A corpus-based study on the inventory and grammaticalisation of compound indefinite pronouns in Croatian

Roman Fisun, Björn Hansen

The class of series-forming indefinite pronouns in Croatian is, like in other Slavonic languages, affected by ongoing grammaticalisation processes. In addition to the indefinite pronouns recorded in grammar manuals and in studies on indefiniteness, it can be extended to include a number of elements whose degree of grammaticalisation varies considerably. Based on corpus material from the hrWaC (Croatian web corpus) and HNK (Croatian National Corpus), we were able to provide the most complete list of Croatian indefinite markers yet compiled, and to analyse the lexical sources of their grammaticalisation. Furthermore, we propose an approach for calculating the degree of grammaticalisation of indefinite pronouns. For this purpose, fifteen specific sub-criteria were developed on the basis of Lehmann's theory of grammaticalisation, and a mixed method combining both quantitative and qualitative analysis of corpus material was used to evaluate the degree of grammaticalisation according to these sub-criteria.

Keywords: indefinite pronouns; Croatian; grammaticalisation; association measures; corpus linguistics; grammaticalisation parameters

Fisun, Roman • University of Regensburg • roman.fisun@ur.de

Hansen, Björn • University of Regensburg • bjoern.hansen@ur.de

The research is part of the project "Compound indefinite pronouns in Slavonic languages. A contribution to a second-generation semantic map of indefiniteness" funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) (HA 2659/10-1).

Die Welt der Slaven 69 (2024) 2, 293–330 DOI: 10.13173/WS.69.2.293

1. Introduction[1]

Indefinite pronouns in Croatian include elements such as the ones in bold in the following sentences:

(1) *Hoću li* ***išta*** *doznati?*
'Will I learn anything?'

[1] We thank Zrinka Kolaković, Edyta Jurkiewicz-Rohrbacher and Martina Rybová for their valuable comments on an earlier draft of this paper. We further express our gratitude to Krystyna Kupiszewska for proofreading and to Michael Wastl for his assistance with formatting the manuscript.

(2) ***Tko god*** *je vidio njegove slike, bio je zadivljen.*
'Anyone who saw his pictures was impressed.'

(3) *Mi smo šutjeli, ali su* ***ostali*** *vikali.*
'We were silent but the others shouted.' (Barić et al. 1997: 206f.)

From a morphological point of view, we can distinguish between compound indefinite pronouns (CIPs)[2] such as *i-šta* and *tko god* on the one hand, and simple indefinite pronouns such as *ostali* on the other. In this paper, we deal exclusively with the former; we also exclude the closely related category of negators (like *ništa*).[3] We will apply the label 'pronoun' to both nominal forms and pro-adverbs. In this sense, we do not distinguish between the terms 'pronouns' and 'pro-forms' (cf. Bhat 2004). CIPs consist of up to three components: a marker of indefiniteness (like *i, god*), a **k*-root[4] for the ontological category (*što*, *tko*) and—in some cases—a case ending (Marković 2013: 331).[5] We will call the marker of indefiniteness a 'modifier' following Bondareva (2010), who introduced the term *modifikator* for the respective component in Russian. Pre- and postposed indefinite markers are sometimes considered affixes or 'affixoids' (Ermakova 1996; 2000), but the term *modifier* seems more accurate: the indefinite markers of CIPs, as will be shown below, do not form a homogeneous group and only partially have the features of fully morphologised affixes. There are 'series' of indefinite pronouns that contain the same modifier and, consequently, are characterised by the same type of indefiniteness: the series with *i-* encompasses *itko, išta, ikakav, igdje* etc. Following Haspelmath (1997), we analyse such series as a whole instead of examining each combination of a modifier with a **k*-root. We consider only nine Croatian **k*-roots, corresponding to the eight main ontological categories: 'person' (*tko*), 'thing' (*što*), 'property' (*kakav, koji*), 'manner' (*kako*), 'place' (*gdje*), 'amount' (*koliko*), 'reason' (*zašto*) and 'time' (*kada/kad*). To simplify matters somewhat, we assume that each **k*-root (or two **k*-roots in the case of 'property') corresponds to only one ontological category and thus use the terms **k*-root and ontological category synonymously below.

In line with the grammaticographic tradition, Marković (2013: 317) claims that pronouns form a closed word class with a small number of elements (*zatvorena i malobrojna vrsta*). However, with regard to indefinite pronouns this is problematic: besides the so-

[2] The term *compound indefinite pronouns* was inspired by Ermakova (1996) (*sostavnye mestoimenija* in Russian). As a separate group, CIPs have received the most attention in Russian studies. Works fully devoted to this problem that should be mentioned include Ermakova (1996; 2000), PhD theses by Sokolova (2007) and Bondareva (2010), and also studies by Testelec & Bylinina (2005) and Mel'čuk (2012). These scholars have proposed a number of terms for the units described: *compound pronouns* (*sostavnye mestoimenija* by Ermakova), *amalgams* and *quasi-relatives* (*amal'gamy i kvazireljativy* by Testelec & Bylinina), and *K-expressions* (*k-vyraženija* by Mel'čuk).

[3] We omit negative pronouns and prefixes from the literature review.

[4] The term proposed by Isačenko (1965: 163) refers to all modern stems of Slavonic interrogative pronouns that etymologically go back to the same Proto-Slavonic root **kъ-* (cf. also Marković 2013: 357; quoting Skok 1971: 74).

[5] Marković (2013) uses the term *opći pojam*. Ontological categories are discussed in the works Jackendoff (1983) and Haspelmath (1997).

called 'traditional'[6] indefinite pronouns such as *tko god* 'whoever', *netko* 'someone', etc., numerous expressions of indefiniteness consisting of several orthographic words have emerged as a result of ongoing language change.[7] Some of these, like *ma tko* are mentioned in grammar manuals and in studies on indefiniteness. They are said to take on pronominal functions, but they are not necessarily treated as pronouns in the proper sense of the word (e.g. in Barić et al. 1997, see below). One example of such a CIP that is not mentioned in the grammars is *tko zna*. It clearly has an indefinite meaning, which can be seen in the following examples from the parallel corpus InterCorp, in which it is rendered in English with the indefinite pronouns *some* and *anything:*

(4) *Tako čudni da se moglo pomisliti kako ih zajedno drži* ***tko zna kakva*** *tajna.*
'So strange that one thought **some** secret must connect them.'

(5) *Da nije bilo vas* ***tko zna što*** *bi se dogodilo.*
'Without your help, **anything** might have happened here.' (InterCorp v14 – Croatian-English)

In this paper we would like to address the following two research questions: which elements in Croatian can be treated as CIP modifiers, and what is their degree of grammaticalisation? Section 2 briefly examines the treatment of indefinite pronouns in Croatian grammaticography and lexicography; Section 3 contains a corpus-based study that aims to identify the inventory of CIP modifiers; it also discusses the applicability of different queries and statistical measures; in Section 4 we analyse the degree of grammaticalisation of individual Croatian modifiers, while taking into account real language use and its dynamic character. We show that grammaticalisation theory is well-suited for capturing the obvious variability. The study is based on mixed methods, combining corpus studies based on the massive hrWaC 2.2 web corpus with qualitative analyses. In Section 5 we summarise the most important findings.

We would like to underline that due to space limitations we do not deal with the semantics of Croatian CIPs here, except for semantic aspects discussed in connection with the parameters of grammaticalisation (Section 4.2).

2. The state of research

In this section, we will attempt to establish which elements are treated as modifiers in Croatian grammaticography. First, we will discuss the treatment of indefinite pronouns (Croatian 'neodređene zamjenice') in some reference grammars of Croatian.

Among indefinite pronouns, Maretić (1963: 198ff.) lists only elements spelled together and points out two cases of the separate spelling of modifiers with **k*-roots: separation of prefixal indefinite pronouns by prepositions and optional separation of the series with *:god* by encyclical elements. In total, he mentions six preposed and one postposed element:

[6] Some authors refer to the fully grammaticalised indefinite pronouns as "traditional" or "orthodox" (cf. Ermakova 1996; Sokolova 2007; Bondareva 2010).

[7] In Russian linguistics, it was noted as early as in the 19th century that indefinite pronouns are affected by dynamic processes (cf. Lavrov 1983; Sokolova 2007).

- preposed: *gdje:, i:, koje:, ne:, sav:/sva:/sve:, što:*
- postposed: *:god*

The 11th edition of Težak & Babić's grammar (1996: 113f.) counts 9 modifiers of indefinite pronouns and distinguishes within the group between substantival ('iminične') and adjectival ('pridjevne') ones. The proadverbs with the same modifiers are accordingly not taken into account:

- preposed: *bilo, gdje:, i:, koje:, ne:, ma, sav:/sva:/sve:*
- postposed: *:god and god*

The grammar does not give information about the status of these modifiers, but it lists the **k*-roots they can be attached to. It is worth noting that the number of **k*-roots with different modifiers is not consistent.

The description distinguishes between the writing of *god* together and separately with **k*-roots: "nisu složenice, nego dvije reči sa dva akcenta [they are not compounds, but two words with two accents]" (op. cit., 114). The authors emphasise that these word combinations also have a special meaning and cite relevant examples:

(6) ***Tkȍ gȍd*** *ga znâ svȁk ga hvâlī.*
'Everyone who knows him praises him.' (op. cit., 114)

Barić et al. (1997) distinguish personal ('lične'), possessive ('posvojne'), reflexive ('povratne'), demonstrative ('pokazne'), interrogative ('upitne'), relative ('odnosne') and indefinite pronouns ('neodređene zamjenice') relevant for the present study (ibid.). The chapter 'Indefinite pronouns' mentions the following elements (op. cit., 206f):

- preposed: *ne:,*[8] *i:, sva:, gdje:, po:, što:, koje:, pone:*
- postposed: *:god*

Furthermore, they list certain word combinations with **k*-roots used as indefinite pronouns. The elements are written separately.

- preposed: *ma, makar, bud, budi*
- postposed: *god,*[9] *mu drago, ti volja, te volja, hoćeš, hoće*
- preposed or postposed: *bilo*

Finally, there are non-compound indefinite pronouns such as *koji, jedan, drugi, ostali* and *ini.* They do not form series like the elements such as *i:* (*itko, išta,* etc.) mentioned above do. It is interesting to note that in the textbook Silić & Pranjković (2007: 127f.) the

[8] To avoid inconsistencies in the representation of the orthography, we will use the following notation, which diverges from the Croatian standard: a colon (*gdje:*) indicates the spelling of a modifier together with the **k*-root, as in *gdjetko, gdješta*, etc.; a hyphen (*-god*) stands for the rare non-normative hyphenated spelling seen in *tko-god*; the absence of additional signs (*ma*) represents spelling of the series as a separate element, like *ma tko.*

[9] *God* as a separately written unit appears in both groups: as an indefinite pronoun and as a word combination. According to Jozić et al. (2013), variation in spelling is assumed to indicate different meanings.

inventory of modifiers of indefinite pronouns is smaller than in Barić et al. (1997). They mention the following:

- preposed: *ne:, i:, sav:/sva:/sve:, koje:, gdje:*
- postposed: *:god*
- with particles: *ma, god, bilo, mu drago, god mu drago*

In contrast to Barić et al. (1997), the authors give a list of possible modifier-**k*-root combinations (op. cit., 127).

Kunzmann-Müller (2002: 164–168) also refers to the heterogeneity of the category of indefinite pronouns. She sees "Reihen aus Interrogativpronomen und verschiedenen Affixen [series of interrogative pronouns and various affixes]" as the core of the category (op. cit., 164) and mentions 8 elements (see below); the separately spelled *god* belongs to this group as well. This element is placed in a separate section from *:god*. Furthermore, Kunzmann-Müller mentions "idiomatisierte Ausdrücken in der Position vor oder hinter dem Pronomen [idiomatised expressions in the position before or after the pronoun]" (op. cit., 168):

- affixes: *ne:, pone:, i:, koje:, što:, gdje:, :god, god*
- preposed idiomatised expressions: *ma, makar, bilo*
- postposed idiomatised expressions: *mu (ti, vam) drago*

Finally, we briefly present the treatment of indefinite pronouns in the handbook of Croatian syntax by Katičić (1986). The author does not dedicate a separate chapter to indefinite pronouns but discusses how selected pronouns behave in interrogative sentences. Katičić (op. cit., 138ff.) discusses the element *ne:* treating *nešto, netko, nečiji* and *nekakav* as indefinite pronouns and the forms *negdje* and *nekuda* as indefinite adverbs. He further mentions the elements *:god, god* and *i:*.

Marković (2013: 332) discusses pronouns from a morphological perspective. He mentions four elements that he calls indefinite prefixes (*ne:, i:, pone:, sva:*), and three preposed (*bilo, ma, makar*) and three postposed particles (*bilo, god, mu drago*).

The definiteness–indefiniteness distinction in Croatian has received some attention (e.g. Marković 2002), often in relation to adjective declension (e.g. Znika 1987; Hansen 2004a). However, there are only very few works that deal specifically with indefinite pronouns. Progovac (1990) provides a formal syntactic analysis of *bilo* as compared to the English *any*. She calls both 'free choice items'. In her squib, the author argues that the two readings, as bound (narrow scope existential) or as free (wide scope universal), depend on the distance of the trigger. The works of Progovac (1991, 1994) additionally deal with the modifier *i:*, treated as a negative polarity item. Progovac's work is also quoted in Haspelmath's (1997) broad typological study on indefinite pronouns. Among the rich cross-linguistic data presented, Haspelmath mentions the Croatian pronouns with the following modifiers:

- preposed: *bilo, bogzna:, i:, ne:*
- postposed: *bilo, mu drago*

Šarić (2002: 190–193) discusses some indefiniteness markers in the broader context of quantification. She analyses the indefinite pronouns *ijedan, itko, netko, bilo koji, bilo tko, bilo*

što, išta as logical existential quantifiers with the meaning 'at least one' (*bar jedan*) and the negation pronouns *nijedan, nitko, ništa* as their counterparts. (op. cit., 191).

