Transcribing Historical Population Sources Written in Cyrillic: Methodological Challenges in Training HTR Models for Romanian Parish Registers in Transylvania

Daniela Mârza

Babeş-Bolyai University, Centre for Population Studies, Cluj-Napoca, Romania, elena.marza@ubbcluj.ro

Abstract. The large-scale digitization of archival holdings has created new opportunities for historical population research, but effective access to handwritten sources remains limited due to the absence of reliable automatic transcription tools. This paper presents the training of a Handwritten Text Recognition (HTR) model for the automatic transcription and transliteration of Romanian parish registers written in Cyrillic characters, a category of sources that constitutes a substantial yet difficult-to-access component of modern Romanian documentation. Focusing on Orthodox parish registers from Transylvania dating from the early nineteenth century, the paper combines a historical overview of Romanian Cyrillic writing with a methodological discussion of transliteration and transcription practices, followed by an empirical assessment of trials conducted using the *Transkribus* platform. The results obtained so far reveal high Character Error Rates and demonstrate that the standard HTR training workflow is insufficient for producing a functional automatic solution for this type of material. The main obstacles arise from the lack of orthographic standardization, the structural mismatch between the Cyrillic alphabet and the Romanian language, graphic polysemy, abbreviations, superscriptions, irregular spacing, and significant variation in handwriting. These difficulties show that transcription in this context cannot be fully automated and must instead be approached as a hybrid, semi-automatic process that integrates HTR, rule-based transliteration, lexical validation, and sustained human intervention.

https://doi.org/10.24193/RJPS.2025.2.05
© Centre for Population Studies

By documenting both the progress achieved and the limitations encountered, this article contributes to ongoing debates in digital humanities and historical demography regarding the applicability of artificial intelligence to complex historical sources. It argues that, despite current constraints, even imperfect automatic transcriptions can significantly enhance accessibility and research efficiency, provided their use is methodologically transparent and critically informed.

Keywords: automatic transcription, handwritten text recognition (HTR), Romanian Cyrillic, parish registers, historical demography, digital humanities, Transkribus

1. Introduction

Recent advances in digital technologies have opened new possibilities for developing applications that support the automatic transcription of parish registers, thereby facilitating faster data extraction and analysis. In this context, the Center for Population Studies at Babeş-Bolyai University in Cluj[1] is involved in a series of projects to expand the historical population database it has built and manages, the Historical Population Database of Transylvania[2] . The most recent project, *From parish registers to digital infrastructures: the development a HTR solution for automatic transcription of civil status church books* focuses on developing tools for the automatic transcription of parish registers from Transylvania, handwritten in Romanian (using both Latin and Cyrillic scripts), Hungarian, German, and Latin. One of the main challenges of this initiative is developing an application capable of transcribing registers written in Romanian using Cyrillic characters.

The training of this model is primarily aimed at developing a tool for transcribing, more precisely, transliterating, historical demographic sources, particularly Romanian parish registers written in Cyrillic characters. At the same time, as a language model, its applicability extends to any handwritten document in Romanian using the Cyrillic script. Because most collection inventories held at the county branches of the National Archives have not yet been digitized, it is impossible to estimate the exact volume of such documents. However, the use of the Cyrillic alphabet until the mid-19th century in Moldova and Wallachia, as well as in Romanian communities in Transylvania (Grama 1989), points to a very large body of sources for which no automatic transcription solution currently exists. The future development of a tool capable of efficiently transcribing these materials is therefore of great

[1] https://csp.centre.ubbcluj.ro/

[2] https://hpdt.ro/

importance, as it will significantly facilitate access to documents as they are digitized by the National Archives[3] and eliminate the need for specialized knowledge of Cyrillic paleography.

This article aims to share with the scientific community the experience gained to date in training a language model for the transcription of parish registers written in Romanian using the Cyrillic alphabet, with particular emphasis on the difficulties and challenges encountered in this process. It includes a brief overview of the historical circumstances that led to Romanian, a Romance language, being written for centuries in Cyrillic characters, followed by a discussion of the specific features of this transcription task and the methods employed to date. This theoretical section is followed by the specific case of the transcription of parish registers in Transylvania, detailing the methodology employed, the results obtained to date, and the difficulties encountered.

1.1. Romanian language written with cyrillic characters

The use of the Cyrillic alphabet for writing Romanian emerged from a historical and cultural context shaped by the integration of medieval Romanian territories into the Byzantine-Slavic sphere of influence. Although Romanian is a language of Latin origin, its written tradition developed primarily within an Eastern Orthodox environment, in which Church Slavonic functioned for centuries as the language of worship, administration, and learned culture. Along with the adoption of Church Slavonic, Romanian communities also adopted the Cyrillic alphabet, which gradually became the principal means of recording both Slavonic and Romanian texts (Boroianu 1971). This adoption was a direct consequence of integration into the Orthodox religious and cultural tradition. Within this context, Cyrillic writing provided a stable framework for the development of Romanian written culture at a time when the Latin alphabet was no longer in common use in Orthodox Southeast Europe (Bianu and Cartojan 1940).

