Project presentation

A New Project – Croatian Web Dictionary MREŽNIK

Lana Hudeček
Institute of Croatian Language and Linguistics
Republike Austrije 16, 1000 Zagreb, Croatia
lhudecek@ihjj.hr

Milica Mihaljević
Institute of Croatian Language and Linguistics
Republike Austrije 16, 1000 Zagreb, Croatia
mmihalj@ihjj.hr

Summary

A new project Croatian Web Dictionary financed by the Croatian Science Foundation is presented. The main goal of the project is to create a monolingual corpus-based dictionary of Standard Croatian compiled in accordance with contemporary findings of computational linguistics. The authors explain the importance of such a project, its methodology, and expected results. As entries in Mrežnik will be connected with many other databases from the Institute of Croatian Language and Linguistics an overview of these databases and other resources is given.

Key words: corpus-based dictionary, Croatian language, e-lexicography, web dictionary

Introduction

Contemporary lexicography is primarily e-lexicography[1] and that will undoubtedly be its future as the range of possibilities of new digital technologies cannot be compared with the limited possibilities of printed dictionaries. Since 1990s e-lexicography has developed rapidly and this led to the appearance of the first online dictionaries (before that, dictionaries were stored on CDs and DVDs). This development was followed by research in e-lexicography and corpus linguistics. There are many online dictionaries of many languages in Europe, e.g. elexiko (http://www.owid.de/wb/elexiko/start.html) of the Institute of German Language, Wielki słownik języka polskiego (http://www.wsjp.pl/) of the Institute of Polish Language, Swedish online dictionary (http://spraakbanken.gu.se/karp), etc. It is important to note that in some Slavic countries (e.g. Poland) the

[1] More about this see in the paper Jermen, Kraus, Starčević Stančić (2015) and Štrkalj Despot, Möhrs (2015).

development of an online dictionary of the national language is considered a project of national importance.

In Poland, for instance, such a dictionary has been created in several phases starting in 2006, from 2008 with the support of the Polish Ministry of Science and Higher Education, and from 2013 to 2018 with the support of the National Programme for the Development of Humanities (Narodowy Program Razwoju Humanistyki). Among Slavic languages, Croatian is one of the few languages without a scientifically compiled online dictionary of the national language.

In Croatia, the Croatian Language Portal (HJP) exists, with an online dictionary which is the result of collaboration between Novi Liber and Srce (http://hjp.novi-liber.hr/). However, this dictionary was not compiled as an online dictionary, but an online version of the already printed dictionary published by the publishing house Novi Liber, and sold in the printed version for the last 15 years. There are some other monolingual dictionaries of Croatian language available digitally, such as Prvi školski rječnik hrvatskoga jezika (The First Dictionary of the Croatian Language) by Ankica Čilaš Šimpraga, Ljiljana Jojić, and Kristian Lewis (2008) and Veliki rječnik hrvatskoga jezik (The Big Dictionary of the Croatian Language) by Ljiljana Jojić et al. (2015), but they are not corpus-based, based on the principles of e-lexicography, normative, and publically available. Apart from the aforementioned dictionaries, the following Croatian language lexical resources are available: Wječnik (https://hr.wiktionary.org/wiki/)2, CroWN, the Croatian Wordnet (http://meta-share.ffzg.hr/repository/browse/croatian-wordnet), Meta-Net.HR (http://ihjj.hr/metafore), Struna (Croatian Special Field Terminology of Croatian Repository) (http://struna.ihjj.hr), Croatian Lexicographic Heritage Portal (http://crodip.ffzg.hr/), and Croatian Terminology Portal (http://nazivlje.hr/), which searches Struna, dictionaries, glossaries, and lexicons from the Lexicographic Institute Miroslav Krleža3. There are also many multilingual terminological databases including Croatian as one of the languages (e.g. EMITEL – e-Encyclopaedia of Medical Physics and Multilingual Dictionary of Terms, http://www.emitel2.eu, Multilingual Archival Terminology, http://www.ciscra.org/mat/mat/termlist/l/Croatian, Meta-Share, http://meta-share.ffzg.hr/repository/search/, Microsoft Terminology Collection https://www.microsoft.com/Language/en-US/Search.aspx, etc.

From this short overview, we can conclude that although there are many language resources for Croatian available online there is no monolingual corpus-based dictionary of Standard Croatian compiled in accordance with contempo-

[2] *Wječnik* is the Croatian version of *Wiktionary*. *Wiktionary* is a collaborative international project to produce a free-content multilingual dictionary. It aims to describe all words of all languages. It is designed as the lexical companion to *Wikipedia*. *Wiktionary* is a wiki, which means that everybody can edit it.

