Towards Innovative Study of Cyrillic Alphabet through Contemporary Information Technologies. Challenges and New Perspectives

Marco Scarpa[1][0000-0002-9320-5990], Desislava Paneva-Marinova[2][0000-0001-5998-687X]

[1] University of Messina, Department of Ancient and Modern Civilizations,
13, Giovanni Palatucci Avenue, 98168, Messina, Italy
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
8, G. Bonchev, Str., Sofia, 1113, Bulgaria
marco.kmnc@gmail.com , dessi@cc.bas.bg

Abstract. This study aims to present the challenges and explore new opportunities offered by contemporary information technologies for the innovative study of the Cyrillic alphabet. The focus is on the concepts and activities, whose primary goal is the compilation of a digital Repositorium of data from primary medieval epigraphic and manuscript sources dating from the 9th to the 14th centuries. This repository will provide new services for determining the origin, defining, and systematizing the South Slavic manuscripts preserved in various libraries, from the perspective of their paleographic and codicological features. By leveraging modern digital technologies incl. Computer Vision and AI Techniques in the systematization of target objects, the study seeks to foster a fresh understanding of the history of the Cyrillic alphabet.

Keywords: Digital Humanities, Cyrillic Paleography, South Slavic Manuscripts, Paleographic Analysis.

1 Introduction

It is estimated that in 2020, some 255 million people across 15 countries used the Cyrillic alphabet, that is approximately about 3.3% of the global world population, primarily for writing Slavic languages (Arefiev, 2022). As is well known, the Cyrillic script originated in Bulgaria at the end of the 9th century, as a work of the disciples of Sts Cyril and Methodius. It was based on the Glagolitic alphabet — devised by their teachers for the mission in Moravia and Pannonia — and the Greek alphabet, which had long been known in those lands. For this reason, one of Bulgaria's most significant contributions to European and global cultural heritage is the preservation of the work of Sts Cyril and Methodius, along with the creation and diffusion of the Cyrillic alphabet and Slavic literacy.

Writing and the alphabet serve as vehicles for transmitting a people's language and culture. As language and culture evolve, so too does the alphabet that conveys them.

An evolution occurring across both time and space, focusing primarily on the phonetic correspondence between speech and its written form and representation with the phenomena of the disappearance of certain graphemes (for example, the nasal vowels in the medieval Serbian and Croatian context) or the modified use of others (as seen in the alternation of nasal vowels or the 'yers' in the Middle Bulgarian tradition).

Writing, however, also undergoes evolution in its graphic form. In this case, too, this occurs across space, inviting a differentiated approach to East and South Slavic palaeography—and, in the latter case, particular attention to the specific features of the Serbian and Bulgarian contexts. It also unfolds over time, as evidenced by the using of palaeography for manuscript dating.

The classic manuals of Cyrillic Slavic palaeography (for example: (Sreznevskii, 1885), (Karskii, 1928), (Shtepkin, 1967)) outline the principal phenomena of this graphic evolution. However, no comprehensive study of the evolution of writing exists, especially considering that these manuals are over a century old and were authored by Russian scholars. On the one hand, the known manuscript material has expanded considerably during this century and become far more accessible; on the other hand, a study from the South Slavic perspective appears to be particularly important.

2 South Slavic Cyrillic Manuscripts Research Supported by Digital Technologies

In this context, and to address these needs, we have conceived the research project *"Development of the Cyrillic Alphabet from the 9th to the 14th Century in the South Slavic Lands: Research and Digital Presentation"*. The project is led by the Cyrillo-Methodian Research Center at the Bulgarian Academy of Sciences, together with the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences. It is funded for this three-year period by the Bulgarian National Science Fund and continues a trajectory begun with the project *"Fourteenth Century South Slavonic Scribes and Scriptoria (Paleographic Attribution and Online Repertorium)"*.

Our project aims to present the challenges and explore new opportunities offered by contemporary information technologies for the innovative study of the Cyrillic alphabet. The focus is on the project's concepts and activities, whose primary goal is the compilation of a digital Repertorium of data from primary medieval epigraphic and manuscript sources dating from the 9th to the 14th centuries. This digital resource will be made available to the scholarly community as a reliable tool for determining the origin, defining, and systematizing the South Slavic manuscripts preserved in various libraries, from the perspective of their paleographic and codicological features. To achieve this ambitious goal, the project concentrates on discovering new data and integrating it with existing material. By leveraging modern digital technologies in the systematization of this corpus, the project seeks to foster a fresh understanding of the history of the Cyrillic alphabet.

The study is based on data concerning South Slavic Cyrillic manuscript practices documented in the Balkans from the 9th to the 14th centuries. This information will be

processed, analyzed, and made publicly accessible via a dedicated online platform, offering new insights into the graphic forms through which Cyrillic script evolved during this period. The collection and classification of the necessary data for the reliable identification of authentic medieval Cyrillic writing practices, the project aims to generate new fundamental knowledge about the structure and activities of the known literary centers and scriptoria where South Slavic manuscripts were produced and used, as well as the development of the alphabet itself within both chronological and geographical frameworks.