Wonisch (2012: 129–134) gives an overview of the inventory of indefinite pronouns from a general Slavonic perspective. He distinguishes between prefixes and particles that can form indefinite series in Bosnian, Croatian, Montenegrin and Serbian (op. cit., 131):

- prefixes: *ne:, pone:, i:, sva:/sav:/sve:, koje:, gdje:, što:*
- particles: *god* (and additionally "synthetic" *:god*), *ma, bilo, mu drago*

After having browsed grammar handbooks and a couple of linguistic studies, we checked the online lexicographic resource Hrvatski Jezični Portal (HJP) and the normative dictionary Školski rječnik hrvatskoga jezika (ŠRHJ). The following table contains information about all the modifiers found either in the grammar handbooks or in the lexicographic resources. We try to indicate cases where the assignment of part of speech goes with specific **k*-roots through alignment in the corresponding cells.

Tab. 1: The coverage of CIP modifiers in selected grammaticographic and lexicographic sources

Modifier	Literature	**k*-root (HJP)	PoS (HJP)	**k*-root (ŠRHJ)	PoS (ŠRHJ)
Preposed					
bilo:	/	kakav	pronoun	/	/
bilo	TB, B, SP, Mk, KM, H	/	/	kako, što, tko	particle
bogzna:	H	što, kako	adverb	/	/
gdje:	Mt, TB, B, SP, KM, W	tko, što, gdje, kad (lemma: gdje-)	first part of complex adverbs and pronouns	kad	adverb
				tko, što	indefinite pronoun
				koji	pronoun
		čiji, koji (separate lemmas)	indefinite pronoun		
i:	Mt, TB, B, SP, Mk, KM, H, W	tko, šta, koji, kakav	pronoun adjective + pronoun	tko, šta, koji, kakav	indefinite pronoun
				gdje, kada, kako	adverb
		gdje, koliko, kad, kada, kako	adverb		
koje:	Mt, TB, B, SP, Mk, KM, W	šta, tko	indefinite pronoun	šta, kakav	indefinite pronoun
		kakav	adjective	kako, kud, kuda	adverb
		kako, kud, kuda	adverb		

Modifier	Literature	*k-root (HJP)	PoS (HJP)	*k-root (ŠRHJ)	PoS (ŠRHJ)
ne:	Mt, TB, B, SP, Mk, KM, H, W	tko, što, koji, kakav	indefinite pronoun	tko, što, kakav	indefinite pronoun
				kako, gdje, koliko, kada	adverb
		kako, gdje, koliko, kad, kada	adverb		
po:	B	koji	pronoun+ adverb	koji	indefinite pronoun
pone:	B, Mk, KM, W	tko	pronoun	tko, što, ki	indefinite pronoun
		što	pronoun/ adverb	kad, gdje	adverb
		(ki – not a k-root)	pronoun + adjective		
		kad, gdje	adverb		
ma	TB, B, SP, Mk, KM, W	tko, što, koji, kakav, gdje, koliko, kada, kad, kako, kuda	conjunction as part of general pronouns and adverbs	gdje, kad, kako, koliko, što, tko	particle
makar	B, Mk, KM	tko, koliko	conjunction as part of general pronouns and adverbs	/	/
malo[10]	/	što, koji	adverb used with relative pronoun	/	/
bud	B	koji	pronoun	/	/
		što	adverb		
budi	B	kakav	pronoun	/	/
sva:	Mt, TB, B, SP, Mk, W	tko, čiji, šta	pronoun	tko, čiji, šta, ki, kakav	indefinite pronoun
		ki	indefinite pronoun	kako	adverb
		kakav	adjective		
		kad	adverb		
		kako	particle/adverb		
što:	Mt, B, KM, W	šta	indefinite pronoun	/	/

[10] *Malo* is not mentioned in the grammatical descriptions and specialised works on indefiniteness but does occur in the HJP. It was only added to the table after its extraction from the web corpus (see below).

Modifier	Literature	*k-root (HJP)	PoS (HJP)	*k-root (ŠRHJ)	PoS (ŠRHJ)
Postposed					
bilo	B, SP, Mk, H, W	/	/	gdje, kad, koliko	particle
:god	Mt, TB, B, SP, KM, W	tko	pronoun	tko, što kakav	indefinite pronoun
		što	indefinite pronoun	gdje, kad, kako	adverb
-god[11]	/	tko, što, kakav, gdje	second part of general pronouns or adverbs	/	/
god	TB, B, Mk, KM, W	/	/	/	/
hoće	B	/	/	/	/
hoćeš	B	/	/	/	/
mu drago	B, SP, Mk, KM, H, W	/	/	/	/
te/ti volja	B	/	/	/	/

Literature: B – Barić et al. 1997; H – Haspelmath 1997; KM – Kunzmann-Müller 2002; Mk – Marković 2013; Mt – Maretić 1963; SP – Silić & Pranjković 2007; TB – Težak & Babić 1996; W – Wonisch 2012

It can be seen that there is a high degree of variation in the attribution of word classes; some elements are treated as indefinite pronouns, others as members of open word classes like adverbs or adjectives. It is also noticeable that the classification depends on the *k-root; cf. *svatko* > pronoun, *svaki* > indefinite pronoun and *svakakav* > adjective. There is major variation as to the interpretation of modifiers like *ma*, which is treated as a conjunction in dictionaries, and as a particle in Silić & Pranjković (2007). Many of the postposed modifiers are not mentioned in HJP, especially those that consist of more than one orthographic word, such as *te volja.* As examples (4, 5) with the collocate *tko zna* and Haspelmath's form *bogzna* show, the group of CIPs that are written separately seems to be larger than assumed in the literature. As our abridged review of Croatian grammaticography and lexicography shows, there is a great deal of disagreement about the scope of the category of indefinite pronouns. The lists of indefinite markers offered in the literature do not follow any clear, consistent and explicit criteria.

[11] *-god* is not mentioned in the grammatical descriptions and specialised works on indefiniteness but does occur in the HJP. It was only added to the table after its extraction from the web corpus (see below).

3. The corpus-based study

3.1. The hrWaC v2.2 corpus

Extraction from large corpora offers a robust method that can ensure the detection of the largest possible number of non-codified modifiers. In our study, we primarily relied on the massive hrWaC 2.2 web corpus. Additionally, the results were compared with material from the Croatian national corpus (HNK). The choice of corpus was based on the following considerations. We assumed that modifiers are the results of ongoing grammaticalisation processes, which first emerge in colloquial language. The relative frequency of some CIPs is quite low. Accordingly, the language corpus used should i) contain colloquial language as far as possible, ii) be quite large, and finally iii) have adequate tools for analysis and extraction of data. Unfortunately, neither the Hrvatski nacionalni korpus (HNK) nor the Hrvatska jezična riznica (HJR) fulfil criteria i) and ii). Both corpora are rather small and essentially represent standard written language. Colloquial and spoken languages are quite weakly represented (mainly in the form of fictional texts): the corpora do not have retrievable spoken sub-corpora, the HJR contains no spoken data and in the HNK the proportion of spoken sources is not indicated (Ćavar & Brozović Rončević 2012). The importance of corpus size is illustrated with two CIPs, *itko* and *kojetko*, as shown in Table 2. While the analysis of frequent, codified CIPs like *itko* based on these corpora is not problematic, no clear conclusions about the functions or even about the existence of the CIP *kojetko* are possible on the basis of only one (HNC) or six (HJR) entries (note that this pronoun is mentioned in the Hrvatski jezični portal). The relative frequencies of *kojetko* in the HJR and hrWaC are very similar; in contrast, its absolute frequency in the hrWaC is 10 times higher and allows its properties to be analysed.

	HNK (2,559,160 words)	HJR (84,536,657 words)	hrWaC 2.2 RFTagger (1,405,794,913 words)
itko (in HJP + ŠRHJ)	2040 (9.41 ipm)	1788 (17.57 ipm)	55777 (39.9 ipm)
kojetko (only in HJP)	1 (0.00 ipm)	6 (0.06 ipm)	70 (0.05 ipm)

Tab. 2: Frequencies of compounds of the modifiers *i:* and *koje:* with word forms of *tko*[12]

The deciding factor for our choice of the hrWaC as the main source was that web corpora contain language use that is oriented towards colloquial language in the form of user-generated content like comments and forums. The question of how far the use of web corpora can influence research results and whether such results can even be considered meaningful has already been discussed in the literature. Benko (2017) analysed the lexical coverage of the Czech and Slovak (traditional) national corpora (SYN and prim, respectively) as compared with large web corpora of the Aranea project by determining what proportion of lexis from medium-sized lexicographic dictionaries was covered in samples of traditional and web-based corpora of the same size. Benko claims that although traditional corpora "are slightly 'better' within the range of their size, this advantage can be

[12] Regular expressions for [lc="i(tko|ko|kog(a)?|kom(e|u)?|kim(e)?)"] and [lc="koje(tko|ko|kog(a)?|kom(e|u)?|kim(e)?)"].

outperformed by the sheer size of larger corpora" (op. cit., 48). He argues that the differences between balanced corpora and web corpora with more than 2 billion tokens are not significant for capturing (new) lexicographic units (op. cit., 48). He comes to the conclusion that "web corpora should not be considered 'inferior', but rather 'different'" (op. cit., 43). Jurkiewicz-Rohrbacher & Hansen & Kolaković (2017) used the example of clitic climbing in Serbian, Croatian and Bosnian to show how web corpora can successfully be used to find rare structures.

3.2. Extracting the modifiers: queries and measures

In this section we provide some more technical details on the extraction of modifiers from the hrWaC 2.2 and HNK v.30. As expected, the web corpus provided a much larger number of modifiers, with all series found in the HNK also present in the hrWaC. The hrWaC is accessible via the corpus management and query systems Sketch Engine and NoSketch Engine, whereas the HNK is accessible only via NoSketch Engine. The two systems provide functions and tools to extract CIPs, such as 'word list', 'collocations', 'word sketch' and 'N-grams'. We mainly used the first two functions.[13] A corpus-based extraction of modifiers has to take into account the following parameters of variation. First, as shown in Section 2, modifiers can precede the **k*-root (*netko*), follow it (*tkogod),* or occupy either position (*bilo tko, tko bilo*). Second, a characteristic feature of CIPs is the possibility of different spellings of the modifier with the **k*-root: together, separately, or hyphenated. Individual modifiers may have a standardised spelling: *gdjetko, tkogod, makar tko*. However, the data retrieved from the hrWaC show a high variability in spelling in real written texts. A codified (in the sense of being recorded in dictionaries or reference books) modifier can be variously written. For example, the hrWaC contains the spelling variants *bilotko* (361 entries), *bilo-tko* (4 entries) and *bilo tko* (36.585 entries). CQL queries treat the hyphenated spelling as one orthographic word form in search and analysis. Thus, in this respect, the task involves the extraction of two types of modifiers, which can appear before and after **k*-roots: those written i) as one or ii) as two orthographic words. We searched for modifiers of the first type separately for each **k*-root using the 'word list' function: for instance, lemmas corresponding to the regular expressions **tko* for preposed and *tko** for postposed series were used for the ontological category of 'person'. After that, all the elements that were found with an absolute frequency of more than 5 were analysed. The first preliminary selection criteria were the ability of the items to form series with different ontological categories and the presence of an indefinite meaning, operationalised as the possibility of substituting a given item with a codified indefiniteness marker in some contexts.

To find modifiers of the second type, the collocation[14] search function was used. In the orthographic sense (and also in the analysed corpora), separately written CIPs are a combination of at least two word forms. The **k*-root can be conceptualised as the node of the collocation and the modifier as its collocator. Better results could be obtained by applying

[13] Because of the characteristics of CIPs, the use of the functions 'word sketch' and 'N-grams' proved ineffective.

[14] In corpus linguistics, collocations are understood as "a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon" (Evert 2005: 17).

the collocation search function to concordances with individual *k-roots. Thus, the concordance queries to find modifiers of the second type have the simplest possible structure, as, for example, for the ontological category of person: [lemma="tko"].

The collocates of all *k-roots were analysed separately, and the preliminary selection criteria for modifiers were the same as for items with compound spelling. Although modifiers may consist of several orthographic words (e.g. *bog zna*) and may additionally be shifted to the left by prepositions (e.g. *bog zna zbog čega*), searches for collocations with a distance of more than 1 token from the node resulted in too much data noise. Therefore, the decision was made to search for candidates for collocations—word forms—placed at a distance of –1 for preposed and 1 for postposed modifiers.

(No)Sketch Engine allows results to be sorted by several statistical measures of association: T-score, MI, MI3, log likelihood, min. sensitivity, logDice, MI.log_f.[15] However, in morphological terms CIP modifiers are highly heterogeneous. Furthermore, pre- and postposed CIPs differ as to their morphosyntactic structure in general. Apart from units used exclusively with *k-roots and characterised by a low frequency in the corpus, there are also modifiers that have homonymous autosemantic equivalents with quite high frequencies (*bilo, malo, mnogo* etc.). All these factors make it impossible to choose only one universal association measure for finding modifiers belonging to all source groups. Figures 1 and 2 show scatter plots illustrating the ranking lists of collocation strength between modifiers (or their components, such as *znati* for *vrag će znati*) and the lemma *tko* for different (No)Sketch Engine association measures. We included modifiers that can be written separately, which are discussed in the next section. At the bottom of each plot we show which modifiers are missing from the first thousand ranking positions.[16] As may be seen, no single association measure can catch all the modifiers, and only an analysis of all ranking lists can retrieve a maximally complete list of modifiers. While some modifiers such as the preposed *bilo, malo* or *god* rank high according to all measures, the least frequent modifiers such as *bud, budi, neznano, makar, mnogo* from the preposed group and *treba, valja, bilo* from the postposed one are missing in the output. In addition, the plots show that searches for only one *k-root, *tko*, do not detect all the possible modifiers: the modifier *neznano* or non-normative spelling *hočeš* could only be found in combination with other *k-roots. Furthermore, as will be shown below (cf. Figure 7), none of the association measures can give a reliable indication of which modifiers are more bonded to their *k-roots, i.e. show the degree of their grammaticalisation (see Section 4.3).

