Croatian Web Dictionary Mrežnik: One year later - What is different?

Lana Hudeček,[*] Milica Mihaljević[†]

[*]Institute of Croatian Language and Linguistics
Republike Austrije 16, Zagreb, Croatia
lhudecek@ihjj.hr

[†]Institute of Croatian Language and Linguistics
Republike Austrije 16, Zagreb, Croatia
mmihalj@ihjj.hr

Abstract

The authors compare the lexicographic experience of compiling a paper desk dictionary *Školski rječnik hrvatskoga jezika* (2012, *School Dictionary of the Croatian Language*, *ŠR*) with the compilation of *Hrvatski mrežni rječnik – Mrežnik* (*Croatian Web Dictionary – Mrežnik*, *M*) which is now in progress. They focus on the new insights brought to the lexicographic work by the use of Sketch Engine giving examples from both dictionaries.

1. Introduction

In the Institute of Croatian Language and Linguistics, the *Croatian Web Dictionary – Mrežnik* is being compiled within the research project IP-2016-06-2141 financed by the Croatian Science Foundation. It is a four-year project and the work on the project started on 1st March 2017. The dictionary consists of three modules: the module for adult native speakers of Croatian with 10,000 entries, the module for elementary school children with 3,000 entries, and the module for foreigners with 1,000 entries.[1] The dictionary is based on two Croatian corpora: *the Croatian Web Corpus hrWaC* (http://nlp.ffzg.hr/resources/corpora/hrwac/)[2] and *the Croatian Language Repository* (CLR; riznica.ihjj.hr)[3]. The lexicographers select freely data from the corpus (the dictionary is corpus-based and not corpus-driven) as well as from other Croatian dictionaries, websites, and other resources. The dictionary is corpus-based due to the normative aspect of the dictionary and the unrepresentativeness of the two available Croatian corpora. The corpora differ in size and the source of the texts: *hrWaC* is much bigger and consists of texts mostly from newspapers, forums, blogs, etc. written in the journalistic and/or colloquial style. *CLR* is smaller and consists mostly of older texts often written in the literary style. The dictionary work is supported by Sketch Engine[4], a corpus query system used to support the analysis of the lemmas. The compilation of the dictionary is based on Word Sketches[5] specially adapted to the needs of the project, which are based on a developed Sketch Grammar[6] and the application of the GDEx[7] module for finding appropriate examples in the corpus. Some categories were added to Word Sketches while sometimes regular expressions were used, e.g. with interjections and conjunctions where Sketch Grammar didn't give adequate results, (see Table 1).

To support the preparation of the dictionary text the TLex software package, a professional software application for compiling dictionaries is used. TLex is adapted to the needs of the project as entry fields in TLex have been designed according to the dictionary entry model developed by the editors of the dictionary, i.e. the authors of the paper. An important characteristic of *Mrežnik* is a system of links within a module (synonyms, antonyms, masculine/feminine pairs, derivatives) and to other databases outside *Mrežnik,* i.e. links to repositories which will be created as a part of this project and compiled simultaneously with the dictionary (*Conjunction Repository*, *Repository of Idioms*, *Repository of Ethnics and Ktetics*, *Male/Female Portal*) as well as with repositories which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics. The result of the *Mrežnik* project will be a free, monolingual, hypertext, searchable, online dictionary of standard Croatian. The focus of this paper will be on the module for adult native speakers.[8]

[1] More about the project and the structure of the three modules see in Hudeček and Mihaljević, (2017a, 2017b, 2017c).

[2] *hrWaC* is a Croatian web corpus made up of texts collected from the Internet, i.e. from the .hr top-level domain. The corpus was created in January 2014 with the total size over 1.2 billion words. The current version of the corpus (v2.0) contains 1.9 billion tokens and is annotated with the lemma, morphosyntax, and dependency syntax layers.

[3] *Croatian Language Repository* is a project which started in 2005, and the corpus consists of Croatian literature, non-fiction, scientific publications and university textbooks, school books, literature translated by outstanding Croatian translators, journals and newspapers, books from the pre-standardization period of Croatian language that are adapted to standard Croatian.
The *Croatian Language Repository* corpus was processed using ReLDI tagger with Word Sketches version 1.4 by Nikola Ljubešić. It has 101,782,863 tokens and 85,273,724 words.

[4] All terms used in this paper are defined in the *Glossary* on the *Mrežnik* website ihjj.hr/mreznik/page/pojmovnik. More on Sketch Engine see in Kilgarriff et al (2014: 7-36).

[5] Croatian Word Sketches give Croatian collocations categorized by grammatical relations. Croatian Word Sketches do not have a reference yet but as they are adapted from the Slovenian model. For Slovenian Word Sketches see Krek (2006).

[6] See Kilgarriff et al (2010: 372–379).

[7] See Kilgarriff et al (2008).

[8] More about the module for children see in Hudeček and Mihaljević (in print) and more about the module for foreigners see in Hudeček et al (2018).

