pejorative)' > *ručerd-ar(a)* 'a very big hand (very pejoratively)'.[7]

1.3. On the corpora

The data for this study come from two very different general corpora of Croatian: hrWaC (v2.2., Ljubešić & Klubička 2014) and the Croatian National Corpus (HNK

3 *Evaluative* is a cover term for augmentatives, diminutives, pejoratives, melioratives, and some other types of modificational uses (Grandi & Körtvélyessy 2015, Bauer 1997: 538, Buljan 2024). Nouns that clearly interpret as (pejorative) augmentatives are here subscripted with aug(pej); where a quantitative (aug) or qualitative (pej/mel) reading is implausible but the noun still reads as evaluative (e.g. as purely expressive word-play, see examples 2, 3 and 11), it is subscripted with eval.

4 In Babić's semasiologically organized volume (2002), these extended forms appear in the same section as *-ar(a)* simply because affixes ending in the same letter string are discussed together. However, Babić does not exclude the possibility of a historical relatedness between such forms in principle (2002: 69). For more, see Buljan (2024).

5 Some such suffixations have made an appearance in work based on non-standard varieties; e.g. *kladara* appears as an evaluative/playful take on the more conventional *kladionica* 'a betting shop' (*kladiti se* 'to bet') in Košćak's study on Croatian slang (2018).

6 Baayen's measure of productivity cannot be used to compare data sourced from two different corpora (for the mathematical/logical background see Baayen & Lieber 1991, Baayen 2009).

7 I take attestation in the corpus or in dictionaries as evidence of the (non-)existence of intermediate stages of suffixation. Since **rupenda* 'a big hole' is absent from hrWaC and general dictionaries, I assume the derivation of *rupendara* from *rupa* with *-endar(a)* (Buljan 2024). Also, the precise contextual reading of *ručerda* and *ručerdara* may differ from the interpretations given here (cf. Buljan 2024).

v3.0, Tadić 2005, 2009).[8]

HNK and hrWaC are non-homogeneous collections of different kinds of content. For HNK we have information about the source (e.g. Official Gazette), register (e.g. fictional prose) or functional characteristics (e.g. informational texts) of the texts included. HrWaC content is marked for the year of crawling, part of webpage (heading or text), website and web domain (websites ranging from those of various government agencies to those of cat food producers). This information is incomparable and neither corpus includes metadata necessary for a targeted automatic search of suffixations according to criteria like register, genre, style, topic, medium etc. Nonetheless, we could reasonably conclude that HNK and hrWaC are considerably different in the proportions of content representing different "text-types" or "registers" (see 2.1).[9]

HNK is a fairly well-balanced corpus of Standard Croatian. HNK's first version (30 Mw) includes: 1. Information-centered texts: newspapers, magazines, non-fictional prose (22 200 000 tokens or 76%); 2. Fictional prose (6 900 000 tokens or 22%); 3. Mixed texts: fictional/non-fictional prose, speeches (900 000 tokens or 2%). The second version (v2.5) grew to 104.3 Mw through incorporation of material from *Vijenac*, a bi-weekly newspaper featuring topics in culture, science and arts. The content added to the latest version (v3.0; 234 Mw) comes from the Croatian Official Gazette (Štefanec, pc).

The Croatian Web Corpus hrWaC 2.2 (1.4B tokens) includes texts automatically crawled from the .hr top-level domain (Ljubešić & Klubička 2014). It is unbalanced and lacks a structured breakdown of content comparable to that of HNK v1.0. Still, we learn from Figure 1 that the two biggest content contributors are the websites forum.hr, blog.hr, which together account for 25.7% of hrWaC's token-size. Also included are company, government, news and magazine websites but there is a conspicuous absence of the Croatian literary canon.[01]

Although we only have a coarse understanding of their content, it is a fair assessment that HNK underrepresents highly informal spontaneous everyday language, while hrWaC is not strong on the language of the academia, high culture and literature.

8 The analysis of hrWaC data presented here was already completed when I became aware of the launch of a new, much bigger and genre-annotated web corpus of Croatian, the so-called CLASSLA-web.hr corpus (Ljubešić & Kuzman 2023).

9 HrWaC also contains various non-standard and dialectal content, but this is impossible to control for given the anonymity of much web-crawled content.

01 A detailed breakdown is available via SketchEngine.

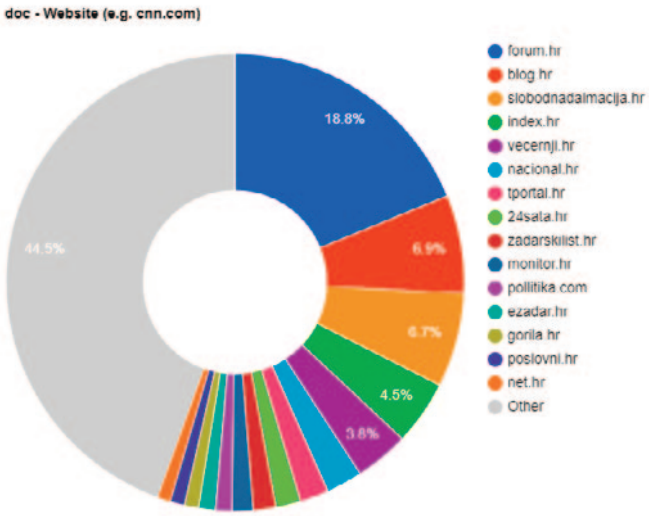

Figure 1. Websites sourced for hrWaC (size-ranked by token coverage)

