COLLOCATIONS IN THE CROATIAN WEB DICTIONARY – *MREŽNIK*

Lana HUDEČEK, Milica MIHALJEVIĆ
Institute of Croatian Language and Linguistics

Hudeček, L., Mihaljević, M. (2020): Collocations in the Croatian Web Cictionary – Mrežnik. Slovenščina 2.0, 8(2): 78–111

DOI: https://doi.org/10.4312/slo2.0.2020.2.78-111

The *Croatian Web Dictionary – Mrežnik* project aims to create a free, monolingual, easily searchable, hypertext, born-digital, corpus-based dictionary of the Croatian standard language. Collocations play an important role in *Mrežnik*. At the outset of the *Mrežnik* project, the concept of collocations and their presentation was modelled after the *elexiko* project. However, this concept was modified during the project on the basis of corpus analysis. This paper will outline the presentation of collocations of headwords of different word classes. Some important issues connected with collocations in *Mrežnik* are collocation extraction methods, collocations as a means of differentiating meanings and extracting new meanings, the use of stylistic and terminological labels in collocations, and the relationship of collocations with normative and pragmatic notes, definitions, and subentries.

Keywords: collocations, Croatian, e-dictionary, *Mrežnik*, born-digital dictionary

1 INTRODUCTION

Collocations have received a great deal of attention in recent years. This is not surprising, as they can be considered “the building blocks of language and … fundamental units of language” (Sinclair, 2004, p. 213). They constitute a major challenge for linguists, lexicographers, native speakers, dictionary users, and language learners alike. The challenge for the linguist is how to define them and differentiate them from other multiword expressions.[1]

[1] On collocations in Croatian (cf. Blagus Bartolec, 2014); on collocations in Croatian for non-native speakers (cf. Ordulj, 2018).

Collocations are also an important part of a dictionary entry. The *Oxford Collocations Dictionary* defines collocations as “the way words combine in a language to produce natural-sounding speech and writing” (McIntosh, 2018, p. V). A narrower view is that collocations are an unpredictable combination of lexical units, i.e. “a combination that cannot be produced based on the regular syntactic or semantic properties of the units involved” (Granger, 2012, p. 216). For lexicographic purposes, collocations can be defined as “a recurrent combination of words, where one specific lexical item (‘the node’) has an observable tendency to occur with another (the collocate), with a frequency far greater than chance” (Atkins and Rundell, 2008, p. 223).

Automated procedures for extracting collocations have been developed and are continually being improved. This means that lexicographers can very quickly obtain large quantities of collocational information, which facilitates dictionary compilation; however, this also poses difficulties for the lexicographer, as collocations entail a number of methodological problems and force the lexicographer to take certain decisions.

Collocations also present a challenge for dictionary users, who often “cannot be sure where to find collocations; a universal format, be it with regard to placement or typography, has yet to be realized” (Durkin, 2016, p. 37).[2] Yet, Durkin (ibid.) notes that born-digital dictionaries do not have space restrictions as print dictionaries, which means that collocations can be provided in the entries of both components of the collocaton.[3]

As linguists from the Institute of Croatian Language and Linguistics provide language advice daily, we know that many user questions are connected with multiword expressions. Although native speakers tend to use collocations more intuitively, some language advice also relates to collocations, e.g. in

[2] “Cobuild6 and LDOCE5, for example, give collocations a separate status in the microstructure, listing (and, if necessary, explaining) them in a self-contained box (...). Thus, users can locate the data immediately without looking through the entire entry” (Durkin, 2016, p. 37). Cobild6 is the sixth edition of the Cobuild dictionary and LDOCE5 is the fifth edition of the *Longman Dictionary of Contemporary English*.

[3] However, in born-digital dictionaries, there is still the risk of information death, wherein the user is overwhelmed by the abundance of information on the screen; thus, the choice of what to include and how to present it in the dictionary interface still remains.

the administrative style. This was the reason a special project on Croatian collocations was launched – the *Croatian Collocation Database* (CCD).[4] Special attention was paid to collocations in the *Mrežnik* project, which is the focus of this paper, for the same reason. Some entries in *Mrežnik* are linked to collocations in the *Croatian Collocation Database* (cf. Hudeček and Mihaljević, 2019a).

1.1 *Mrežnik*

Croatian is still one of a few national languages that does not have a freely available online corpus-based dictionary compiled according to the rules of contemporary e-lexicography or systematic research on e-lexicography; this was the reason for starting the *Croatian Web Dictionary – Mrežnik* project (cf. Hudeček and Mihaljević, 2017a; Hudeček and Mihaljević, 2017b; Hudeček, 2018). *Mrežnik* is a four-year project (1st March 2017 – 28th February 2021) financed by the Croatian Science Foundation. The result of the project will be a free, corpus-based, born-digital, monolingual, easily searchable, hypertext, normative online dictionary of the Croatian standard language. It will become the central meeting point of all language resources compiled at the Institute of Croatian Language and Linguistics, and will thus become a long-term project after the initial four-year period.

Mrežnik is a hypertext dictionary, as its entries and sub-entries are interconnected, as well as linked with entries in databases created within the framework of the *Mrežnik* project[5], as well as with databases being created by project collaborators or other Institute members within the framework of other

[4] "The *CCD* is primarily based on traditional lexicographic and lexicological settings of multiword lexical units (...), so that the main plan is to put together in one database the most common Croatian multiword lexical units by defining their semantic types and context of use. The database will be a useful source to be included in other more advanced MWE sources (Croatian and international) for the development of tools that enable the extraction of MWEs on the basis of their semantic and lexical features (...)" http://ihjj.hr/kolokacije/english/about/.

[5] The databases created in parallel with the creation of the dictionary are: a language advice database (http://jezicni-savjetnik.hr/), language advice for schoolchildren (http://hrvatski.hr/savjeti/), a conjunction database with a description of groups of conjunctions and their modifications, a database of explanations of the origins of idioms (http://hrvatski.hr/frazemi/), a database of ethnics and ktetics (http://hrvatski.hr/etnici-i-ktetici/).

projects (cf. Hudeček and Mihaljević, 2019a). In addition to the module for adult native speakers of Croatian, the dictionary includes a module for schoolchildren and a module for non-native speakers of Croatian (cf. Mihaljević, 2018). *Mrežnik* is based on the *Croatian Web Repository Online Corpus*[6] and the *Croatian Web Corpus*.[7] As it is a corpus-based and not a corpus-driven dictionary, *Mrežnik* takes all other available print and web sources into account in addition to these two corpora. This means that, while the collocations are primarily based on Word Sketches[8] and the aforementioned corpora, other collocations can be added to the dictionary even if they are not attested in the corpora, but the compiler intuitively knows that they are commonly used in Croatian and can be found on the web. The reason for this approach is that there is currently no representative corpus of the Croatian language, and the aim is for the collocations to be representative of the Croatian (standard) language and not of the available corpora.

In order to present the approach to collocations in *Mrežnik*, the paper focuses on the problem of collocation extraction for *Mrežnik* and the compilation of the collocational blocks for different word classes. Furthermore, it also shows how collocations help differentiate between meanings of polysemous words, when and how pragmatic and normative notes explaining the usage of collocations are added, when stylistic and terminological labels are used, and how collocations help the lexicographer differentiate between the meanings of quasi-synonyms and recognize meanings not yet recorded in Croatian dictionaries.

2 MULTIWORD EXPRESSIONS IN *MREŽNIK*

Collocations are multiword expressions and in order to differentiate them from other multiword expressions, a brief overview of multiword expressions and the approach to them in *Mrežnik* will be provided. According to Atkins and Rundell (2008, p. 167), multiword expressions (MWE) “are a central part of the vocabulary of most languages, and need to be accounted for in the dictionary... All fixed and semi-fixed phrases are important, and worth recording during the analysis process of dictionary writing.”