Our study of 19 preposed and 7 postposed modifiers shows that in general, the best result is obtained with the association measure logDice (A plots in Figures 1 and 2). Min. sensitivity (B) and log likelihood (C) also give good results, but even so they did not make

[15] Detailed mathematical descriptions of all the measures can be found in "Statistics Used in Sketch Engine" (n.d.).

[16] The analysis described here was conducted as part of a project devoted to the analysis of CIPs in five Slavonic languages. For the other languages, corpora from the TenTen-family were used, which are available only via Sketch Engine. Sketch Engine has a crucial output limitation that is absent from NoSketch Engine: only the first thousand entries on any list (collocations, N-grams, etc.) can be viewed.

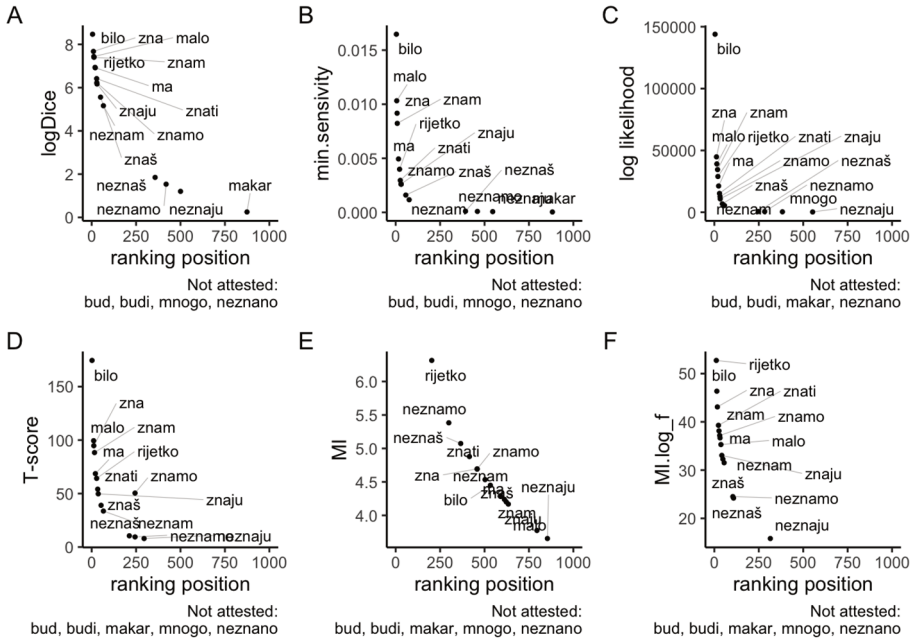

Fig. 1: Ranking of preposed modifiers with the **k*-root *tko* in the collocator lists by association measure

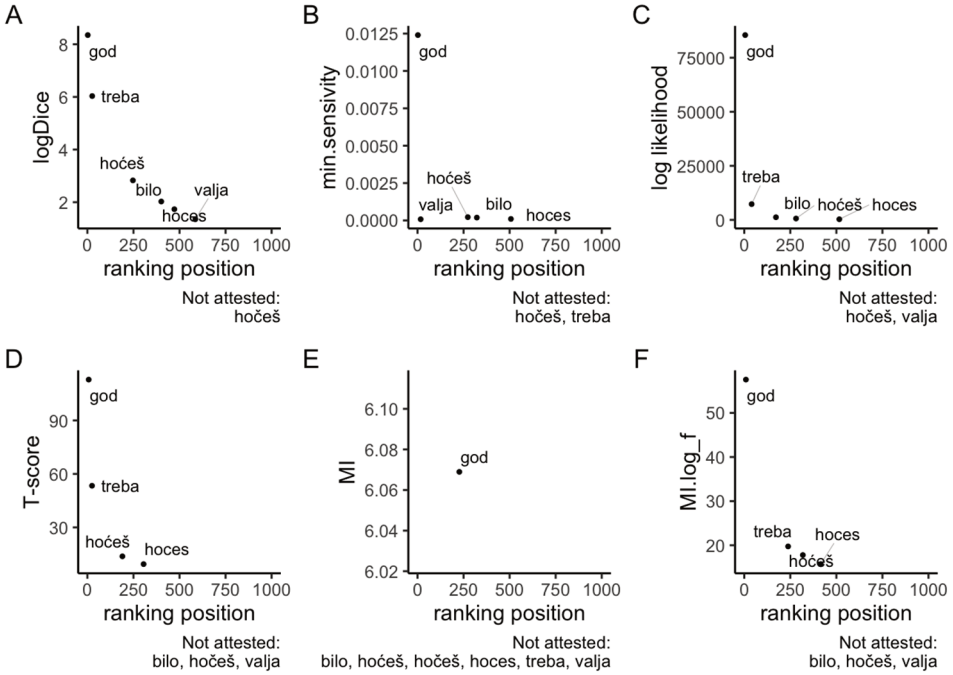

Fig. 2: Ranking of postposed modifiers with the **k*-root *tko* in the collocator lists by association measure

it possible to find all modifiers. In the output of MI.log_f (F), the more frequent modifiers found are generally ranked higher, but the less frequent modifiers remain undetected. For the preposed series alone, logDice, min. sensitivity and log likelihood also show the best results; for the postposed series, the largest number of modifiers can be extracted using logDice. In contrast, the MI measure (E plots), which prefers infrequent collocators (cf. Killgarriff & Kosem 2012: 40), was able to capture only a few modifiers. In addition, preposed modifiers rank lower in the output of MI than in the case of other association measures. According to the plots, T-score (D plots) and MI have no advantage over other measures for modifier extraction. In view of the above, to extract the modifiers as described in the next section, we worked with collocator outputs of all the main ontological categories, sorted by all association measures.

3.3. Results: the inventory of CIP modifiers in Croatian

The inventory list discussed in this section contains modifiers with different degrees of grammaticalisation and morphologisation. It also includes some elements in which indefinite semantics seem to overlap with an evaluative meaning, as in the case of *treba* and *valja*. Modifiers that can synchronically be considered to consist of multiple word forms (such as *mu drago*) were included only unsystematically. We came across several spelling variants that are considered incorrect from a normative perspective, such as *ne:/nie:/nje:*; as expected, forms written without diacritic signs (*hoces*) also occur. We would like to emphasise that we do not treat these examples as data noise but, on the contrary, as material found in user-generated content (forum, comments etc.) that is close to spoken language. As before, no modifiers with an absolute frequency lower than 5 were considered. In total, the corpus-based inventory contains 23 preposed and 9 postposed modifiers, while the grammatico-graphic and lexicographic source-based list (presented in Table 1) comprised 14 and 4 units respectively.[17] The results of the modifier extraction are presented in Table 3.

We excluded some modifier-like elements from the list because they differ in two essential points: they lack an indefinite meaning and/or they do not form series. For example, the list does not contain the elements *dru(g):* (attested in the corpora only as a part of the CIP *drugdje*) or *pokad:* (only in *pokadkad*[18]). According to Sketch Engine tools, elements such as *neznaju/neznaš/neznamo/ne znaju/ne znaš/ne znamo* could be considered good candidates for indefinite pronoun modifiers. However, the analysis of uses in the corpus showed that in almost all cases (as in 9 below) they occur as a part of the source construction without semantic shift and morphologisation: cf. Figure 5, which shows that the degree of grammaticalisation of this series is the lowest and differs substantially from that of other modifiers on our list.

Due to space limitations, we cannot discuss each modifier in detail; we therefore restrict ourselves to some general observations. First, it is worth noting that all modifiers mentioned in the grammar handbooks and other works were found in the corpora. This corroborates the robustness of our corpus analysis. Second, our corpus search has brought to

[17] Assuming a similar grouping of different spellings of modifiers, such as *bilo:/bilo*.

[18] The corpora also contain *pokadšto*. However, it apparently cannot be considered a CIP, as its meaning 'sometimes' (ontological category of 'time') does not coincide with the ontological category of the **k*-root *što*.

preposed modifiers		postposed modifiers	
attested in Table 1	new	attested in Table 1	new
bilo/bilo-/bilo: 'be$_{ptcp}$' Bog zna/bog-zna-/ bogzna: 'God.knows' bud-/bud:/bud/budi-/ budi:/budi 'be$_{imp}$' gdje: 'where' i: 'and' koje: 'which' ma/ma: 'but' makar 'though' malo/malo: 'few' ne:/nie:/nje 'not' po:[20] 'on' pone:[21] 'on.not' što: 'what' sva:/sve:/sav: 'all'	mnogo/mnogo: 'a.lot' ne-znam-/neznam/ neznam: 'I.do.not.know' neznamo/ne znamo/ne znaš/neznaš/ne znaju/neznaju 'we/you/they.do.not. know' neznano 'unknown' pogdje:[22] 'on.where' rijetko 'rarely' tko zna/tko-zna- 'who knows' vrag zna/vrag će znati 'devil knows/will know' znaš/znaš-/znamo/ znamo- 'you/we.know'	:god/-god/:god̵ 'suitable' bilo/-bilo/:bilo 'be$_{ptcp}$' hoćeš/hočeš/hoces 'you.want' mu drago/:mudrago 'pleasant.for.him' te volja/ti volja 'you.like'	:što 'what' treba 'be.necessary$_{3sg}$' valja 'be.necessary$_{3sg}$' :već 'already'

Tab. 3: Modifiers extracted from the corpora: new and already attested in Table 1

the fore three groups of CIPs hitherto not mentioned in the literature. The first comprises several modifiers containing the verb *znati: Bog zna, neznam, neznaju, neznaš, neznamo, neznano, tko-zna, vrag zna, znamo, znaš.* Another type that needs further semantic analysis is CIPs with the quantifiers *mnogo* and *malo*,[19] and the third is modifiers with modals like *treba* and *valja*. Here are some examples extracted from the hrWaC which illustrate the usage of the CIPs that have not hitherto been noted in the literature:

mnogo/mnogo:

(7) *Odgovor na ovo pitanje u* ***mnogočemu*** *ovisi o kvaliteti i opsegu samog rada.*
'The answer to this question depends in many respects on the quality and the scope of the work.' (hrWaC 2.2)

[19] The series with the modifier *malo* can be semantically analysed as pronouns with general existential reference (cf. Padučeva 2016).

[20] The ability of *po:* to form series is questionable. There are 27,805 examples of the lemma *pokoji* in the corpus, but only 7 examples of *pokad(a)* and 2 examples of *pogdje*. *Po:* can be combined with the modifier *gdje* as *pogdje:* (which is a separate modifier on the list), with an additional meaning of 'irregularity', a discussion of which, however, is beyond the scope of this article.

[21] *Pone*: can be seen as a variant of the modifier *ne:* with the additional prefix *po-*.

[22] This modifier can apparently be considered a variant of the modifier *gdje:* with the additional prefix *po-*. Their semantic relationship, however, needs special study.

ne-znam-/neznam/neznam:

(8) *Danas popodne sam iz nekog* ***ne-znam-kojeg*** *razloga odlučio prošetati gradom Labinom.*
'I don't know exactly why, but today in the afternoon I decided to take a stroll in Labin.' (hrWaC 2.2)

neznaju/neznaš/neznamo/ne znaju/ne znaš/ne znamo:

(9) […] *i sigurno je jedno da će kompletna Mala Rakovica biti dio grada Samobora u skoroj budućnosti, ili ćemo se valjda odcijepiti i pripojiti* ***ne znamo kome****.*
'[…] one thing is sure, in the foreseeable future all of Mala Rakovica will be part of Samobor or we will separate and join with—we don't know—whom.' (hrWaC 2.2)

neznano

(10) *U zbilji su oni dugoprstići u drugoj sferi gdje je razvidno da* ***neznano kojim*** *smjerovima odlazi novac od transfera talentiranih igrača u inozemstvo* […].
'In reality they are thieves in the other sphere where it is obvious that the money from the transfer of talented players is going abroad through some unknown channels.' (hrWaC 2.2)

pogdje:

(11) *Opera Rigoletto prvi je dio poznate Verdijeve „latinske trilogije", koja se* ***pogdjegdje*** *još nazivlje i „romantična trilogija".*
'The opera Rigoletto is the first part of Verdi's famous "Latin Trilogy", in some sources also called the "Romantic Trilogy".'

rijetko

(12) *Nažalost,* ***rijetko tko*** *uspijeva putem hrane zadovoljiti dnevne potrebe ovih nutrijensa* […].
'Unfortunately, hardly anyone manages to get the necessary daily amount of these nutrients through food.' (hrWaC 2.2)

tko zna/tko-zna

(13) *Sjediš pored* ***tko-zna-koga****, vruće je, piša ti se, a vozač staje samo u Severinu (na pola puta).*
'You are sitting next to God knows who, it is hot, you have to pee, and the driver stops only in Severin (half-way there).' (hrWaC 2.2)

vrag zna/vrag će znati

(14) *Slikaju valjda i sad nekakve sjene, samo ne sjene kuća, nego* ***vrag će znati čega****.*
'Now they are probably painting some shadows, not of houses but of God knows what.'[23] (hrWaC 2.2)

[23] Literally: 'The devil will know what'. Slavonic languages contain a number of such "taboo intensifiers" (cf. Kehayov 2009), which take the form of modifiers of the type 'X knows what', where X can stand for *vrag* 'devil', *bog* 'God' in Croatian, *господь* 'Lord', *дьявол* 'devil', *хуй* 'dick', *хрен* 'dick' (literally 'horseradish') in Russian; *кат* 'hangman', *батька* 'father' in Ukrainian; *pies* 'dog', *cholera* 'cholera', *diabli* 'devils', *bogowie* 'gods' in Polish; *pánbůh* 'Lord God' in Czech, etc.

znaš/znamo/znaš-/znamo-[24]

(15) *Možda bi mogla razmisliti o brijanju* ***znaš čega****.*
'Perhaps you should think about shaving the "you know what".' (hrWaC 2.2)

:što

(16) *U javnosti se to donedavna smatralo razumljivim* ***kadšto*** *i hvalevrijednim dok je danas upravo to kanda najupitnije.*
'Until recently this was considered understandable in public, sometimes even commendable, but today it seems really questionable.' (hrWaC 2.2)

treba

(17) *Već i vrapci na granama pjevaju o stotinama tisuća lažnih glasača u Hrvatskoj kojima je lako manipulirati izbore i izabrati* ***koga treba****.*
'It is all over town that there are hundreds of thousands of fake voters in Croatia who easily manipulate the elections and vote for the appropriate candidates.' (hrWaC 2.2)

valja

(18) *Ne čupaj kukolja' – okreni se i pogledaj* ***što valja****, pogledaj pšenicu.*
'Don't pull out the weeds, turn around and have a look at the nice things, look at the wheat.' (hrWaC 2.2)

:već

(19) *Ako postaneš project manager, koordinator ili* ***štoveć****, to je valjda najunosnije.*
'If you become a project manager, coordinator or whatever, that probably pays best.' (hrWaC 2.2)

4. Analysis within the framework of grammaticalisation theory

The observed variation in spelling indicates that CIPs are located on a scale ranging from free word combinations to word forms. This is explained by the fact that the class of CIPs is undergoing certain dynamic processes. We will show that the development of fully lexical elements into CIP modifiers is a typical case of grammaticalisation understood as "the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status, e.g. from a derivative formant to an inflectional one" (Kuryłowicz 1965: 69). The first part of the definition refers to 'primary' and the second to 'secondary grammaticalisation'. The development of CIPs belongs to the former type. For an overview of the current state of the art in grammaticalisation research we refer the reader to the *Oxford Handbook of Grammaticalization* by Narrog & Heine (2011). The codified Croatian indefinite pronouns of the series *ne-, bilo* and *i-* have already been analysed in relation to their lexical sources in Haspelmath (1997). However, there has hitherto been no description of CIPs in relation to their degree of grammaticalisation.