2. METHODOLOGY

2.1. On semantic categories and text type categories

Two types of semantic categories are generally used in WF. Ideally, whenever the meanings of morphologically complex words can be described relationally – as a function of the meaning of the base/motivating word (Pounder 2000, Mengel 2009) – they should be so described. 'Evaluative' is one such relational category. The derivative modifies the meaning of the base, usually by moving its implicit property up/down some quantitative (size)/qualitative (affect) scale. Still, we often find complex words described ontologically as 'nouns for people', 'nouns for animals', etc. Nouns like *preživač* 'an animal that ruminates' < *preživati* 'to ruminate' can thus be described relationally as 'one (animal) that Vs' (cf. Babić 2002: 97) or as an 'animal noun' (cf. Barić et al. 1997: 315). Similarly, a noun like *zmijara* 'a big snake' < *zmija* 'a snake' can be described relationally as 'a big entity (animal)' or as 'an animal noun'. My primary focus here is on *-ar(a)*'s suspected "evaluative turn" and less so on the myriad, usually minor and sometimes difficult-to-identify relational meanings among non-evaluative suffixations (note the relationally diverse kinds of place nouns like 'place to store N', 'place producing N', 'place made of N'). Therefore, my de-

scriptive framework will include 'evaluative' as the only relational category,[11] the remaining data will be classed – for simplicity – into ontological categories like 'people' (*vračara*), 'animals' (*muzara*), 'objects' (*pepeljara*), etc.[21] The latter will subsequently be collapsed into the joint 'non-evaluative' category when testing for significance the observed difference between the two datasets in the proportions of evaluative to non-evaluative meanings (section 3.2).

Finally, to discuss potential links between corpus structure and semantic structure, I will take inspiration from Biber & Conrad's (2009: 5) definition of registers, and from Egbert et al.'s (2015) framework for usage-based web-register classification (modified somewhat based on a comparable classification in Ljubešić & Kuzman 2023).[31] Registers are seen as text varieties fundamentally based on situational characteristics (like topic or communicative purpose), whose pervasive linguistic features are functionally associated with such contexts (e.g., face-to-face conversation has more 1st and 2nd person pronouns than academic prose).[41] Above all, registers are defined by purposes like "narrating events" (as in travel blogs, personal blogs, historical articles, etc.), "describing information" (as in research articles, legal terms, course materials, etc.), "expressing opinions" (as in opinion blogs/articles, reviews, letters to editors), "using facts to persuade" (as in commercial websites, descriptions of sales items, promoting events), "explaining instructions" (as in how-to manuals, recipes), "expressing oneself lyrically" (as in lyrics, poems, prayers), and include discussion/QA forums, as platforms for interactive discussions on various topics in the form of comments (Egbert et al. 2015: 1825, Ljubešić & Kuzman 2023). This part of

11 I should point out that evaluative nouns are also ontologically diverse and may refer to objects (examples 2, 3), abstract entities (example 1), places (*London* > *Londonjara*), people (*komunjara*), etc.

21 Interestingly, in Babić *komunjara* 'a diehard communist; a communist pejoratively' is described ontologically, as 'a noun for male or female persons' and is only parenthetically glossed as depreciative (p. 130). However, the noun is arguably relationally evaluative – a pejorative derivation from *komunist* 'a communist'.

31 Egbert et al's (2015) and Ljubešić & Kuzman's (2023) classifications do not match in every detail. Notably, Egbert et al. class novels and short stories into the narrative register together with dry, matter-of-fact kinds of informational prose like news reports, historical articles, and reserve the "lyrical register" for lyrics, poems and prayers. In contrast, Ljubešić & Kuzman class all "literary" texts into the same – as they call it – prose/lyrical genre. Neither system would be delicate enough for a study of the distribution of evaluative morphology within literary texts specifically. This would be an interesting future research project, since it has already been shown that morphological evaluatives are more likely to be found in comedies, the types of texts closer to spoken interaction, and in dialogic sections of narrative texts (Grandi 2003, in Gaeta 2015).

41 In Biber & Conrad's system, registers and genres are simply different perspectives on studying situational varieties. Assuming a genre perspective entails exploring linguistic features used to structure complete texts, like conventional openings or closings of a business letter (2009: 16). The terms *register* and *genre* have had a long and troubled history (for some discussion see Kuzman & Ljubešić 2023), but many now use them interchangeably (cf. Egbert et al. 2015).

my analysis will remain preliminary and informal. Since it is supremely difficult to distinguish registers in practice (especially in the slippery landscape of computer-mediated communication, cf. Giltrow & Stein 2009), I will employ a working classification of registers loosely organized around prevailing communicative purposes like those listed above (see Table 4). Second, I will restrict this analysis to hapaxes for feasibility. Since different tokens of the same noun may belong to different registers, the alternative would require analyzing a sizable corpus of richly contextualized tokens of all the nouns in the two databases.[51] The decision to focus on hapaxes only is also partly justified by the fact that most novel suffixations are likely to be found among hapaxes – at least in the much bigger hrWaC corpus.[61]

2.2. On quantification

I established the type frequency of *-ar(a)*'s meanings, viz. in how many different nouns each meaning category occurred. Since some nouns are ambiguous (*betonjara* 'a building made of concrete' or 'a factory producing concrete'), multiple tokens of each suffixation were examined to capture any additional meanings. Specifically, all tokens were examined, up to a random 300-token sample for nouns whose frequency exceeds 300. Nouns with two different meanings, like *betonjara*, contributed two counts of – in this case – locative meanings. Thus, in column 2 of Table 1 the type frequency of meanings is higher than the type frequency of nouns (in parentheses). The same procedure was applied regardless of whether the ambiguous form qualifies as one polysemous noun with several related senses (cf. the metonymically linked *ludara* 'a crazy head' vs. 'a crazy person' > *lud* 'crazy') or as different homonymous nouns (*kožara* 'a thick-skinned winter apple' < *koža* 'skin' vs. *kožara* 'a tannery (leather manufacturing plant)' < *koža* 'leather, hide').[71] This methodological move facilitates quantification and significance testing and is justified insofar as my interest is not in polysemy/homonymy.

