STATE-OF-THE-ART ON MONOLINGUAL LEXICOGRAPHY FOR CROATIA (CROATIAN)

Kristina Š. DESPOT, Lana HUDEČEK, Tomislav STOJANOV
Institute for the Croatian Language and Linguistics

NIKOLA LJUBEŠIĆ
Institut “Jožef Stefan”; University of Ljubljana, Faculty of Computer and Information Science


DOI: http://dx.doi.org/10.4312/slo2.0.2019.1.65-77.

In this minireview, the state of the art of the Croatian monolingual lexicography is presented. A brief overview and classification of all existing lexicographic resources is provided in the first part of the minireview, followed by somewhat more detailed insight into the existing Croatian monolingual dictionaries and monolingual lexicographic projects, orthography dictionaries, and dictionary writing systems used.

Keywords: Croatian lexicography, e-lexicography, monolingual dictionaries

1 INTRODUCTION

Croatian lexicography has a remarkably long and rich tradition,¹ which is more than five centuries long, and which has always been an important part of European lexicography, built on the tradition of Latin dictionaries (Štrkalj Despot and Möhrs, 2015, p. 329). However, after the important change of lexicographic paradigm brought by the digital era, the discontinuation of this tradition became

very apparent, primarily in the number and quality of e-dictionaries compared to other European, and even other Slavic languages (Ibid, p. 330).

2 AN OVERVIEW OF EXISTING LEXICOGRAPHIC RESOURCES (BASED ON THEIR FORMAT)

Here, we will provide an overview of the current state-of-the-art of the Croatian e-lexicography based on Štrkalj Despot and Möhrs (2015, p. 335), which propose the classification of the contemporary Croatian lexicographic accomplishments following de Schryver (2003), and Klosa (2013). They have classified the contemporary Croatian lexicography as follows:

1) Corpus driven dictionaries: e. g. Croatian Frequency Dictionary (Moguš, Bratanić and Tadić, 1999), Dictionary of Marulić’s Judita (Moguš, 2001) etc.

2) Corpus based dictionaries: e. g. Dictionary of the Croatian Language for Schools, (Birtić et al., 2012) etc.

3) Dictionaries that are available in the closed digital forms (CD/DVD): First Croatian School Dictionary (Čilaš Šimpraga, Jojić, and Lewis, 2008), Big Dictionary of the Croatian Language (Jojić et al., 2015);

4) Open access network dictionaries: Croatian linguistic portal;² The MREŽNIK project;³ Open access lexical databases (including terminological ones): CroWN,⁴ MetaNet.HR,⁵ e-Glava,⁶ Database of Collocations,

---

² (http://hjp.novi-liber.hr/)
³ http://ihjj.hr/projekt/hrvatski-mrezni-rjecnik-mreznik/70/
⁴ (http://meta-share.ffzg.hr/repository/browse/croatian-wordnet)
⁵ MetaNet.HR (http://ihjj.hr/metafore) is the project of the Institute for the Croatian Language and Linguistics, led by K. Š. Despot. The project functions as an expanded version of Croatian FrameNet. This database lists hierarchically organized conceptual metaphors and metonymies decomposed into source-target relations among cognitive primitives, image schemas, and semantic frames, which are further decomposed into semantic roles (see Despot et al., In press).
⁶ E-Glava (Valency Database of Croatian Verbs) (http://valencije.ihjj.hr/) is the project of the Institute for the Croatian Language and Linguistics, led by M. Birtić. It follows German valency dictionary – e-VALBU as linguistic and lexicographical model. It currently contains verbs belonging to the semantic class of psychological verbs with their meaning analysis and valency patterns (see Birtić et al. 2018).

3 GENERAL MONOLINGUAL DICTIONARIES (INCLUDING SCHOOL DICTIONARIES)

In Croatia, general monolingual dictionaries are published either by public institutions (Institute for the Croatian Language and Linguistics, Lexicographic Institute Miroslav Krleža), or by private publishing houses.

The Institute for the Croatian Language and Linguistics (IHJJ) is the central Croatian research institution for the research of the Croatian language, and as such, it is responsible for the implementation and realization of the long term national philological projects of a strategic importance. It is best known and recognized for its work within national monolingual modern and historical lexicography. It was founded in 1948. with the main goal to finish the big Dictionary of the Croatian or Serbian Language (known as Academy Dictionary), a monumental lexicographic masterpiece that consists of 300,000 entries published in 23 books.