[24] The pronouns of this series can express a weakly definite or definite reference. Three groups of such contexts can be distinguished: an expression of definiteness without additional meanings, use as a euphemism (as in the given example), and allegorical use. Definite and weakly definite uses of indefinite pronouns have been described on Russian material in Fisun (2016).

Even for Russian, where the class of CIPs is best researched, certain modifiers remain more or less undescribed (cf. Sokolova 2007). The description of the grammaticalisation of CIPs in Croatian and other Slavonic languages is an essential contribution both to the general theory of grammaticalisation and to the understanding of indefinite pronouns.

4.1. Sources of grammaticalisation

All the modifiers found can be classified according to their lexical sources. Haspelmath (1997: 130–141) offers a cross-linguistic classification consisting of six types, which we will expand in order to classify the Croatian modifiers. Examples of modifiers are given in their standard spelling. In Croatian, there are representatives of four of the types of source constructions suggested by Haspelmath:

i) CIPs of the 'dunno' type derive from constructions with the original meaning 'I don't know'. This group of modifiers includes six preposed elements: *Bog zna; ne znam; ne znaju/ne znaš/ne znamo; neznano; tko zna; vrag zna.*
ii) The 'want/pleases' type contains modifiers with the source meaning 'you want'. Four postposed modifiers of this type were found: *te/ti volja; :god; hoćeš; :mudrago/mu drago.*
iii) Source constructions of the 'it may be' type contain a form of the verb 'to be' in their structure. In Croatian, this type includes *bilo*, which can stand both before and after the **k*-root, and the preposed modifier *budi*.
iv) In Croatian, three focus particles *makar*, *ma*, *:već* and the associated with them *i:* 'and, also, even' have developed into modifiers.

In the Croatian corpus material, no modifiers were found that derive from source constructions with the meaning 'no matter' and 'or', which Haspelmath postulates for some other languages. However, the former meaning may be expressed by grammaticalised series, e.g., with the modifier *god.*

Furthermore, it turns out that Croatian has some modifiers that are not included in Haspelmath's classification:

v) There are five modifiers that have developed from **k*-roots. *Što* can be pre- or postposed; *gdje:; koje:* and *pogdje:* are preposed.
vi) Two modifiers that derive from expressions of necessity (*treba* and *valja*) were found.
vii) Variants of the modifier *znaš/znamo* developed from predicates with the meaning 'known'. As mentioned above, CIPs with this modifier can express a weakly definite or even definite reference (cf. Fisun 2016), see the example of 15 above. Their negated variants belong to the 'dunno' type.
viii) The modifiers *malo:/malo; mnogo:/mnogo* and *rijetko* form a group of quantitative expressions. They seem to combine quantification with a subtle type of indefinite semantics. The quantifier *sva:* can also occur as a modifier of this pronominal series with interrogatives.
ix) Croatian, like most other Slavonic languages, has an indefinite series derived from the negative particle: *ne:; pone:*. In addition to elements with the normative spelling of the codified *ne:* series, the much less frequent variants *nie:/nje:* were also found in the corpus. A characteristic feature of Croatian as compared to other Slavonic languages is

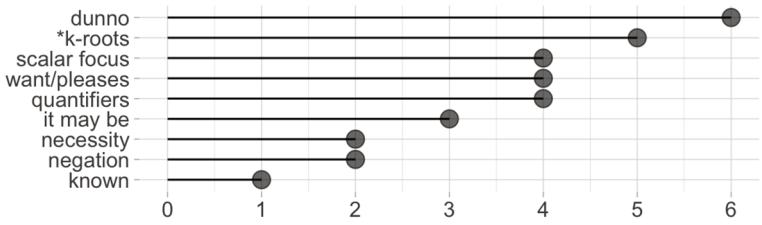

Fig. 3: Number of modifiers in each source group

the presence of a related modifier that is composed of the negative particle and the preposition *po*.

We kept the modifier *po:* out of this classification because, on the one hand, we could not find an etymological analysis clearly linking it to the verbal prefix and, on the other hand, we could not find modifiers derived from similar sources in other Slavonic languages. Figure 3 summarises the sources of grammaticalisation of Croatian CIPs and shows the number of modifiers in each group.

4.2. Degree of grammaticalisation

We assume that the degree of grammaticalisation of indefinite serial pronouns generally corresponds to the degree of grammaticalisation of their indefiniteness markers, i.e. their modifiers. Thus, below we use 'modifier grammaticalisation degree' and 'series grammaticalisation degree' synonymously. In order to characterise the degree of grammaticalisation of individual modifiers, we apply the well-known approach developed by Lehmann (2015 [1982]). Lehmann's model sparked intensive research, which subsequently led to the modification of the model, but without questioning of the basic assumptions. According to Lehmann, a lexeme, as a free autonomous linguistic sign, passes into the class of linguistic signs bound to other units in the course of grammaticalisation, and becomes a grammatical formative at the end of the process (op. cit., ix). The nature of the process therefore causes the loss of autonomy of the language sign. Three aspects of autonomy are postulated: 'cohesion' (the binding of a language sign to other signs or its relation to them, increases with grammaticalisation), 'variability' (the sign's manipulability and mobility) and 'weight' (scope that enables the formation of oppositions to other signs). The variability and weight of the language sign decrease with grammaticalisation (op. cit., 130). According to Lehmann, all three aspects of autonomy mentioned interact with paradigmatic and syntagmatic features of the language signs, resulting in six grammaticalisation parameters: 'integrity', 'paradigmaticity', 'paradigmatic variability', 'structural scope', 'bondedness' and 'syntagmatic variability'. These parameters were intensively discussed in subsequent studies. They can be said to have been essentially confirmed in the current linguistic literature (cf. Norde 2012; Wiemer 2014; Hansen 2017; Cuyckens 2018): "A final advantage of using a rigid set of criteria such as Lehmann's parameters is that it offers clear arguments to either accept or reject a given change as a legitimate example of (de)grammaticalization" (Norde 2012: 105). In the extensive literature on grammaticalisation it has been pointed out that this phenomenon is a matter of degree (cf. Hansen 2004b); although there are some proposals as to how to measure that degree, so far, no consensus has been reached. For instance, the model proposed in Correia Saavedra (2021) is not applicable to CIPs as a) it deals with

Parameter according to Lehmann (2015)	Our operationalisation
1.Integrity	a. semantic shift, b. decategorialisation, c. phonological erosion
2. Paradigmaticity	a. indefiniteness, b. combinability with ontological categories
3. Paradigmatic variability	a. uniqueness, b. modifier frequency
4. Structural scope	a. syntactic condensation, b. coordinative constructions, c. violation of dependency relations between modifier and *k-root
5. Bondedness	a. spelling of modifier and *k-root, b. autonomy, c. separability, d. limited distribution of modifiers
6. Syntagmatic variability	position of modifier within the CIP

Tab. 4: Lehmann's parameters and our operationalisation for modifiers

highly grammaticalised elements and b) it is based on a clear-cut identification of a lexical source of the grammatical counterpart.

In our analysis of grammaticalisation we will apply a mixed-method approach that combines corpus data with qualitative analyses. This means that we will not rely exclusively on corpus studies. As a matter of fact, corpus linguistics and grammaticalisation studies have "tended to work in blissful ignorance of each other" (Mair 2011: 239) for a long time. The strength of corpora is that they provide authentic data in their original context, which makes it possible to study the bridging contexts that are so important in early stages of grammaticalisation. As Mair points out, corpus-linguistic methodology is particularly well-suited to addressing frequency-related research questions (op. cit., 242f). As we will see, however, there are other central parameters of grammaticalisation that can only be studied qualitatively, because the hrWaC lacks semantic annotation. This holds true for example for the analysis of semantic changes. In the following, we discuss two types of parameters, which are not to be confused with each other: the six grammaticalisation parameters proposed by Christian Lehmann, and the 15 sub-criteria that we will call 'grammaticalisation criteria', which are our operationalisation of Lehmann's parameters in the practical analysis of CIPs. An overview of the correlation between Lehmann's parameters and our practical sub-criteria is provided in Table 4. As may be seen, the parameters have varying numbers of practical sub-criteria. We distinguished 15 sub-criteria in total. However, they have different weights because they describe different aspects of the grammaticalising element. As will be shown, each criterion and parameter forms a kind of scale. In practice, however, we simplified these scales to a 'yes-no' distinction.

Each modifier from the corpus-based list (Table 3) was evaluated according to the practical grammaticalisation criteria and was annotated with '+' if the modifier showed

stronger grammaticalisation according to the sub-criterion in question, and '–' if it did not. Some modifiers cannot be evaluated in terms of sub-criteria 1b, 1c and 4a and were therefore annotated with a '0'. The decategorialisation (i.e. loss of inflection) of modifiers such as *i:*, *mnogo*, *mak.ar*, *:već* etc., whose source lexical elements are non-inflected words, is not subject to assessment. Similarly, we lack reliable etymological data to unambiguously evaluate the degree of syntactic condensation of the modifiers *po:* and *pone:*, and the phonological erosion of *:god*. Thus, our approach is based on a cumulation of features that are measured on a binary scale (feature present vs feature absent).

The results of the analysis of all presented modifiers according to our operationalisations of the parameters are summarised in Table 5. In the next sections, we will explain these operationalisations in relation to CIPs and present the necessary frequency data from hrWaC in more detail. This will be illustrated mainly on the basis of the four modifiers *bogzna, tko zna, malo, treba* and additionally some others.

4.2.1. Integrity: semantic shifts, morphological and phonological erosion

According to Lehmann (2015: 126), "[…] integrity of a sign is its possession of a certain substance which allows it to maintain its identity, its distinctness from other signs, and grants it a certain prominence in contrast to other signs in the syntagm". It seems sensible to distinguish between semantic and formal integrity. Semantic integrity concerns the ways in which the linguistic sign gains new, more grammaticalised (and hence more abstract) layers of meaning and sheds old ones. Lehmann (2015: 136) sees the loss of semantic integrity primarily in the process of 'bleaching' of meaning, of desemanticisation. Hansen (2001: 403f), however, uses the case of the grammaticalisation of Slavonic modal constructions to show that the polysemanticisation (in the sense of development of new functions) of the grammaticalising element is also part of this process. Phonological integrity concerns the reduction of phonological words and the transformation of words into affixes.

1a Semantic shifts: Modifiers may show substantial semantic changes compared to their lexical sources or to homonymic autonomous word forms. A clear case of loss of semantic integrity is offered by the modifier *tko zna*, which has fully lost its original semantics of interrogativity, as may be seen in this example from InterCorp:

(20) *Odjednom je glas,* ***tko zna zašto,*** *izgubio onu punoću i samopouzdanje.*
'For **some** reason, the richness and confidence of the other voice waned sharply.' (InterCorp v14)

More subtle semantic changes can be noted for example in the case of the modifier *malo* as in (21), which in the English equivalent has a scalar negative reading 'almost no one':

(21) *proživjela sam stvari k.ao* ***malo tko*** *od mojih vršnjak.a.*
'I've experienced something **almost no one** my age ever has.' (InterCorp v14)

1b Decategorialisation: In the course of grammaticalisation the source of the modifier may lose morphosyntactic integrity, in the sense of loss of morphosyntactic properties characteristic of lexical or other less grammaticalised forms (Kuteva & Heine 2012).

	i:	mnogo	pone:	ne:	pogdje:	po:	gdje:[25]	koje:	malo	sva:	rijetko	ma	bilo (pre)	makar	Bog zna	što:	budi	neznano	vrag zna	tko zna	ne znam	ne znaju/-š/-mo	znaš/-mo	:god	bilo (post)	mu drago	:već	te/ti volja	:što	hoćeš	treba	valja
Freq	275792[26]	9634	205559	4243846	31	27816	1677	21810	23834	957821	14086	6510727	366589	928	5616	215528	81529	57	98	24357	136314	34791	37728	16011	1648430	372	517	941	86331	12250[32]	34048[33]	21734
1a	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	−	−	+	+	+	+	+	+	+	−	−
1b	0	0	0	0	0	0	0	+	0	+	0	0	+	0	+	+	+	0	−	−	−	−	−	+	+	+	0	−	+	−	−	−
1c	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	0	−	−	−	−	−	−	−	−
2a	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	−	−	+	+	+	+	+	+	+	−	−
2b	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
3a	+	−	+	+	−	+	−	−	−	+	−	−	+	−	−	−	+	−	−	−	−	−	−	−	+	−	+	−	−	−	−	−
3b	+	+	+	+	−	+	+	+	+	+	+	+	+	−	+	+	−	−	−	+	+	+	+	+	+	−	−	−	−	+	+	+
4a	+	+	0	+	+	0	+	+	+	+	+	+	+	+	−	+	+	+	−	+	+	−	−	+	+	+	+	+	+	+	−	−
4b	+	+	+	+	+	+	+	+	−	+	−	−	−	+	−	+	+	+	+	−	−	−	−	+	+	+	−	−	+	−	−	+
4c	+	+	−	−	+	−	+	+	+	+	+	−	+	−	−	+	−	+	−	+	+	−	−	+	+	+	−	+	+	−	+	−
5a	+	+	+	+	+	+	+	+	+	+	−	+	+	−	+	−	+	−	−	−	−	−	−	+	−	−	+	−	+	−	−	−
5b	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−	−
5c	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	+	−	+	−	+	−	−	−
5d	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−
6	+	+	+	+	+	+	+	+	+	−	+	+	−	+	+	−	+	+	+	−	−	−	+	+	−	+	+	+	−	+	+	+

Tab. 5: Results of modifier analysis by proposed operationalisation

[25] Separate spelling is also possible. Such examples are however extremely rare compared to the homonymic conjunction and cannot be counted automatically.