A final note, since my aim is to stake out *-ar(a)*'s semantic space on the corpus-based evidence of how contemporary speakers of Croatian are actually using the suf-

51 This would be impossible for many tokens from hrWaC whose original texts/websites no longer exist.

61 On various terms that suggest "novelty of a word" – viz. *hapax, nonce-formation, neologism* and *occasionalism,* and their non-overlapping uses see Bauer (2001), Hohenhaus (2005), Mikić Čolić (2021), Buljan (2024). On how corpus size correlates with the likelihood that hapaxes are indeed novel, rather than old but rare words see Baayen & Renouf (1996), Plag et al. (1999), Gaeta & Ricca (2015) and below.

71 I will ignore readings representing postderivational semantic extensions, like the metonymic development of 'staff' readings of locatives, as in the hypothetical *Cijela kožara dobila je očekivanu povišicu* 'the whole tannery got a much-expected raise'.

fix, I will unapologetically include in my counts even those data that others would ostracize to the creative "fringes" of WF (like examples 1–3).[81]

3. RESULTS AND DISCUSSION

3.1. General descriptives

Table 1. Some descriptive statistics for the two databases of *-ar(a)* nouns sourced from HNK and hrWaC

Corpus (1)	TypF of meanings (2)	Hapax (3)	Dictionary-attested hapaxes (4)
HNK (30M)	105 (95 nouns)	20	9/20 (45.0%)
hrWaC (1.4B)	462 (362 nouns)	115	8/115 (7.0%)

According to column (2) of Table 1, the suffix appeared in 105 distinct readings of 95 formally distinct *-ar(a)* nouns in HNK and in 462 distinct readings of 362 formally distinct *-ar(a)* nouns in hrWaC. This discrepancy is fairly unremarkable with hrWaC being 46.7 times the size of HNK (column 1).

More interesting is the difference in hapaxes (columns 3 and 4). While the 20 hapaxes sourced from HNK account for 21.1% of the 95 distinct nouns in *-ar(a)*, of 362 distinct nouns in hrWaC as many as 115 (31.8%) are hapaxes. Moreover, almost half the hapaxes in HNK are old, dictionary-attested words that probably just did not have room to recur (column 4). Among them, the following figure in the Croatian Encyclopedic Dictionary (HER, 2004) and/or the 12-volume Dictionary of Croatian or Serbian by the Yugoslav Academy of Arts and Sciences (AR, 1880-1976): *ludara* 'a type of fungus, Rubroboletus satanas', *tintara* 'an ink container (of head)', *opajdara* 'a rantipole', *blatnjara* 'a mud house', *kućara*, obsolete augmentative for 'a house'. Some of the dictionary-absent ones occur with some frequency online, which suggests that they may enjoy some familiarity in "the speech community". They include, among others, *pekmezara* 'a jam factory', *jeftinjara* 'something cheap' or *Kenjara* (< *kenjati* 'to take a shit'), the pejorative nickname for a quadrant of Diocletian's Palace in Split where people illegally dump garbage/defecate. In sum, considering – conservatively – only dictionary attestations, almost half HNK's hapaxes are not new

81 Such coinages tend to be neglected especially in studies of productivity. As a technical term in WF, productivity is highly controversial. Among others, some refuse counting as instances of "productive WF" any words created "deliberately" for special, short-term effect, especially if this involves any form of "rule-breaking" (Schultink 1961; Aronoff 1976; Bauer 1983; 2001; Dressler 2000; but see Munat 2007; Miller 2014; Buljan 2024).

at all; with those attested online (N = 17), the not-so-new nouns account for a staggering 85.0% of HNK's hapaxes. By comparison, only 8 (7.0%) of hrWaC hapaxes belong to the established norm/lexicon, the rest are novel, still unregistered suffixations. Dictionaries feature, among others, *lokardara* 'a fishing net for mackerels', *mazutara* 'a mazut plant', *paščara* 'a dog house', *platnara* 'a linen production facility'. Since hrWaC consists entirely of web-crawled content, I assumed that a hrWaC hapax is somewhat familiar if, regardless of its dictionary-attestation, it occurred online in at least one more example beyond that found in hrWaC. This was the case with 51 hapaxes (44.3%). All in all, there is a clear difference in the number of hapaxes that are presumably not new in HNK 45.0% (or 85.0%) vs. hrWaC 7.0% (or 44.3%). This discrepancy is unsurprising given that HNK is far smaller than hrWaC, but it does suggest that the text types/registers included in the two corpora may not be equally welcoming of certain types of lexical innovation. This will be more obvious when I later return to the semantics of hapaxes.

Whether the truly novel suffixations, especially those from hrWaC, are likely to survive beyond their original context of use is a different issue. Many are evaluative and quite likely ephemeral, cf. examples (1–3), and (4 –5):

(4) *… i kod nas ima idijota,* ***klinjara*** *koje si još vrte Zeitgeist i nisu ni uzeli brigu da pogledaju prave dokaze*
'There are idiots around here too, kids$_{aug/eval}$ who're still into the Zeitgeist and never bothered to examine real evidence'

(5) *Da se ti nebi povukao ka Sanader a poslije završio u bajbuku. Hvala bogu da vise necemo slusati o tom* ***majmunjari***
'You aren't going to retreat like Sanader, only to end up in the cooler. Thank goodness we won't be hearing about that ape$_{aug/eval}$ any more …'

3.2. *Semantic structure of -ar(a) in HNK and hrWaC*

Tables 2 presents the semantic breakdown of *-ar(a)* in the two complete databases.