The project Croatian Normative Dictionary, led by L. Hudeček in the period from 2006 to 2013, resulted in the printed monolingual dictionary co-published with Školska knjiga in 2012. - School Dictionary of the Croatian Language (Birtić et al., 2012), which contains 30,000 entries. This dictionary is aimed at advanced users in their secondary education, but it serves as a

---

7 The Croatian Valency Lexicon (CROVALLEX 2.0008) (http://theta.ffzg.hr/crovallex), compiled by N. Mikelić Preradović at the Faculty of Humanities and Social Sciences, Zagreb University, is the formal description of valency frames of Croatian verbs. It contains roughly 1,800 most frequent verbs, which were selected from the Croatian frequency dictionary. The structure is based on the Valency Lexicon of Czech Verbs (VALLEX 1.0) developed at the Institute of Formal and Applied Linguistics (ÚFAL), in Prague 2004 (Žabokrtský and Lopatková, 2007).

8 (http://struna.ihjj.hr)

9 (http://hrana.ffzg.hr)

10 (http://crodip.ffzg.hr)

11 (http://hjp.znanje.hr)

12 (http://nazivlje.hr)
general monolingual dictionary as well. It is the first Croatian general monolingual corpus based dictionary. The conception of the dictionary and the lexicographic workflow were determined and led by Lana Hudeček and Milica Mihaljević (see Birtić et al., 2012, Introduction). The workflow was not organized around alphabetically ordered entries, but around semantically or grammatically closely related groups of words (such as colours, week days, seasons, pronouns, conjunctions, etc.), which resulted in the consistent lexicographic processing. In this dictionary, in addition to a thoroughly compiled grammatical unit, definition, and phraseological unit, a lot of attention has been paid to the systematicity of semantic, morphological and grammatical definitions, relationships between words (synonymy, antonymy, homonymy) and the appropriateness of examples and idioms.

As a predecessor to this dictionary, in 2008, the First Croatian School Dictionary (Čilaš Šimpraga, Jojić, and Lewis) consisting of 2,500 entries was published by IHJJ and Školska knjiga. This dictionary is aimed to be used by 1st and 2nd elementary school graders as a preparation and practice for future usage of real dictionaries. Dictionary consists of everyday words that are known to the children as well as of less known words from their handbooks. Its abecedarium was compiled on the lexical material of the school program for the 1st and 2nd grade of elementary school. The structure of the article is simple, offering simple definitions with illustrations. It can therefore serve as a dictionary for foreigners as well. The dictionary is accompanied by a DVD, on which digital version may be used together with lexicographic games.

In the past five years, IHJJ has also become known for its numerous online lexicographic resources, described in the previous chapter.

The biggest Croatian lexicographic project is currently ongoing at IHJJ (with L. Hudeček as its PI), and is funded by the Croatian Science Foundation – Online Croatian Dictionary (MREŽNIK) – (see Hudeček, 2018; Hudeček, and Mihaljević, 2018). The aim of the MREŽNIK project is to compile the first digital born, free, monolingual, corpus-based, hypertextual, easily searchable, online dictionary of the Croatian standard language with three modules: for adult native speakers, for school children, and for foreigners. The dictionary is based on two Croatian corpora: the Croatian Web Corpus hrWaC (http://nlp.ffzg.hr/resources/corpora/hrwac/, Ljubešić and Klubička, 2014), and the Croatian
Language Repository (CLR; riznica.ihjj.hr, Brozović et al., 2018). The lexicographers select data from the corpus as well as from other Croatian dictionaries, websites, and other resources. The dictionary workflow is supported by Sketch Engine (Kilgariff et al., 2004). The compilation of the dictionary is based on Word Sketches specially adapted to the needs of the project, which are based on a developed Sketch Grammar and the application of the GDEx module for finding appropriate examples in the corpus. To support the preparation of the dictionary text the TLex software package is used. The bulk of the project will be finished in 2021.

Lexicographic Institute Miroslav Krleža (LZMK) was founded in 1950, and is focused on and largely appreciated for encyclopaedic lexicography. The first general dictionary produced by this institute (and co-published with Školska knjiga in 2000) was the *Dictionary of the Croatian Language*, edited by J. Šonje. Since this dictionary is produced by the public institution, it is obviously supported from the public funds.13


In 2006, on the basis of the *Dictionary of Foreign Words* (Anić and Goldstein, 1998, 2000), *Encyclopaedic Dictionary, Large Dictionary of the Croatian Language*, and some other resources, Novi Liber created the *Croatian Language Portal*, an open access searchable portal with more than 110,000


