Converting a Paper Card-file Corpus of Croatian Church Slavonic Texts into a Machine Readable Corpus[1]

Lucija Turkalj, Vida Vukoja

Keywords: Croatian Church Slavonic; lexicography, paper card-file, data capture, intelligent character recognition (ICR)

Croatian Church Slavonic (CCS) is a major constituent of the Croatian diachronic language diasystem. A representative set of CCS texts (11th/12th–18th centuries) was excerpted into a parallel paper card-file corpus. Recently, it has been converted into machine-readable and -searchable data by document-capture and data-extraction software which, due to its adaptability, proved to be a convenient tool for managing the 3 languages, 4 scripts, and circa 20 handwriting styles in the corpus.

1. In 863 Cyril and Methodius arrived in Great Moravia and began their mission of evangelizing and spreading literacy in the newly created Slavic language and script, known today as the Old Church Slavonic (OCS) language and Glagolitic script. Although the vast majority of OCS texts were translated from Greek, texts originally written in OCS are also found, as well as traces of translation from Latin. After the Cyrillo-Methodian mission was destroyed in the political turmoils of the day, some of the brothers' disciples arrived in Croatian lands, where OCS came under the heavy influence of Croatian vernacular and gave birth to another language system: Croatian Church Slavonic (CCS). We know that if a text was translated directly into CCS (and not inherited from the OCS corpus), it was originally a Latin (and not Greek) text. As the primarily liturgical language, CCS had a privileged status within the Croatian diglossic diasystem during the period from the 11th/12th until the 17th century.

2. Between late 1950s and early 1990s a corpus of CCS texts was established, primarily for the purpose of compiling the *Rječnik crkvenoslavenskoga jezika hrvatske redakcije (Dictionary of the Church Slavonic Language of the Croatian Redaction*; DCRCS 2000 & 2015). The corpus is a referential, representative, parallel (with

[1] This work has been supported in part by Croatian Science Foundation under the project 2462.

excerpted Greek or Latin texts), parsed handwritten paper card-file containing only written texts (as expected for bookish idioms such as OCS and CCS).

There are two main card-files. The first one (informally, the sources card-file) is based on codex and text sources containing approximately 400 000 cards with between 1 400 000 and 2 100 000 tokens. A basic version of every text selected to be a member of the corpus is recognized and excerpted as the main text. If a particular text has secondary versions, as most often is the case, then those additional versions are noted on parallel cards. The second main card-file (informally, the azbuka card-file) is based on the azbuka sequence of the lemmas. Also, there are three auxiliary card-files: the CCS lemmas card-file with approximately 18 100 cards, the Greek parallels card-file with approximately 60 000 cards, and the Latin parallels card-file with approximately 200 000 cards (Vukoja 2012, 214-6; Vukoja 2014, 1223-7).

Even though all the texts incorporated in the CCS corpus were originally written in the Glagolitic script, they were excerpted in Old Cyrillic transliteration due to a decision made at the Fourth International Congress of Slavicists. (The DCRCS lemmas, both in DCRCS itself and on the cards of the corpus, are also written in the Old Cyrillic script.) Of course, Greek parallel texts are written in the Greek alphabet and Latin parallel texts are written in the Latin script.

3. It is not an easy task to convert into a digital form the texts of a paper card-file corpus comprising 3 languages, 4 scripts, some 20 different handwritings, and a huge number of parallel cards[2] along with the main text cards. The first step in that direction was taken in 2009 when the sources card-file was scanned and converted into a raster image format, appearing as a sequence of card images (Vukoja 2014, 1232; Magdić 2015, 172-3). The main phase of converting the images into machine-readable text commenced in 2014 and will be completed in 2016. Recognition of the card inscriptions is done using ABBYY FlexiCapture, software for scalable document imaging and data extraction which recognizes and reconstructs documents and allow for reviewing and correcting excerpted texts. Scanned cards, graphically improved by the software, are grouped into smaller units related by context and position in the source and recognized by matching with their respective document-type definitions (with additional data and recognition options, e.g., creation of character sets and dictionary lists). Occasional card misarrangments within the card-file are eliminated in order to

[2] For different procedures concerning digitizing Church Slavonic corpora cf. Ribarov & Camuglia 2003, Ribarov 2004.

allow proper content binding (chaining) of word-forms, texts, and parallel card units.