3 Towards Design and Development of a Digital Repertorium of South Slavonic Manuscripts and Copyists from the Period 10th–14th Century

The *"Repertorium of South Slavonic Manuscripts and Copyists (10th–14th c.)"* is a development of the catalogue (The Repertorium of …, n. d.), (Scarpa et al., (2023) created in the *"Fourteenth Century South Slavonic Scribes and Scriptoria (paleographic attribution and online Repertorium)"* project (14th Century South Slavonic Scribes and …, n. d.), (Luchev et al., 2024a), (Luchev et al., 2024b). It is an online platform providing new knowledge into the graphic forms through which Cyrillic writing developed in the period 10th–14th century. It collects and classifies the data necessary for the reliable identification of authentic medieval Cyrillic writing practices. In doing so, it facilitates the study of the Cyrillic alphabet's development from both chronological and geographical perspectives, while also contributing to a deeper understanding of the structure and activities of the known literary centers and scriptoria where South Slavic manuscripts were produced and used. The Repertorium will provide services for structuring, processing, management, protecting, and visualizing South Slavic manuscript data. Its aim is to ensure effective access to the vast amounts of knowledge about digitized Cyrillic heritage, enabling high adaptability, effective interaction, and multiple uses.

A key digital tool designed to support the analysis of medieval Cyrillic handwriting within a codicological, and paleographical framework is the *Cyrillic Paleography Toolkit (CyPaT)*. Developed based on a standardized descriptive model, the CyPaT is fully integrated with the *Repertorium of South Slavonic Manuscripts and Copyists (10th–14th c.)*. CyPaT provides a comprehensive environment for description, processing, and comparative study of manuscript data. Through direct interaction with digitized images, CyPaT enables detailed examination of script features such as letter proportions, stroke composition, and page layout.

The following section presents an advanced approach for the classification of Cyrillic letters using computer vision and Artificial Intelligence technologies as implemented in CyPaT.

3.1 Letter Classification Using Computer Vision and AI Techniques

One very important factor in affective manuscript analysis for researchers is the ability to analyze every letter and detail within a manuscript to draw accurate conclusions. In

the case of extensive manuscripts containing thousands of letters and graphic elements, completing this task manually is quite impossible. This is precisely where the technologies come to support this process, significantly reducing the need for manual work in this process.

The current objective is to train a convolutional neural network (CNN) capable of guessing / identifying the scribe, allograph type, and handwriting style associated with given letter / letters from certain manuscripts. Achieving this task requires the construction of a relatively large, high-quality dataset of letters (letterforms), which includes information about the manuscript, scribe, handwriting type, and allograph classification for every single letter.

Naturally, the manual creation of such a large dataset cannot be done, and this is where Computer Vision (CV) techniques provide support. By employing CV algorithms such as adaptive thresholding, aria filtering and morphological operations, we aim to automatically identify the bounding boxes of every letter in a manuscript. Researchers can then annotate and edit a limited set of these bounding boxes through a specific user interface (Fig 1).

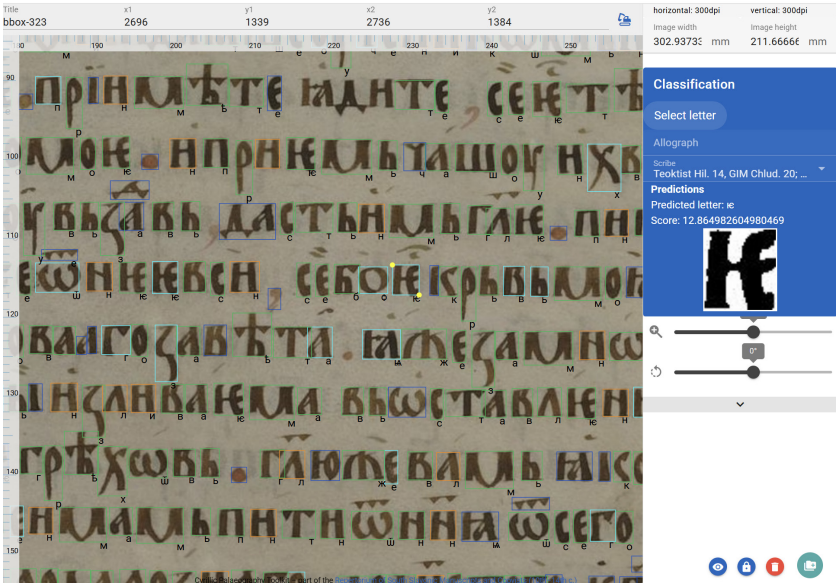

Fig. 1. Annotating and editing functionality.

An automated process based on CV image comparison is used to identify all other bounding boxes that match the already annotated letters, thereby significantly increasing the number of samples in the dataset. Researchers can validate and, if necessary, correct the results of the image matching using a dedicated interface specially designed for this purpose (Fig. 2).