[6] http://riznica.ihjj.hr/index.hr.html

[7] http://nlp.ffzg.hr/resources/corpora/hrwac/

[8] https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/

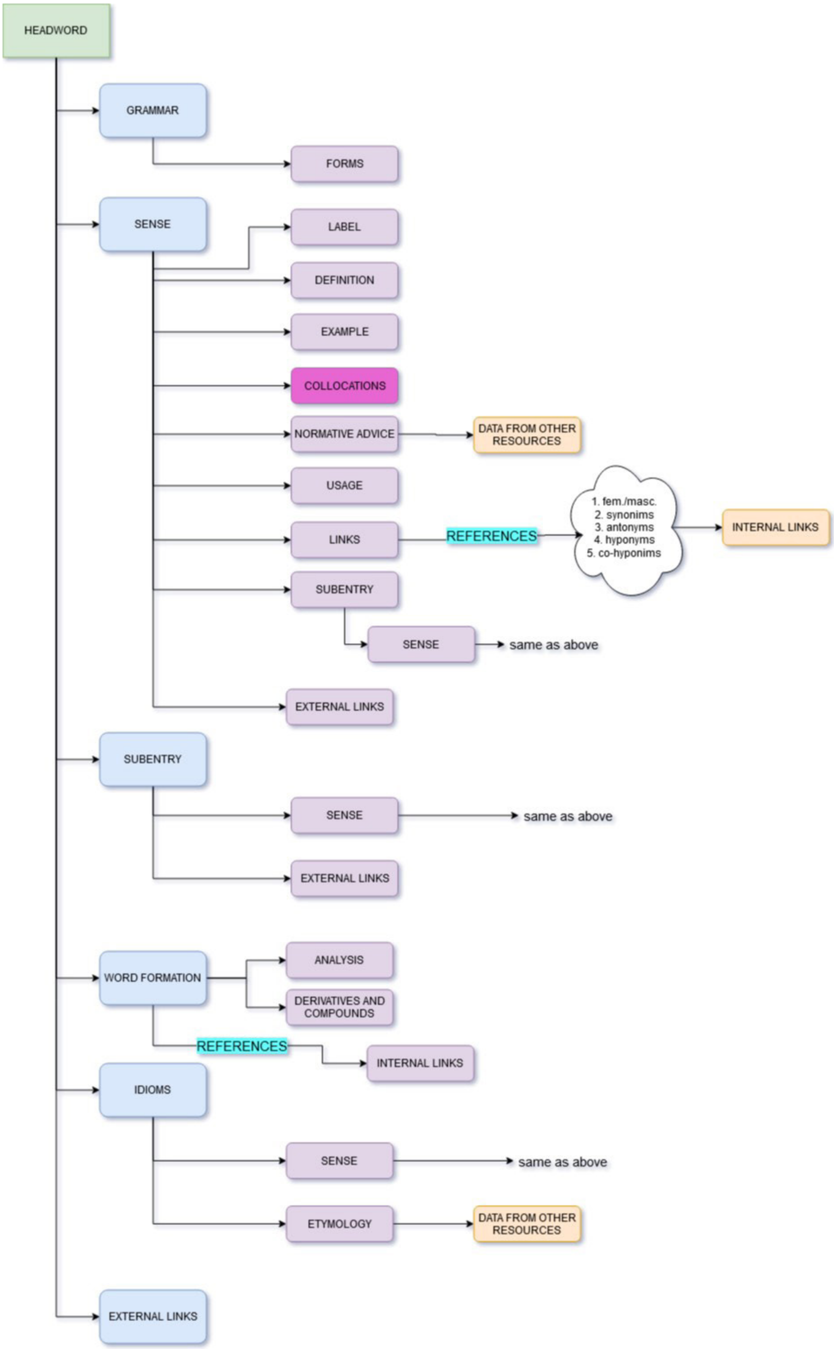


Figure 1: The microstructure of *Mrežnik*.

In the dictionary microstructure of *Mrežnik*, shown in Figure 1, multiword expressions can be presented in subentries (as headwords are always single words), the idiom section (which includes similes, catch phrases, quotations, and proverbs), and the collocational section. We briefly present the approach to multiword expressions used in the subentries and the idiom section before shifting the focus to collocations.

The subentries present terms or phrases the meaning of which cannot be derived from the sum of their constituent parts, e.g. *majčina dušica* (*majčina* = 'mother's', *dušica* = 'little soul', *majčina dušica* = 'thyme'), or when at least one word has a change in meaning, e.g. *morski pas* (*morski* = 'sea', *pas* = 'dog', *morski pas* = 'shark'). However, some frequent terms that can be derived from the sum of their constituent parts are also presented as subentries, especially if they can be linked to the *Struna* terminological database,[9] e.g. the subentries for the entry *broj* ('number') are the mathematical terms *prirodni broj* ('natural number'), *redni broj* ('ordinal number'), *glavni broj* ('cardinal number'). As the terms *redni broj* and *glavni broj* are also linguistic terms, they are also linked to *Hrvatska školska gramatika* ('Croatian School Grammar').[10] However, rare and lesser-known multiword terms are not always treated as subentries in *Mrežnik*. Some of the less frequent terms are provided in the collocational block, in which case they are not accompanied with a definition. The subentries also present some phrases, e.g. the entry *trokut* ('triangle') includes the subentries *ljubavni trokut* ('love triangle'), *ljubavni četverokut* ('love rectangle'), and *Bermudski trokut* ('Bermuda triangle').

The idioms section is compiled by a specially trained phraseologist. Idioms are linked to the database of explanations of the origins of idioms (*Frazemi. Hrvatski u školi*).[11] Some idioms are connected with the articles from the journal *Hrvatski jezik* that provide their etymology (section *Od A do Ž* 'from A to Z').[12]

[9] http://struna.ihjj.hr

[10] http://gramatika.hr

[11] http://hrvatski.hr/frazemi/

[12] https://hrcak.srce.hr/hrjezik

3 COLLOCATIONS IN *MREŽNIK*

Collocations are presented in the collocational block and in sentence examples. There are two basic criteria for choosing good sentence examples in *Mrežnik*: a) they contain a frequent collocation; b) they contain a typical syntactic construction. Some of the collocations provided in the collocational block are also illustrated through sentence examples in the example field.

Not all frequent collocations provided by Word Sketches are included in the final entries in *Mrežnik*. This is because of the difference between statistical collocation, i.e. "any combination of two or more words that is statistically relevant, and a collocation that is deemed relevant for inclusion in a dictionary" (Kosem et al., 2018, p. 991).[13] "Frequent but collocationally unremarkable" (Sinclair, 2002, p. 47) collocations have been excluded from *Mrežnik*. Moreover, due to the nature of *Mrežnik* (standard language dictionary, dictionary for general users, students, and non-native speakers), and especially due to the unrepresentativeness of the existing Croatian corpora, there are many other reasons for excluding statistically relevant collocations from *Mrežnik*: certain collocations are either offensive or inappropriate in polite conversation in standard Croatian, are relevant only to non-standard Croatian, or are not relevant for the general user. It is up to the lexicographers to decide how to select (only) relevant collocations. In addition to choosing suitable candidates, the lexicographers have to decide how and where to indicate collocations, as they can be entered under of the collocational base (semantically more autonomous word) or the collocate (semantically more dependent element), or both.

3.1 Extracting collocations for *Mrežnik*

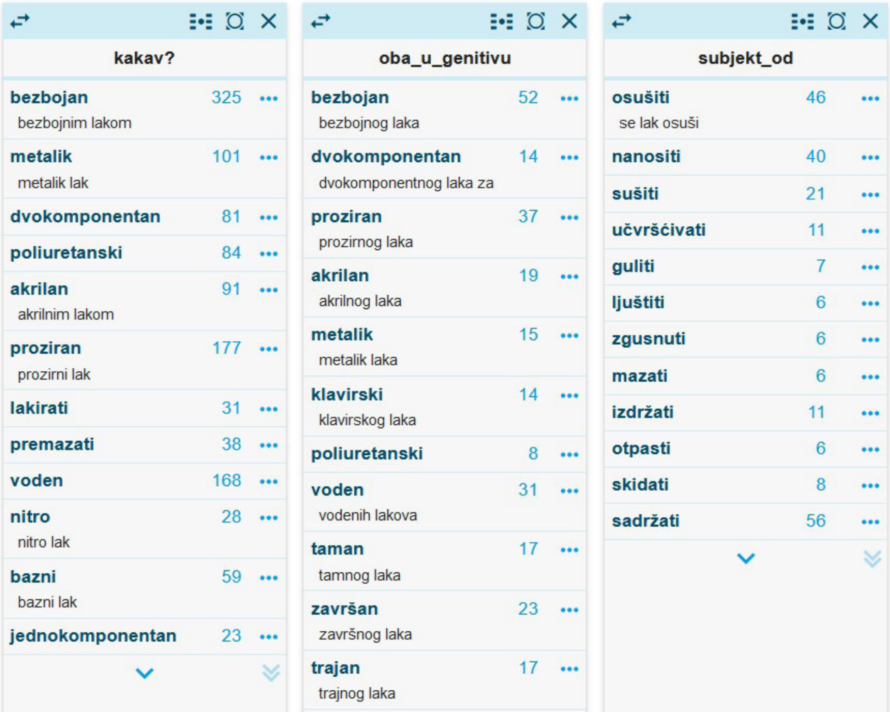
Collocations for the entries in *Mrežnik* are obtained in two ways:

1. Data is extracted from the corpora using the Sketch Engine web tool (cf. Kilgarriff et al., 2004), which allows the display of lemma/word context through Word Sketches (Kilgarriff and Rundell, 2002, pp. 811–815),[14] which are calculated using

[13] Kosem et al. (2018, p. 991) stress that not all statistically relevant collocations are worth 'showing' to dictionary users.

[14] "The Word Sketch processes the word's collocates and other words in its surroundings. It can be used as a one-page summary of the word's grammatical and collocational behaviour. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb,

the sketch grammar developed for Croatian within the *Mrežnik* project.[15] Collocations can be sorted by absolute frequency or logDice score (typicality of the collocation), per syntactic categories. Searches in Word Sketches can be limited to a selected part of speech, e.g. *lak* can be both a noun ('polish') and an adjective ('easy') in Croatian. Figure 2 shows a part of the Word Sketch for the noun *lak*.



kakav?

bezbojan	325	bezbojnim lakom
metalik	101	metalik lak
dvokomponentan	81	
poliuretanski	84	
akrilan	91	akrilnim lakom
proziran	177	prozirni lak
lakirati	31	
premazati	38	
voden	168	
nitro	28	nitro lak
bazni	59	bazni lak
jednokomponentan	23	

oba_u_genitivu

bezbojan	52	bezbojnog laka
dvokomponentan	14	dvokomponentnog laka za
proziran	37	prozirnog laka
akrilan	19	akrilnog laka
metalik	15	metalik laka
klavirski	14	klavirskog laka
poliuretanski	8	
voden	31	vodenih lakova
taman	17	tamnog laka
završan	23	završnog laka
trajan	17	trajnog laka

subjekt_od

osušiti	46	se lak osuši
nanositi	40	
sušiti	21	
učvršćivati	11	
guliti	7	
ljuštiti	6	
zgusnuti	6	
mazati	6	
izdržati	11	
otpasti	6	
skidati	8	
sadržati	56	

Figure 2: Partial Word Sketch for the noun *lak* ('polish').

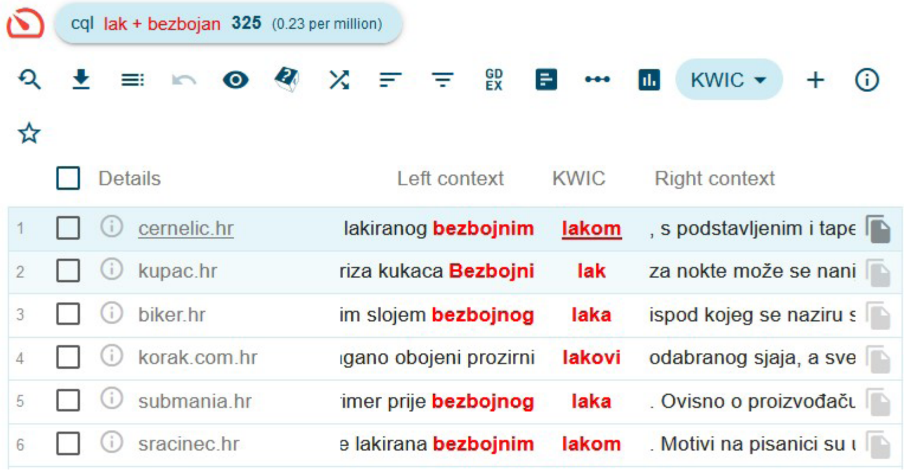
The structure is *adjective* + *noun* in the first column, *noun* + *noun* (both in the genitive case) in the second, and (subject)[16] *noun* lak + *verb* in the third.

words that modify the word etc. The words which will be included in the analysis are defined by rules written in the sketch grammar" https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/.