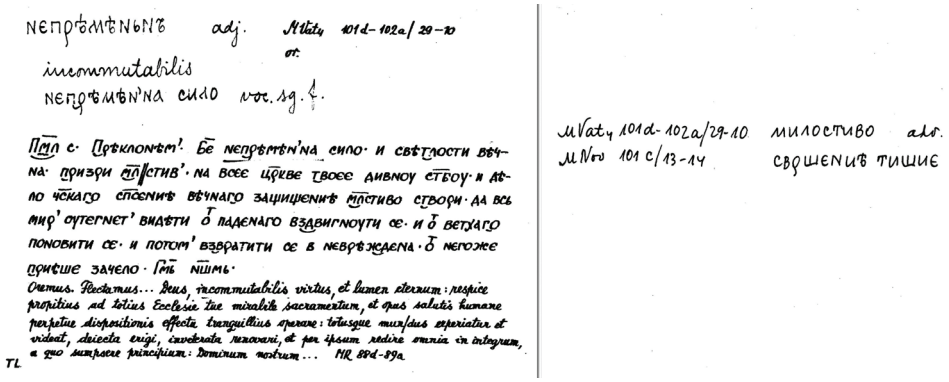

Fig. Card samples

Particular attention is paid to the recognition of the card inscriptions in the following four respects: (a) the main version of the CCS text in question (in Old Cyrillic script), (b) the CCS token that a particular card is dedicated to, which appears in two forms, proper and lemmatized (both in Cyrillic script) and accompanied with its morphological description (in Latin script, also carefully recognized), c) the Greek and/or Latin parallel lexeme of the CCS lemma (in Greek or Latin script), and (d) the source location data of the excerpt (in Latin script). As Greek and/or Latin parallel text in the card is often written in cursive handwriting—thus practically unreadable to a machine—it is added manually to its CCS counterpart text as confirmed in reliable printed or electronic editions. Due to the ink-based paper-card production technology, considerable age of the cards, variable text area size, and offset in relation to borders and other information areas, the inscriptions are of uneven readability. Therefore, and in order to lessen information redundancy and to ease recognition, an effort was made to select cards with as clearly readable text as possible. The ligature ties under letters were also removed. The image area containing the token with lemma and its morphological description was extracted from each card unit, then arranged in tabular form in order to make creation of the document-type definitions easier and minimize the need for additional editing and adjustment of document-type

[3] Binarization, segmentation and morphological preprocessing of images were done by Dragutin Čulinović using Wolfram Mathematica.

definitions once created. If needed, the position of some forms in the document is checked afterwards[3].

The outcome of the machine-reading process described above is recognized and verified for the bulk of the paper-card corpus, i.e., CCS texts in their main version with appropriate Greek and Latin parallel texts are aligned, together with tokens in their proper form with morphological description and in their normalized lemmatized form, and exported into XML file format with preserved structural, morphological and metadata information.

4. Even in its paper-card form, the annotated CCS corpus is an indispensable tool for research on the CCS idiom as well as the prime source for all scholars who for various reasons need to consult CCS texts. Converted to XML and imported into a database, the corpus is ready to be used for a range of purposes—not only easier compilation of the DCRCS entries and historical linguistic research but also the creation of various indices and preparation of digital critical editions of particular CCS texts and so forth.

References

DCRCS 2000 = *Rječnik crkvenoslavenskoga jezika hrvatske redakcije*. Vol. I. (a — vrêdь). 2000. Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta.

DCRCS 2015 = *Rječnik crkvenoslavenskoga jezika hrvatske redakcije*. Vol. II. (vrêdьnь1 — zapovêdnica). 2015. Zagreb: Staroslavenski institut.

Magdić, Antonio. 2015. *Kako je digitalizirana građa Staroslavenskoga instituta*. Hrvatsko glagoljaštvo u europskom okružju / Badurina Stipčević, Vesna; Požar, Sandra; Velčić, Franjo (ur.). Zagreb: Staroslavenski institut. 167–206.

Ribarov, Kiril. 2004. *Towards Intelligent Written Cultural Heritage Processing—Lexical processing.* In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. 1845–1850. URL: http://aclweb.org/anthology/L/L04/ (27.12.2015.)

Ribarov, Kiril & Monia Camuglia. 2003. *Incorporation of Old Church Slavonic Card Files into a Corpus.* In *Scripta & e-Scripta* 1, Sofia: Institute of Literature, Bulgarian Academy of Sciences.

Vukoja, Vida. 2012. *O korpusu Rječnika crkvenoslavenskoga jezika hrvatske redakcije i njegovu odnosu prema korpusima hrvatskoga jezika.* In *Filologija* 59: 207–229.

Vukoja, Vida. 2014. *The Corpus of the Croatian Church Slavonic Texts and the Current State of Affairs Concerning the Dictionary of the Croatian Redaction of Church Slavonic Compiling. Proceedings of the XVI EURALEX International Congress: The User in Focus.* 15–19 July 2014, Bolzano/Bozen. Andrea Abel, Chiara Vettori i Natascia Ralli (eds.). Bolzano/Bozen: EURAC Research, EURALEX. 1221–1235.