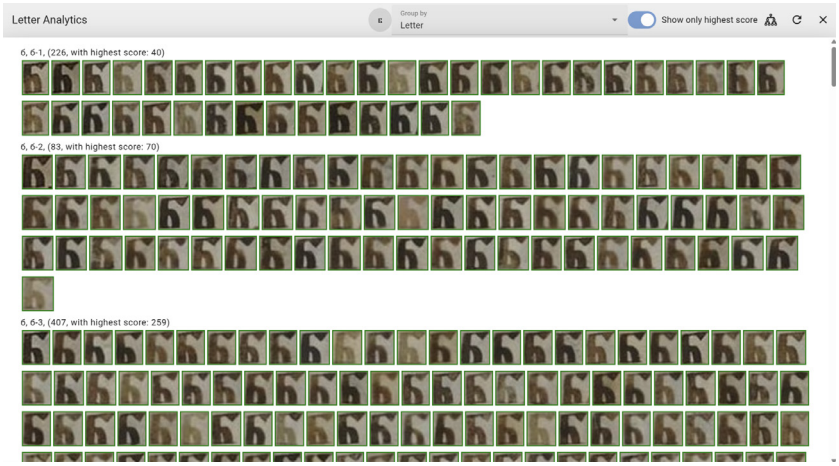

Fig. 2. Interface for validating and correcting the results.

The last three steps are repeated until a sufficiently large number of samples is obtained. Before initiating the convolutional neural network training, a crucial step, known as data augmentation, is performed. This means that for every identified letter we will create a set of samples which will be slightly different (they will be slightly rotated, repositioned by some pixels, etc.). This step is important, as it increases the amount of data available to CNN, thereby improving the accuracy of the results.

The final stage of the process involves training and validating the results.

Although this process is still ongoing and remains a work in progress, the preliminary outcomes are highly promising. The team is confident that the approach is both appropriate and likely to prove successful and valuable. Furthermore, the final dataset is intended to be published as an open-source resource, which could be used for further research of Slavic manuscripts studies.

4 Conclusions and Future Works

The work carried out during the previous project, as well as in the first months of this new phase, has demonstrated the fruitful potential of collaboration between philologists and palaeographers on the one hand, and computer scientists on the other. This collaboration continues not only in shaping scientific approaches that offer new perspectives on the issues, but also in developing the necessary technical and digital tools, and in creating web platforms that make both the research results and the tools themselves accessible to the scholarly community.

Acknowledgements.

This research was carried out with partial funding from the Bulgarian Scientific Fund under research project No КП-06-Н90/9/11.12.2024 “Development of the Cyrillic alphabet from the 9th to the 14th century in the South Slavic lands: research and digital presentation”.

References

14th Century South Slavonic Scribes and Scriptoria (n.d.). *Fourteenth Century South Slavonic Scribes and Scriptoria (palaeographic attribution and online repertorium)*. https://kopisti14.kmnc.bg/

Arefiev, A. L. (2022). Kirilitsa v geolingvisticheskom prostranstve [Cyrillic in the geolinguistic space]. *Vestnik Rossiyskoy akademii nauk*, *92*(3), 238–245.

Karskii, E. F. (1928). *Slavyanskaya kirilovskaya paleografiya* [Slavonic Cyrillic palaeography]. Leningrad.

Luchev, D., Goynov, M., Paneva-Marinova, D., Pavlov, R., & Rangochev, K. (2024a). Repertoire of medieval South Slavic manuscripts and scribes in research context. *Chuzhdoezikovo Obuchenie – Foreign Language Teaching*, *51*(1), 75–89. https://doi.org/10.53656/for2024-01-09

Luchev, D., Goynov, M., Paneva-Marinova, D., Pavlov, R., Rangochev, K., & Zlatkov, L. (2024b). Digital tools for medieval South Slavic manuscripts research. In *Proceedings of the International Conference on Computer Systems and Technologies 2024 (CompSysTech '24)* (pp. 102–108). ACM. https://doi.org/10.1145/3674912.3674946

Scarpa, M., Riparante, M., & Paneva-Marinova, D. (2023). Online database of 14th-century South Slavonic manuscripts. Research results and perspectives. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *13*, 39–44. https://doi.org/10.55630/dipp.2023.13.3

Shtepkin, V. N. (1967). *Ruskaya paleografiya* [Russian palaeography]. Moscow.

Sreznevskii, I. I. (1885). *Slavyano-ruskaya paleografiya XI–XIV v.: Lektsii v Stankpeterburg university v 1865–1880 g.* [Slavonic-Russian palaeography XI–XIV c.: Lectures at St. Petersburg University in 1865–1880]. St. Petersburg.

The Repertorium of South Slavonic Manuscripts and Copyists (n.d.). *Digital catalog of South Slavonic Manuscripts and Copyists (14th c.)*. https://kopisti.kmnc.bg/bg

Received: March 10, 2025
Reviewed: April 14, 2025
Finally Accepted: June 10, 2025