Categories added to Word Sketches					Regular expressions used to search the corpus	
Verb (V)	Preposition (Pr)	Pronoun (P)	Adjective (Adj)	Adverb (Adv)	Interjection (I)	Conjunction (C)
V + *se*, *se* + V	Pr + N, Pr + Adv, V + Pr, N + Pr, Adj + Pr	P + Par, Par + P	Par + Adj	Par + Adv, Adv + Par	Adv + I, V + I, I + Pr, I + N, N + I, I + *i*, I + I, I + P, Par + I	Adv + C, Par + C, C + Adv, C + C

(N = noun, Par = particle)

Table 1: Addition to Word Sketches by categories

2. The Road from *ŠR* to *Mrežnik*

The project team of *Mrežnik* consists of experienced lexicographers most of whom have already compiled the *School Dictionary of the Croatian Language* (*ŠR*), published in 2012, a normative dictionary consisting of 30,000 entries. It is based on a corpus of elementary and high school textbooks from which the lexicographers manually composed the alphabetical list of entries. It was written consulting *CLR*, i.e. all entry words were checked in the corpus for examples and collocations but examples and collocations were not taken from the corpus. It was written in the Softlex dictionary compilation program. Our expectations were that the work on *Mrežnik* would be similar to the work on *ŠR* and that its modified methodology and maybe even some definitions could be used in *Mrežnik*. However, starting work with *Word Sketches* and taking over examples from the corpus gave us new insights into lexicographic work.

2.1. Comparing *ŠR* and *Mrežnik*

The difference between *ŠR* and *Mrežnik* can easily be shown by comparing a prototype entry in both dictionaries. This is the structure of the entry *nastavnik* (teacher) in *ŠR*:

ACCENTUATED HEADWORD nástāvnīk
WORD CLASS *im. m.*
SELECTED ACCENTUATED FORMS OF THE HEADWORD <G nástāvnīka, V nástāvnīče; *mn.* N nástāvnīci, G nástāvnīkā>
DEFINITION osoba koja vodi nastavu
COLLOCATIONS COMPOSED BY THE LEXICOGRAPHER [~ *hrvatskog jezika*; *sveučilišni* ~]

This is the structure of the entry *nastavnik* in *Mrežnik*:

HEADWORD nastavnik
ACCENTUATED HEADWORD nástāvnīk
WORD CLASS *im. m.*
ALL ACCENTUATED FORMS OF THE HEADWORD
(GA nástāvnīka, DL nástāvnīku, V nástāvnīče, I nástāvnīkom; *mn.* NV nástāvnīci, G nástāvnīkā, DLI nástāvnīcima, A nástāvnīke)

1. MEANING – DEFINITION Nastavnik je odrasla osoba bez obzira na spol ili muškarac koji vodi nastavu u srednjoj školi ili na fakultetu.[9]
EXAMPLES FROM THE CORPUS *U tijeku ponavljanja godine studija studenti su se dužni uključiti u izvođenje nastavnog procesa za one predmete za koje im uredno pohađanje nastave nije potvrđeno potpisom predmetnog nastavnika, te trebaju uredno izvršavati nastavne obveze.*
Kako bi se solidarizirali s kolegama u Sindikatu učitelja i srednjoškolskih nastavnika, Ribićev sindikat spreman je odreći se isplate prosvjetnog dodatka od 2,3 posto čim odluka o ukidanju koeficijenata 3, 5, 7 i 9 stupi na snagu.
Profesionalni put Anđelko Klobučar započeo je kao srednjoškolski nastavnik u Zagrebu, a od 1958. do 1963. godine bio je glazbeni suradnik Dubrava-filma, skladajući mnogobrojne glazbe za igrane, kratke igrane, crtane i dokumentarne filmove, od kojih su mnoga od njegovih ostvarenja postala antologijskim primjerima žanra.
COLLOCATIONS FROM WORD SKETCHES INTRODUCED BY QUESTIONS
Kakav je nastavnik? (What is the teacher like?) dežurni, gostujući, honorarni, predmetni, srednjoškolski, strukovni *razg.*, sveučilišni, visokoškolski
Što nastavnik može? (What can the teacher do?) osmišljavati (program, radionicu), predavati, pripremati (nastavu, predavanja), sudjelovati (u projektu, u radu), voditi (aktivnosti, program, radionice)
Što se s nastavnikom može? (What can we do with the teacher?) javiti mu (da je učenik bolestan, da učenik ima problem), olakšati mu rad, omogućiti mu (napredovanje, praćenje, rad), preporučiti ga (za napredovanje, za zaposlenje), zaposliti ga
Čega je tko nastavnik? (What does the teacher teach?) engleskoga, fizike, informatike, hrvatskoga, matematike
Koordinacija (Coordination)*:* nastavnik i mentor, pedagog i nastavnik, profesor i nastavnik, roditelj i nastavnik; nastavnici i nenastavno osoblje, nastavnici i ravnatelj; odnosi se samo na muškarca: nastavnica i nastavnik

[9] Using male forms for both male and female and only for male is a common phenomenon in Croatian. This is stressed in all definitions in the module for adult users (odrasla osoba bez obzira na spol ili muškarac – an adult regardless of his sex or a male). As such a definition would be too complicated for school children in the module for schoolchildren this is explained in a special note. More about this see in Mihaljević (in print).

