Ivana Filipović Petrović, PhD
Linguistics Research Institute of the Croatian Academy of Sciences

Report
on Elexis Transnational Research Visit Grant at Trier Center for Digital Humanities
(Trier, Germany, February 25 – March 8, 2019)

Travel grant: Call 1

Project title: (Retro)digitisation and online publication of the *Croatian Dictionary of the Literary Language*

Introduction

The project proposal that I submitted to the call for Elexis grants for research visits included the plan of (retro)digitisation and online publication of the *Croatian Dictionary of the Literary Language* (CDLL), as well as computer processing of the corpus on which this dictionary is based. Therefore, the goals of my visit were gain the knowledge, competence and (if possible) lexicographical tools and infrastructure for this project. By the end of my visit, it was clear that not only the goals were accomplished, but so much more: I gained a valuable experience of working with experts in the field of digital humanities and with sophisticated lexicographical tools as well as a social and emotional experience of a two-week living in a foreign country. The hosts, Trier Center for Digital Humanities, especially Dr Vera Hildenbrandt and Li Sheng contributed to all mentioned experiences in a best possible way.

In this report I will present the workflow of my visit, i.e. the projects, dictionaries, tools and all useful insights that had an impact to the achievement of the goals. In addition, I will single out some possible solutions for the further process of (retro)digitization of the *Croatian Dictionary of Literary Language*.

Digitization of primary sources: OCR4all

The first goal was the digitization of the primary sources for the CDLL, in order to build an integrated, fully searchable corpus of 400 sources. For this task, I was introduced in tools and methods of digitization, especially to the open-source tool OCR4all which is based on a deep learning algorithm. Given that we have 400 sources for CDLL stored in .JPEG format, I brought them with me on the external disc. OCR4all works in a way that I upload the source in the .JPEG format and set the digitization of the first few pages in motion. After that, the next step is checking the result and correcting misreading in order to train the program and achieve better results on the next pages of the source. In the case of my first test source, the collection of poems of Croatian writer Tugomir Alaupović, the result on the first three pages was not good enough, due to the poor quality of images as well as the presence of diacritical letters and some special characters in the source. But after the training of a model, the result was significantly better: the accuracy of recognition increased to 98, 7 %. In order to improve the result, it is necessary to train the new model based on the corrections. When it comes to the literary (primary) sources for a dictionary, the result must be 99, 95 %, which can be achieved by OCR4all in combination with proofreading. Furthermore, the result is fully-digitized text, which means that search possibilities are increased (unlike the image digitization that we had), as well as the possibilities for a representation of references in the on-line version of a dictionary. In addition, during my visit I was introduced to the methods and tools for compiling the *Mittelhochdeutsches Wörterbuch* by Dr Niels Bohnert. Among other interesting things (lemma list, dictionary writing system, online version), my attention was drawn by the database of the sources used for this dictionary as well as the list of bibliographic references. Actually, it was a pure example of the database that should be provided for CDLL sources, and that is why it is important to make a full-text digitization of the sources. In conclusion, this form of digitization should be seriously considered when digitizing the corpus of the sources for CDLL. At this point, the OCR lectures and training was completed, and we moved along to the other steps in the process of (retro)digitization.

Encoding in XML according to TEI standards and Dictionary Writing System

The second issue from my project proposal, dictionary writing system refers to the new volumes of CDLL which need to be written. Given that every dictionary has its own requirements and peculiarities, very often the solution is developing an in-house dictionary writing system for particular purpose. Also, some dictionary projects use an XML editing tool and make some tailor-made solutions, i.e. they adopted it for lexicography. The TCDH has experience in the development of a dictionary writing system (TAReS: A Webbased System for Editing, Producing, and Publishing Dictionaries in Distributed Offices) and also solid experience with an XML editor such as Oxygen. I have a very good experience with Lexonomy, given that this open-source platform for writing and publishing dictionaries is currently in use on another project that I am working on (*Online Croatian Dictionary of Idioms*). In order to choose or build suitable DWS it is important to know the structure of a dictionary, but also it would be useful to have some knowledge about the mark-up language such as XML.