[26] The forms *isto* and *ista* were excluded due to noise from homonyms. Marginal cases where a modifier was spelled separately were not counted.

[27] The lemmata *mak, mač* and the forms *matko, mako, masta* were excluded due to noise from homonyms.

[28] The lemma *stok* and the word form *stoko* were excluded due to noise from homonyms.

[29] The lemma *budist* was excluded due to noise from homonyms.

[30] The forms *stabilo, kobilo* were excluded due to noise from homonyms.

[31] The forms *kosto* and *košto*, which are misidentified in the hrWaC as lemmata, were excluded due to noise from homonyms.

[32] Examples with the complementiser *da* after the series were excluded.

[33] To reduce the noise of the source construction, the infinitive forms and lemmata *on|onolikol|svil svelsav* before the series, and infinitive forms (distance 0–3) and the complementiser *da* after the series were excluded.

[34] The exclusions are the same as in the case of *treba*.

Decategorialisaton can be detected by screening the corpus-based list of modifiers in Table 3. The majority of modifiers from the list have lost some or all inflectional forms typical of the lexical source element; for example, in the case of *bogzna* the morphological forms **Bog je znao, *bogovi će znati* are not attested. In contrast, there are modifiers that do retain inflectional forms and thus do not completely lose morphosyntactic integrity. This holds for some modifiers related to the verb *znati*: *neznam/neznaš/neznamo/neznaju*.

1c Phonological erosion: The loss of a sign's integrity can manifest as phonological erosion. Since in our work the grammaticalisation processes should be verifiable with the help of the web corpus, only those cases can be taken into account in which this phonological erosion is reflected in the orthography. In Croatian, only the modifier *:god* can be considered to show signs of phonological erosion. However, according to the analysis in Anstatt (1996: 43), the modifier *:god* can be linked to the verb **goditi* as well as to the corresponding noun **godъ*. We therefore decided to apply zero annotation for this case.

4.2.2. Paradigmaticity: indefiniteness and combinability with ontological categories

By the 'paradigmaticity' of the sign, Lehmann understands "the formal and semantic integration both of a paradigm as a whole and of a single subcategory into the paradigm of its generic category" (Lehmann 2015: 141). It increases in the course of the grammaticalisation process. CIPs participate in paradigmatic relations of two types. First, different modifiers form a paradigm whose members determine the function or the set of different indefinite functions of CIPs such as *specific-known, specific unknown* or *non-specific* series (cf. Haspelmath 1997; van der Auwera & van Alsenoy 2013). Second, an indefinite series itself represents a paradigm with an indefinite marker and different **k*-roots. These two paradigms are relevant for CIPs and can be summarised as two criteria that can be correlated with Lehmann's parameter: 'Indefinite meaning component' and 'Combinability with ontological categories'.

2a Indefiniteness: We do acknowledge that the detection of indefinite semantics requires the intuition of the researcher. Our operationalisation of this parameter is the translation of the CIP by a non-compound indefinite pronoun in at least some of its usage contexts in the parallel corpus InterCorp. Examples (20–21) above show the possibility of substituting a compound pronoun with a codified non-compound one in another language in at least certain contexts (in this case, English). A specific feature of less grammaticalised CIPs as compared to fully grammaticalised indefinite pronouns is that they frequently contain non-pronominal semantic components in addition to pronominal meaning invariants. We often encounter meanings related to evaluativity, as seen in the examples (22) and (23) below (on the relation between CIPs and evaluativity, see Rybová in prep.). CIPs with *treba* have an evaluative meaning only and cannot usually be substituted by any codified indefinite marker.

(22) *Ni ja baš ne plešem* ***bogzna kako***.
'I'm not **much good** as a dancer myself, really.' (InterCorp v14)

(23) *Uhvatite me* ***kako treba***.
'Get hold of me **properly**.' (InterCorp v14)

The degree of these CIPs' paradigmatisation is thus lower than for codified indefinite pronouns, but higher than for their lexical sources.

2b Combinability with ontological categories: The second sub-criterion excludes modifiers that form paradigms with only a limited number of ontological categories.[35] The data were retrieved through concordance analysis in the hrWaC. For the purpose of our research, we decided that modifiers capable of forming indefinite series with at least three

	i:	mnogo	pone:	ne:	pogdje:	po:	gdje:	koje:	malo	sva:	rijetko	ma	bilo (pre)	makar	Bog zna	što:	budi	nezнano	vrag zna	tko zna	nc znam	nc znaju/-š/-mo	znaš/-mo	:god	bilo (post)	mu drago	:već	te/ti volja	:što	hoćeš	treba	valja
person	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+
thing	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+
property	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+
manner	+		+	+		+		+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+		+	+	+
place	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
amount	+		+	+		+			+	+	+	+	+	+	+			+	+	+	+	+	+	+	+	+	+	+		+	+	+
reason	+					+			+			+	+	+	+		+	+	+	+	+	+	+		+					+	+	
time	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+		+	+	+	+	+

ontological categories (one third of the number of **k*-roots corresponding to the main ontological categories) would be considered grammaticalised according to this criterion. The combinability of modifiers with ontological categories according to the hrWac is given in Table 6.

The ability of the indefinite marker to form series with different **k*-roots was one of the conditions for extracting modifiers from the corpus. Therefore, in the proposed Croatian list all modifiers (excluding the postposed variant of *što*) fulfil this criterion, although not to the same extent: for instance, *bog zna* (in different orthographic variants) is combined with all the main categories, the codified postposed modifier *:god* fails to occur only with the category 'reason' (*zašto*), while the preposed *gdje:* is not combined with three of the categories: 'manner' (*kako*), 'amount' (*koliko)* and 'reason' (*zašto*).

4.2.3. Paradigmatic variability: uniqueness and frequency

"The paradigmatic variability of a sign is the possibility of using other signs in its stead or of omitting it altogether" (Lehmann 2015: 131). It is the degree of freedom that the speaker has in choosing a linguistic sign (op. cit., 146). It decreases through grammaticalisation, leading to an increase in obligatoriness. Lehmann distinguishes two aspects within this parameter: transparadigmatic and intra-paradigmatic variability. The former concerns the

[35] The eight ontological categories with their corresponding **k*-roots are listed in Section 1.

freedom of the speaker to choose the paradigm itself, while the latter refers to the degree of freedom in choosing between units within a paradigm. We have to point out that in Croatian, indefiniteness is not an obligatorily marked category; i.e. it does not have the status of a grammatical category. Therefore, the speaker always has the option to use other signs or to omit the expression of indefiniteness altogether; cf. a similar situation with modals (Hansen 2001). This means that the degree of transparadigmatic variability of all indefinite markers remains high and differences between modifiers with respect to this parameter cannot be assessed. However, practical criteria can still be proposed for evaluating intraparadigmatic variability. For the group of CIPs in question, we worked out two criteria: modifier uniqueness and increase in frequency.

3a Modifier uniqueness: Modifiers may grammaticalise either individually or as part of a construction that allows more or less variation of the source lexical material. Less intraparadigmatic variability is shown by those modifiers that are unique. For example, the Croatian modifier *bilo* can be considered unique because no other construction with an active participle of a verb has developed into a modifier. Other unique modifiers are: *i:, pone:, ne:, po:, budi, sva:* and *:već*. Non-unique modifiers form groups based on semantically and morphosyntactically similar constructions. Such constructions allow the speakers to choose the modifier according to their own extra-grammatical preferences. The construction 'X knows + **k*-root' is an example of joint grammaticalisation: it has produced several similar compound series present in all Slavonic languages, including in Croatian. The following modifiers are thus considered non-unique: *Bog zna/bog-zna-/bogzna/bogzna:, tko-zna, vrag zna/vrag će znati, znaš/znamo.* Other such groups are: *gdje:, pogdje:, koje:, što:, :što* (**k*-roots); *malo, mnogo, rijetko;*; *ma, mak̨ar*; *ne znam, ne znaju/-š/-mo, neznano, znaš/-mo*; *:god, mu drago, te/ti volja, hoćeš* and *treba, valja.*

3b Frequency: Limited paradigmatic variability in grammaticalised modifiers is, on the other hand, manifested in the increase in frequency of a grammaticalised series. It has been confirmed in the literature that higher levels of grammaticalisation are associated with an increase in frequency (cf. Bybee 2003). The token frequencies of the series were obtained via CQL queries of the hrWaC corpus: the search was performed using only the contact combination 'modifier + 8 main ontological categories'; **k*-root spellings without diacritical marks, quite frequent in user-generated content, were taken into account; however, cases of preposed CIPs separated by prepositions were not considered due to data noise, which cannot be excluded for most modifiers without a manual check. An example query for the modifier *gdje:* is given in example (24):

(24) [lc="gdje(tko|ko|kog(a)?|kom(e|u)?|kim(e)|što|sta|sto|cega|cemu|cim(e)?|čeg(a)|čem(u)?|čim(e)?|koj(i|a|e|oj|u|om|ih)?|kojeg(a)?|kojem(u)?|kojim(a)?|kakav|kakv(e|a|i|ih|u|oj)?|kakvog(a)?|kakvom(u|e)?|kakvim(a)?|kako|gdje|kolik(o)?|kolik(e|a|i|ih|u|oj)?|kolikog(a)?|kolikih|kolikom(u|e)?|kolikim(a)?|zašto|zasto|kada|kad)"]

In some cases, additional exclusions were used in the CQL queries to avoid frequent homonymic units. All frequency data can be found in Table 5. The numbers have not been verified manually. In most cases, they contain noise and also include part of their source constructions. We deem modifiers that have more than 1000 entries in the web corpus to be frequent.

4.2.4. Structural scope: syntactic condensation, coordinative constructions and violation of dependency relations

Lehmann defines the 'structural scope' of a linguistic sign as "the structural size of the construction which it helps to form" (Lehmann 2015: 152). Structural scope decreases during grammaticalisation. This parameter by Lehmann is the most discussed in the literature: other scholars note that in the course of grammaticalisation, not only a narrowing of the scope but also, conversely, its broadening may be observed. In the case of modifiers, a condensation of the syntactic scope takes place as postulated by Lehmann: only **k*-roots can be in the syntactic scope of the stronger grammaticalised modifiers.

4a Syntactic condensation: "The shrinking of structural scope in the course of grammaticalization ends at the stem level" (op. cit., 155). The 'Syntactic condensation' criterion is fulfilled by many modifiers on the corpus-based list, as it was also considered in preselection. Whereas highly grammaticalised modifiers can only modify the pronominal **k*-root (e.g. *i-tko*), the string *bog zna* can have either a narrow scope, as in:

(25) *Što se susjeda tiče, nisam ja baš primjetila da ovdje ni druge mlađe žene jedva čekaju da jedna drugoj zasjednu na kavu ili za* ***bog zna*** **[*koga*]** *peku kolače.*
'As for the neighbours, I didn't realise that here also the other young women did not wait for invitations to coffee or bake pies for God knows whom.'

Or as a free word combination of the source construction, it can govern subordinate clauses, as in:

(26) *Što ti vrijedi truditi se, kada samo* ***Bog zna*** **[*koga*** *će prvog pokupiti s ovog svijeta*].
'What is the use of trying when only God knows whom he will take first from this world.' (hrWaC 2.2)

Note that some CIPs allow the insertion of isolated clitic elements. In this case, the criterion is still considered fulfilled; cf.:

(27) *Pa Srbija ne želi Kosovo, međutim neće mu dati nezavisnost bez cijene, u tom je kvaka, ostale priče o autonomiji i* ***Bog ti zna što*** *samo služe za brisanje očiju dijelu ljudi koji su odgojeni na Kosovskom mitu, iza kulise je najobičnija trgovina Srbije sa Zapadom.*
'And Serbia does not want Kosovo; however, it will not allow independence without a cost, that's the point, all the discussions about autonomy and stuff like that are only intended to dry the eyes of those people who grew up with the Kosovo myth, behind the scenes there is the usual trade between Serbia and the West.' (hrWaC 2.2)

4b Coordinative constructions: The narrowing of the syntactic scope of modifiers to the stem, i.e. their convergence with affixes, is accompanied by the reduction of the frequency of constructions where one modifier is combined with several coordinative **k*-roots, seen in (28).