[15] The corpora were processed using ReLDI tagger with Word Sketches version 1.4 by Nikola Ljubešić within the *Mrežnik* project. The team members checked Word Sketches and suggested some additions and alterations (cf. Hudeček and Mihaljević, 2018b, pp. 106–107).

[16] Although the column is marked as subject, syntactic analysis shows that in many cases the collocate is the object of the collocation, e.g. *nanijeti lak* ('apply polish').

The selected columns are the most typical for the entry *lak*. However, other columns of Word Sketches are analysed by lexicographers as well. Concordances of these collocations are analysed with the option *get a random sample*. A partial concordance of the noun *lak* is shown in Figure 3.



cql lak + bezbojan 325 (0.23 per million)

	☐	Details	Left context	KWIC	Right context
1	☐	ⓘ cernelic.hr	lakiranog **bezbojnim**	**lakom**	, s podstavljenim i tape
2	☐	ⓘ kupac.hr	riza kukaca **Bezbojni**	**lak**	za nokte može se nani
3	☐	ⓘ biker.hr	im slojem **bezbojnog**	**laka**	ispod kojeg se naziru s
4	☐	ⓘ korak.com.hr	ıgano obojeni prozirni	**lakovi**	odabranog sjaja, a sve
5	☐	ⓘ submania.hr	imer prije **bezbojnog**	**laka**	. Ovisno o proizvođaču
6	☐	ⓘ sracinec.hr	e lakirana **bezbojnim**	**lakom**	. Motivi na pisanici su u

Figure 3: Concordance (random sample) of the noun *lak*.

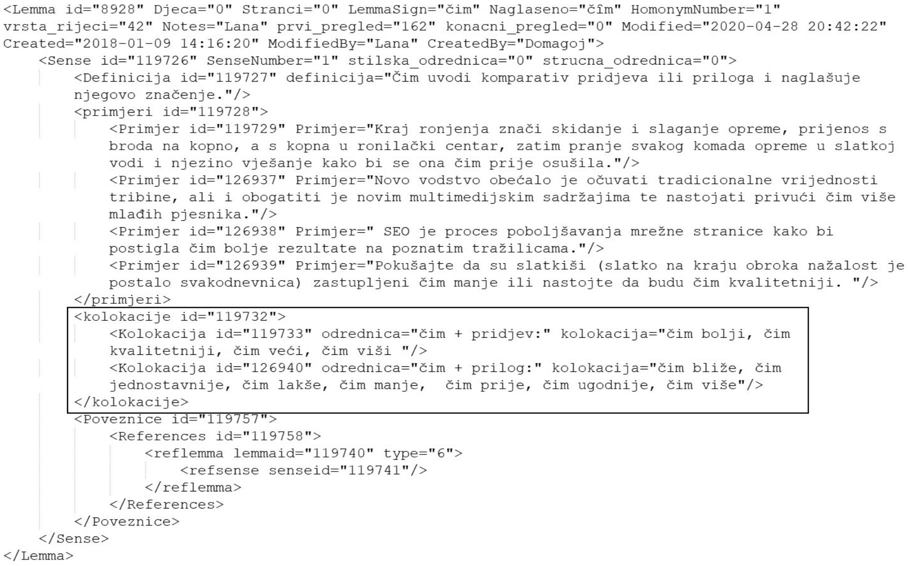
2. A random sample of approximately 300 examples is checked in the *hrWaC* and *Repository* corpora as some collocations the lexicographers know to be typical are not found via Word Sketches due to the unrepresentativeness of the corpus.

3.2 The collocational block in *Mrežnik*

Mrežnik is compiled in the TLex dictionary-writing system;[17] Figure 4 shows a simple (one meaning and just a few collocations) entry (the particle *čim* 'as soon as') in XML. A frame marks the part showing collocations.

The concept of collocations and their presentation was initially modelled after the example of *elexiko* (Haß, 2005; Storjohann, 2005). Thus, we began developing the model for collocations with the questions introduced in *elexiko* (Klosa, 2015, p. 36; Haß, 2005, p. 118). However, while working with the Croatian corpora, we modified the *elexiko* model in accordance with our language material. Collocations consist of a keyword (the headword or the subentry in our case) and a collocate. The same collocation is often listed in two entries,

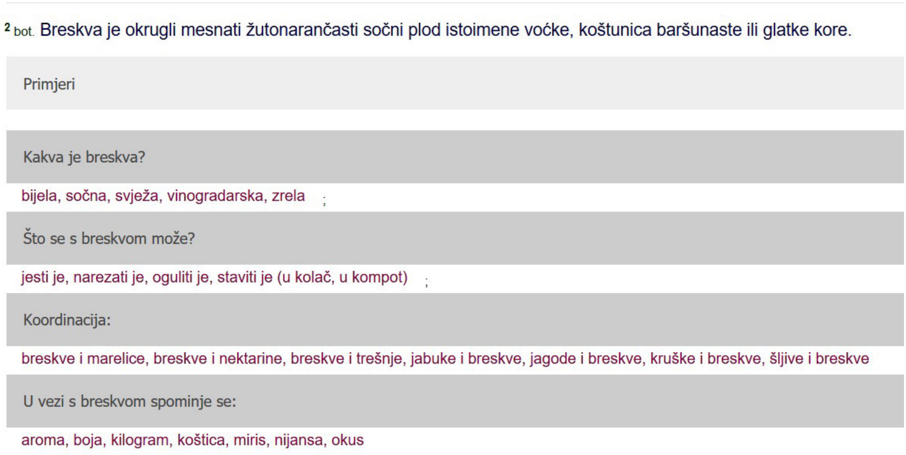
[17] https://tshwanedje.com/tshwanelex/



```
<Lemma id="8928" Djeca="0" Stranci="0" LemmaSign="čim" Naglaseno="čîm" HomonymNumber="1"
vrsta_rijeci="42" Notes="Lana" prvi_pregled="162" konacni_pregled="0" Modified="2020-04-28 20:42:22"
Created="2018-01-09 14:16:20" ModifiedBy="Lana" CreatedBy="Domagoj">
    <Sense id="119726" SenseNumber="1" stilska_odrednica="0" strucna_odrednica="0">
        <Definicija id="119727" definicija="Čim uvodi komparativ pridjeva ili priloga i naglašuje
        njegovo značenje."/>
        <primjeri id="119728">
            <Primjer id="119729" Primjer="Kraj ronjenja znači skidanje i slaganje opreme, prijenos s
            broda na kopno, a s kopna u ronilački centar, zatim pranje svakog komada opreme u slatkoj
            vodi i njezino vješanje kako bi se ona čim prije osušila."/>
            <Primjer id="126937" Primjer="Novo vodstvo obećalo je očuvati tradicionalne vrijednosti
            tribine, ali i obogatiti je novim multimedijskim sadržajima te nastojati privući čim više
            mlađih pjesnika."/>
            <Primjer id="126938" Primjer=" SEO je proces poboljšavanja mrežne stranice kako bi
            postigla čim bolje rezultate na poznatim tražilicama."/>
            <Primjer id="126939" Primjer="Pokušajte da su slatkiši (slatko na kraju obroka nažalost je
            postalo svakodnevnica) zastupljeni čim manje ili nastojte da budu čim kvalitetniji. "/>
        </primjeri>
        <kolokacije id="119732">
            <Kolokacija id="119733" odrednica="čim + pridjev:" kolokacija="čim bolji, čim
            kvalitetniji, čim veći, čim viši "/>
            <Kolokacija id="126940" odrednica="čim + prilog:" kolokacija="čim bliže, čim
            jednostavnije, čim lakše, čim manje,  čim prije, čim ugodnije, čim više"/>
        </kolokacije>
        <Poveznice id="119757">
            <References id="119758">
                <reflemma lemmaid="119740" type="6">
                    <refsense senseid="119741"/>
                </reflemma>
            </References>
        </Poveznice>
    </Sense>
</Lemma>
```

Figure 4: An entry in XML (particle *čim* 'as soon as').

e.g. *crvena jabuka* ('red apple') under *crven* ('red') and under *apple* ('jabuka'). The structure of the collocational block is divided into two fields. Figure 5 shows the demo version of the collocational block for one of the meanings of the headword *breskva* ('peach').



2 bot. Breskva je okrugli mesnati žutonarančasti sočni plod istoimene voćke, koštunica baršunaste ili glatke kore.