The Cyrillic alphabet used in Old Romanian writing was largely derived from the Slavonic alphabet, itself based on the Greek majuscule model. Its graphic inventory comprised approximately 40-43 letters, exceeding the phonetic requirements of the Romanian language (Boroianu 1971). This surplus resulted in an unstable grapheme–phoneme relationship: the same sound could be represented by multiple letters, while a single letter could carry different phonetic values depending on context and local writing traditions.

[3] https://descopera.arhivelenationale.ro/.

In its adaptation to the Romanian language, the Cyrillic alphabet was employed in only a partially functional manner, retaining signs with primarily etymological or traditional value rather than clear phonetic correspondence. At the same time, certain letters were reinterpreted or used conventionally to represent sounds specific to Romanian. This situation accounts for the heterogeneous character of Romanian Cyrillic writing and the difficulties involved in the analysis and transcription of early texts (Bianu and Cartojan 1940).

Romanian Cyrillic writing did not constitute a unified or rigid system but rather a set of graphic practices in continuous evolution. Abbreviations, ligatures, superscript letters, and the use of titla are common features, employed to save space or accelerate the writing process (Bianu and Cartojan 1940; Dragnev and Gumenâi 2003).

The Cyrillic alphabet was used to write Romanian for approximately five centuries, from the earliest known Romanian texts in the sixteenth century until its official replacement by the Latin alphabet in the second half of the nineteenth century. This long period of use explains why the majority of medieval, pre-modern, and early modern Romanian documents are written in Cyrillic (Boroianu 1971).

Documentary evidence further indicates that, even in Transylvania, Orthodox Romanians in rural areas continued to use the Cyrillic alphabet predominantly until relatively late, around 1860, despite being familiar with the Latin alphabet and aware of the Latin origin of their language (Grama 1989). Romanian Cyrillic script was employed mainly in ecclesiastical contexts, within bishoprics, deaneries, and parishes, where it circulated alongside ecclesiastical Latin. From a graphic perspective, it developed under the influence of contemporary Latin handwriting, particularly Mercantile cursive, adapted to the Cyrillic alphabet (Vîrtosu 1968).

2. Methods and techniques used in transliterating the Cyrillic alphabet into the Latin alphabet

Recent research on the automatic transliteration of the Cyrillic alphabet into the Latin alphabet for historical documents converges on the view that this task extends well beyond simple graphemic conversion. Instead, it must be understood as a complex process situated at the intersection of historical variation in writing systems, linguistic ambiguity, and the inherent limitations of automatic text recognition technologies. The literature consistently emphasizes the distinction between transliteration, focused on formal correspondence and reversibility, and transcription (practical or interpretative), which prioritizes

readability and accessibility for contemporary readers (Vakulenko 2024). This conceptual distinction has direct implications for the design of automated systems, the criteria used for their evaluation, and the ways in which their outputs are employed in historical research.

Developing an automatic transliteration solution for converting Cyrillic into Latin scripts presents numerous challenges. Some arise from structural asymmetries between the two writing systems: historical Cyrillic features a richer graphemic inventory than the Latin alphabet and includes signs without direct equivalents, as well as characters with diacritical, numerical, or symbolic functions. As a result, strict transliteration often requires compressing multiple Cyrillic graphemes into a single Latin letter or introducing artificial digraphs and diacritics. Both strategies compromise one-to-one correspondence and may introduce additional ambiguities. Moreover, while specialized diacritics are common in scholarly transliteration systems, their use remains limited in administrative and digital environments, where restricted ASCII character sets are generally preferred (Iliev 2013; Vakulenko 2024).

These challenges are further compounded by graphic polysemy and contextual dependence. In many Cyrillic traditions, a single grapheme may assume different values depending on its position within a word, its graphical environment, or prevailing writing conventions, while identical graphic sequences may allow for distinct phonetic or morphological interpretations. As a result, transliteration alone cannot resolve ambiguity without recourse to interpretative mechanisms, transforming what might appear to be a technical conversion into a fundamentally hermeneutic act (Cristea et al. 2023).

Such constraints are intensified by historical and regional variation. Cyrillic alphabets are not static systems but historical constructs subject to change, including the disappearance of letters and shifts in phonetic values. Moreover, differences among national and regional traditions (Bulgarian, Serbian, Russian, etc.) mean that a transliteration system appropriate for one period or area may prove inadequate for another. Any conversion process that ignores temporal stratification therefore risks projecting anachronistic values onto the source text (Iliev 2013; Cristea et al. 2023).