(28) *Bajdo nije zaslužio da se, zbog* ***tko zna*** **[*kojih*]**$_{*k\text{-}root}$ ***i***$_{and}$ **[*čijih*]**$_{*k\text{-}root}$ *interesa, njegova nogometna genijalnost baca u drugi, treći plan.*
'Bajdo does not deserve to have his footballing genius overshadowed because of someone else's interests.' (hrWaC 2.2)

More grammaticalised modifiers do not permit such constructions:

(29) *Pa cijelo društvo je ovisno* ***o ne[kome]***k-root **ili**or **o ne[čemu]***k-root / *ne[kome]***k-root **ili**or **[čemu]***k-root.*
'The whole of society depends on somebody and something.' (hrWaC 2.2)

To analyse this parameter, the corpus was queried by contact combinations of modifiers and the construction '**k*-root *i/ili *k*-root'. Hyphenated and compound spelling of the modifier and the **k*-roots was considered. Other modifiers placed before and after the constructions were excluded. For concordances with more than 100 occurrences, samples of 100 contexts were analysed and the results were extrapolated to the population. Our analysis shows that in general, coordinative constructions with modifiers are not very frequent. No coordinative constructions were found in the corpus for the following modifiers: *gdje:, i:, koje:, makar, ne:, budi, po:, odne:, pone:, što:, sva:,* preposed *bilo, :što, :god, mu drago* and *pogdje:*. Only single examples of coordinative constructions were attested for *mnogo, valja, vrag zna, neznano*. In our annotation, the modifiers listed above therefore fulfil this sub-criterion. The other modifiers in our list, however, showed varying numbers of examples with coordinative constructions in the hrWaC and therefore do not satisfy this sub-criterion. The highest percentage of coordinative constructions in a CIP series in the corpus was observed for the modifier *:već* (13 in absolute terms or 2.51% of the total 517 uses). The highest absolute number of occurrences, 555, was found for the modifier *ne znam*; however, this represents only 0.4% of its total frequency. Figure 4 shows modifiers (in their most common form) for which coordinative constructions have been attested and these constructions' percentage of the total frequency of modifiers:

4c Violation of dependency relations: This criterion is fulfilled by the modifiers that deviate outward from the syntactic patterns of their source constructions. For example, the modifier *treba* loses its original association with the modal verb *trebati*, which is characterised by different types of valency structures, e.g. with infinitives (*treba raditi*), or with nominal phrases (*treba mi telefon*). The modifier combines with different types of **k*-roots,

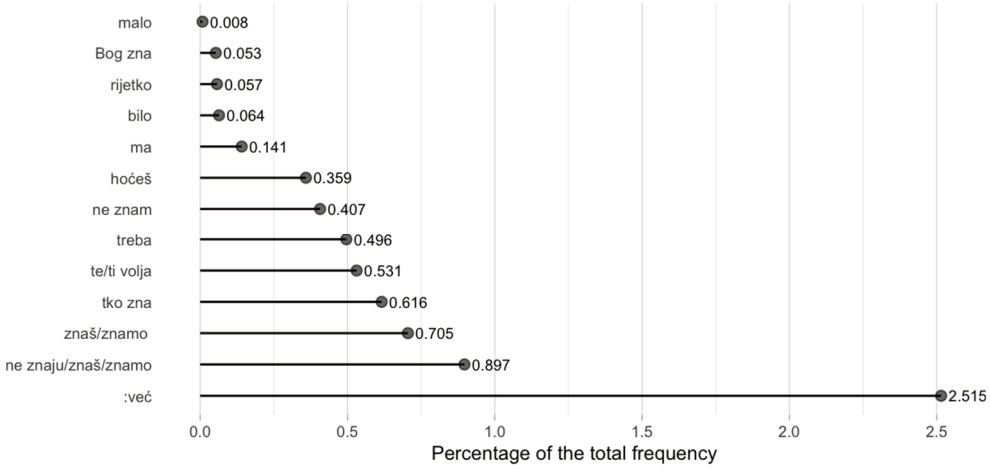

Fig. 4: Modifiers with attested coordinative constructions

as in the example (23) above. Similarly, the modifiers *bogzna/bog zna* and *tko zna* have a different valency structure than their source: whereas the lexical verb *znati* governs the accusative case or a complement clause, the modifiers do not assign any case (cf. forms like in (28)).

4.2.5. Bondedness: spelling, autonomy, separability and distribution of modifiers

Bondedness is the intimacy with which a sign is connected with another sign to which it bears a syntagmatic relation (Lehmann 2015: 157). The 'bondedness' or syntagmatic cohesion of an indefinite marker is the degree of cohesion between the modifier and the interrogative. Bondedness increases in the course of grammaticalisation. In the process, a language unit passes through the bondedness scale: from free lexical item to clitic morpheme, from clitic to bound morpheme or affix, and finally from independent affix to integral part of another morpheme (op. cit., 157). In a practical analysis of CIPs, however, it is usually quite unclear to which level of the Lehmann scale a given modifier should be assigned. For evaluation of the degree of grammaticalisation according to this parameter, Mel'čuk's (1997, Chapter 5) word form criteria can be helpful. For CIPs four specific sub-criteria can be determined: 'spelling of the modifier and the **k*-root', 'autonomy', 'separability' and 'limited distribution of modifiers'.

5a Spelling: The tendency to write modifiers together with the adjacent word in Slavonic languages, where bound morphemes are normally written together with the **k*-root, provides evidence that these modifiers tend to be placed further on the bondedness scale. In Croatian, CIPs can be written (normatively and not) with the **k*-roots in three ways: separately (like *mak.ar tko*, *tko treba),* hyphenated (like *tko-zna-gdje*, *tko-zna-tko*) or together (like *itko*, *gdjetko*). The hyphenated spelling is not normative for the codified Croatian series; it is, however, attested for all less-grammaticalised series. In our analysis, we do not distinguish between hyphenated spelling and compound spelling: if the spelling of a given series is normative or proven at least one hundred times in the hrWaC, we consider this sub-criterion to be fulfilled. The *bogzna* series may be cited as an example. Its frequencies in the corpus with interrogatives of the eight main ontological categories are *bogzna* (3229), *bog zna* (840) and *bog-zna-* (132). In the case of this modifier, we can also observe the tendency for *bog* to be spelled in lower case more often, while in the source construction the capital letter is preferred. There are, however, many exceptions to this pattern.

5b Autonomy: According to Mel'čuk, the main property of the word-form (free lexical element) is autonomy. First, he distinguishes between strong and weak autonomy. Strong autonomy characterises those linguistic signs that can form a complete utterance that does not include other signs, i.e. can be separated by pauses (Mel'čuk 1997: 158), like the bold sequence in (30):

(30) *Možda oni misle da su ustvari mene iskorištavali.* ***Tko zna.***
'They might think that they really exploited me. Who knows.' (hrWaC 2.2)

We assumed that modifiers lack autonomy. In our corpus-based modifier list, we find both modifiers that can occur autonomously and those that cannot form complete utterances, as illustrated in the original example (31) taken from the hrWaC. For this criterion, a

question test can be applied, as seen in constructed example (32). This example shows the possibility of splitting a sentence: the **k*-root is separated from the modifier to form an interrogative part, while the modifier forms an isolated, complete utterance signifying uncertainty.

(31) *Naravno kako je to krški teren jezero se sigurno neće napuniti iz prve tako da će se i nakon toga rafting nastaviti.* ***Koliko*** *dugo??* ***Bog zna****.*
'Of course, as this is karst terrain, the lake will definitely not fill up at first, so the rafting will continue after that. How long? God knows.' (hrWaC)

(32) *Nešto sam čula, ali izbrbljala je to moja susjedaMarica, pa* ***bog zna koliko*** *će biti istine.*
[**Koliko** će biti istine? **Bog zna**.]
'I heard something, but my neighbour Marica blurted it out, so God knows how much truth there will be. [How much truth will there be? God knows.]' (hrWaC)

One example of the more grammaticalised modifiers that do not form complete utterances on their own is *malo.* It lacks strong autonomy.

(33) ***Malo tko*** *nakon njegovih predavanja ostane ravnodušan.*
[Tko nakon njegovih predavanja ostane ravnodušan? ***Malo**.]
'After his lectures hardly anyone remains indifferent.' (hrWaC 2.2)

Modifiers that can form isolated utterances have to retain their indefinite semantics. Thus, the autonomous uses of *malo* 'little', *bilo* 'was' or *gdje* 'where' have no bearing on this criterion, since in these cases the elements do not show semantic equivalence with the corresponding modifiers. *Malo*, *bilo* and *gdje* therefore fulfil the sub-criterion.

5c Separability: According to Mel'čuk, elements that remain after strongly autonomous word forms have been excluded show so-called 'weak autonomy': other elements can be inserted between the separable sign and the sign to which it relates without changing or losing the semantic relationship between them. Mel'čuk emphasises that different degrees of separability exist depending on the status of the separating elements (Mel'čuk 1997: 164). The highest degree of separability occurs in the CIPs in example (34), whose modifiers can be separated from **k*-roots by autonomous word forms; they do not fulfil this criterion.

(34) *Prošlo ljeto na Jadranu su bili i Ecclestoni i Benetton, Carolina, Schumacher, Spielberg i* ***bog zna još koliko*** *turbo lovatora* […]
'Last summer Eccleston, Benetton, Carolina, Schumacher, Spielberg and God knows how many of the super-rich visited the Adriatic Sea […]' (hrWaC 2.2)

The criterion is considered fulfilled by modifiers that cannot be separated from their **k*-roots by autonomous word forms. The insertion of a preposition between a preposed modifier and a **k*-root is not regarded as a violation of the sub-criterion as such a variation in the position of prepositions is a feature of both uncodified (35) and most codified (36) series.

(35) a. *Dođemo u gostionu,* ***bog zna u koju.***
'We are going to a restaurant, to any restaurant.' (hrWaC 2.2)

b. *Galatasaray nije* ***u bog zna kakvoj*** *formi.*
'Galatasaray is not in the best shape.' (hrWaC 2.2)

(36) a. *Bez obzira na prepreke* ***u bilo kojem*** *pravcu, one se mogu otkloniti.*
'No matter the obstacles in any direction, they may be remedied.' (hrWaC 2.2)

b. *Jednako veliki problem je kad zapnemo,* ***bilo u kojem*** *stanju, dobrom ili lošem, jer promjena je sila koja pokreće.*
'A similarly big problem occurs when we get stuck in any kind of state, a good or a bad one, because change is a driving power.' (hrWaC 2.2)

5d Limited distribution: For this sub-criterion, a sign's ability to combine with word forms of different classes is evaluated. If the variability of the possible word form classes is broader, the distributive variability of the sign is greater. The more grammaticalised modifiers are therefore supposed to have a very limited distribution and to connect exclusively with interrogatives. Therefore, in the present analysis, the modifiers were tested for the criterion 'used exclusively as a modifier'. Our list includes some modifiers (*ne:, pone:, :god*) that in the contemporary language, in the specified semantic function, can only be associated with **k*-roots (only as modifiers of CIPs).[36] Their distributive variability is thus substantially restricted. To illustrate, according to this sub-criterion, by combining not only with **k*-roots, but also with nouns or adjectives, *malo* has a lower degree of grammaticalisation:

(37) ***Malo ljudi****$_{noun}$* *zna da su Males i Kerum prijatelji.*
'Only few people know that Males and Kerum are friends.' (hrWaC 2.2)

(38) *Noćenje je bilo* ***malo prohladno****$_{adjective}$*.
'The night was a bit cold.' (hrWaC 2.2)

4.2.6. Syntagmatic variability: position within the CIP

6 Position within the CIP: "The syntagmatic variability of a sign is the ease with which it can be shifted around in its context. In the case of a grammaticalized sign, this concerns mainly its positional mutability with respect to those constituents with which it enters into construction" (Lehmann 2015: 167). Paradigmatic variability decreases through grammaticalisation. The degree of positional freedom of all CIP modifiers is quite low: with the possible exceptions mentioned in the previous section, they are placed immediately before or after their **k*-root. A 'fixed position within the CIP' can thus serve as a criterion that the modifier has no positional freedom. A position may be called fixed if the modifier cannot change its location with respect to the **k*-root, cf. ***bogzna*** *što* vs **što* ***bogzna****, što****god*** vs *****god****što,* ***gdjetko*** vs **tko****gdje****.* In contrast, modifiers with attested variation include ***bilo*** *što* vs *što* ***bilo****,* ***ne znam*** *tko* vs *tko* ***ne znam****.*

[36] The prefixes *ne:* and *pone:* used with verbs have no relation to the expression of indefiniteness and should thus be considered homonymous on the synchronic level.

4.3. Results of the grammaticalisation analysis

After annotating the modifiers in terms of the sub-criteria, a grammaticalisation score was calculated for each of Lehmann's parameters: a maximum possible score of 1 indicates complete grammaticalisation and 0 indicates no grammaticalisation. Since we used different numbers of sub-criteria for different parameters, the values of sub-criteria within their corresponding parameter differ: from 0.25 within parameters that have 4 criteria to 1 for the single criterion of "Syntagmatic variability". Some modifiers that often have an overall high level and a long history of grammaticalisation were annotated with a zero, as shown in section 4.2. In order to avoid automatic interpretation of such cases as a violation of sub-criteria when calculating the grammaticalisation scores, the decision was made to evaluate the criterion as fulfilled also in these cases. After calculating the scores for each parameter, also the means of the six parameters, that is, the grammaticalisation scores for all series were calculated. This overall score describes the general degree of grammaticalisation of a modifier. The grammaticalisation degree was classed as follows: score up to 0.33 – weak, from 0.34 to 0.66 – medium, from 0.67 to 1 – strong.