Primjeri

Kakva je breskva?

bijela, sočna, svježa, vinogradarska, zrela ;

Što se s breskvom može?

jesti je, narezati je, oguliti je, staviti je (u kolač, u kompot) ;

Koordinacija:

breskve i marelice, breskve i nektarine, breskve i trešnje, jabuke i breskve, jagode i breskve, kruške i breskve, šljive i breskve

U vezi s breskvom spominje se:

aroma, boja, kilogram, koštica, miris, nijansa, okus

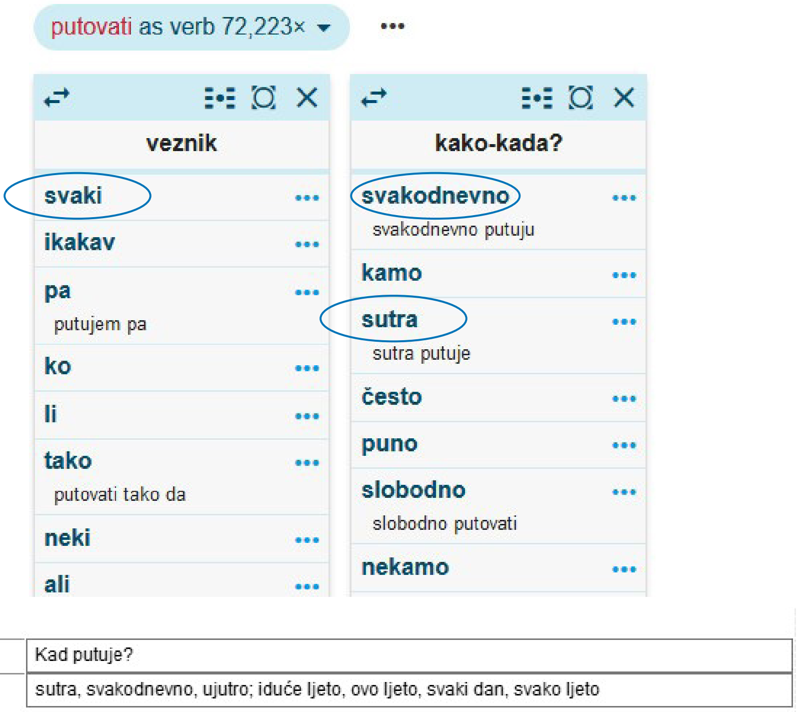
Figure 5: Demo version of the collocational block of the entry *breskva* ('peach').

As is apparent from Figure 5, each collocational field in the collocational block consists of two subfields (determinant and collocates). Determinants can be:

1. Questions, e.g. *Kakva je breskva?* ('What is a peach like?'), *Što se s breskvom može?* ('What can one do with a peach?'). There is a limited number of questions for each word class. However, if needed, the editors can add more questions. These questions usually mirror grammatical relations, e.g. the answer to the question *What is x like?* is typically an adjective, sometimes a noun in the genitive case, and less often the construction *noun + preposition* za ('for') + *noun* or a semi-compound.

2. Introductory phrases, e.g. *Koordinacija:* ('Coordination'), *U vezi s x spominje se*: ('Mentioned in connection with x'), *U imenima* ('In names').

3. Grammatical formula (usually used with grammatical words), e.g. *usklik + imenica u dativu:* ('interjection + noun in the dative').



putovati as verb 72,223×

veznik
- svaki
- ikakav
- pa
  putujem pa
- ko
- li
- tako
  putovati tako da
- neki
- ali

kako-kada?
- svakodnevno
  svakodnevno putuju
- kamo
- sutra
  sutra putuje
- često
- puno
- slobodno
  slobodno putovati
- nekamo

Kolokacija:

odrednica	Kad putuje?
kolokacija	sutra, svakodnevno, ujutro; iduće ljeto, ovo ljeto, svaki dan, svako ljeto

Figure 6: Comparison of Word Sketch columns and a collocational field (the verb *putovati* 'to travel').

The selected collocates of the headword follow in the second subfield. They are provided in alphabetical order and not by frequency. This is illustrated with a comparison of the Word Sketch columns *kako-kada* ('how-when') and *veznik* ('conjunction') with the collocational field *kada* 'when' in *Mrežnik*, as shown in Figure 6. The *veznik* column includes many words that are not conjunctions, some of which are relevant for this collocational field. The comparison shows which collocates from the Word Sketch have been selected for the presentation in this field in *Mrežnik*. It is an evidence that collocations cannot be extracted mechanicaly from Word Sketches and must be carefully selected by the lexicographer.

Collocates are also occasionally grouped into grammatical and/or semantic groups, with the groups being separated by a semicolon.

3.3 Collocations of different word classes

The editors developed a set of collocational questions, introductory phrases, and/or grammatical formulas for each word class after analysing a sample dataset (cf. Hudeček and Mihaljević, 2018b). These were modified and new questions added if needed. Each word class presented different collocational problems. The collocational questions and introductory phrases always appear in the same order. An overview of typical collocational questions and phrases for each word class is provided in the following sections.

3.3.1 Nouns

Table 1 shows the collocational questions and phrases[18] for nouns.

[18] The questions and introductory phrases always appear in the same order, although not all of them are used for every headword. This is why different headwords were used to illustrate the collocations in Table 1.

Table 1: *Collocational questions and introductory phrases – nouns*

Croatian		**English**	
Question or introductory phrase	**Example collocations**[19]	**Question or introductory phrase**	**Literal translation of Croatian collocations**
Kakav je x?	**mašta**: bolesna, bujna, neiscrpna, neobuzdana, pokvarena	What is x like?	**imagination**: sick, vivid, inexhaustible, unrestrained, corrupt
Što x ima?	**list**: bazu, peteljku, plojku, žilice	What does x have?	**leaf**: base, petiole, plate, veins
Što x može?	**Crkva**: osuđivati, priznavati, pozivati, slaviti, učiti, upozoravati	What can x do?	**Church**: condemn, acknowledge, invoke, celebrate, teach, warn
Što se s x može?	**mašta**: buditi je, pobuditi je, razbuktati je,	What can one do with x?	**imagination**: awaken, inflame, kindle
Koordinacija:	**mašta**: mašta i fantazija, mašta i kreativnost, mašta i stvarnost, mašta i volja	Coordination:	**imagination**: imagination and fantasy, imagination and creativity, imagination and reality, imagination and will
U vezi s x spominje se:	**mašta**: plod, proizvod, tvorevina, zaljubljenik	Mentioned in connection with x	**imagination:** fruit, product, creation, lover
U imenima:	**duh**: Koko i duhovi (novel)	In names:	**ghost:** Koko and the ghosts[20]

Descriptive and possessive adjectives that answer the first collocational question *What is x like?* are alphabetized separately and separated with a semicolon, as shown in Table 2.

Table 2: *Collocates of the word mjenjačnica ('exchange office')*

	Croatian	**English**
	mjenjačnica	**exchange office**
	Kakva je mjenjačnica?	What is an exchange office like?
descriptive adjective	obližnja, ovlaštena, povoljna, privatna, zatvorena	nearby, authorized, affordable, private, closed
possessive adjectives	supetarska, trogirska	from Supetar, from Trogir

[19] The table provides collocations for one meaning of each headword only. Most of the headwords have more than one meaning.

[20] A famous Croatian novel for children.

While modeling the collocational block for nouns, the editors had to answer these questions:

• **which collocations to include.** When choosing collocations, the editors take into account corpus data (provided by Word Sketch) and evaluate the suitability of collocations for inclusion in the collocational block of *Mrežnik*. For example, the collocations *brkata konobarica* ('mustachioed waitress'), *sisata konobarica*, *prsata konobarica* ('large-breasted waitress'), *alkoholizirani maturant*, *pijani maturant* ('drunk secondary-school graduate') would not be considered as suitable collocations for *Mrežnik* although they have the highest logDice score in the column *kakav?* (What is x like?). Namely, any collocations that might insult anybody on the basis of their age, sex, race, sexual orientation, nationality, religion, etc. have been excluded from *Mrežnik* (cf. Hudeček and Mihaljević, 2018b, p. 109).

• **coordination.** The coordination field lists elements connected with *i* ('and'), *te* ('as well as'), *ili* ('or'), *odnosno* ('namely'), and /. Coordination presented the following problems:

1) how to differentiate between the following two cases:

a) *X* belongs simultaneously to two groups connected by a coordinator, e.g. *nogometaš i sportaš* ('a footballer and an athlete'), *nastavnik i pedagog* ('a teacher and an educator'), *profesorica* i *prevoditeljica* ('a teacher and a translator'), *književnica i prevoditeljica* ('a writer and a translator'), *vaterpolist i reprezentativac* ('a water polo player and a member of the national team').

b) Two groups are linked by a coordinator, e.g. *nogometaši i košarkaši* ('footballers and basketball players'), *učenici i nastavnici* ('students and teachers'), *vaterpolisti i košarkaši* ('water polo players and basketball players').

The lexicographer distinguishes between the two groups on the basis of an analysis of Word Sketch and concordances. The solution used in *Mrežnik* is to separate the two groups according to the opposition singular/plural and placing a semicolon between them.

2) How to differentiate between these two cases:

a) the noun refers only to men;

b) the noun (especially in the plural) refers to both men and women.

These two groups were separated by a semicolon and introduced by the introductory phrase *odnosi se samo na muškarce* (‘refers only to men’) or *odnosi se samo na muške osobe* (‘refers only to male persons’). This is illustrated by the coordination of the word *vaterpolist* (‘water polo player’). The difference between the two introductory phrases is that the phrase *odnosi se samo na muškarce* is used when it applies only to men (e.g. *liječnik* ‘doctor’) and the phrase *odnosi se samo na muške osobe* applies to boys as well as men (*nogometaš* ‘footballer’). This is shown in Table 3.

Table 3: *Coordination of the noun vaterpolist (‘water polo player’)*

Koordinacija	Coordination
vaterpolist i reprezentativac; vaterpolisiti i košarkaši; *odnosi se samo na muške osobe*: vaterpolisti i vaterpolistice	water polo player and member of the national team; water polo players and basketball players*, refers only to male persons*: water polo players and water polo players (f.)

In the coordination field, collocations can refer to the same person, e.g. *vaterpolist i reprezentativac* (‘water polo player and the member of the national team’), and to two groups of sportsmen, e.g. *vaterpolisti i košarkaši* (‘water polo players and basketball players’). The collocation *vaterpolisti i vaterpolistice* (‘male water polo players and female water polo players’) refers only to *muške osobe* (‘male persons’).

3) The order of nouns connected by a coordinator had to be determined. After trying out all possibilities,[21] we decided to make the headword the first member of the coordinated phrase. The exception to this are set phrases the order of which is fixed or much more common, e.g. *lijevi i desni* (‘left and right’).

[21] The possibilities were: the collocation is copied in the form it occurs in the Word Sketch with the possibility of repeating the same elements but in a different order (e.g. *vaterpolist i reprezentativac* but possibly also *reprezentativac i vaterporist*), the collocation is listed in the order the elements occur more often but without repeating the elements (e.g. only *vaterpolist i reprezentativac*), the collocation is listed in the order in which the headword appears in the second place (*reprezentativac i vaterpolist*).