Table 2. TypF of *-ar(a)*'s meanings in hrWaC and HNK

	Semantic category		hrWaC		HNK	
			TypF	%	TypF	%
1	Evaluative		**157**	**34.0**	6	5.7
2	Non-evaluative	Location	147	31.8	**47**	**44.8**
3		Object	67	14.5	24	22.9
4		Person	28	6.1	6	5.7
5		Food/drinks	17	3.7	6	5.7
6		Plant	16	3.5	7	6.7
7		Animal	16	3.5	6	5.7
8		Body	7	1.5	1	1.0
9		Abstract	7	1.5	2	1.9
	Total		462	100.0	105	100.0

Location is the most TypF meaning but only in the more conservative HNK. HNK results actually corroborate Babić's intuition that "almost one half of all *-ar(a)* nouns are locative" (2002: 130). The only other category that reaches double digits are 'object' nouns like *kupusara* metaphorical for 'a voluminous book with little valuable content' < *kupus* 'cabbage'. The meagre six evaluatives are *komunjara*, *gužvara* 'a big crowd' < *gužva* 'a crowd',[91] *sunjara* 'a hot, scorching sun' < *sun(ce)*[02] 'the sun', the obsolete *kućara* 'a cottage, shed$_{aug}$' < *kuća* 'a house' (AR V: 731), *miljara* 'a thousand$_{eval}$' < *milja* 'a thousand' and *kafanara* 'a coffeehouse $_{eval}$' < *kafana* 'a coffeehouse'.[12]

In hrWaC, evaluatives are dominant. Locatives are close, but rank second nonetheless. This result most certainly flies in the face of traditional description (Babić 2002, Barić et al. 1997).

Although HNK and hrWaC are so obviously different in the proportion of evaluatives to non-evaluatives, this was checked for significance using a Chi-Square Test of Independence. In the HNK, observed frequencies indicated that 6 nouns were eval-

91 *Gužvara* appeared more frequently in its more common meaning 'a type of pie with crumpled pastry'.

02 Brackets here suggest base shortening, a common practice in WF (Babić 2002: 35; cf. Mel'čuk's deletion 1932: 493).

12 Admittedly, *kafanara* (a hapax) might also be the genitive plural inflection of the non-target agentive *kafanar* 'a coffee house owner'. The example could be read either way: *Upravo su oni glavni gosti kavana. Stoga, zagrebačka društvenost je dakle društvenost kafanska i gostionička (...)Bez kafanara nema kod nas društvenosti "jer" glavni naš društveni faktor je alkohol, vino.* 'They are the main patrons of coffee houses. Zagreb social life is, therefore, of the coffeehouse and tavern "brand" (…) Without coffee houseseval/coffee house owners, there is no social life "because" our main socializing factor is alcohol, wine'.

uative, in contrast to 99 non-evaluatives (Expected frequencies: Evaluatives = 30.2, Non-Evaluatives = 74.8). In hrWaC, 157 nouns were evaluative, while 305 were not (Expected frequencies: Evaluative = 132.8, Non-Evaluative = 329.2). The Chi-Square Test yielded significant results ($\chi^2(1) = 33.377$, $p = .000$, Phi = .24). The association between corpus and meaning structure proved to be significant, with a moderately sized effect.

The difference between HNK- and hrWaC-sourced data is even starker in the subset of hapaxes (Table 3). In hrWaC, evaluative hapaxes are almost 2.5 times as frequent as locative hapaxes, while in HNK locatives prevail (at 45%).[22]

Table 3. TypF of meanings among *-ar(a)* hapaxes in hrWac and HNK

	Semantic category		hrWaC		HNK	
			TypF	%	TypF	%
1	Evaluative		**56**	**48.7**	2	10.0
2	Non-evaluative	Location	23	20.0	**9**	**45.0**
3		Person	9	7.8	2	10.0
4		Object	13	11.3	1	5.0
5		Plant	5	4.3	1	5.0
6		Food/drinks	4	3.5	4	20.0
7		Animal	3	2.6	0	0.0
8		Body	2	1.7	1	5.0
	Total		115	100.0	20	100.0

To save space, I can only flesh out these dry figures with a handful of evaluatives and locatives. The only two evaluative hapaxes in HNK are *kućara* and *kafanara*. Evaluative hapaxes from hrWaC include (a selection): *zmijara* < *zmija* 'a snake', *zakonjara* < *zakon* lit. 'law', used in slang as a secondary interjection (Ameka 2006) meaning 'cool', *uspješničara* < *uspješnica* 'a success story', *uhljebara* < *uhljeb* 'one who is employed (usually in the public sector) through connections, not on merit', *tipkovnjara* < *tipkovnica* 'a keyboard', *Snejpara* < *Snejp* (professor Severus Snape), *pusara* (example 1), *majmunjara* (example 5) and many more. Locative hapaxes in hrWaC include mainly nouns for commercial locatives like *vijčara* 'a bolt and screw factory' < *vijak* 'a bolt/screw', *vlažara* 'a humidifying chamber' < *vlažiti* 'to humidify', *mrežara* 'a (fishing) net factory' < *mreža* 'a (fishing) net', or "made-of" locatives, like *kartonjara* 'a makeshift shelter made of cardboard' < *karton* 'cardboard' or *crijepara* 'a house with a tile roof' < *crijep* 'a tile'. The locative hapaxes in HNK include *čipsara* 'a potato chip factory' < *čips* 'a potato chip', and the earlier mentioned *pekmezara* and *Kenjara*.