14 The school edition of this dictionary has also been printed in 2015, and based on its material a monumental Encyclopaedic Dictionary of the Croatian Language was published in 2003.
entries (containing basic grammatical information, definitions, examples, as well as phraseological, onomastic and etymological information). *The Croatian Language Portal* is currently the only open access online dictionary of the Croatian language.\(^{15}\) Despite the fact that its content has not been updated, it is very popular and widely used. According to the Croatian part of the survey on dictionary use in Europe results,\(^{16}\) cca 50\% of all the examinees confirmed it as their first and only dictionary choice, obviously because it is currently the only online freely available resource.\(^{17}\)

Školska knjiga publishing house (founded in 1950) is a private Croatian publisher with important participation in the market of school dictionaries and orthographic dictionaries. Školska knjiga published its first general monolingual dictionary in 2015 (*The Great Dictionary of the Standard Croatian Language*, ed. Lj. Jojić) containing more than 120,000 entries on 1800 pages. This dictionary has its digital edition in form of the desktop and mobile applications, which is accessible only by purchasing the printed edition. It has been financially supported by the Ministry of Science, Education and Sports of the Republic of Croatia, and the Croatian Academy of Sciences and Arts.

Currently, there are no general monolingual dictionaries of Croatian for foreign learners, but within the Mrežnik project, a module is envisaged to accommodate these needs.

### 4 CROATIAN ORTOGRAPHY DICTIONARIES

As is the case with the majority of Slavic languages, the Croatian language has a long and rich tradition of production and wide usage of ortography dictionaries. There are four Croatian ortography dictionaries currently in use, all of which are an integral part of an orthography handbook (Anić and Silić, 2001; Badurina, Marković and Mićanović, 2008; Babić and Moguš, 2011; Jozić et al., 2013). They consist primarily of lemmas causing possible ortographic

\(^{15}\) In 2015 this portal was taken over by Znanje publishing house.

\(^{16}\) The results of a survey on dictionary use in Europe, focusing on general monolingual dictionaries are presented in Kosem et al., 2018. The survey covers close to 10,000 dictionary users in nearly thirty countries. It was completed by 9,562 respondents, over 300 respondents per country on average.

\(^{17}\) Mrežnik project is currently the only publicly funded project of building an online dictionary, and its content will also be freely available by the end of the project in 2021.
problems. The Orthography handbook produced and published by the Institute for the Croatian Language and Linguistics (Jozić et al., 2013.) is recommended for use in elementary and secondary education by the Ministry of Science and Education of the Republic of Croatia. Its spelling dictionary is freely available through a web page since 2013.\footnote{http://pravopis.hr/rjecnik} It is widely used and the most popular online linguistic resource in Croatia. Since 2015, Babić and Mo-guš (2011) spelling dictionary is freely accessible through desktop and mobile applications.\footnote{www.rjecnici.hr}

Judging from the number of orthography handbooks during the history, and especially from the number of their editions and printed copies in present days, their popularity and market potential, even their national symbolism, is valued higher than that of general monolingual dictionaries.\footnote{In the survey, for general monolingual dictionaries national symbolism is rated the lowest among other seven characteristics (see Kosem et al. 2018).}

5 DICTIONARY WRITING SYSTEMS

For many years, the only dictionary writing system used in Croatia was Softleks\footnote{https://softleks.hr/wp}– a software developed by a small private company in early 90’s. Softlex is a blend of a custom NoSQL database optimized for storing dictionary data and a WYSIWYG text editor to enable quick and user-friendly enter of complex dictionary data. At the time of its implementation, 16 years before the first XML standard, it was very advanced, but by not being constantly upgraded it soon became a bottleneck in development of a national lexicography resulting in a rather slow process of our inclusion into modern e-lexicographic trends.

The majority of the dictionaries presented here is primarily manually compiled with a very limited corpus support. Despite the fact that the connection with corpus seems not to play an important role to dictionary users (among 11 users’ criteria of importance, two of them related to corpora were rated at the 5th and 7th place in the survey), it is is a conditio sine qua non for modern lexicographers.

Methodological turning point in the Croatian lexicography is the compilation of digital born The Online Croatian Dictionary (Mrežnik), see ch. 3. It is
being compiled in TLex DWS, which has been customized to the project needs. SketchEngine is being used as a CQS. This information is integrated with TLex program, in which lexicographic processing is done on all levels. For the morphological level descriptions, the project uses the inflectional morphological lexicon hrLex.\(^{22}\) When the lexicographic processing is finished, all the data will be exported to a web application, as well as to the CLARIN repository. This will enable MREŽNIK to be widely used both by public and for language technology applications that require access to the full data.