4) Collocations are listed in the order of the coordinator used: *i* (‘and’), *te* (‘and’), *ili* (‘or’), *ni* (‘neither’), *niti* (‘nor’), /, *odnosno* (‘rather’); collocations with each new coordinator are separated by a semicolon.

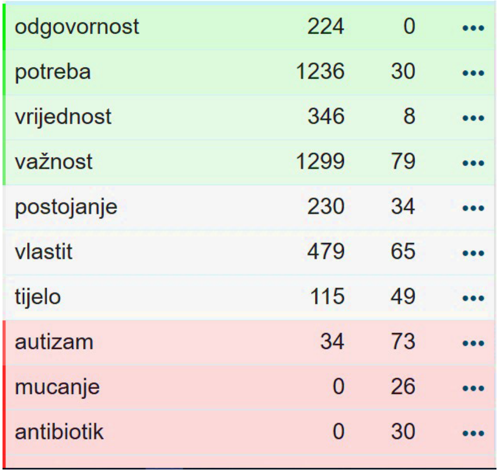
• **proper names.** Although one can argue that proper names are not collocations, they were included in the collocational field. As proper names occur quite frequently in Word Sketches, this information could be useful for the user,[22] e.g. the word *list* (‘leaf’) occurs most often in the names of newspapers (*Večernji list*, *Jutarnji list*, etc.); *Večernji list* and *Jutarnji list* have the highest score in Word Sketches for the lemma *list*. After analysing all name categories occurring in Word Sketches, it was decided to include the following categories in the collocational block: place names, names of organizations and events, names of holidays, and names of commemorations.

The occurrence of the headword in names often revealed facts that were commented upon in the pragmatic note, e.g. *jučer* (‘yesterday’) in the meaning ‘past time’, *danas* (‘today’) ‘present time’, and *sutra* (‘tomorrow’) ‘future time, time to come’ are used very often in the names of various events, e.g. *Razvoj turizma u Kninu: danas i sutra* (‘The development of tourism in Knin: today and tomorrow’). The term *svjesnost* (and not the synonymous term *svijest*) often occurs in proper names. The word *svjesnost* ‘awareness’ (as opposed to its (quasi-)synonym *svijest*) occurs most often in the names of days, weeks, or months dedicated to something, often an illness, disability, or disorder, e.g. *Međunarodni dan svjesnosti o mucanju* (‘International Stuttering Awareness Day’).[23] The Word Sketch Difference for the grammatical relation *noun + preposition o* (‘about’) + *noun* of the lemmas *svijest* (90,010 occurrences in the corpus) and *svjesnost* (8,621 occurrences in the corpus) is shown in Figure 7. The numbers in the second column indicate the frequency of collocates of *svijest*, while those in the second column indicate the frequency of collocates

[22] This is based on our experience in giving language advice. Users sometimes ask for advice in choosing an appropriate name for an event or a document title. This is a question of combining word elements appropriately, not of encyclopedic knowledge. Single-word proper names are never entry words in *Mrežnik*, but they are sometimes provided in the pragmatic note (e.g. the personal names *Jagoda* and *Višnja* in the entries *jagoda* [‘strawberry’] and *višnja* [‘sour cherry’]). This is especially useful in the module for non-native speakers of Croatian.

[23] For more on the meaning of the terms *svijest* and *svjesnost*, cf. Vrgoč and Mihaljević (2019).

of *sujesnost,* e.g. *svijest o odgovornosti* ('awareness of responsibility'), *svijest o potrebi* ('awareness of a need') vs. *svjesnost o autizmu* ('awareness of autism'), *svjesnost o mucanju* ('awareness of stuttering').



odgovornost	224	0	•••
potreba	1236	30	•••
vrijednost	346	8	•••
važnost	1299	79	•••
postojanje	230	34	•••
vlastit	479	65	•••
tijelo	115	49	•••
autizam	34	73	•••
mucanje	0	26	•••
antibiotik	0	30	•••

Figure 7: Partial Word Sketch Difference for lemmas *svijest* and *svjesnost*.

Names are introduced by the introductory phrase *U imenima*: ('in names'). The class to which the name belongs (e.g. film, novel, event) is provided in brackets if the name is not self-explanatory, e.g. for *Hrvatski slavistički kongres* ('Croatian Slavic Studies Congress'), the word *kongres* is not provided as an explanation as it is a part of the name; for *Bravo maestro* ('Well done Maestro'), the word *film* is provided in brackets.

• **the grammatical form of collocates.** Although collocational questions and answers are mostly in the singular, sometimes the plural was required. This is the case if the collocation implies more than one person or thing, e.g. *What can x do? okupljati se* ('bring together'). Singular and plural collocates are separated by a semicolon (as shown in Table 3).

• **terminological and stylistic labels.** Terminological and stylistic labels are used in the collocational field in some cases. This is especially true for collocations that are only used in the colloquial style or that do not belong to the standard language. Granger and Paquot (2012, p. 165) stress that non-native writers can be seriously misled by the presentation of collocations, as they are not provided with any help to decide which collocations are the most appropriate in academic writing. This is often also true for native speakers and in all styles of writing.[24]

Style labels are used when the collocate is stylistically marked or does not belong to the standard language, e.g. one of the answers to the question *Što stomatologinja može?* ('what can a dentist do') is *pokrpati zub* ('mend a tooth') marked by the label *žarg*. ('jargon'), as this collocation does not belong to the standard language.

• **dividing groups of collocates.** Collocates are sometimes grouped and divided by a semicolon according to syntactic and semantic criteria, e.g. the answer to the question *What is x like?* can be an adjective, a compound (consisting of two nouns, sometimes hyphenated) or a phrase that has the structure *headword + noun in the genitive*. These groups are separated by a semicolon as shown in the collocational field *Kakva je čistačica?* of the entry *čistačica* in Table 4.

Table 4: *Collocates of the noun čistačica[25] ('cleaning lady')*

Kakva je čistačica?	What is a cleaning lady like?
dežurna, obična, školska, vrijedna, zaposlena, x-godišnja; čistačica spremačica, teta čistačica *hip*.	on call, ordinary, school, hardworking, employed, x-year-old; cleaning lady and housekeeper, aunty cleaning lady

Table 4 shows two groups of collocations answering the question *Kakva je čistačica?* ('What is a cleaning lady like?'):

- adjectives, e.g. *vrijedna čistačica* ('hardworking cleaning lady');
- nouns, e.g. *čistačica spremačica* ('cleaning lady and housekeeper'), *teta čistačica* ('aunty cleaning lady').[26]

As the age of a person often occurs in the corpus, this is indicated by the construction *x-godišnji/x-godišnja* ('x-year-old').

[24] This statement is supported by our experience in giving language advice, teaching Croatian to students of electrical engineering and journalism (native speakers of Croatian), and editing Croatian texts written by native speakers, as well as by Hudeček (2020) and Blagus Bartolec (2017).

[25] For more on masculine and feminine professional nouns in *Mrežnik*, see Hudeček and Mihaljević (2019b).

[26] *Teta čistačica* is a hypocoristic way young children address cleaning ladies at school or in kindergarten. This is indicated with the label *hip*. ('hypocoristic').

3.3.2 Verbs

Verbal collocations are very complex as they depend on the syntactic characteristics of the verb (reflexive, impersonal, transitive, intransitive), verbal valence, and the semantic characteristics of the verb. Each semantic class of verbs (e.g. verbs of motion, psychological verbs, etc.[27]) has partly different collocational questions. Collocational questions are divided according to the sentence elements that answer them: subjects, objects, adverbials. Questions are different for imperfective, perfective, and reflexive verbs, as well as for animate and inanimate subjects.

1. Questions for the subject are shown in Table 5.

Table 5: *Collocations denoting the subject*

	Imperfective		Perfective	
	Croatian	**English**	**Croatian**	**English**
animate	Tko x?	Who x?	Tko može x?	Who can x?
	trčati: atletičar, konj	**run**: athlete, horse	**upoznati:** polaznik, student	**get to know**: attendant, student,
inanimate	Što x?	What x?	Što može x?	What can x?
	svijetliti: krijesnica, lampa	**shine**: firefly, lamp	**pasti**: bomba, jabuka	**fall**: bomb, apple

2. Questions for the object are shown in Table 6.

Table 6: *Collocations denoting the object*

	Imperfective verbs		Perfective verbs		Reflexive verbs	
	Croatian	**English**	**Croatian**	**English**	**Croatian**	**English**
Direct object	Što se x?	What is x?	Što se može x?	What can be x?		
	čitati: knjiga, tekst	**read:** book, text	**dati:** glas, odgovor,	**give:** a vote, a response		
Indirect object	Čemu x? Komu x?	To/at whom/what can one x?	Komu se može x?	To/at whom can one x?	Komu se x? Čemu se x?	To/at whom/ what can one x?
	mahati gomili, obožavateljima,	**wave to**: the crowd, the fans	**mahnuti**: konobarici, navijačima	**wave**: the waitress, the fans	**smijati se:** prijatelju, šali	**laugh**: a friend, a joke

[27] The semantic classification of verbs is based on the classification made in the project *e-Glava*. More on the project *e-Glava* see Birtić et al. (2017) and on the classification of verbs see Brač and Bošnjak Botica (2015).

3. The questions for adverbial collocations depend on the semantic class of the verb (e.g. motion verbs have different questions than static verbs) and on the adverbial class as shown in Table 7 (only imperfective verbs are shown). Perfective verbs have modified questions, e.g. imperfective verb: *Kako se x?*, perfective verb: *Kako se može x?*; imperfective verb: *Kad se x?*, perfective verb: *Kad se može x?*, etc.