22 This difference could not be tested for significance due to one low cell count (2 evaluatives/HNK).

Table 4 shows how the hapaxes (regardless of meaning) distribute according to my working categories of register (section 2.1). Interestingly, hrWaC is the most hapax-intensive in the one register where HNK – for obvious reasons (cf. section 1.3) – has none, viz. the interactive discussion/QA forum.

Table 4. Distribution of hapaxes per register in hrWaC and HNK

	Register	hrWaC		HNK	
		Frq	%	Frq	%
1	Interactive, discussion: Discussion forum, QA forum	**51**	**44.3**		
2	Opinion: Opinion blogs/articles, reviews, advertisements	18	15.7	2	10.0
3	Informational description/explanation (objective/factual): research articles, legal terms, course materials, encyclopedic entries, news/sports reports etc.	19	16.5	**8**	**40.0**
4	Narrative (narrating events for entertainment or informational purposes): personal blogs, magazine articles, travel blogs, historical articles	25	21.7	**9**	**45.0**
5	Literary: fictional prose, poetry, song lyrics, etc.	2	1.7	1	5.0
	Total	115	100.0	20	100.0

Before exploring in some detail the semantics/register intersection, let me illustrate how examples from one meaning category (location) manifest in different registers:

Interactive discussion (forum)

(6) *Dal je i u toj **fensinjari** napojnica obavezna …*
'Is tipping a must in that **fancy place** of yours'

Opinion (blog)

(7) *Jer kad imaš dvadeset i nešto, a odrastao si u ovom "sustavu vrijednosti" koji duhovno nije ništa drugo do žalosna sklepotina od **kartonjara** i gdjekoje zahrđale limenke poput onih na periferiji pakla zvanog Ciudad de Mexico … normalno je da se grebeš za lovu svim sredstvima*.
'Coz, when you're twenty some, and you've grown up in this "value system", a spiritually pitiful ragbag of **cardboard shelters** and rusty sheet metal constructions like those on the outer rim of the hell called Ciudad de Mexico … it's normal that you should be all about how to get your hands on money'

Information/explanation

(8) ***Mrežara** nam već tri mjeseca radi u tri smjene i tako će, s obzirom na narudžbe, biti do kraja godine*.
'Our **net production line** has been running in three shifts for three months now, and – with the orders in the pipeline, so it will remain until year-end'

Narrative

(9) *Tada ga koristi komorska uprava, a kasnije vojska. Nakon toga u njemu je **Platnara**, a 1849. godine i nautički kabinet*.
'At the time it was used by the chamber directorate, later by the military. Thereafter, it housed **a linen production facility**, and in 1849 the maritime chamber'

Literary

(10) *Miha Rabar bi se svako jutro prije zore šuljao oko Tučine* ***slamnjare****, a kada bi zarudjelo, on se skupi uza zid i nagnuvši pažljivo glavu napinjao bi oči gledajući u sobnu polutminu*.
'Every morning before dawn, Miha Rabar would sneak around Tuča's **straw roofed house**; when the dawn broke, he's press himself against the wall, poke his head out carefully and, straining his eyes, peer into the semi-darkness of the room'

Tables 5a (on HNK) and 5b (on hrWaC) chart the distribution of the semantic categories of hapaxes per register. Allowing for the very small number of HNK hapaxes altogether, and for the fact that half are not even novel nouns, it is rather unsurprising that the busiest intersection in the "soberly conservative" HNK is between location, on one side, and the informational (3 locative hapaxes) and the narrative (5 locative hapaxes) registers, on the other (Table 5a).

Table 5a. Crosstabulation of semantic category x register for HNK hapaxes

	Semantic categ.		Register					
			Interact.	Inform.	Narrat.	Opinion	Literary	Total
1	Evaluative				1	1		2
2	Non-evaluative	Locative		**3**	**5**		1	**9**
3		Person			1	1		2
4		Object		1				1
5		Animal						
6		Plant		1				1
7		Food/drinks		3	1			4
8		Body			1			1
	Total			8	9	2	1	20

More interesting is Table 5b. It shows that HrWaC hapaxes are not only overwhelmingly evaluative (row 1), but the majority of them come from interactive discussion/QA forums, meaning notwithstanding (column 1).

Table 5b. Crosstabulation of semantic category x register for hrWaC hapaxes

	Semantic categ.		Register					
			Interact.	Inform.	Narrat.	Opinion	Literary	Total
1	Evaluative		**27**	1	14	14		**56**
2	Non-evaluative	Locative	8	7	4	2	2	23
3		Person	7		2			9
4		Object	6	2	3	2		13
5		Animal		3				3
6		Plant	1	3	1			5
7		Food/drinks	1	3				4
8		Body	1		1			2
	Total		**51**	19	25	18	2	115

In other words, the forum has the highest concentration of largely novel (cf. section 3.1) nouns which quite expectedly (given the nature of the register) are of the evaluative kind. In this register, people are most at ease, quick to express qualitative and quantitative evaluation of others, and to engage in creative word-play as in the purely ludic example (11).

