OCR and TEI for the production of ELTeC – Würzburg Training School, 16-17 April 2018

Jelena Andonovski
andonovski@unilib.rs
University Library
"Svetozar Marković"
Belgrade, Serbia

PAPER SUBMITTED: 23 June 2021
PAPER ACCEPTED: 07 November 2021

As the basic task of the Action CA16204 D-Reading is corpus preparation, from the beginning of the project the most important issue was to define methods for text processing, annotation of corpus material and metadata creation. With these aims in mind, the project coordinators organized the first workshop for the Action partners. The two-day workshop was held on April 16 and 17, 2018, at the University of Würzburg, Germany (Figure 2).[1] The organizers were Leonard Konle and Fotis Jannidis from the University of Würzburg, while lecturers were Leonard Konle and Christian Reul, also from the University of Würzburg, and Lou Burnard, an internationally recognised expert in digital humanities, particularly in the area of text encoding and digital libraries. The workshop was attended by 11 participants from 10 countries: Jelena Andonovski from the University of Belgrade, Serbia; Alex Ciorogar from UBB Cluj-Napoca, România; Simon Gabay from the University of Neuchatel, Switzerland; Meliha Hadžić from the Burch University, Sarajevo, Bosnia and Herzegovina; Magdalena Krol from the Institute of Polish Language, Polish Academy of Sciences, Poland; Ioana Lionte from the University "Alexandru Ioan Cuza", Iași, Romania; Anna Rehorkova from the Charles University, Prague, Czech Republic; Floriana Sciumbata from the University of Trieste, Italia; Anna Maria Sichani from the Kings Digital Lab, London, Great Britain; Adeliana Silva from the Nova University of Lisbon, Portugal; Andrejka Zejn from the Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia (Figure 1).

The aim of the workshop was to introduce participants to methods of text processing and annotation.[2] Methods for Optical Character Recognition (OCR) were presented, as well as the available software for that purpose, and then it was explained how to encode texts in the electronic format chosen

[1] About the training school

[2] Training material

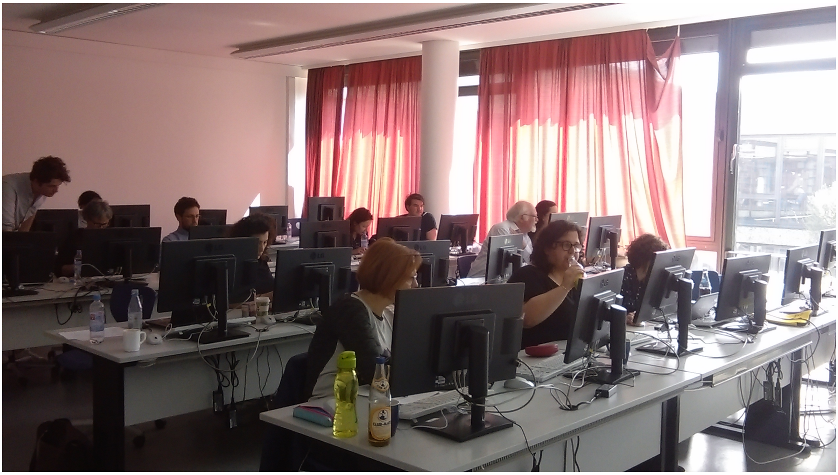

Figure 1. The second day of the workshop

for ELTeC,[3] that is TEI/XML. The first day of the workshop, April 16, was dedicated to OCR. First of all, lecturers presented the basic characteristics of OCR in general, and then some software packages, Abbyy FineReader[4] and an open-source tool OCR4all.[5] Thomas M. Breuel from the University of Kaiserslautern/DFKI, Xerox, Google, currently Nvidia, presented OCR4all and pointed out some of its basic characteristics:

1. It was primarily created to digitally explore very early printed texts;
2. It was prepared to be an open source tool;
3. It was made to be understandable and adaptable for users lacking technical experience;
4. It is independent from the software platform;
5. It is based on some open-source tools (the central part is the OCRopus tool based on Python, which enables preprocessing, layout segmentation, character recognition and model training).

The first day ended with the lecture given by Anna Řehořková from the Institute of the Czech National Corpus, who shared the Institute's experiences in digitizing material for the Czech National Corpus. During the day,

[3] About ELTeC at DH2019

[4] Abbyy FineReader

[5] OCR4all

Figure 2. Participants of the workshop enjoying Würzburg

there was one lunch break and one coffee break, and in the evening, a dinner was organized for all participants, during which they could exchange personal experiences in digitization and corpus preparation.

The second workshop day, April 17, was dedicated to corpus annotation and metadata creation. The lecturer was Lou Burnard. At the beginning, he presented the XML Editor oXygen and its characteristics, then TEI/XML structure for the text encoding and at the end TEI header[6] for metadata creation. After that, he introduced participants to the specially prepared ELTeC encoding Schemas. He explained the method of ELTeC encoding Schemas' creation according to TEI P5 Guidelines, the ODD chaining technique - One Document Does it all.[7] In this way three levels of text encoding were prepared:

1. **Level 0** (eltec-0) – basic TEI structure for text encoding in ELTeC corpus;
2. **Level 1** (eltec-1) – additional elements for encoding (for example, text annotation for lyrics);
3. **Level 2** (eltec-2) – linguistic and semantic annotation of texts, at the level of individual tokens and segments.

[6] TEI header

[7] ODD chaining technique - One Document Does it all

After the lunch break participants had a practical work session, during which they worked with a concrete example: they encoded previously OCRed text using TEI/XML and checked its correctness with the XML Editor oXygen using XML Schemas eltec-0 and eltec-1.

In order to prepare metadata, participants were introduced to the TEI Header. At that moment, the structure of metadata for ELTeC corpus was not yet precisely defined. The workshop was thus an excellent place to discuss the header structure, having in mind that participants were librarians, literary scholars, as well as researchers working in digital humanities. Many different opinions could be heard about metadata creation, which data are important, which have to be mandatory etc. At the end of the workshop an assignment was prepared for participants, to complete when they return home: to prepare one text for the ELTeC corpus according to the guidelines they were given at the workshop.