Povezuje se s (Is often connected with): dežurstvom, edukacijom, izobrazbom, kompetencijom, mobilnošću, obrazovanjem, osposobljavanjem, plaćama, usavršavanjem

SYNONYM FOR THE 1ST MEANING (LINK TO THE ENTRY) SINONIM: professor 4

STYLISTIC LABEL (COLLOQUIAL STYLE) [2] *razg*.

2. MEANING – DEFINITION Nastavnik je odrasla osoba bez obzira na spol ili muškarac koji vodi nastavu u višim razredima osnovne škole.[10]

EXAMPLES FROM THE CORPUS: *S učenicima će raditi mladi voditelji iz Saveza, daroviti informatičari, studenti i srednjoškolci koji su na natjecanjima postizali najbolje rezultate, kao i nastavnici informatike u osnovnim školama.*

Osnovnoškolski nastavnik (49) iz Slavonskoga Broda, osumnjičen za spolnu zloporabu djeteta mlađeg od 15 godina i iskorištavanje djece za pornografiju, pušten je nakon dva tjedna iz pritvora.

COLLOCATIONS FROM WORD SKETCHES INTRODUCED BY QUESTIONS

Kakav je nastavnik? dežurni, osnovnoškolski

Što nastavnik može? osmišljavati (program, radionicu), predavati, pripremati (nastavu, predavanja), sudjelovati (u projektu, u radu), voditi (aktivnosti, program, radionice)

Što se s nastavnikom može? javiti mu (da će učenik izostati s nastave), olakšati mu (posao, rad), omogućiti mu (napredovanje, praćenje, rad), preporučiti ga za što, zaposliti ga

Čega je tko nastavnik? engleskoga, fizike, informatike, hrvatskoga, matematike

Koordinacija: odgajatelj i nastavnik, profesor i nastavnik, roditelj i nastavnik, pedagog i nastavnik, učitelj i nastavnik; odnosi se samo na muškarca: nastavnica i nastavnik

SYNONYMS FOR THE 2ND MEANING (LINKS TO THE ENTRIES): SINONIMI: učitelj (predmetni učitelj 1), **profesor 4**

3. MEANING – DEFINITION [3] Nastavnik je odrasla osoba bez obzira na spol ili muškarac koji komu prenosi kakva znanja ili ga poučava kakvim vještinama.

EXAMPLES FROM THE CORPUS *Rukovoditelj letenja mora biti punoljetan, mora imati letačko iskustvo od najmanje 30 sati letenja i 30 uzlijetanja i slijetanja kao zapovjednik jedrilice nakon izdavanja dozvole pilota jedrilice, a kada lete učenici piloti mora imati važeće ovlaštenje nastavnika letenja.*

Rukovoditelj tehničkih poslova i nastavnik padobranstva u Aeroklubu Borovo Branislav Mišić izjavio je da se mladim padobrancima u prvih 10 - tak skokova padobran otvara automatski te da je takav slučaj trebao biti i s poginulim mladićem.

COLLOCATIONS FROM WORD SKETCHES INTRODUCED BY A QUESTION *Čega je tko nastavnik?* crtanja, gitare, kuhanja, letenja, padobranstva

SYNONYMS FOR THE 3RD MEANING (LINKS TO THE ENTRIES): SINONIM: učitelj 3

SUBENTRY strukovni nastavnik

DEFINITION Strukovni nastavnik organizator je i voditelj strukovno-teorijske nastave te praktične nastave i vježba u strukovnim školama; svoje je temeljno obrazovanje stekao na nekome od nepedagoških fakulteta, a pedagoške kompetencije stekao je dopunskim pedagoško-psihološkim i didaktičko-metodičkim obrazovanjem.

EXAMPLES FROM THE CORPUS *Agencija za strukovno obrazovanje i obrazovanje odraslih, od 26. do 28. ožujka 2018. godine u Opatiji, organizira „Dane strukovnih nastavnika".*

Može se zaključiti da je neophodno osigurati odgovarajuću izobrazbu i kontinuirano usavršavanje strukovnih nastavnika u pedagoško-didaktičkom, ali i strukovnom području.

FEMALE PAIR OF SUBENTRY (LINK TO THE ENTRY) ŽENSKO nastavnica (strukovna nastavnica), učiteljica (strukovna učiteljica 2)

SYNONYM FOR THE MEANING OF SUBENTRY (LINK TO THE ENTRY) SINONIM učitelj (strukovni učitelj)