[3] More about this see in the paper Jermen, Kraus, Starčević Stančić (2015).

rary standards of computational linguistics. For the purpose of creating such a dictionary, detailed research in e-lexicography is required. The final and primary goal of this project is to develop a corpus-based Croatian Web Dictionary. The logo of the project and of the dictionary is:

Figure 1. Logo of the project

Methodology

The Croatian Web Dictionary will include dictionary entries which have accentuated entry words, grammatical definitions, and grammatical blocks, together with accentuated word forms, detailed definitions with appropriate examples for individual meanings, the most frequent collocations and idioms, synonyms and antonyms. The dictionary will be normative in nature, which is indicated by that fact that the project will include the writing of three hundred short linguistic advice entries (up to several sentences long). Entry words and collocations which are not recommended for use will be linked from the basic text of the e-dictionary to linguistics advice entries. Some very common anglicisms will also be connected with the portal *Bolje je hrvatski* (Better in Croatian) compiled also by team members in which Croatian words and phrases are suggested for some commonly used anglicisms. Three thousand dictionary entries will include definitions for elementary school children, and one thousand dictionary entries will include definitions for learners of Croatian as a foreign language. Dictionary for schoolchildren will be illustrated and definitions from that dictionary will also be used on the website *Hrvatski u školi* (Figure 2).

Figure 2. A crossword puzzle based on *Prvi školski pravopis* on the website *Hrvatski u školi*

Conjunction dictionary entries will include descriptions of conjunction groups and modifiers. Ktetic and ethnic dictionary entries will include links to the ethnic and ktetic repository, and some most common terms will be linked to the terminological database Struna (struna.ihjj.hr) also compiled at the Institute of Croatian Language and Linguistics. During the compilation of the online dictionary, two types of activities will take place: 1. activities connected with the development of computer and computational linguistic prerequisites for the compiling of an online dictionary and 2. activities connected with lexicographic data processing. The dictionary will be written using the TLex software package, a professional software application for compiling dictionaries adapted to the needs of the project (designing the entry fields according to the dictionary entry model developed by the editors of the dictionary). SketchEngine will be used to search the corpora. The dictionary will be based on these two corpora: the Croatian Web Corpus hrWaC (http://nlp.ffzg.hr/resources/corpora/hrwac/) and Croatian Language Repository (riznica.ihjj.hr). The obvious problem with the methodology of MREŽNIK is that MREŽNIK will be based on the large unbalanced corpus the Croatian Web Corpus and on Riznica (which is a smaller corpus containing many texts from literature and from older periods). However, MREŽNIK will be only corpus-based and not corpus-driven and the lexicographers will select freely data from the corpus as well as from other Croatian dictionaries. The compilation of the dictionary will be based on designing word "sketches" (WordSketches) for each corpus separately, the prerequisite of which is a developed grammar sketch (SketchGrammar), the application of the GDEx module for finding appropriate examples in the corpus, checking individual entries using a morphological lexicon (https://www.clarin.si/repository/xmlui/handle/11356/1056) and exporting data from TLex in order to use it in Web applications and repositories clarin.si and GitHub.

Project results

Thus, the result of the MREŽNIK (Croatian Web Dictionary) project will be a free, monolingual, hypertext, searchable, online dictionary of Standard Croatian with ten thousand dictionary entries compiled during the four-year period. The dictionary entries will contain links to repositories which will be created as a part of this project and compiled simultaneously with the dictionary as well as with repositories which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics. The project will include:

1. Compiling a dictionary of ten thousand dictionary entries (with accentuated entry words and accentuated word forms in the grammatical block, with detailed definitions (also definitions for schoolchildren and definitions for foreigners), examples, antonyms, synonyms, collocations and idioms, male-female relations, pragmalinguistic explanations, etc. At the end of the project, the dictionary will be available online on the domain rjecnik.hr.

2. Developing repositories and connecting them with the basic dictionary: The Linguistic Advice Repository of (300 linguistic advice entries), The Conjunction Repository (for all conjunctions in the dictionary), The Idiom Repository (50 entries), The Ethnics and Ktetics Repository (300 inhabitant names and adjectives).

biciklista > biciklist

U hrvatskome standardnom jeziku imenice koje završavaju na *-ist* muškoga su roda i pogrešno ih je izgovarati i zapisivati s *-a* na kraju kao imenice ženskoga roda. Takve su, primjerice, imenice *aktivist, alpinist, biciklist, daltonist, harfist, idealist, iluzionist, kroatist, okulist, optimist, perfekcionist, pesimist, pijanist, šahist.* Pogrešno je *ići okulisti , vidjeti harfistu, razgovarati o šahisti,* a pravilno je *ići okulistu, vidjeti harfista, razgovarati o šahistu.*