Furthermore, encoding in XML will be needed in (retro)digitization process as well, after the digitization of printed volumes. So considering all that, it was very useful for me to learn how to encode in XML for various reasons. I was introduced to the rules of this markup language as well as the TEI guidelines for the encoding of dictionaries. Given that the first 12 volumes of CDLL, that were published between 1985 and 1990, require detailed revision according to the consistent lexicographical treatment established for the new volumes,[1] my practical work on encoding in XML started on the entries from the volume 13 (2013). By the end of my visit, I was able to successfully encode different types of entries from CDLL. I will also be able to use this knowledge if we are going to use GROBID Dictionaries, which is a tool for structuring dictionaries, for converting the data from PDF into the TEI XML, at least on the PDFs of two new volumes that are made from Word.

In conclusion, the benefits of learning how to encode in XML according to TEI are twofold: it will serve as a guide when choosing and building DWS for completing CDLL, and also as the data for an online version of the *Dictionary*.

[1] The *old* volumes of this dictionary suffer from some serious imperfections: the citations often don't match to the source reliably due to the fact that they were rewritten from the index cards into the *Dictionary*, and not directly from the source. Also, the lexicographical treatment of the figurative meaning is not thoroughly implemented. Beside this, bibliographical references in the first 12 volumes are consisted only from the last name of the author, which is insufficient for the contemporary user. All these problems are solved in the new volumes and the *Dictionary* is methodologically significantly improved.

Online publishing

Finally, the third step in (retro)digitization of a legacy dictionary – it's online publication requires the building of a database (with the text and tagging as an output of encoding) and the development of a graphical user interface. In this field, TCDH has some remarkable solutions, for example the interface for the *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm* or the *Goethe-Wörterbuch*. Besides that, in the Center is currently under development an open-source tool for the online publication of dictionaries. This so called Dictionary Viewer will provide an infrastructure that allows publishing a dictionary with its accompanying pretexts like prefaces, abbreviations indexes etc. Also, the *Wörterbuchnetz* is a dictionary network where a user can search for words in 28 German dictionaries (monolingual, bilingual, historical, general and special purpose). All this achievements in the field of online publishing of the dictionaries are very significant and progressive and it was very useful and stimulating for me to be at the heart of them.

Conclusion

At the end of my second week in Trier I started to draw some conclusions about the process of (retro)digitization of legacy dictionaries and what it requires. Moreover, I was thinking about the state of the art on the projects of this kind: projects concerning historical dictionaries, especially the ones that are not finished yet and they started a hundred years ago. The common situation for many of them is the same: you have the material in some form (printed dictionaries, printed sources, maybe scans or images, Word and PDF documents, handwritten index cards). Also, you have a two or three lexicographers with a linguistic background that are busy and overwhelmed with the senses, figurative meanings, examples of use and consistent lexicographical treatment. And the third thing that you have is a desire for technical improvement.

What can you do? This extraordinary group of experts in TCDH taught me and showed me what you can do and I will use these insights for CDLL. When it comes to (retro)digitization of printed volumes, first step is image-digitization of origins. Second step is post-processing of digital images via OCR of this scanned origin and proofreading or double keying of the text. Next step is the analysis of an entry structure in order to start with the markup, i.e. encoding. The last step is using encoded data (XML files) for the building of a database and finally the online publication. In addition, if you have the sources for the dictionary, which should be digitized as well, you should apply the first two steps and then create a database of the sources.

Some of these steps can be realized by open-source tools. For example, in the case of CDLL we will consider OCR4all, Sketch Engine option *create your own corpus* for our sources and Lexonomy as DWS as well as online publishing, given that I already have an experience of configuring on this platform. Although, it should be pointed out that you still need experts from several fields in order to achieved all this: an expert in computer science and an expert in digital humanities. For two wonderful weeks, I enjoyed the company of some of the best.

Special thanks to: Dr Vera Hildenbrandt, Li Sheng, Dr Niels Bohnert, Dr Hanna Büdenbender, Anja Hennemann and Felix Thielen.