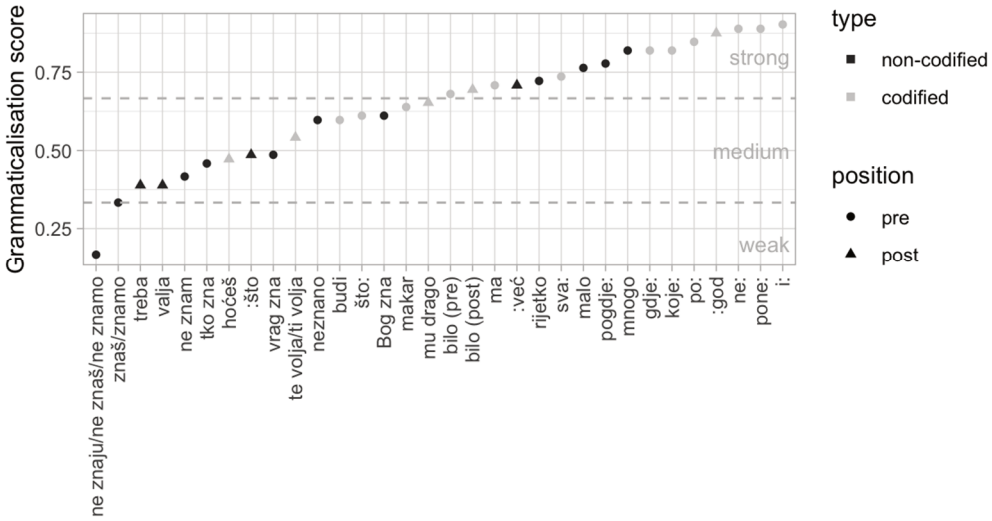

Fig. 5: Grammaticalisation score of codified and non-codified modifiers

Figure 5 summarises the information on the general grammaticalisation score and the grammaticalisation degree of codified and non-codified modifiers. As expected, a number of non-codified modifiers (e.g. preposed *znaš/znamo* and postposed *valja, treba*)[37] rank lower than codified modifiers: they are less grammaticalised. We assume that this is why they have not been codified, i.e. reported in reference works such as grammar books and dictionaries. The preposed *ne znaju/ne znaš/ne znamo* have barely any signs of grammaticalisation. However, not all of the non-codified modifiers found by us are weakly grammati-

[37] In this case we treat only the modifiers mentioned in the grammars presented in section 2.1 as codified.

calised: some of them (*mnogo, pogdje:* resp. *:već*) have relatively high grammaticalisation scores. Only one modifier that was marginally mentioned in the grammars and thus notionally belongs to the codified group has a degree of grammaticalisation lower than 0.5: *hoćeš*. Figure 5 also clearly shows that preposed modifiers are higher on the grammaticalisation scale than postposed ones, irrespective of their codification status: for example, only one new postposed modifier, *:već*, has a score higher than 0.5. Postposed modifiers have a lower degree of grammaticalisation, scoring lower primarily on the sub-criteria separability, spelling, frequency and uniqueness. In general, the formation of postposed series seems to be a less productive model, both quantitatively and qualitatively. It is interesting to note that in the case of modifiers that can stand both before and after a **k*-root—*bilo* and *što*—the postposed variants are particularly marginal in number, and *:što* shows very limited compatibility with ontological categories. A smaller number of postposed modifiers is observed not only in Croatian, but also in Russian, Ukrainian, Polish and Czech. Although in all these languages old, codified and strongly grammaticalised postposed series are present, new postposed series tend not to have a high degree of grammaticalisation.

According to our analysis, preposed *i:, mnogo, ne*: and postposed *:god, bilo, mu drago* are the most grammaticalised modifiers. The least grammaticalised are the numerous variants of modifiers deriving from the predicate *(ne) znati* in the preposed group, and *valja* and *treba* in the postposed group.

A summary of the degree of grammaticalisation is given in Table 7. The average grammaticalisation score is given in brackets. As can be seen, the grammaticalisation degree of preposed and postposed modifiers varies from medium to strong. For all criteria apart from 'syntagmatic variability', the grammaticalisation degree of postposed series is slightly weaker than that of the preposed ones.

Next, we would like to discuss the individual weight of each parameter. Croatian modifiers obtain medium scores on the grammaticalisation parameter 'integrity'. 'Phonological erosion' has the weakest representation: only the modifier *:god* could possibly meet this criterion. 27 of the total 32 modifiers showed semantic shifts and only 10 of them had undergone morphosyntactic decategorialisation (cf. Table 5). This is also consistent with the data presented in Norde (2012).

In terms of the 'paradigmaticity' parameter, the modifiers are strongly grammaticalised. The sub-criterion 'Indefiniteness' is fulfilled by 27 modifiers. The sub-criterion 'combinability with ontological categories' is fulfilled by all 22 preposed and 8 of the postposed

	Integrity	Paradig-maticity	Paradigmatic variability	Structural scope	Bondedness	Syntagmatic variability
pre	medium (0.55)	strong (0.91)	medium (0.54)	strong (0.67)	medium (0.58)	strong (0.74)
post	medium (0.48)	strong (0.83)	medium (0.39)	medium (0.63)	medium (0.36)	strong (0.78)
all	medium (0.53)	strong (0.89)	medium (0.5)	medium (0.66)	medium (0.52)	strong (0.75)

Table 7: Grammaticalisation degree of modifiers in Croatian

series. As may be seen from Tables 5 and 6 above, the modifier *:što* violates this criterion: only the combinations *gdješto*[38] and *kadšto* were found in the hrWaC corpus.

Grammaticalisation in terms of 'paradigmatic variability' can be characterised as medium; however, postposed modifiers have noticeably lower values. Although only 7 preposed and 2 postposed modifiers meet the criterion of 'uniqueness', 23 modifiers in total show an increase in frequency.

Grammaticalisation degree measured as 'structural scope' is strong for preposed and medium for postposed modifiers. The sub-criterion 'Syntactic Condensation' is met by a large number of both pre- and postposed modifiers (24). The sub-criterion 'Violation of dependency relations' between modifier and **k*-root was observed in only 19 series. Furthermore, we have observed a lack of 'coordinative constructions' with a single modifier and multiple **k*-roots for 19 series.

In general, the modifiers we found have a medium level of grammaticalisation in terms of the 'bondedness' parameter. Stronger grammaticalisation is seen for the sub-criteria 'separability' (23 modifiers) and absence of 'autonomy' (22 units). When we consider the sub-criteria 'spelling' (17) and 'limited distribution' (4), the modifiers are less grammaticalised. The latter sub-criterion is fulfilled only by the preposed modifiers *ne:, pone:* and *pogdje:* and by the postposed *:god*. The CIPs show a strong degree of grammaticalisation in terms of the 'Syntagmatic variability' parameter. A fixed 'position of modifiers within the CIP' is a feature of 24 modifiers.

Figure 6 summarises the data on the average grammaticalisation score and degree of codified (mentioned in section 2.1) and new modifiers. The data show that the average

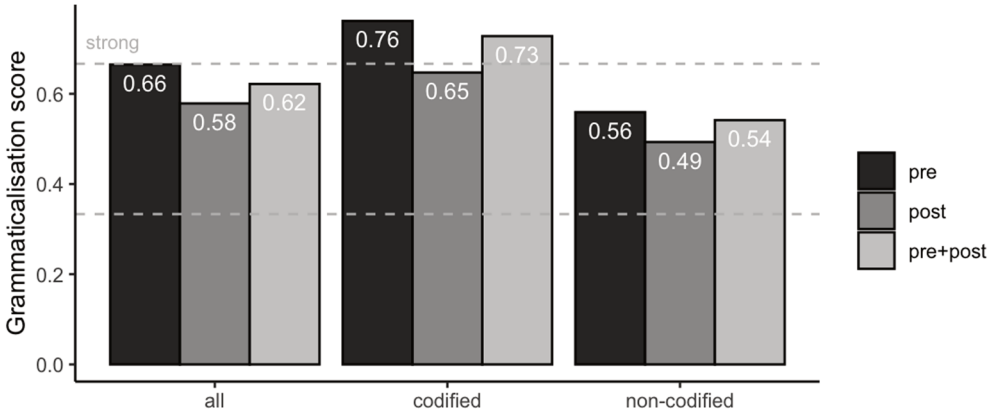

Fig. 6: Grammaticalisation score of codified vs new compound series

[38] Both *što:* and *gdje:* can be modifiers. The separation of homonymic cases is sometimes quite problematic. An example of a case where *gdje* unambiguously corresponds to the ontological category 'place' and *što:* is a modifier is found in the following:

(i) *Kad se pogledaju ti srednjovjekovni koncepti, ima mjesta sumnji da se sukobljavaju dva srednjovjekovlja, ne baš posve prevladana a* ***gdješto*** *i povampirena* […].
'When one looks at these mediaeval concepts, there is room for doubt that two medievalisms are colliding, not fully overcome and in some places rising from the dead […].' (hrWaC 2.2)

grammaticalisation score of the codified modifiers is in all cases significantly higher than that of the non-codified ones extracted from the corpus: it is approximately 1.3–1.4 times higher for the codified modifiers than for the uncodified ones. The grammaticalisation degree of all the codified modifiers can be considered strong (although taken separately, the score of postposed codified series is on the border between strong and average grammaticalisation), while the grammaticalisation degree of new modifiers is medium. The average degree of grammaticalisation of all the preposed modifiers was slightly stronger (score of 0.66) than that of all postposed ones (score of 0.58). Similar results are seen for codified and new modifiers taken separately: a slightly larger difference was observed between codified modifiers, where preposed modifiers were 1.14 times more grammaticalised than postposed ones.

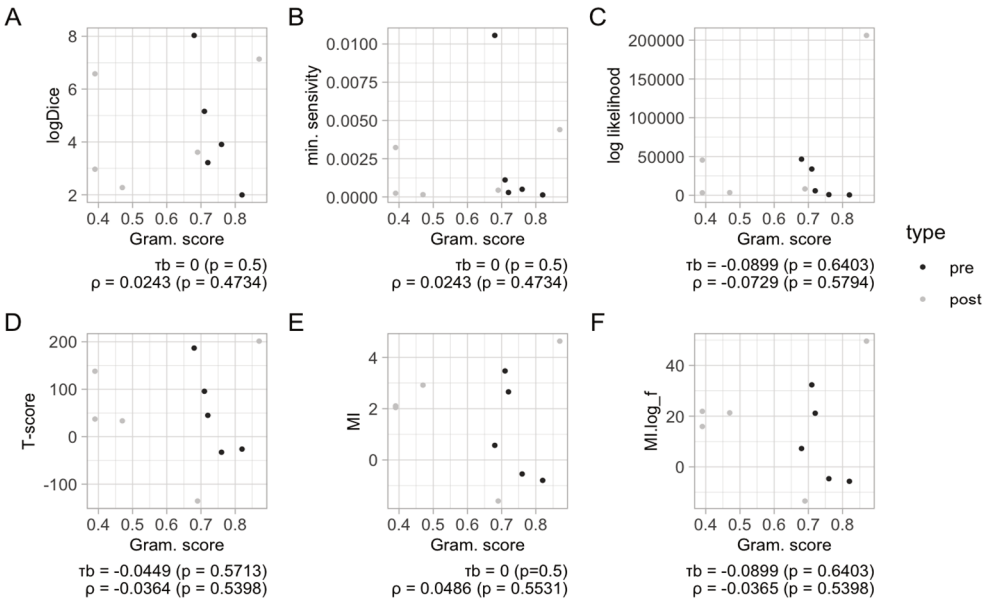

Fig. 7: Relationships between association measures and grammaticalisation scores of some modifiers

The relationship between the association measures of modifiers and their **k*-roots and the grammaticalisation score is an interesting question. To establish it, we took 10 one-word modifiers that can be written separately from their **k*-roots and identified in the collocation output*: bilo* (preposed), *ma, malo, mnogo, rijetko, bilo* (postposed), *god, hoćeš, treba, valja* (both). Figure 7 shows scatter plots representing the relationships between 6 association measures computed by Sketch Engine for the combinations 'modifier + **k*-roots' and our modifier grammaticalisation scores. The plots do not provide any evidence about a possible correlation between any of the association measures and the grammaticalisation score of modifiers. Similarly, Kendall's (Kendall 1938) and Spearman's (Spearman 1904)

rank correlation[39] coefficients do not show a significant correlation in any of the cases either: the results of applying the τ and ϱ tests with the corresponding p-values are also shown in Figure 7. We thus have good reason to assume that there is no significant correlation between collocation association measures of modifiers and their **k*-roots and the degree of grammaticalisation of modifiers, and that lexical association measures are not a reliable tool for measuring the degree of grammaticalisation.

5. Conclusion

As in other Slavonic languages, in Croatian the class of indefinite series-forming pronouns is affected by ongoing grammaticalisation processes. In addition to codified indefinite pronouns it includes a number of elements whose degree of grammatcalisation varies considerably. Hitherto, there have been no attempts in the literature to systematise the inventory of possible indefinite markers and to describe their degree of grammaticalisation, which is the purpose of this paper. The analysis of corpus material from the hrWaC and HNK allowed us i) to propose the most complete list of Croatian modifiers as yet, containing 23 preposed and 9 postposed modifiers, and ii) to analyse the lexical sources of their grammaticalisation according to Haspelmath's approach (1997). It turned out that on the one hand, not all types of source constructions reported by Haspelmath develop into modifiers in Croatian. On the other hand, the study on Croatian also revealed source types not present in Haspelmath (1997). These findings are a valuable contribution not only to Croatian linguistics, but also to the general theory of indefinite pronouns and grammaticalisation.

Furthermore, the paper proposes an approach for evaluating the degree of grammaticalisation of the modifiers. For that purpose, fifteen specific sub-criteria were developed on the basis of Lehmann's theory of grammaticalisation, making it possible to measure the modifiers in terms of the parameters 'integrity', 'paradigmaticity', 'paradigmatic variability', 'structural scope', 'bondedness' and 'syntagmatic variability'. For the evaluation according to these sub-criteria, a mixed method combining both quantitative and qualitative analysis of corpus material was proposed. The grammaticalisation scores thus calculated allowed us to measure the degree of CIP grammaticalisation and to compare different series. According to our data, the preposed modifiers *i:, mnogo, ne*: and postposed modifiers *:god, bilo, mu drago* are the most grammaticalised. It turned out that, as expected, the grammaticalisation degree of new compound series is generally much lower than that of codified series: in most cases, grammarians rely on their intuition to choose the more grammaticalised series for description. However, exceptions have also been found: the series *mnogo, rijetko* and *:već*, which are not codified in Croatian, nevertheless have a rather high level of grammaticalisation. In addition, some series such as *što:*, although described in grammars, have quite a low grammaticalisation score. The analysis also allowed us to conclude that in general, preposed modifiers are more grammaticalised than postposed ones. Croatian modifiers have a higher degree of grammaticalisation when characterised using the parameters 'para-

[39] These correlations were used because the values of most association measures belong to logarithmic scales and their relationships with the grammaticalisation score are not linear. According to Croux & Dehon (2010), both correlations are robust and efficient, with some superiority of Kendall's τ. As may be seen, in our case the results of the two tests were quite similar.

digmaticity', and 'syntagmatic variability', with the overall degree of grammaticalisation of serial indefinite pronouns varying from medium to strong.

