european lexicographic infrastructure

(Retro)digitisation and online publication of the Croatian Dictionary of Literary Language

Ivana Filipović Petrović

The Institute for Linguistic Studies, Croatian Academy of Sciences

Introduction

This project involves the (retro)digitisation and online publication of the Croatian Dictionary of Literary Language (CDLL), as well as computer processing of the corpus on which the dictionary is based.

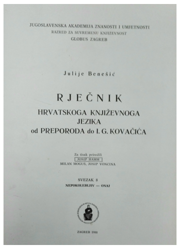
The aim is to improve the work on the remaining volumes of the CDLL and make the dictionary available in open access via online publishing, which will be a significant contribution to Croatian lexicography.

About the CDLL

Full title: *Dictionary of the Croatian Literary Language from the Revival to Ivan Goran Kovačić*

Raw material: 150 cardboard boxes with handwritten index cards

Lexicographer who compiled the collection of index cards: Julije Benešić (1883–1957)



JUGOSLAVENSKA AKADEMIJA ZNANOSTI I UMJETNOSTI
RAZRED ZA SUVREMENU KNJIŽEVNOST
GLOBUS ZAGREB

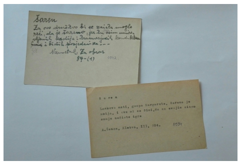
Julije Benešić

RJEČNIK
HRVATSKOGA KNJIŽEVNOGA
JEZIKA
od PREPORODA do I. G. KOVAČIĆA

Za tisak priredili
JOSIP HAMM
MILAN MOGUŠ, JOSIP VONČINA

SVEZAK 8
NEPOKOLEBLJIV — ONAJ

ZAGREB 1988



The main characteristics of the CDLL:

- ❑ based on citations from the literary works written by 113 Croatian writers during a one hundred year period between 1835 and 1945
- ❑ a unique socio-historical record of the Croatian language in the late 1800s and early 1900s
- ❑ a descriptive dictionary: it includes many idiosyncratic uses which are often left out of dictionaries, especially prescriptive dictionaries that are still predominant in Croatian lexicography

Timeline

Volume	Publication year	Content
1-12	1985-1990	A-R
13	2013	S-Sr
14	2017	St-Š
15	In progress	T-Ul
16-18	Future plan	Um-Ž

The first twelve volumes are the legacy of Julije Benešić and the editors who published his work thirty years after his death.

The other six volumes (two published, one in progress and three planned in the future) are based on Benešić's raw material. A team of three lexicographers at the Institute for Linguistic Studies of the Croatian Academy of Sciences has been working on the dictionary since 2008.

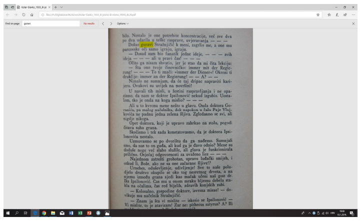
Given that the old volumes contain certain lexicographic inconsistencies and that they are not very user-friendly in some cases, in the new volumes the microstructure of the dictionary has been improved on several levels.

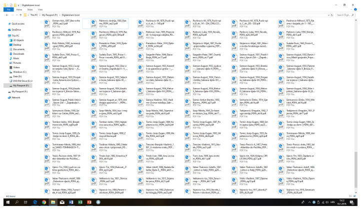
Problems: where we are at the moment

1. A partially searchable corpus made up of individual PDF documents

❑ A corpus of 400 literary works is only partially searchable due to the fact that OCR was not done properly.

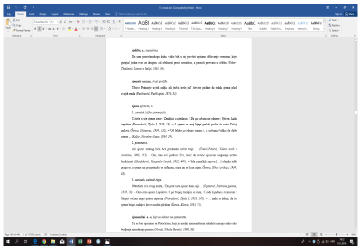
❑ Each of the 400 works is searchable separately, in an individual PDF document, and there is no integral corpus.





2. Inadequate DWS

The CDLL is compiled using Microsoft Word: lexicographers copy entries from index cards into a Word document.



3. Low availability

New volumes are published in a small number of copies, and the first twelve volumes are no longer available in bookstores and very rarely in libraries.

Goals: where we want to be

The 14 printed volumes should be digitised and published online with some necessary corrections.

Future work on the next four volumes (15-18) requires an integral corpus, a convenient dictionary writing system and infrastructure for online publishing.

The first step in accomplishing these goals is a two-week visit to the **Trier Centre for Digital Humanities**, whose main focus is (retro)digitisation and online publication of legacy dictionaries.

The benefits of the methodological and technical improvements are twofold:

1. The usefulness of this diachronic corpus goes beyond Benešić's dictionary: it can be used for various studies of the history of language.
2. It would be the first and only descriptive dictionary of Croatian available in open access.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.