(11) *Normalno kupljen, sa naručenzi i draj monate čekanjzi je oko hundert nojn und nojncih* ***kunjara****…*
'Bought through regular channels, with zee order and drei monate ze veiting, it's around hundert neun und neunzig kuna$_{eval}$'

There is no doubt that the setting of computer-mediated communication (CMC) and our perceptions of how such communication works are affecting our communication behavior. Above all, its "metaphorical and technical sense of distance" lends us more freedom than we enjoy in monitored face-to-face communication. Beyond that, the general uncertainty about message uptake may also drive us to becoming ever more creative – if only to attract attention (Giltrow & Stein 2009: 13). Nevertheless, one should be careful not to give CMC too much credit as it most certainly does not have exclusive rights to evaluative morphology (or WF innovation in general). Evaluative morphology is also at home in slang and face-to-face interaction regardless of medium, in the less controlled written varieties like dialogic parts of literary narratives, or comedies (Gaeta 2015). This only means that if we could compare hrWaC-sourced *-ar(a)* nouns with, for instance, *-ar(a)* nouns sourced from some (as yet non-existent) corpus consisting entirely of informal, face-to-face spoken interaction, the results would most certainly not diverge so dramatically.

It may well be that not every affix is as sensitive to the effects of register as *-ar(a)* turned out to be. But the results reported in this study still contribute to over four decades of research which showed that register – difficulties of definition and classification aside – may be one of the most powerful predictors of linguistic variation (Biber & Conrad 2009, Egbert et al. 2015), including in WF (Baayen 1994, Plag et al 1999, Gaeta 2015). Clearly, the results also suggest that much more register-oriented work is warranted in future for a fuller understanding of our (changing) WF behavior, especially if we use that behavior to build our bottom-up hypothesis about its cognitive representation.

4. CONCLUSION

In this paper I set out to compare the semantic profile of the multifunctional Croatian suffix *-ar(a)* based on data sourced from two very different general corpora of Croatian, HNK and hrWaC. Its locative function (traditionally recognized as central) remains unchallenged in the conservative HNK, but yields to evaluatives in the more "liberal" web-based hrWaC. If hapaxes are any good as indicators of WF productivity/creativity (in the non-technical sense), we must conclude that hrWaC is much more receptive of innovation, especially in the evaluative arena. This is not only because it is considerably larger than HNK (guaranteeing that its many hapaxes will indeed be novel words), but because it is home to texts of a most interactive/informal nature. Despite the preliminary nature of my conclusions and the possibility that I may have not found "the most natural" seams between register types, the data unambiguously show that our hypotheses about language users' grammar may be considerably affected by corpus composition. Admittedly, other affixes may not be equally sensitive to the textual composition of corpora, but we should nonetheless keep an eye out on register as a potential factor behind linguistic variation – especially if we use corpus data to generate usage-based hypotheses about grammar.

REFERENCES:

1. Ameka, Felix K. (2006), ″Interjections″, in: Keith Brown (ed.), *Encyclopedia of language & linguistics*. 2nd edn., Elsevier, Oxford, 743–746.
2. Aronoff, Mark (1976), *Word-formation in Generative Grammar*, MIT Press, Cambridge Mass.

3. Baayen, Harald (1992), ″Quantitative aspects of morphological productivity″, in: Geert Booij, Jaap van Marle (eds.), *Yearbook of morphology 1992*, Kluwer, Dordrecht, 109–149.
4. Baayen, Harald (1994), "Derivational productivity and text typology", *Journal of Quantitative Linguistics* 1(1), 16–34. https://doi.org/10.1080/09296179408589996
5. Baayen, Harald (2009), ″Corpus linguistics in morphology: Morphological productivity″, in: Anke Lüdeling, Merja Kytö (eds.), *Corpus linguistics. An international handbook*, *Volume* 2, De Gruyter, Berlin, 899–919.
6. Baayen, Harald, Rochelle Lieber (1991), ″Productivity and English derivation: a corpus-based study″, *Linguistics*, 29, 801–844.
7. Baayen, Harald, Antoinette Renouf (1996), ″Chronicling the Times: productive lexical innovations in an English newspaper″, *Language*, 72, 69–96.
8. Babić, Stjepan (2002), *Tvorba riječi u hrvatskome književnome jeziku*, 3rd improved edn., HAZU; Nakladni zavod Globus, Zagreb
9. Barić, Eugenija, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, Marija Znika (1997), *Hrvatska gramatika*. 2nd revised edn, Školska knjiga, Zagreb
10. Barlow, Michael, Suzanne Kemmer (eds.) (2000), *Usage-based models of language*, CSLI, Stanford
11. Bauer, Laurie (1983), *English word formation*, CUP, Cambridge
12. Bauer, Laurie (1997), ″Evaluative morphology: In search of universals″, *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 21(3), 533–575.
13. Bauer, Laurie (2001), *Morphological productivity*, CUP, Cambridge
14. Biber, Douglas, Susan Conrad (2009), *Register, genre, and style*, CUP, Cambridge
15. Blumenthal-Dramé, Alice (2012), *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*, De Gruyter, Berlin
16. Bogunović, Irena, Jasmina Jelčić Čolakovac, Mirjana Borucinsky (2022), ″The Database of English words and their Croatian equivalents″, [Baza]
17. ″Brdar, Mario (2016), ″Why Modrić and Real rather than Real and Modrić? On the order of proper names under coordination″, *Jezikoslovlje*, 17(1-2), 377–395.
18. Buljan, Gabrijela (2023a), ″Značenja i oblici hrvatskog sufiksa -AR-A: korpusna studija″, *Fluminensia*, 35(1), 27–59.