PRAGMATIC NOTE[11] Riječi *učitelj*, *nastavnik* i *profesor* drukčije su određene u zakonu danas (*Zakon o odgoju i obrazovanju u osnovnoj i srednjoj školi* 2008.) nego što je to bilo prije, pa to dovodi do nedosljedne uporabe tih riječi. Danas prema Zakonu u osnovnoj školi rade učitelji, a u srednjoj i na fakultetu nastavnici. Značenje se tih riječi usklađeno sa Zakonom dosljedno nalazi npr. u natječajima za posao te na mrežnim stranicama škola. Međutim, značenje je tih riječi često drukčije upotrebljava u publicističkome i razgovornome stilu, pa se često govori i o osnovnoškolskim nastavnicima/profesorima i srednjoškolskim profesorima. U srednjim strukovnim školama rade strukovni učitelji, koji se često zovu i strukovni nastavnici. Zbrku povećava i to što učitelji mogu napredovati u zvanje učitelja mentora i učitelja savjetnika, a nastavnici u zvanje profesora mentora i profesora savjetnika. Na fakultetu rade sveučilišni nastavnici, koji mogu imati znanstveno-nastavno i umjetničko-nastavno zvanje docenta, izvanrednoga profesora i redovitoga profesora ili nastavno zvanje predavača, višega predavača ili profesora visoke škole. Dakle, profesori su redoviti profesori, izvanredni profesori i profesori visoke škole. Od radnih se mjesta i znanstveno-nastavnih i umjetničko-nastavnih zvanja razlikuju titule koje se dobivaju završetkom određenoga

[10] We decided to separate meanings 1 and 2 although the definitions differ only in emphasizing different levels of education for pragmatic reasons and the difference in synonyms. The first meaning is the official term used in documents (laws, certificates, diplomas, etc.), and the second is a colloquial and historic term which doesn't occur in contemporary official documents. It is synonymous to the official term *učitelj* (or *razredni učitelj*).

[11] In the pragmatic note the difference between the usage of the words *učitelj*, *nastavnik*, and *profesor* are explained as these words have similar meanings which vary according to the formality of style and are used differently e.g. in legal documents and newspapers. Also, they are used differently now than they were ten years ago so the meaning of these words also varies according to the time when the text was written.

fakulteta. Onaj tko završi učiteljski fakultet, dobiva titulu diplomirani učitelj/diplomirana učiteljica. Onaj tko završi nastavnički fakultet po bolonjskome procesu, dobiva titulu mag. edu., odnosno magistar edukacije, ali onaj tko je diplomirao prije uvođenja bolonjskoga procesa, dobivao je titulu profesora, pa danas u školama radi još mnogo profesora. Učenici osnovne škole danas se najčešće svojim učiteljima obraćaju riječju *učitelj*, a učenici srednje škole i studenti svojim se nastavnicima obraćaju riječju *profesor*. Odnos među riječima *učitelj*, *nastavnik* i *profesor* odražava se i na odnos među riječima *učiteljica*, *nastavnica* i *profesorica*.

By comparing these two entries we can conclude:

1. the entries in *Mrežnik* are much longer

2. while the entry for the headword *nastavnik* in *ŠR* has only these fields: word class, selected grammatical forms, definition, two collocations, the structure of the entry for the same headword in *Mrežnik* is as follows: word class, all grammatical forms, three different definition with separate examples and collocations (introduced by questions or phrases *What is xx like?*, *What can xx do?*, *What can we do with xx?*, *What does xx teach?*, *Coordination:*, *Often connected with:*), some collocations having a stylistic label (*razg*. – colloquial), subentry with examples, synonyms connected with particular meanings (links), male/female pairs connected with particular meanings (links), pragmatic note, word formation information and derivatives. If the derivatives are separate entries in the dictionary, they are connected via hyperlink to their entry and if not, they are only listed.

There are two answers to the question why the structure of the same entry in these two dictionaries differs so much:

1. As *Mrežnik* is a web dictionary more data than in the printed dictionary could be given: all accentuated noun forms, information on the male/female pair (as we have experience that this is the data that users require very often[12]), word-formation information and derivatives (as this is the data very often required by students), questions for collocations based on the model used in elexiko[13] and a pragmatic note.

2. Using Word Sketches we found new meanings not recorded in Croatian dictionaries. Every new meaning has examples from the corpus, which led us to the awareness that often there are useful pragmatic comments which could be inserted into the dictionary. In the example above while using Word Sketches and corpus examples we became aware of the complex net of relations between the two sets of words: *učitelj*, *nastavnik*, *professor* (male teacher) and *učiteljica*, *nastavnica*, *profesorica* (female teacher) as these terms are used differently (and sometimes inconsistently even in the same document) in legal documents, school practice, newspapers, and everyday speech. We interviewed a few schoolteachers to get a clearer view of how these terms are used in schools. For this reason, a short pragmatic note was introduced into all six of the above-mentioned entries.

The relation between male and female pairs is complicated by the fact that the male form can mean *male* (in the dictionary definitions this is expressed by *muškarac* if it applies only to adults or *muška osoba* if it applies to male children as well) but also male and female (especially in the plural) regardless of gender. This is expressed by the formula *osoba bez obzira na spol* or *muška osoba bez obzira na spol* (person regardless of gender or adult person regardless of gender).

2.2. Collocations and examples

As collocations and examples are extracted from the corpus this brings us to another difference between *ŠR* and *Mrežnik*. This will be shown on the example of the feminative of the entry *nastavnik* (teacher) *nastavnica* (female teacher).

Comparing these two entries we can see that in *ŠR* the entry for the word *nastavnik* closely corresponds to the entry for the word *nastavnica* and once having compiled the entry *nastavnik* we could compile *nastavnica* in a matter of minutes. On the other hand, in *Mrežnik* the entry *nastavnica* corresponds to the entry *nastavnik* in definitions and questions asked for collocations but differs considerably in most of the examples and collocations (see Table 2).