Table 7: *Collocations denoting adverbials*

	Croatian		**English**	
adverbial	**question**	**example**	**question**	**example**
of manner	Kako (se može) x?	**mahati:** bijesno, nervozno	How can one x?	**wave**: angrily, nervously
of place	Gdje x? (static verbs)	**ljetovati:** u kampu, na Pagu	Where x?	**spend the summer**: in a camp, on Pag
	Kamo x? (verbs of motion)	**putovati:** kući, izvan grada	To where x?	**travel**: home, out of the city
	Kuda x? (verbs of motion)	**putovati**: diljem svijeta, kroz Neum	Which way x?	**travel:** across the world, through Neum
of time	Kad x?	**svijetliti**: noću, trajno	When x?	**shine**: at night, permanently
of reason	Zbog čega x?	**putovat**i: zbog posla, zbog zabave	Why x?	**travel**: for work, for fun
of company	S kim x?	**putovati:** s klubom, s prijateljima,	With whom x?	**travel:** with a club, with friends
of means	Čime se x?	**mahati**; krilima, pištoljem	With what x?	**wave**: wings, a gun
of frequency	Koliko često x?	**putovati**: često, tjedno	How often x?	**travel**: often, weekly

Tables 6 and 7 show the complexity of verbal collocations. Coordination also often occurs with verbs: *voljeti i ljubiti* (love and love/kiss), *voljeti i mrziti* (love and hate).

3.3.3 Adjectives

The most common question introducing adjectives is the question *Što je x?* (What is x?). We list the nouns answering this question in the following order: animate, inanimate, abstract. These three noun groups are divided by semicolons. Collocational questions and introductory phrases for the adjective *loš* (bad) are provided in Table 8.

Table 8: *Collocational questions and introductory phrases – adjective loš ('bad')*

Croatian	Example	English	Example
Što je loše?	čovjek; navike	What is bad?	person; habits
Koliko je što loše?	jako, iznimno	To what degree is something bad?	very, extremely
Koordinacija:	loš i nekvalitetan; dobar ili loš	Coordination:	bad and of low quality; good or bad

Terminological labels are used only to distinguish between different meanings of the collocate, e.g. the entry *crven* ('red') features the question *Što je crveno?* 'What is red?', the answers to which are e.g. *div* astr. ('giant', astronomy), *karton* sp. ('card', sports), *patuljak* astr. ('dwarf', astronomy), *vjetar* med. ('wind', medicine). Collocates of the adjective *crven* ('red') are given in Table 9.

Table 9: *Collocates of the adjective crven ('red')*

Što je crveno?	What is red?
boja, haljina; div *astr*., karton sp., krvna zrnca, patuljak *astr*., vjetar *med.*	color, dress; div *astr*., card *sp*., blood cells, dwarf *astr*., wind *med.*

3.3.4 Adverbs

Collocations of adverbs formed in Croatian from the neutral form of adjectives by conversion (e.g. *jako* formed from the neutral form of the adjective *jak*) are introduced by the questions *Što se može x?* ('What can be done in a x manner?') and *Koliko je što x?* ('To what degree is something x?'). However, in other adverb groups, collocations are introduced by the introductory phrase *uz glagole:* ('with verbs:'), e.g. the adverbs *gdje* ('where')*, kuda* ('where to)*, kamo* ('which way'), and *uz prijedloge* ('with prepositions'), e.g. *jako blizu* ('very near'). Table 10 shows the collocational questions and introductory phrases for adverbs on the example of *loše* ('badly').

Table 10: *Collocational questions and introductory phrases for loše ('badly')*

Croatian		English	
Question and phrases	**Examples**	**Question and phrases**	**Examples**
Što se može loše?	**loše**: biti plaćen, igrati	What can be done badly?	**badly:** be paid, play
Koliko je što loše?	**loše**: katastrofalno, veoma	To what degree is something bad?	**badly:** disastrously, very
Uz pridjeve:	**besmrtno**: neozbiljan, zaljubljen	With adjectives:	**immortally**: frivolous, in love
Koordinacija:	**loše**: dobro i lose; loše ili nikako	Coordination:	**badly**: well and badly; badly or not at all

3.3.5. Numbers

Collocational questions, introductory phrases, and grammatical formulas differ for cardinal and ordinal numbers, and in Table 11 we provide prototype collocational questions for both groups. Although one can argue that no collocations need be given with numbers and that numbers are not collocational words at all, based on our experience with students and providing language advice, we believe that some prototype collocations with numbers can also be useful (from a semantic and a syntactic point of view), e.g. *prvo mjesto* ('first place'); *sedam patuljaka* ('seven dwarfs'), *sedam dana* ('seven days'); *dvanaest mjeseci* ('twelve months'); *pet do devet* ('five to nine'), *pet na dan* ('five a day'), *pet od šest* ('five out of six'), etc. Table 11 shows collocations of cardinal and ordinal numbers.

Table 11: *Collocational questions and introductory phrases – numbers*

	Croatian		English	
	Question	**Example**	**Question**	**Example**
glavni brojevi ('cardinal numbers')	Čega je x?	pet prstiju	What do we have x of?	five fingers
	x + prijedlog + N	pet na dan	x + preposition + y	five a day
	Koordinacija:	pet i šest	Coordination:	five and six
redni brojevi ('ordinal numbers')	Što je x?	peti mjesec	What can be x?	fifth month (May)
	Koordinacija:	peti ili šesti	Coordination:	fifth or sixth

Some collocations with numbers motivated the inclusion of a normative note, e.g. *drugi najbolji* ('second best') is a very common collocation in the corpus but should be replaced by *drugi* ('second') in standard Croatian, as *drugi najbolji* is considered a pleonasm and literal translation from English.

3.3.6 Interjections

Collocations of interjections are mostly introduced with syntactic formulas and the introductory phrases *Koordinacija:* (‘koordination’) and *U imenima:* (‘in names’), as shown in Table 12.

Table 12: *Collocational questions and introductory phrases – interjections*

Croatian		**English**	
glagol + x:	reći bravo	verb + x:	say bravo
x + prijedlog + :	bravo za orkestar	x + preposition + noun:	bravo to the orchestra
x + imenica u vokativu:	bravo dečki	x + noun in the vocative:	bravo (well done) boys
Koordinacija:	ajme i jao	Coordination:	oh my and wow
U imenima:	Bravo Maestro (film)	In names:	Well Done Maestro (film)

3.3.7 Pronouns

Collocational questions depend on the pronoun category (personal pronoun, possessive pronoun, demonstrative pronoun, relative pronoun, etc.). Table 13 shows some collocational questions and introductory phrases for personal and possessive pronouns.

Table 13: *Collocational questions and introductory phrases – pronouns*

	Croatian		**English**	
personal pronouns	**Koordinacija:**	**ja**: (i) ja i ti; (ili) ja ili on/ona	Coordination:	**I/me:** you and I; I or he/she
possessive pronouns	**Što je x?**	**moj:** djetinjstvo, mišljenje	What is x?	**my:** childhood, opinion
	Koordinacija:	moj i tvoj, ti i tvoj…	Coordination:	mine and yours, you and your…
	U imenima:	Naši i vaši (serija)	In names:	Ours and Yours (TV series)

Possessive pronouns have the same question *Što x može*? (‘What can x do?’) as adjectives. Possessive pronouns sometimes function as nouns, e.g. one of the meanings of the prounoun *naši* (‘our’). In this case, collocations can be the same as typical collocations of nouns, e.g. *Što naši mogu?* (‘What can ours do?’) *biti poraženi*, *izgubiti, pobijediti, slaviti, trijumfirati* (‘be beaten, lose, win, celebrate, triumph’).

3.3.8 Conjunctions

Typical collocations of conjunctions are introduced by the phrase *U vezničnim skupinama:* ('in conjunction groups'), e.g. *ali* ('but?'): *ali ipak* ('but still')*, ali isto tako* ('but the same'). Reduplicated conjunctions such as *ili…ili* ('either…or') have syntactic formulas such as *Uz glagole:* ('with verbs') and *Uz prijedloge u prijedložnim izrazima:* ('with prepositions in prepositional phrases'), e.g. *ili dati ili uzeti* ('either give or take'), *ili ostati ili otići* ('either leave or stay'), *ili izvan čega ili unutar čega* ('either outside or inside of something'), etc.

3.3.9 Particles

There is no unique collocational model for particles. The collocational field is adapted to each collocation, e.g. modifiers are introduced by introductory phrases stating the word class which follows the modifier, e.g. *Uz pridjeve:* ('with adjectives') *čim bolji, čim veći* ('as good as possible, as big as possible'), *Uz priloge*: ('with adverbs') *čim bliže, čim jednostavnije* ('as close as possible, as simple as possible').

3.3.10 Prepositions

Prepositions are the only word class for which no collocations are provided in *Mrežnik*, as they are considered a non-collocational word class. The reason for this is that word combinations like *iz daljine* ('from afar'), *iz inata* ('out of spite') are provided in examples under different meanings as shown in Table 14.

Table 14: *Meanings and examples of the preposition iz ('from')*

Croatian		**English**	
Definition	**Example**	**Definition**	**Example**
Iz označuje da tko ili što izlazi ili potječe odakle	Krenuli smo iz Kutine u 6 sati.	Iz ('from') indicates that somebody or something leaves or originates from somewhere.	We left Kutina at 6 o'clock.
Iz označuje da tko ili što pripada određenomu razdoblju.	Crkveni je namještaj uglavnom iz doba baroka i klasicizma.	Iz ('from') indicates that somebody or something belongs to a certain period.	The church furnishings are mostly from the Baroque and Classicist period.
Iz označuje da je što uzrok čemu drugom.	Turci su, nemajući što izgubiti, zaigrali iz inata.	Iz ('out of') indicates that something is the reason for something else.	The Turks, having nothing to lose, played out of spite.

3.4 The role of collocations in determining and differentiating meanings

Work on *Mrežnik* confirms Firth's (1957, p. 11) famous slogan "You shall know a word by the company it keeps". Namely, the meaning of words is "determined by their grammatical and lexical environment (syntagmatic relations like colligation and collocation) as well as by the situation in which they are used (style, pragmatics)" (Altenberg and Granger, 1996, p. 22). Collocations for each word class in *Mrežnik* helped the lexicographers distinguish meanings, provide precise definitions, and list useful pragmatic and normative notes. For example, in the analysis of the antonymous adjectives *dobar* ('good') and *loš* ('bad'), closely connected meanings were defined as shown in Table 15. Other meanings in which these adjectives are not antonymous are not provided in this table. The table only provides collocations answering the question *What is x?*.