	N	G	D	A	L	I
jednina	biciklist	biciklista	biciklistu	biciklista	biciklistu	biciklistom
množina	biciklisti	biciklista	biciklistima	bicikliste	biciklistima	biciklistima

NE	aktivista	alpinista	biciklista	daltonista	harfista	idealista	okulista
DA	aktivist	alpinist	biciklist	daltonist	harfist	idealist	okulist

Figure 3. An entry from the database of language advice *Jezični savjeti* (http://jezicni-savjetnik.hr/)

3. Connecting the basic dictionary with other online resources which are currently being developed at the Institute of Croatian Language and Linguistics: *The Verb Valence Repository*, *The Collocation Repository*, *The Croatian Terminology Repository* (Struna) (Figure 4), *The Croatian Metaphor Repository*, website *Bolje je hrvatski* (Figure 5).

4. Compiling a reverse dictionary. Although the reverse dictionary is planned for the last year of the project as it has to contain the complete word list of ten thousand words, a pilot reversed dictionary has already been compiled by Josip Mihaljević using a test word list. This dictionary will be a working tool for all lexicographers on the project and it will become available online at the end of the project (Figure 6).

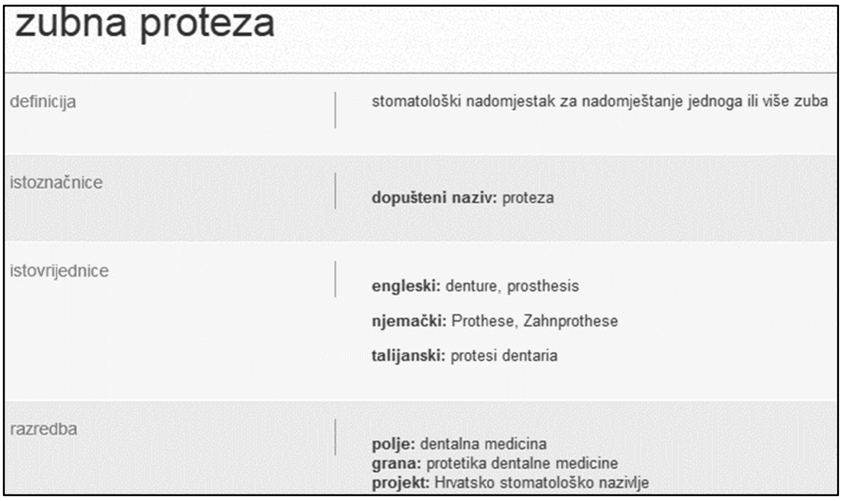

Figure 4. An entry from the terminological database *Hrvatsko strukovno nazivlje STRUNA* (http://struna.ihjj.hr/naziv/zubna-proteza/13383/#naziv)

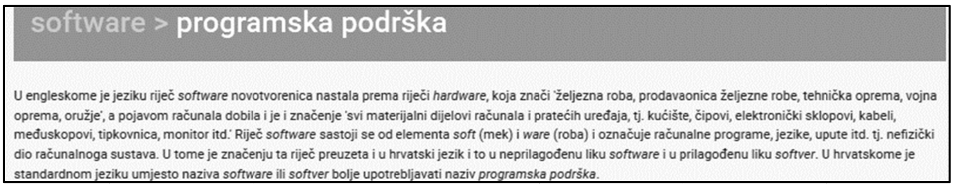

Figure 5. An entry from the database *Bolje je hrvatski* (http://bolje.hr/)

Figure 6. Pilot version of the reverse Croatian dictionary

5. Extensive research on e-lexicography, training, and dissemination of acquired knowledge as well as a contribution to the area of e-lexicography, which did not receive the attention it deserves in the Croatian scientific community so far, is needed.

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

Literature
Blagus Bartolec, G.; Hudeček, Lana; Jojić, Ljiljana; Kovačević, Barbara; Lewis, Kristian; Matas Ivanković, Ivana; Mihaljević, Milica; Miloš, Irena; Ramadanović, Ermina; Vidović, Domagoj. (2012). *Školski rječnik hrvatskoga jezika*. Institut za hrvatski jezik i jezikoslovlje – Školska knjiga. Zagreb.
Hudeček, L. Jozić, Ž., Lewis, K., Mihaljević, M. (2016). *Prvi školski pravopis*. Institut za hrvatski jezik i jezikoslovlje.
Jermen, N., Kraus, C., Starčević Stančić, I. (2015). Lexicography and Encyclopaedistics in the Digital Environment. Infuture 2015. The Future of Information Sciences. E-Institutions Openness, Accessibility, and Preservation, Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, 65-75.
Štrkalj Despot, K., Möhrs, C. (2015). Pogled u e-leksikografiju. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 41(2), 329-353.