19. Buljan, Gabrijela (2023b), ″Neke misli o nastanku augmentativnog/evaluativnih značenja hrvatskog sufiksa –ara″, *Suvremena lingvistika*, 49(95), 1–27.
20. Buljan, Gabrijela (2024), *Aspects of innovation in Croatian word-formation: A corpus-based study of suffixes* -ara, -ana and -stan, Faculty of Humanities and Social Sciences, Osijek
21. Bybee, Joan L. (2006), ″From usage to grammar: The mind's response to repetition″, *Language*, 82(4), 711–733.
22. Bybee, Joan L. (2010), *Language, usage and cognition*, CUP, Cambridge
23. Clark, Eve (1982), ″A case study of innovation in the child's lexicon″, in: Eric Wanner & Lila R. Gleitman (eds.), *Language acquisition: The state of the art*, CUP, Cambridge, MA, 390–425.
24. Costa, Marcella (2017), ″Augmentatives in Italian and German: From contrastive analysis to translation″, in: Maria Napoli, Miriam Ravetto (eds.), *Exploring intensification: Synchronic, diachronic and cross-linguistic perspectives,* John Benjamins, Amsterdam, 353–370.
25. Daničić, Đuro, Matija Valjavec, Pero Budmani, Tomo Maretić, Stjepan Musulin, Slavko Pavešić, eds. (1880–1976), *Rječnik hrvatskoga ili srpskoga jezika*, JAZU, Zagreb
26. Divjak, Dagmar, Catherine L. Caldwell-Harris (2015), ″Frequency and entrenchment″, in: Eva Dabrowska, Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*, De Gruyter, Berlin, 53–75.
27. Dokulil, Miloš (1968), ″Zur Theorie der Wortbildung″, *Wissenschaftliche Zeitschrift der Karl-Marx-Universität, Gesellschafts- und Sprachwissenschaftliche Reihe*, 17(2-3), 203–211.
28. Dressler, Wolfgang U. (2000), ″Extragrammatical vs. marginal morphology″, in: Ursula Doleschal, Ana M. Thornton (eds.), *Extragrammatical and marginal morphology*, Lincom, München, 1–10.
29. Dressler, Wolfgang U., Lavinia Merlini Barbaresi (1994), *Morphopragmatics: diminutives and intensifiers in Italian, German and other languages*, De Gruyter, Berlin
30. Egbert, Jesse, Douglas Biber, Mark Davies (2015), ″Developing a bottom-up, user-based method of web register classification″, *Journal of the Association for Information Science and Technology*, 66(9), 1817–1831.
31. Ellis, Nick E. (2006), ″Language acquisition as rational contingency learning″, *Applied Linguistics*, 27(1), 1–24.

32. Ellis, Nick C. (2012), ″Frequency-based accounts of SLA″, in: Susan Gass, Alison Mackey (eds.), *Handbook of Second Language Acquisition*, Routledge, London & New York, 193–210.
33. Filko, Matea (2020), *Unutarleksičke i međuleksičke strukture imeničkoga dijela hrvatskoga leksika*, Doktorski rad, Filozofski fakultet u Zagrebu, Zagreb
34. Gaeta, Livio (2015), ″Evaluative morphology and sociolinguistic variation″, in: Nicola Grandi, Lívia Körtvélyessy (eds.), *Edinburgh Handbook of Evaluative Morphology*, Edinburgh University Press, Edinburgh, 121–133.
35. 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. Volume* 2, De Gruyter, Berlin, 842–858.
36. Giltrow, Janet, Dieter Stein, eds. (2009), *Genres in the Internet. Issues in the theory of genre*, John Benjamins, Amsterdam/Philadelphia
37. Givón, Talmy (1979), *On understanding grammar*, Academic Press, New York
38. Grandi, Nicola, Lívia Körtvélyessy (2015), ″Introduction: why evaluative morphology?″, in: Nicola Grandi, Lívia Körtvélyessy (eds.), *Edinburgh handbook of evaluative morphology*, Edinburgh University Press, Edinburgh, 3–21.
39. Hohenhaus, Peter (2005), ″Lexicalization and institutionalization″, in: Pavol Štekauer, Lieber Rochelle (eds.), *Handbook of Word-Formation*, Springer Verlag, Dordrecht, 353–373.
40. Hopper, Paul, Sandra A. Thompson (1980), ″Transitivity in grammar and discourse″, *Language*, 56, 251–299.
41. Hummel, Martin (2015), ″The semantics and pragmatics of Romance evaluative suffixes″, in: Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, Franz Rainer (eds.), *Word-Formation. An International Handbook of the Languages of Europe. Volume* 2, De Gruyter, Berlin, 1528–1545.
42. Jojić, Ljiljana, Ranko Matasović (eds.) (2002–2004), *Hrvatski enciklopedijski rječnik* (HER), vols. 1–12, EPH d.o.o. and Novi Liber, Zagreb
43. Kendall, Tyler (2011), ″Corpora from a sociolinguistic perspective″, *Revista Brasileira de Linguística Aplicada*, 11(2), 361–389.
44. Kiršova, Mirjana (1999), *Nomina loci u savremenom srpskom jeziku*, Univerzitet Crne Gore, Podgorica
45. Košćak, Nikola (2018), *Šrajbenzi spiku: Stilovi hrvatske žargonske i žargonizirane proze 1990-ih i 2000-ih*, Stilistika.org, Zagreb