Table 15: *Collocates for different meanings of loš ('bad') and dobar ('good')*

	Definition		**Collocates**	
loš ('bad', 'wrong')	Loš je koji ima negativne osobine ili neželjena svojstva.	Bad is that which has negative characteristics.	čovjek, kvaliteta, strana, stvar, vrijeme	person, quality, side, thing, time
	Loš je koji nije onakav kakav treba biti, koji ne ispunjava očekivanja.	Bad is that which is not as it should be, that which does not fulfil expectations.	igra, ocjena, odnos, rezultat, situacija, stanje, start	game, rating, relationship, result, situation, condition, start
	Loš je koji obavještava o nečemu lošem ili najavljuje loše.	Bad is that which reports on something bad or predicts something bad.	najava, vijest, znak	announcement, news, sign
	Loš je koji nije ispravno utemeljen i logičan.	Bad is that which is not correctly founded or logical.	zaključak	conclusion
	Loš je koji ne donosi korist, koji nema rezultate.	Bad is that which does not bring profit or results.	poslovanje, plan, (poslovni) potez	business, plan, (business) move

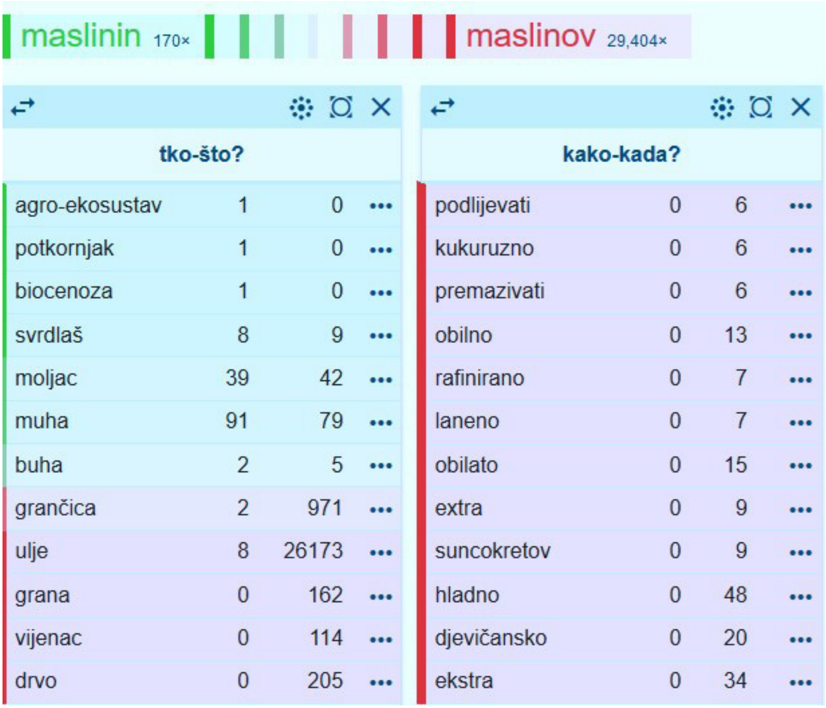
	Definition		Collocates	
dobar ('good')	Dobar je koji ima pozitivne osobine ili poželjna svojstva.	Good is that which has positive characteristics.	čovjek, igrač, odnos, prijatelj, stvar, vino, vrijeme	person, player, relationship, friend, thing, wine, time
	Dobar je koji je onakav kakav treba biti, koji ispunjava očekivanja.	Good is that which is as it should be, which fulfils expectations	dan, film, igra, momčad, rezultat	day, movie, game, team, result
	Dobar je koji obavještava o nečemu dobromu ili najavljuje dobro.	Good is that which reports on something good or predicts that something good will happen.	najava, vijest, znak	announcement, news, sign
	Dobar je koji je ispravno utemeljen i logičan.	Good is that which is not correctly founded or logical.	ideja, izbor, način, primjer, rješenje	idea, choice, way, example, solution
	Dobar je koji ne donosi korist, koji ima rezultate.	Good is that which does not bring profit or results.	posao, praksa, suradnja	work, practice, collaboration

Collocations led to the identification of new subentries as yet unrecorded in Croatian dictionaries, e.g. *ljubavni trokut, ljubavni četverokut* ('love triangle', 'love rectangle'). Collocations also motivated the lexicographers to introduce new meanings as yet unrecorded in Croatian dictionaries, e.g. two meanings of *phonology* in Table 16. A similar distinction was made in the meanings of *morfologija* ('morphology')*, sintaksa* ('syntax')*, tvorba riječi* ('word formation'), etc.

Table 16: *Collocates for two meanings of fonologija ('phonology')*

Definition		Collocates	
Fonologija je grana gramatike koja proučava glasove kao razlikovne jezične jedinice	Phonology is a branch of grammar concerned with sounds as distinctive units.	dijakronijska, generativna, opća, povijesna	diachronic, generative, general, historical
Fonologija je sustav glasova kao razlikovnih jezičnih jedinica i njihovih međuodnosa.	Phonology is the system of sounds as distinctive units and their interrelations.	čakavska, praslavenska, štokavska	Čakavian, proto-Slavic, Štokavian

Collocations sometimes helped differentiate between meanings of similar words, e.g. the adjectives *maslinin* and *maslinov*. Both of these adjectives are derived from the noun *maslina* ('olive'), have approximately the same meaning *koji se odnosi na maslinu* 'relating to an olive', and are considered synonyms. However, the Word Sketch Difference in Figure 8 shows that most of the collocates of these two adjectives differ.



tko-što?			
agro-ekosustav	1	0	•••
potkornjak	1	0	•••
biocenoza	1	0	•••
svrdlaš	8	9	•••
moljac	39	42	•••
muha	91	79	•••
buha	2	5	•••
grančica	2	971	•••
ulje	8	26173	•••
grana	0	162	•••
vijenac	0	114	•••
drvo	0	205	•••

kako-kada?			
podlijevati	0	6	•••
kukuruzno	0	6	•••
premazivati	0	6	•••
obilno	0	13	•••
rafinirano	0	7	•••
laneno	0	7	•••
obilato	0	15	•••
extra	0	9	•••
suncokretov	0	9	•••
hladno	0	48	•••
djevičansko	0	20	•••
ekstra	0	34	•••

Figure 8: Partial Word Sketch Difference for *maslinin* and *maslinov*.

The adjective *maslinin* (170 occurrences in the corpus) mostly occurs with nouns denoting a parasite: *potkornjak* ('bark beetle')*, svrdlaš* ('borer')*, moljac* ('moth')*, muha* ('fly')*, buha* ('flea'), or with those denoting biological terms *agroekosustav* ('agroecosystem'), *biocenoza* ('biocenosis'). On the other hand, the adjective *maslinov* (29,404 occurrences in the corpus) occurs with nouns denoting parts of the plant, e.g. *grančica* ('twig'), *grana* ('branch'), *drvo* ('tree'), or products made from the plant, e.g. *ulje* ('ulje'), *vijenac* ('wreath'). This resulted in different definitions for these adjectives as shown in Table 17.

Table 17: *Meanings of the adjectives maslinin and maslinov*

Headword	Definition	Collocations	Definition	Collocations
maslinin	Maslinin je koji se odnosi na maslinu.	agroekosustav, biocenoza; potkornjak, svrdlaš, moljac, muha, buha	*Maslinin* is that which relates to olives.	agroecosystem, biocenosis; bark beetle, curculio, moth, flea, fly
maslinov	Maslinov je koji je napravljen od masline.	ulje, vijenac	*Maslinov* is that which is made from olives .	oil, wreath
	Maslinov je koji je dio masline (stabla)	grančica, grana, drvo, list	*Maslinin* is part of an olive tree.	twig, branch, tree, leaf

Similar difference in collocations and meanings can be inferred from the Word Sketch Differences for the adjectival pairs *trešnjin/trešnjev* (adjectives derived from *trešnja* ‘cherry’), *višnjin/višnjev* (adjectives derived from *višnja* ‘sour cherry’), etc.

3 CONCLUSION

Mrežnik is the first normative born-digital corpus-based dictionary of standard Croatian. It is based on the two existing Croatian corpora, the *Croatian Web Repository* and the *Croatian Web Corpus*, neither of which are representative of the Croatian standard language. This is why other available print and web sources are sometimes consulted[28] and why the approach in the dictionary is corpus-based instead of corpus-driven. This also means that no statistical threshold could be used. For practical lexicographic reasons, multiword expressions in *Mrežnik* are presented in three categories: in subentries, in the collocational block, and in the idiom block. Due to this structure, collocations are defined in a broader sense and include MWEs of grammatical words and proper names, i.e. all relevant data provided by Word Sketch that is not included in a subentry or the idiom block was included in the collocational block.

Each word class, with the exception of prepositions, exhibits different collocational relations and has different collocational questions and phrases. Coordination is the one collocational relation that has the widest range and appears

[28] This is especially true for rare words and neologisms not recorded in the corpora, e.g. *koronavirus* (Coronavirus).

in all word classes that display collocational relations. In terms of word classes, verbs show the widest and most complex range of collocational relations.

In dealing with the collocational block in *Mrežnik*, the editors had to answer the following questions: Which collocational questions and introductory phrases should be included for each class or subclass of words?; Which collocations should be included in *Mrežnik*?; When should stylistic labels be included in the collocational block?; When should terminological labels be included in the collocational block?

The analysis of collocations from Word Sketches motivated the lexicographers to form pragmatic and normative notes, which can be helpful to users. This analysis also helped differentiate between meanings or quasi-synonyms, and contributed to the inclusion of new meanings not yet recorded in Croatian dictionaries. The research conducted for the *Mrežnik* project also confirms Michael Rundell's statement: "A high percentage of useful collocations occur in one of four key grammatical relations" (Rundell, 2010). Table 18 contains the four most typical syntactic structures of collocations in *Mrežnik*.