A further finding is that we have not been able to establish any correlation between the grammaticalisation score of modifiers and the collocational strength of their association with **k*-roots.

As far as we know, there have been no previous attempts to measure the degree of grammaticalisation of specific indefinite series or indefinite pronouns in a language as a whole. Thus, the proposed approach can be applied to other languages, and provides new data for cross-linguistic comparison.

Author contributions

Sections 1 and 5 were produced collectively by the authors. Section 2 was authored by Björn Hansen, who also undertook the overall editing of the manuscript. Section 3 was written by Roman Fisun. Section 4 was developed through particularly close collaboration: the general method for measuring grammaticalisation was proposed by Roman Fisun, while the analysis of Croatian corpus data was conducted by Björn Hansen.

Abbreviations

3 – third person; CIP – compound indefinite pronoun; HJP – Hrvatski Jezični Portal; imp – imperative; ptcp – participle; sg – singular; ŠRHJ – Školski rječnik hrvatskoga jezika.

Language corpora used

HJR (Hrvatska jezična riznica): http://riznica.ihjj.hr (retrieved 30 Sep 2022).
HNK (Hrvatski nacionalni korpus) v.30: http://filip.ffzg.hr/cgi-bin/run.cgi/first_form (retrieved 30 Sep 2022).
hrWaC 2.2 (RFTagger): https://www.sketchengine.eu (retrieved 30 Sep 2022).
InterCorp: http://www.korpus.cz (retrieved 30 Sep 2022).

Literature

Anstatt, Tanja. 1996. *‚Zeit': Motivierungen und Strukturen der Bedeutungen von Zeitbezeichnungen in slavischen und anderen Sprachen*. München: Sagner.

Barić, Eugenija et al. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Benko, Vladimír. 2017. Are web corpora inferior? The case of Czech and Slovak. In Kupietz, Marc & Witt, Andreas & Bański, Piotr & Tufiş, Dan & Cristea, Dan & Váradi, Tamás (eds.), *Proceedings of the workshop on challenges in the management of large corpora and big data and natural language processing (CMLC-5+BigNLP 2017) including the papers from the Web-as-Corpus (WAC-XI) guest section*, 43–48. Birmingham, Mannheim: Institut für Deutsche Sprache.

Bhat, Darbhe N. S. 2004. *Pronouns* (Oxford Studies in Typology and Linguistic Theory). Oxford: Oxford University Press.

Bondareva, Galina A. 2010. *Sostavnye mestoimenija v russkom jazyke*. (Diss. kand.) Voronež: Voronežskij gosudarstvennyj universitet.

Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In Joseph, Brian & Janda, Richard (eds.), *The Handbook of Historical Linguistics*, 602–623. Oxford: Blackwell. DOI: 10.1002/9780470756393.ch19.

Ćavar, Damir & Brozović Rončević, Dunja. 2012. Riznica: The Croatian language corpus. *Prace filologiczne* 63. 51–65.

Correia Saavedra, David. 2021. *Measurements of grammaticalization: Developing a quantitative index for the study of grammatical change*. Berlin, Boston: De Gruyter Mouton. DOI: 10.1515/9783110753073.

Croux, Cristophe & Dehon, Catherine. 2010. Influence functions of the Spearman and Kendall correlation measures. S*tatistical Methods and Applications* 19. 497–515. DOI: 10.1007/s10260-010-0142-z.

Cuyckens, Hubert. 2018. Reconciling older and newer approaches to grammaticalization. *Yearbook of the German Cognitive Linguistics Association* 6(1). 183–196. DOI: 10.1515/gcla-2018-0009.

Ermakova, Ol'ga P. 2000. Vzaimodejstvie dvux sistem častej reči (mestoimennoj i znamenatel'noj) pri obrazovanii sostavnyx nominacij. In Kleszczowa, Krystyna & Selimski, Ludwig (eds.), *Słowotwórstwo a inne sposoby nominacji*, 147–152. Katowice: Gnome.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. (Dissertation.) University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Fisun, Roman. 2019. Zusammengesetzte Indefinitpronomen im Ukrainischen. In Macjuk, Halina, Mytnik, Irena & Novikova, Olena (eds.), *Mova v suspil'stvi: Semantyka, syntaktyka, prahmatyka*, 85–99. Warschau: Wydawnictwo UMCS.

Fisun, Roman S. 2016. Ob opredelënnosti neopredelënnyx mestoimenij. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta: Serija Russkaja filologija* 5. 158–170.

Hansen, Björn. 2001. *Das Modalauxiliar im Slavischen: Semantik und Grammatikalisierung im Russischen, Polnischen, Serbischen/Kroatischen und Altkirchenslavischen*. München: Sagner.

Hansen, Björn. 2004a. Eine korpuslinguistische Studie zur Dynamik der Adjektivdeklination im Serbischen/Kroatischen. In Hansen, Björn (ed.), *Linguistische Beiträge zur Slawistik XI*, 31–45. München: Sagner.

Hansen, Björn. 2004b. The grammaticalization of the analytical imperatives in Russian, Polish and Serbian/Croatian. *Die Welt der Slawen* 49. 257–274.

Hansen, Björn. 2017. What happens after grammaticalization? Post-grammaticalization processes in the area of modality. In Van Olmen, Daniël & Cuyckens, Hubert & Ghesquière, Lobke (eds.), *Aspects of grammaticalization: (Inter)subjectification and directionality*, 257–280. Berlin: Mouton de Gruyter.

Haspelmath, Martin. 1997. *Indefinite pronouns* (Oxford Studies in Typology and Linguistic Theory). Oxford: Oxford University Press.

HJP – *Hrvatski Jezični Portal.* https://hjp.znanje.hr/ (retrieved 30 Sep 2022).

Isačenko, Aleksandr V. 1965. O sintaksičeskoj prirode mestoimenij. In *Problemy sovremennoj filologii: Sbornik statej k semidesjatiletiju V. V. Vinogradova*, 159–166. Moskva: Nauka.

Jackendoff, Ray S. 1983. *Semantics and cognition*. Cambridge (MA): MIT Press.

Jozić, Željko et al. 2013. Višerječnice. In Jozić, Željko et al. (eds.), *Hrvatski pravopis*. Zagreb. http://pravopis.hr/pravilo/viserjecnice/29/ (retrieved 14 Nov 2022).

Jurkiewicz-Rohrbacher, Edyta & Hansen, Björn & Kolaković, Zrinka. 2017. Web Corpora – the best possible solution for tracking rare phenomena in underressourced languages: Clitics in Bosnian, Croatian and Serbian. In Bański, Piotr et al. (eds.), *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing,* 49–55. Mannheim: Leibniz-Institut für Deutsche Sprache. https://ids-pub.bsz-bw.de/files/6243/10.+Jurkiewicz_Kolakovic_Hansen_Web_Corpora_2017.pdf (retrieved 30 Sep 2022).

Katičić, Radoslav. 1986. *Sintaksa hrvatskoga književnog jezika: Nacrt za gramatiku*. Zagreb: Globus.

Kehayov, Petar. 2009. Taboo intensifiers as polarity items: Evidence from Estonian. *Language Typology and Universals* 62(1–2). 140–164. DOI: 10.1524/stuf.2009.0009.

Kendall, Maurice G. 1938. A new measure of rank correlation. *Biometrika* 30 (1/2). 81–93. DOI: 10.1093/biomet/30.1-2.81.

Kilgarriff, Adam & Kosem, Iztok. 2012. Corpus tools for lexicographers. In Granger, Sylviane & Paquot, Magali (eds.), *Electronic Lexicography*, 31–55. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199654864.003.0003.

Kunzmann-Müller, Barbara. 2002. *Grammatikhandbuch des Kroatischen unter Einschluß des Serbischen*. Frankfurt am Main u. a.: Peter Lang.

Kuryłowicz, Jerzy. 1965. The evolution of grammatical categories. *Diogenes* 13(51). 55–71. DOI: 10.1177/039219216501305105.

Kuteva, Tania & Heine, Bernd. 2012. An integrative model of grammaticalization. In Wiemer, Björn, Wälchli, Bernhard & Hansen, Björn (eds.), G*rammatical replication and grammatical borrowing in language contact*, 159–198. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110271973.159.

Lavrov, Leonid V. 1983. Značenie i upotreblenie pronominal´nyx obrazovanij tipa *kto (čto) ugodno*. In *Grammatičeskaja semantika russkogo jazyka*, 21–26. Vologda.

Lehmann, Christian. 2015. *Thoughts on grammaticalization*. Berlin: Language Science Press. DOI: 10.26530/oapen_603353.

Mair, Christian. 2011. Grammaticalization and corpus linguistics. In Narrog, Heiko & Heine, Bernd (eds.), *The Oxford Handbook of Grammaticalization*, 239–50. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199586783.013.0019.

Maretić, Tomo. 1963. *Gramatika hrvatskoga ili srpskoga književnog jezika.* Zagreb: Matica Hrvatska.

Marković, Ivan. 2002. Nešto o određenosti/neodređenosti u hrvatskome. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 28(1). 103–150.

Marković, Ivan. 2013. *Uvod u jezičnu morfologiju*. Zagreb: Disput.

Mel´čuk, Igor' A. 1997. *Kurs obščej morfologii*. Bd. 1. Moskva, Wien: Jazyki russkoj kul´tury & Izdatel´skaja gruppa "Progress".

Narrog, Heiko & Heine, Bernd (eds.). 2011. *Oxford Handbook of Grammaticalization*. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199586783.001.0001.

Norde, Muriel. 2012. Lehmann's parameters revisited. In Davidse, Kristin & Breban, Tine & Brems, Lieselotte & Mortelmans, Tanja (eds.), *Grammaticalization and language change: New reflections,* 73–110. Amsterdam: John Benjamins.

Padučeva, Elena V. 2016. Mestoimenija slaboj opredelennosti (serija na *koe*-; serija na *ne*-; *odin*). In Plungjan, Vladimir A. (ed.), *Materialy dlja proekta korpusnogo opisanija russkoj grammatiki.* http://rusgram.ru/Местоимения_слабой_определенности (retrieved 30 Sep 2022).

Progovac, Ljiljana. 1990. Free-choice *bilo* in Serbo-Croatian: existential or universal? *Linguistic Inquiry* 21(1). 130–35.

Progovac, Ljiljana. 1991. Polarity in Serbo-Croatian: Anaphoric NPIs and pronominal PPIs. *Linguistic Inquiry* 22(3). 567–72.

Progovac, Ljiljana. 1994. *Negative and positive polarity: A binding approach*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511554308.

Rybová, Martina. (In prep.) *Die Beziehung zwischen Evaluativität und Indefinitheit am Beispiel der zusammengesetzten Indefinitpronomina im Tschechischen.* (Promotionsschrift.) Regensburg: Universität Regensburg.

Šarić, Ljiljana. 2002. *Kvantifikacija u hrvatskome jeziku*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Silić, Josip & Pranjković, Ivo. 2007. *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Zagreb: Školska knjiga.

Skok, Petar. 1971–1974. *Etimologijski rječnik hrvatskoga ili srpskoga jezika*. Zagreb: Jugoslavenska Akademija Znanosti i Umjetnosti.

Sokolova, Svetlana V. 2007. *Dinamičeskie processy v sisteme mestoimennyx slov sovremennogo russkogo jazyka*. (Diss. kand.) Moskva: MGU.

Spearman, Charles. 1904. The proof and measurement of association between two things. *American journal of psychology* 15. 72–101. DOI: 10.2307/1412159.

ŠRHJ – *Školski rječnik hrvatskoga jezika.* https://rjecnik.hr/ (retrieved 30 Sep 2022).

Statistics Used in Sketch Engine (n.d.). https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf (retrieved 30 Sep 2022).

Tatevosov, Sergej G. 2002. *Semantika sostavljajuščix imennoj gruppy: Kvantornye slova*. Moskva: IMLI RAN.

Testelec, Jakov G. & Bylinina, Elizaveta G. 2005. O nekotoryx konstrukcijax so značeniem neopredelennyx mestoimenij v russkom jazyke: Amal´gamy i kvazireljativy. Report for the seminar "Theoretical semantics", IPPI RAN. http://www.rsuh.ru/binary/1787534_99.1322270635.82662.pdf (retrieved 15 Jul 2016).

Težak, Stjepko & Babić, Stjepan. 1996. *Gramatika hrvatskoga jezika: Priručnik za osnovno jezično obrazovanje*. Zagreb: Školska knjiga.

Van der Auwera, Johan & van Alsenoy, Lauren. 2011. Mapping indefiniteness: Towards a Neo-Aristotelian approach. In Kitis, Eliza & Lavidas, Nikolaos & Topintzi, Nina & Tsangalidis, Tasos (eds.), *Selected Papers from the 19th International Symposium on Theoretical and Applied Linguistics, Thessaloniki April 2009*, 1–14. Thessaloniki.

Wiemer, Björn. 2014. *Quo vadis* grammaticalization theory? Why complex language change is like words. *Folia Linguistica* 48(2). 425–468. DOI: 10.1515/flin.2014.015.

Wonisch, Arno. 2012. *Das Pronominalsystem des Bosnischen-Bosniakischen, Kroatischen und Serbischen.* Münster u. a.: LIT.

Znika, Marija. 1987. O upotrebi određenih i neodređenih pridjevnih oblika. *Jezik* 34(4). 101–106.