46. Körtvélyessy, Lívia (2009), ″Productivity and creativity in word-formation: A sociolinguistics perspective″, *Onomasiology Online*, 10, 1–22.
47. Kuzman, Taja, Nikola Ljubešić (2023), ″Automatic genre identification: a survey″, *Language Resources and Evaluation*. 10.1007/s10579-023-09695-8
48. Ljubešić, Nikola, Filip Klubička (2014), ″{bs,hr,sr}WaC - Web Corpora of Bosnian, Croatian and Serbian″, *Proceedings of the 9th Web as Corpus Workshop* (WaC-9)", Gothenburg, Sweden: Association for Computational Linguistics, 29–35.
49. Ljubešić, Nikola, Taja Kuzman (2023), ″CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation″, *Machine Learning and Knowledge Extraction*, 5, 1149–1175.
50. Mattiello, Elisa (2008), *An introduction to English slang: A description of its morphology, semantics and sociology*, Polimetrica, Milan
51. Mel'čuk, Igor (1932), *Aspects of the theory of morphology*, De Gruyter, Berlin
52. Mengel, Swetlana (2009), ″Wortbildungsbedeutung″, in: Sebastian Kempgen, Peter Kosta, Tilman Berger, Karl Gutschmidt (eds.), *The Slavic Languages/An International Handbook of their structure, their history and their investigation. Band 1/Volume 1*, De Gruyter, Berlin, 775–781.
53. Mikić Čolić, Ana (2021), *Neologizmi u hrvatskome jeziku*, Filozofski fakultet, Osijek
54. Miller, Gary D. (2014), *English lexicogenesis*, OUP, Oxford
55. Milosavljević, Stefan, Boban Arsenijević (2022), ″What differentiates Serbo-Croatian verbal theme vowels: content or markedness?″, *Glossa: a journal of general linguistics*, 7(1), https://www.glossa-journal.org/article/id/8535/
56. Munat, Judith (2007), ″Lexical creativity as a marker of style in science fiction and children's literature″, in: Judith Munat (ed.), *Lexical creativity, texts and contexts*, John Benjamins, Amsterdam, 163–185.
57. Plag, Ingo (1999), *Morphological productivity. Structural constraints on English derivation*, De Gruyter, Berlin
58. Plag, Ingo, Christiane Dalton-Puffer, Harald Baayen (1999), ″Morphological productivity across speech and writing″, *English Language and Linguistics*, 3(2), 209–228.
59. Pounder, Amanda (2000), *Process and paradigms in word-formation morphology*, De Gruyter, Berlin
60. Säily, Tanja (2011), ″Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations″, *Corpus Linguistics and*

Linguistic Theory, 7(1), 119–141.
61. Schmid, Hans-Jörg (ed.) (2017), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, De Gruyter, Berlin
62. Schultink, Henk (1961), ″Produktiviteit als morfologisch fenomeen″, *Forum der Letteren*, 2, 110–125.
63. Silić, Josip, Ivo Pranjković (2007), *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*, Školska knjiga, Zagreb
64. Simonović, Marko, Predrag Kovačević (2022), ″Possessive, kind and not so kind: the different uses of the adjectival *-ov* in Serbo-Croatian″, *Annual Review of the Faculty of Philosophy*, 47(3), 87–109.
65. Tadić, Marko (2005), ″Developing the Croatian National Corpus and beyond″, in: Peter Grzybek, (ed.), *Contributions to the science of text and language*, Springer, Dordrecht, 295–300.
66. Tadić, Marko (2009), ″New version of the Croatian National Corpus″, in: Dana Hlaváčková, Aleš Horák, Klara Osolsobě, Pavel Rychlý (eds.), *After half a century of Slavonic natural language processing*, Masaryk University, Brno, 199–205.
67. Taylor, John R. (2015), ″Word-formation in cognitive grammar″, in: Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, Franz Rainer (eds.), *Word-formation. An international handbook of the languages of Europe. Volume* 1, De Gruyter, Berlin, 145–158.
68. Žanić, Ivo (2010), ″Purgerinjosi, tovarinjosi i leginjice – tvorbene inovacije u hrvatskim vernakularima″, in: Mario Brdar, Marija Omazić, Višnja Pavičić Takač, Tanja Gradečak-Erdeljić, Gabrijela Buljan (eds.), *Prostor i vrijeme u jeziku: jezik u prostoru i vremenu*, HDPL - Filozofski fakultet Sveučilišta J. J. Strossmayera, Zagreb/Osijek, 155-164.

DVA LICA ISTOGA SUFIKSA: NEKE MISLI O KORIŠTENJU KORPUSA U UPORABNIM PRISTUPIMA ANALIZI TVORBE RIJEČI

U ovom se radu uspoređuje semantički profil jednog multifunkcionalnog derivacijskog sufiksa temeljen na analizi podataka prikupljenih iz dvaju općih digitalnih korpusa hrvatskog jezika. Primarna je motivacija za ovaj rad utvrditi mogu li naši zaključci o značenju afiksa ovisiti o korpusu odabranom kao izvor empirijskog materijala. To je pitanje od ključne važnosti posebno kada se tvorbi riječi pristupa iz perspektive uporabnih modela jezika. Ako gramatiku smatramo kognitivnom organizacijom našeg jezičnog iskustva (Bybee 2006) i ako posegnemo za velikim, općim digitalnim korpusima kao odrazima toga iskustva, moramo biti svjesni da bismo upotrebom podataka iz različitih korpusa mogli doći do različitih hipoteza o internaliziranoj gramatici. U ovom radu iznosimo rezultate semantičke analize hrvatskog imeničkog sufiksa *-ar(a)* u kontroliranijem Hrvatskom nacionalnom korpusu v3.0 i u liberalnijem mrežnom korpusu hrWaC v2.2, u kojima su različite funkcije toga sufiksa pokazale dominantnima. To ne znači, naravno, da bi se slične nepodudarnosti nužno utvrdile analizom nekih drugih afiksa i nipošto ne umanjuje vrijednost upotrebe korpusa u tvorbenim istraživanjima. No ipak nas upozorava da svoje rezultate trebamo tumačiti kao odraz lingvističkog iskustva "zahvaćenog" dotičnim korpusom, ne nužno kao općevažeću istinu o analiziranim afiksima.

Author's address
Adresa autorice

Gabrijela Buljan
J. J. Strossmayer University of Osijek
Faculty of Humanities and Social Sciences
gbuljan@ffos.hr