When it comes to the internet availability, among EU countries, the Eurostat statistics for the year 2016 ranks Croatia at the 21st place.\(^{23}\) According to the social media management agency Hootsuite, the internet penetration rate in Croatia in 2017 is 75%, which is on average higher among the South European countries, but lower in comparison with Western Europe.\(^{24}\)

When it comes to crowdsourcing, it has not been used so far for compiling monolingual dictionaries. However, this might change soon due to results of a 3-year research project SenseHive: Dynamic Crowdsourcing Models for Incremental Construction of Lexico-Semantic Resources, funded by the Croatian Science Foundation (2015–2018), with Jan Šnajder as its PI, which is aimed at developing a comprehensive crowdsourcing methodology for incremental construction of large-scale lexico-semantic resources. The research combines dynamic crowdsourcing, corpus-based models of semantics (distributional semantics and topic models), and active machine learning methods into a comprehensible and language-independent crowdsourcing framework, the SenseHive.\(^{25}\)

8 DICTIONARY USE

The status of lexicography in general in Croatia is relatively satisfying, both in terms of lexicographic theory and practice, especially considering printed dictionaries. There exists a great amount of lexicographic research (E.g. there

\(^{22}\) https://www.clarin.si/repository/xmlui/handle/1135...

\(^{23}\) https://goo.gl/27HPe8 (link to the Eurostat web page on the internet access)

\(^{24}\) https://www.slideshare.net/wearesocialsg/digital-in-2017-southern-europe

\(^{25}\) The project description and the list of project publications may be retrieved at https://www.researchgate.net/project/SenseHive-Dynamic-Crowdsourcing-Models-for-Incremental-Construction-of-Lexico-Semantic-Resources.
are more than 400 publicly available lexicographic papers on Hrčak, an open access portal of Croatian scientific journals), however, it is primarily focused on the lexicographic content (e.g. Nikolić-Hoyt, 2004, Tafra, 2005), while the research on dictionary use is completely neglected.

Dictionary use is unfortunately not systematically taught in primary education nor is it a part of official curriculum for it. It is up to teachers to dedicate one or two school hours to teach children how to use dictionaries, and they usually do this in collaboration with school librarians in 7th or 8th grade of elementary school, when learning about the history of the Croatian language. Lexicography is a part of the curriculum only for the highest grades of secondary education. However, this might change due to a fact that Croatia is going through a national educational reform, and publicly available version of the curriculum for the Croatian language involves training in dictionary use among other learning strategies.26

Surely, earlier and more official introduction to using dictionaries and other basic printed and online linguistic resources would be far more fruitful and would benefit future general lexicographic and linguistic culture.

9 CONCLUSIONS AND OUTLOOK

From the overview given above it can be noticed that the majority of Croatian general monolingual dictionaries (including school dictionaries) are printed dictionaries. The majority of those dictionaries are primarily manually compiled with none, or very limited corpus usage. They are published both by public institutes and by private publishers, and in both cases they are entirely or partially publicly funded. There is no obvious difference in terms of users’ trust to either of those. The ease of access to content seems to play a much more important role. Leading monolingual dictionaries do not seem to be connected with national symbolism, it seems that there is much stronger connection of orthography handbooks with national symbolism.

At the moment, there is only one free-of-charge and web-supported dictionary - the Croatian Language Portal. There is also an application-based The Great

26 As educational tools, dictionaries are specifically mentioned in the first, second, and seventh grade of elementary school. http://www.kurikulum.hr/wp-content/uploads/2016/03/Hrvatski-jezik.pdf
Dictionary of the Standard Croatian Language, which is available only if the printed version is bought.

The biggest Croatian lexicographic project and the only digital born and corpus-based dictionary – the Online Croatian Dictionary Mrežnik- is currently being conducted at the Institute for the Croatian Language and Linguistics. This dictionary is a methodological turning point in the Croatian lexicography leading it towards the inclusion into the newest trends and highest standards of European e-lexicography.

REFERENCES


STANJE ENOJEZIČNE LEKSIKOGRAFIJE: HRVAŠČINA

Pregledno predstavljamo stanje enojezične leksikografije na Hrvaškem. Najprej podamo kratak pregled in klasifikacijo obstoječih leksikografskih virov, v nadaljevanju pa se osredotočimo na obstoječe hrvaške enojezične slovarje, pravopisne slovarje ter slovaropisne sisteme.

Ključne besede: hrvaška leksikografija, e-leksikografija, enojezični slovarji

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna. / This work is licensed under the Creative Commons Attribution-Share-Alike 4.0 International.

https://creativecommons.org/licenses/by-sa/4.0/