This, however, brings us to the new problem of how closely our collocations should correspond to that what we get from Word Sketches. So far we have noticed two major problems:

1. Collocations for male and female agent nouns differ considerably and in a way that shouldn't (because the dictionary has an educational purpose and not only a scientific one) be reflected in a dictionary of the standard language. This will be illustrated by comparing the collocations for *konobar* (waiter) and *konobarica* (waitress) and *pekar* and *pekarica* (male and female baker). If we look for the lemma *konobarica* in Sketch Engine the first collocation is *brkata* (with a moustache) followed by *sisata*, *prsata* (having big breasts), and the first verb having *konobarica* as an object is an impolite verb *zajebavati*. On the other hand, if we check the corresponding male word *konobar* the collocations are quite different: *neljubazan* (impolite), *ljubazan* (polite), *pomoćni* (assistant), *priučeni* (trained on the job, inadequately trained), and the verbs having the highest score are *zamoliti*/*moliti* (ask), *dozvati*/*zovnuti* (call). With the female noun *pekarica* (baker) among the top collocations are *fatalna* (fatal) and *pohotna* (lusty). Even with the word *čistačica* (cleaning lady) *seksi* (sexy) has a very high score. No such adjectives occur with the masculine words *pekar* and *čistač*.

[12] We give language advice on a daily basis by telephone and e-mail and questions about female/male pairs occur very often. That is the reason we have launched a new project *Male and female in Croatian*. More on the project see on ihjj.hr/projekt/musko-i-zensko-u-hrvatskome-jeziku.

[13] See Klosa (ed.) (2011) and Möhrs (2014).

ŠR	**nástāvnica** *im. ž.* <G nástāvnicē; *mn.* N nástāvnice, G nástāvnīcā> žena koja vodi nastavu [~ *hrvatskog jezika*; *sveučilišna* ~]
M	nástāvnica **nastavnica** im. ž. (G nástāvnicē, DL nástāvnici, A nástāvnicu, V nástāvnice, I nástāvnicōm; *mn.* NAV nástāvnice, G nástāvnīcā, DLI nástāvnicama) **[1]** Nastavnica je žena koja vodi nastavu u srednjoj školi ili na fakultetu . *Nastavnica biologije u istoj srednjoj školi Dolores Dobrinkić na posao putuje triput tjedno iz Splita, za što je u siječnju potrošila dvije trećine plaće. Članica ste pokreta »Opus Dei«, što nekako sugerira, s obzirom i na Vaš posao sveučilišne nastavnice, angažman u javnosti.* *Kakva je nastavnica?* mlada, omražena, omiljena, predmetna, srednjoškolska, stroga, sveučilišna, umirovljena, x-godišnja; *Što nastavnica može?* organizirati (nastavu, predavanja, radionice), osmišljavati (program, radionicu), predavati, pripremati (nastavu, predavanja), sudjelovati (u projektu, u radu), voditi (aktivnosti, program, radionice); *Što se s nastavnicom može?* napadati je se, pitati je se što, pozdravljati je; *Čega je tko nastavnica?* biologije, engleskoga, fizike, informatike, hrvatskoga, matematike , povijesti, vjeronauka; *Koordinacija:* mentorica i nastavnica, nastavnica i nastavnik, profesorica i nastavnica, razrednica i nastavnica (čega), ravnateljica i nastavnica, učenice i nastavnice, učiteljice i nastavnice **MUŠKO: nastavnik :1** **SINONIM: profesorica** **[2]** razg. Nastavnica je žena koja vodi nastavu u višim razredima osnovne škole . *Osnovnoškolska nastavnica koja je bez ikakvog jasnog razloga izbacivala s nastave dijete koje je štićenik doma za nezbrinutu djecu, odbijajući priznati da je učenik škole, dobila je zabranu rada od Prosvjetne inspekcije. Pretvorili smo se u tvornice 'biflanja', a mi nastavnici pretvoreni smo u birokrate - istaknula je osnovnoškolska nastavnica Ivana Kovač.* *Kakva je nastavnica?* dežurna, osnovnoškolska; *Što nastavnica može?* organizira (nastavu, predavanja, radionice), osmišljava (program, radionicu), predaje, priprema (nastavu, predavanja), sudjeluje (u projektu, u radu), vodi (aktivnosti, program, radionice); *Što se s nastavnicom može?* napadati je se, pitati je se, pozdravljati je se; *Čega je tko nastavnica?* biologije, engleskoga, fizike, informatike, hrvatskoga, matematike povijesti, vjeronauka; *Koordinacija:* nastavnice i nastavnici, odgajateljice i nastavnice, učiteljice i nastavnice **MUŠKO: nastavnik 2** **SINONIM: učiteljica predmetna učiteljica :1** **[3]** Nastavnica je žena koja komu prenosi kakva znanja ili ga poučava kakvim vještinama. *Pa zbog velikog broja polaznika tečajeva gitare angažirali smo još jednu kolegicu kao drugu nastavnicu gitare.* *Čega je tko nastavnica?* gitare, klavira, pjevanja **MUŠKO: nastavnik 3** **SINONIM: učiteljica 2** **strukovna nastavnica** Strukovna nastavnica organizatorica je i voditeljica strukovno-teorijske nastave te praktične nastave i vježba u strukovnim školama; svoje je temeljno obrazovanje stekao na nekom od nepedagoških fakulteta, a pedagoške kompetencije stekao je dopunskim pedagoško-psihološko i didaktičko metodičkim obrazovanjem. **MUŠKO: nastavnik strukovni nastavnik :1** **SINONIM: učiteljica (strukovna učiteljica 1)** Riječi *učitelj*, *nastavnik* i *profesor* drukčije su određene u zakonu danas…