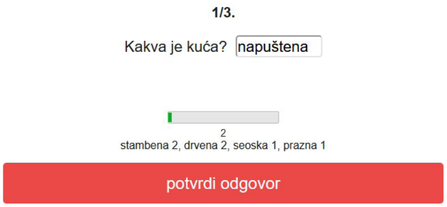
Table 18: *Typical syntactic structures of collocations in Mrežnik*

verb + noun	**maknuti:** posudu, nogu	**move:** a bowl, a leg
adjective + noun	**djevojka:** mlada, slobodna	**girl:** young, single
adverb + verb	**maknuti:** hitno, zauvijek	**remove**: urgently, forever
adverb + adjective	**mali** (comparative *manji*): znatno, jako	**small** (comparative *smaller*): quite, very

Collocations also present a challenge for the gamification of *Mrežnik* (Cf. Mihaljević, 2019a; 2019b), which is in progress at the moment.[29] Games for learning collocations and their relations to different meanings are still in the development phase. The idea is to associate different possible collocates (taken from Word Sketch) to different meanings of a word (taking definitions from *Mrežnik*, e.g. definitions of *kuća* 'house') or to different (similar) words (e.g. *maslinin/maslinov*). Another game provides the collocational question for a

[29] Many educational games for children, non-native speakers, and native speakers have been developed. They mostly focus on orthography, morphology, syntax, and on the lexical level. There are also some games for learning special and old alphabets and for learning idioms. Many language games are available at *Hrvatski u igri*.

word from *Mrežnik* and asks players to find some frequent collocates. A sample of the collocational game is shown in Figure 9.



1/3.

Kakva je kuća? napuštena

2

stambena 2, drvena 2, seoska 1, prazna 1

Figure 9: A collocational game (*Kakva je kuća?* 'What is a house like?').

Hopefully, the model used in *Mrežnik* can be useful for other born-digital dictionaries of Croatian and other (Slavic) languages, especially those that do not yet have a born-digital dictionary and a representative corpus of the national (standard) language.

Acknowledgments

This paper was written as part of the research project Croatian Web Dictionary – Mrežnik (IP-2016-06-2141) financed by the Croatian Science Foundation.

REFERENCES

Dictionaries, databases and digital resources

Croatian Collocation Database. Retrieved from http://ihjj.hr/kolokacije/english (1. 2. 2020.)

Croatian Collocation Database. Retrieved from http://ihjj.hr/kolokacije (8. 2. 2020)

Croatian Special Field Terminology – Struna. Retrieved from http://struna.ihjj.hr/en (30. 8. 2019)

Croatian Web Corpus – hrWaC. Retrieved from http://nlp.ffzg.hr/resources/corpora/hrwac/)

Croatian Web Repository Online Corpus. Retrieved from http://riznica.ihjj.hr/index.hr.html

eLexiko. Retrieved from www.owid.de/docs/elex/start.jsp/

Frazemi. Hrvatski u školi. http://hrvatski.hr/frazemi/

Hrvatska školska gramatika. http://gramatika.hr/

Hrvatski jezik. https://hrcak.srce.hr/hrjezik/

Hrvatski u igri. http://hrvatski.hr/igre/

McIntosh, C. (Ed.). (2018). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Sketch Engine Guide. Retrieved from https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/

Other

Altenberg, B., & Granger, S. (1996). Recent trends in cross-linguistic lexical studies. In B. Altenberg & S. Granger (Eds.), Lexis in Contrast. Corpus-based approaches (pp. 3–50). Amsterdam: John Benjamins Publishing Company.

Atkins, B. T. S., & Rundell, M. (2008). The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Birtić, M., Brač, I., & Runjaić, S. (2017). The Main Features of the e-Glava Online Valency Dictionary. In I. Kosem et al. (Eds.), Electronic lexicography in the 21st century. Proceedings of eLex 2017 Conference, 19–21 September, 2017, Leiden, the Netherlands (pp. 43–62). Brno: Lexical Computing CZ s.r.o.

Blagus Bartolec, G. (2014). *Riječi i njihovi susjedi: Kolokacijske sveze u hrvatskom jeziku*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Blagus Bartolec, G. (2017). Glagolske kolokacije u administrativnome funkcionalnom stilu. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 43*(2), 285–309.

Brač, I., & Bošnjak Botica, T. (2015). Semantička razdioba glagola u bazi hrvatskih glagolskih valencija. *Fluminensia, 27*(1), 105–120.

Durkin, P. (Ed.). (2016). *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in linguistic analysis,* 1–32.

Granger, S., & Paquot, M. (2012). *Electronic Lexicography*. Oxford: Oxford University Press.

Haß, U. (Ed.). (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. (Schriften des Instituts für Deutsche Sprache). Berlin/New York: de Gruyter.

Hudeček, L., & Mihaljević, M. (2017a). A New Project – Croatian Web Dictionary MREŽNIK. In I. Atanassova et al. (Eds.), *The Future of Information Sciences. INFuture2017, Integrating ICT in Society* (pp. 205–213). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences.

Hudeček, L., & Mihaljević, M. (2017b). Hrvatski mrežni rječnik – Mrežnik. *Hrvatski jezik, 4*(4), 1–7.

Hudeček, L. (2018). Izazovi leksikografske obrade u jednojezičnome mrežnom rječniku (na primjeru *Hrvatskoga mrežnog rječnika – Mrežnika*). In T. Salyha (Ed.), *Visnyk of Lviv University: Series Philology, 69*, 29–38.

Hudeček, L., & Mihaljević, M. (2018a). Croatian Web Dictionary Mrežnik: One year later – What is different? In D. Fišer & A. Pančur (Eds.), *Proceedings of the Conference on Language Technologies & Digital Humanities,* Ljubljana (pp. 106–113).

Hudeček, L., & Mihaljević, M. (2018b). *Hrvatski mrežni rječnik – Mrežnik: Upute za obrađivače.* Retrieved from: http://ihjj.hr/mreznik/uploads/upute.pdf (27. 10. 2019)

Hudeček, L., & Mihaljević, M. (2019a). Croatian Web Dictionary – Mrežnik – Linking with Other Language Resources. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 Conference* (pp. 72–98). Leiden: Lexical Computing CZ s.r.o.

Hudeček, L. (2020). Administrativizmi u rječniku (na primjeru Hrvatskoga mrežnog rječnika Mrežnika). In M. Glušac (Ed.), *Zbornik radova sa znanstvenoga skupa Od norme do uporab 2* (pp. 53 –76). Osijek – Zagreb: Filozofski fakultet Sveučilišta Josipa Jurja Strossmayera u Osijeku – Hrvatska sveučilišna naklada.

Kilgarriff, A., & Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications – a Case Study. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the 10th EURALEX International Congress* (pp. 807–818). Copenhagen: University of Copenhagen.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). Lorient: Universite de Bretagne – sud.

Klosa, A. (2015). Wortgruppenartikel in elexiko: Einneuer Artikeltyp im Onlinewörterbuch. *Sprachreport Jg, 31*(4), 34–41.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 989–997). Ljubljana: Ljubljana University Press. Retrieved from https://euralex.org/publications/collocations-dictionary-of-modern-slovene/ (8. 2. 2020)

Mihaljević, J. (2019a). Gamification in E-Lexicography. In P. Bago et al. (Eds.), INFuture 2019: Knowledge in the Digital Age (pp. 155–164). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences.

Mihaljević, J. (2019b). Games for Learning Old and Special Alphabets – The Case Study of Gamifying Mrežnik. In R. Bernardi et al. (Eds.), CLiC-it 2019: Italian Conference on Computational Linguistics. Bari: AILC. Retrieved from http://ceur-ws.org/Vol-2481/paper49.pdf (27. 4. 2020)

Mihaljević, M. (2018). Hrvatski mrežni izvori za djecu i strance. In T. Salyha (Ed.), *Visnyk of Lviv University: Series Philology* (69, pp. 75–89). doi: 10.30970/vpl.2018.69.9298

Ordulj, A. (2018). *Kolokacije u hrvatskom kao inom jeziku*. Zagreb: Hrvatska sveučilišna naklada.

Rundell, M. (2010). *Macmillan Collocations Dictionary: from start to finish*. Retrieved from http://www.macmillandictionaries.com/MED-Magazine/October2010/59-MCD-start-to-finish.htm (27. 4. 2020)

Sinclair, J. (2002). Intuition and annotation – the discussion continues. In K. Aijmer & B. Altenberg (Eds.), *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* (pp. 40–59). Göteborg.

Sinclair, J. M. (2004). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Storjohann, P. (2005). elexiko: A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics, 34,* 55–82.

Vrgoč, D., & Mihaljević, M. (2019). Jesmo li svjesni situacije? Terminološka raščlamba naziva *situational awareness* u vojnome kontekstu. *Strategos, 3*(1), 7–42.

KOLOKACIJE V HRVAŠKEM SPLETNEM SLOVARJU *MREŽNIK*

Cilj projekta *Hrvaški spletni slovar – Mrežnik* je izdelati brezplačni, enojezični, enostaven, hipertekstni, izhodiščno digitalno in korpusno zasnovan slovar standardnega hrvaškega jezika. V *Mrežniku* imajo kolokacije pomembno vlogo. Na začetku projekta so kolokacije in njihova predstavitev temeljile na projektu *elexiko,* kasneje pa je bil na podlagi korpusnih analiz koncept nekoliko prilagojen. V prispevku predstavimo model vključevanja kolokacij pri iztočnicah različnih besednih vrst. Hkrati izpostavimo pomembnejše tematike, povezane s kolokacijami v *Mrežniku*, kot so: metode luščenja kolokacij, vloga kolokacij pri ločevanju med pomeni in prepoznavi novih pomenov, uporaba stilnih in terminoloških oznak pri navajanju kolokacij ter odnosi med kolokacijami in normativnimi in pragmatičnimi informacijami, razlagami in podgesli.

Ključne besede: kolokacije, hrvaški jezik, e-slovar, *Mrežnik*, izvirno digitalni slovar