Velibor Ilić/Lenka Bajčetić/Snežana Petrović/Ana Španović

SCyDia – OCR FOR SERBIAN CYRILLIC WITH DIACRITICS

Abstract In the currently ongoing process of retro-digitization of Serbian dialectal dictionaries, the biggest obstacle is the lack of machine-readable versions of paper editions. Therefore, one essential step is needed before venturing into the dictionary-making process in the digital environment – OCRing the pages with the highest possible accuracy. Successful retro-digitization of Serbian dialectal dictionaries, currently in progress, has shown a dire need for one basic yet necessary step, lacking until now – OCRing the pages with the highest possible accuracy. OCR processing is not a new technology, as many open-source and commercial software solutions can reliably convert scanned images of paper documents into digital documents. Available software solutions are usually efficient enough to process scanned contracts, invoices, financial statements, newspapers, and books. In cases where it is necessary to process documents that contain accented text and precisely extract each character with diacritics, such software solutions are not efficient enough. This paper presents the OCR software called "SCyDia", developed to overcome this issue. We demonstrate the organizational structure of the OCR software "SCyDia" and the first results. The "SCyDia" is a web-based software solution that relies on the open-source software "Tesseract" in the background. "SCyDia" also contains a module for semi-automatic text correction. We have already processed over 15000 pages, 13 dialectal dictionaries, and five dialectal monographs. At this point in our project, we have analyzed the accuracy of the "SCyDia" by processing 13 dialectal dictionaries. The results were analyzed manually by an expert who examined a number of randomly selected pages from each dictionary. The preliminary results show great promise, spanning from 97.19% to 99.87%.

Keywords OCR; Cyrillic; Serbian language; retro-digitization; convolutional neural networks

Contact information

Velibor Ilić
The Institute for Artificial Intelligence Research and Development of Serbia, 21000 Novi Sad, Serbia
velibor.ilic@ivi.ac.rs

Lenka Bajčetić
Institute for the Serbian Language of SASA
lenka.bajcetic@gmail.com

Snežana Petrović
Institute for the Serbian Language of SASA
snezzanaa@gmail.com

Ana Španović
Institute for the Serbian Language of SASA
tesicana@gmail.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.