Table 2: Entries *nastavnica* (female teacher) in *ŠR* and *Mrežnik* (M)

2. As both *hrWaC* and *Riznica* have many newspaper examples, very often collocations reflect news reporting on murder, rape, drugs, etc. not really characteristic for a certain word. This is especially common for female agent nouns, e.g. with many female agent nouns verbs *maltretirati*, *ubiti*, and *silovati* occur (mistreat, kill, rape), e.g. with the word *nastavnica* collocations *pretući*, *gađati*, *napasti* (beat, hit, attack) are the first three results in the row *koga – što*, i.e. the row in which *nastavnica* is in the accusative case. If we compare the same row for the word *djevojka* girl and the word *nastavnik* we get these results: The collocations for the word *djevojka* (girl) with the

highest score are *zaprositi*, *silovati*, *oženiti*, *upoznati*, *ubiti* (ask to marry, rape, marry, meet, kill) and the collocations for *nastavnik* are *educirati*, *obrazovati*, *osposobljavati*, *pozivati* (educate, specialize, invite). While analyzing these results we have to bear in mind the above-mentioned fact that *nastavnik* often refers to a male as well as a female person, especially in legal texts. As the dictionary is not corpus-driven but only corpus-based we avoided collocations which are not characteristic for a certain word or which can be offensive to a general dictionary user or which we feel are only the result of the unrepresentativeness of the corpus in which there are too many journalistic texts. However, such collocations will occur with the verbs *kill*, *rape*, *mistreat*, etc. as they illustrate the basic meaning of these words. The corpus gives also many uninformative collocations as *postati konobarica/konobaricom*, *raditi kao konobarica* (become a waitress, work as a waitress). They are uninformative as they can occur with any professional noun, e.g. *medicinska sestra* / *nastavnica* / *pekarica* / *profesorica* / *učiteljica*, etc.

Working with the corpus and Sketch Engine our approach to synonyms and antonyms has also changed. In *ŠR* synonymous entries resembled each other completely, i.e. they had the same example differing only in the synonymous word and they had same collocations. If a word is used only in the colloquial style it was marked with the label *razg*. and directed to the entry belonging to the neutral standard language (*v.* – see). In *Mrežnik* such words have complete entries, consisting of a definition (or definitions), which is the same as the definition of their synonym belonging to the neutral standard language, examples and collocations (from the corpus, i.e. different from that of their synonyms). The words *stomatolog* and *zubar* were selected to illustrate this point as, although the word *zubar* is often used in Croatian, it is not the official professional term. This can be seen on the page of Croatian Terminological Database *Struna*. Entries *zubar* and *stomatolog* in *Mrežnik* will be linked to the entries in *Struna*.

2.3. Synonyms and antonyms

ŠK	**stomatòlog** *im. m.* <G stomatòloga, V stomatòlože; *mn.* N stomatòlozi, G stomatòlōgā> liječnik koji se bavi stomatologijom; ***sin.***: zubar *razg.*
	zùbār *im. m.* <G zubára, V zùbāru/zùbāre; *mn.* N zubári, G zubárā> *razg. v.* stomatolog
M	stomatòlog **stomatolog** im. m. (GA stomatòloga, DL stomatòlogu, V stomatòlože, I stomatòlogom; *mn.* NAV stomatòlozi, G stomatòlōgā, DLI stomatòlozima, A stomatòloge) Stomatolog je liječnik koji se bavi stomatologijom. *Najnižu cijenu usluga određuje Stomatološka komora a potom je svaki privatni stomatolog usklađuje sa svojim mogućnostima naplate, zarade itd., što ovisi o mnogo čimbenika. U ordinaciji stomatologa dr. Živka Dijana jučer je u Zadru održana edukacijska radionica na kojoj su prezentirane najsuvremenije metode pomlađivanja lica, usana i zubnog mesa filerima Esthelis.* *Kakav je stomatolog?* budući, dežuran, dječji, estetski, izabrani, kvalitetni, nezaposlen, obiteljski, odabrani, poznati, privatni, ugovorni, vrhunski; *Što stomatolog može?* izvaditi zub/živac, preporučiti (terapiju, vađenje zuba, zubni konac); *Što se sa stomatologom može?* izabrati ga, obavijestiti ga (o lijekovima koje tko uzima, o terapiji, o trudnoći), posjetiti/posjećivati ga, preporučiti ga komu, zamoliti ga (da što preporuči/objasni); *Koordinacija:* ginekolozi i stomatolozi, liječnici i stomatolozi, pacijenti i stomatolozi, pedijatri i stomatolozi, stomatolozi i zubni tehničari; stomatolog ili specijalist (dentalne patologije, endodoncije, oralne kirurgije, paradontolog, stomatološke protetike, za plastičnu kirurgiju); *Povezuje se s*: dežurstvom, intervencijom, kongresom, kontrolom, nadzorom, pomoći, posjetom, pregledom, preporukom, savjetom, udrugom, udruženjom, uputom **ŽENSKO: stomatologica :1, stomatologinja :1, zubarica :1** **SINONIM: zubar :1**
	zùbār **zubar** im. m. (GA zubára, DL zubáru, V zùbāru/zùbāre, I zubárom/zubárem; *mn.* NV zubári, G zubárā, DLI zubárima, A zubáre) *razg*. Zubar je liječnik koji se bavi stomatologijom. *Tko želi zdrave zube i čvrste desni, mora zube prati 2 do 3 puta dnevno, svilenim koncem pročistiti prostor između zuba i redovito ići zubaru na kontrolu. U nekoliko trenutaka možete*

	postići osmijeh iz snova, a da pritom ne idete zubaru na zahvate koji koštaju i oštećuju zubnu caklinu. *Kakav je zubar?* besplatan, dežurni, izvrstan, poznati, privatni, socijalni, ugledan, uspješan, vrhunski ; *Što zubar može?* brusiti zub, liječiti desni/zub, otvoriti zub, praviti/raditi navlaku/plombu/protezu/zub, vaditi zub/živac; *Što se sa zubarom može?* bojati ga se, ići k njemu, dolaziti k njemu, imati ga, javljati mu se, nazivati ga se, plaćati mu, posjećivati ga; *Koordinacija:* kirurzi i zubari, zubar i ortodont, zubari i doktori, zubari i frizeri, zubari i ginekolozi, zubari i liječnici, zubari i okulisti; odnosi se samo na muškarce: zubar ili zubarica; *Povezuje se s:* preporuka, strah, usluga **ŽENSKO: stomatologica :1, stomatologinja :1, zubarica :1** **SINONIM: stomatolog :1**

Table 3: Synonymous entries *stomatolog* and *zubar* (dentist) in *ŠR* and *Mrežnik*

All this holds for antonyms as well. Moreover, as there is no need to save space in a web dictionary in polysemic entries synonyms and antonyms are connected to each meaning, i.e. not connected to the lemma even in cases of full synonyms/antonyms.

2.4. Corpus versus system

Working with the corpus often made us choose between the language system and data derived from the corpus. This was a common problem with lemmas which have a low frequency in the corpus. In Croatian the professional noun of masculine gender can in certain contexts (especially used in the plural) refer to persons of both sexes (or of unknown sex) and in other only to males. However, with certain entries and subentries such contexts were difficult to find (e.g. subentry *strukovni nastavnik* professional teacher, teacher in a vocational school).

On the other hand, for some entries, a corresponding female (e.g. *strukovna nastavnica*) or male noun (e.g. *pomoćnik porodničara* male midwife) could not be found in the corpus or had a very low frequency. Nevertheless, we decided to include such entries not only because they reflect the language system but also as many users of our language advice services often ask for information on the formation and usage of masculine/feminine pairs. Often native speakers of Croatian have problems in forming feminine/masculine pairs and due to the changing sociolinguistic context the need to use them occurs. As there are many problems connected with the relations between male and female nouns in Croatian a separate research project *Muško i žensko u hrvatskome jeziku* closely connected with the *Mrežnik* project is also conducted at the Institute.

Some examples in which Croatian native speakers have problems in forming a feminine or a masculine noun of a more frequently used feminine/masculine noun denoting professions are shown in the table below:

Masculine	**Feminine**	**English**
diskdžokej	diskdžokejica	disk-jockey
knjigoveža	knjigoveškinja	bookbinder
krupje	krupjeica	croupier
mornar	mornarica	sailor
ronilac	roniteljica	diver
tekstopisac	tekstopiskinja	songwriter

Table 4: Masculine/feminine pairs native speakers of Croatian find difficult

Some feminine/masculine forms are explained in a special note (advice), e.g. *čitalac > čitatelj, psihologica > psihologinja*. In *ŠR* only some of these examples were marked by the labels *v.* (see) and → (replace with) while examples which do not occur as often were not included in the dictionary.

2.5. Pleonasms and paronyms

The corpus made us aware of the fact that pleonasms occur very frequently. For example, the expression *no međutim* has 1611 occurrences in hrWaC, so the users of *Mrežnik* will be made aware of this very common mistake in a note. Another very common pleonasm has the structure *žena* (woman) + feminine form of the agent noun, e.g. *žena vozačica* (woman + driver in the feminine form). The corpus makes us aware of the fact that this is mostly used in the negative context connected with the

stereotype 'women are bad drivers'. *Mrežnik* offers the advice that instead of the pleonastic construction *žena vozačica* in a stylistically unmarked context only *vozačica* should be used.

The corpus also makes us aware of mistakes connected with the use of paronyms, e.g. words *psihički*, *psihološki*, and *psihologijski* (mental or psychic, psychological); *genski*, *genetički*, and *genetski* (referring to gens, genetics or genesis); *religijski* and *religiozan* (religious or poius, religious) are often confused. The meanings of terms *smrtni list* and *smrtovnica* (both terms can be translated as *death certificate* into English) are often confused or these terms are considered as synonymous although they refer to two different documents as *smrtovnica* is a document issued by a registrar on the basis of *smrtni list* issued by the coroner. The relations between these words are then explained in a pragmatic note.

3. Conclusion

The possibilities of linking different resources and giving different data are enormous and one should not fall into the trap of giving some information only because it is available (as one of the reviewers of the project warned us) and not because the users need it. If we receive certain questions repeatedly from the users of our language advice service or notice it is the topic of discussion on different blogs (e.g. the above-mentioned difference between *učitelj*, *nastavnik*, and *profesor*) we consider that an explanation should be given in the dictionary. A note on language usage will be edited within the dictionary while additional data (verb valency, the etymology of idioms, metaphoric extensions, terminological data, etc.) will be given as additional information on a link.

The work with collocations from Sketch Engine has after only one year broadened our lexicographic views:

1. Often after looking at Word Sketches, we have decided we need to have more than one meaning for a word that had only one meaning in *ŠR*.

2. We have concluded that although a semantic relation (synonyms, antonyms, male/female relation) exists between particular words their collocations can differ considerably. Thus, collocations are given for each meaning of the lemma separately, i.e. synonyms have the same definition but can have different collocations.

3. We couldn't rely on Word Sketches completely and had to make a selection between offered data as many examples were uninformative, not characteristic for the lemma but for the corpus from which they were taken, not polite or biased towards a particular sex, social or ethnical group, etc.

4. We have concluded that data that we got from the corpus and Word Sketches could often be useful and should be included in a pragmatic or a language advice note.

4. Acknowledgments

This paper is written within the research project *Croatian Web Dictionary – Mrežnik.* (IP-2016-06-2141), financed by the Croatian Science Foundation.

5. References

Baza hrvatskih glagolskih valencija – GLAVA. Accessed at http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija/27/ 18 January 2018.

Matea Birtić et al. 2012. *Školski rječnik hrvatskoga jezika*. Zagreb: Školska knjiga – Institut za hrvatski jezik i jezikoslovlje.

Bolje je hrvatski. Accessed at http://bolje.hr/ 18 January 2018.

Hrvatsko strukovno nazivlje – STRUNA Accesed at http://struna.ihjj.hr/ 18 January 2018).

Lana Hudeček and Milica Mihaljević. 2017a. Hrvatski mrežni rječnik – Mrežnik. *Hrvatski jezik*, 4(4):1–7.

Lana Hudeček and Milica Mihaljević. 2017b. A new project – Croatian web dictionary MREŽNIK. In *The Future of Information Sciences. INFuture 2017, Integrating ICT in Society*, pages 205-2013, Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences. Zagreb.

Lana Hudeček and Milica Mihaljević. 2017c. The Croatian Web Dictionary Project – Mrežnik. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 172–192. Lexical Computing CZ s.r.o., Brno – Leiden.

Lana Hudeček et al. 2017. Radionica na Croaticumu – provjera rječničke koncepcije modula za strance na terenu. *Hrvatski jezik*, 4/4: 9–12.

Lana Hudeček and Milica Mihaljević.2018. Normiranje hrvatskoga jezikoslovnog nazivlja. In Hrvatski prilozi 16. međunarodnom slavističkom kongresu, pages 49 – 62. Hrvatsko filološko društvo, Zagreb.

Jezični savjetnik Accessed at http://jezicni-savjetnik.hr/ 20 February 2018.

Adam Kilgarriff et al. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Adam Kilgarriff et al. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII Euralex International Congress*, pages 425–432. Universitat Pompeu Fabra. Barcelona. Institut Universitari de Linguistica Aplicada.

Adam Kilgarriff et al. 2010. A quantitative evaluation of word sketches. In *Proceedings of the XIV Euralex International Congress*, pages 372–379. Fryske Akademy. Leeuwarden.

Annette Klosa. (ed.). 2011. *elexiko. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs*. Narr. Verlag. Tübingen.

Simon Krek and Adam Kilgarriff. 2006. Slovene word sketches. In *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*. Institut Jožef Štefan. Ljubljana. http://www.kilgarriff.co.uk/Publications/2006-KrekKilg-Ljub-SloveneWS.pdf.

Milica Mihaljević. Hrvatski mrežni izvori za djecu i strance. In Zbornik *20 godina kroatistike u Lavovu*. Lavov (in print)

Christine Möhrs. 2014. Landeskundliche Wortschatzübungen auf der Basis von Kollokationen. Zur Nutzung von elexiko für Deutschlehrende. In: *Themenheft »Dateninterpretation und -präsentation in Onlinewörterbüchern am Beispiel von elexiko«*. Deutsche Sprache 4/2014, pp 309-324.

Repozitorij metafora. Accessed at http://ihjj.hr/metafore/ 20 February 2018.