Additional complexity arises from the influence of extralinguistic factors: legal, administrative, political, cultural, and technological, which may subordinate linguistic criteria to external objectives such as international readability, bureaucratic standardization, or digital interoperability. This interplay helps explain the coexistence of multiple competing systems for the same language or alphabet, as well as the resulting normative fragmentation (Vakulenko 2024, Babayev 2025, Zagórski 2015).

In the specific case of Romanian, transliteration challenges are inseparable from the historical evolution of writing practices. The Cyrillic alphabet was used in Romanian territories for more than four centuries, initially for Slavonic liturgical and administrative texts and later for texts in Romanian itself, without the establishment of a standardized Cyrillic norm. Instead, Romanian Cyrillic writing emerged through the gradual adaptation of Slavonic graphemes to a Romance language with a distinct phonological structure (Cristea et al. 2020; Cristea et al. 2023).

The fundamental difficulty lies in the fact that the Cyrillic alphabet was not designed to represent the phonetic system of the Romanian language. Adopted as an external writing tool and only imperfectly adapted, it resulted in an unstable grapheme–phoneme relationship, so that conversion into the Latin alphabet necessarily entails methodological choices that shape the linguistic interpretation of the text (Bianu and Cartojan 1940). Until the early nineteenth century, Romanian written in Cyrillic coexisted with Slavonic, and the shift to the Latin alphabet occurred gradually through a transitional phase in which Cyrillic and Latin characters appeared side by side within the same texts. This transition was neither linear nor uniform but unfolded through a plurality of graphic solutions influenced by authorial practice, typographic conventions, regional variation, and cultural traditions, without the emergence of a single stable norm (Frîncu et al. 2023, Rebeja 2023).

Within this context, the same Cyrillic letter may correspond to different phonetic values, while the same sound may be represented by multiple graphemes. Such instability severely limits the effectiveness of fixed transliteration tables and necessitates contextual or interpretative mechanisms (Cristea et al. 2020; Frîncu et al. 2023). The difficulty is amplified by the use of abbreviations, overwritings, and ligatures, as well as by the presence of letters without a clear phonetic value (in final position or within a word), elements specific to medieval practices, which cannot be mechanically transposed into the Latin alphabet without losses or interpretative interventions (Bianu and Cartojan 1940).

Before the advent of automated solutions, when transcription was carried out manually, Romanian philological practice developed several approaches for rendering Cyrillic texts in the Latin alphabet. On the one hand, strict transliteration, applied letter by letter, seeks to establish a fixed correspondence between Cyrillic signs and Latin characters. Its principal advantage lies in reversibility, as the Latin version can, at least in theory, be reconstructed back into its original Cyrillic form (Boroianu 1971). On the other hand, interpretative transcription relies on linguistic analysis and aims to

reproduce the presumed pronunciation of the language at the time of writing. While this approach enhances readability, it also introduces a degree of subjectivity and depends heavily on the editor's expertise in historical phonetics and dialectology (Dragnev and Gumenâi 2003).

An intermediate and often regarded as more rigorous solution consisted of publishing a facsimile alongside a transliteration or transcription, thereby enabling direct comparison with the original document and limiting editorial intervention (Bianu and Cartojan 1940). In practice, however, the conceptual distinction between transliteration and transcription, though theoretically fundamental, was frequently blurred, resulting in inconsistent and at times contradictory editorial practices in the history of Romanian Cyrillic text editions. From a methodological standpoint, the polysemy of Cyrillic letters constitutes a major obstacle. A single letter may correspond to multiple phonetic values depending on context, while the same sound may be represented by different graphemes. Some signs function both as vowels and semivowels or as components of diphthongs, whereas others serve primarily etymological purposes without a direct counterpart in Romanian pronunciation (Boroianu 1971).

This polysemy is structural in nature, arising from the adaptation of a Slavic graphemic system to a Romance language in the absence of a coherent orthographic reform, a process that produced persistent functional overlaps. It manifests through the multiplicity of phonetic values associated with a single grapheme, contextually conditioned alternations (such as position within the word or graphemic proximity), and traditional non-phonetic conventions inherited from Slavonic practice (Cristea et al. 2020).

Advances in automatic transcription technologies have prompted a series of initiatives aimed at developing solutions for the automatic transliteration of the Cyrillic alphabet into Latin. Recent research has shown that the effectiveness of automatic transliteration is strongly constrained by the performance of earlier stages in the digitization pipeline, including scanning, segmentation, and recognition. To address these challenges, modular architectures have been proposed that integrate image preprocessing, character detection and classification, lexical validation, and transliteration, thereby managing the specific characteristics of historical Cyrillic documents and limiting error propagation. In this context, it is consistently emphasized that any recognition error, whether at the character or word level, is directly transferred to the transliteration output, affecting both the Character Error Rate (CER) and the Word Error Rate (WER) of the final text (Cristea et al. 2023).

The limitations of traditional OCR applied to historical Cyrillic are well documented in studies on the digitization of early printed materials. Commercial OCR solutions often fail to cover the full range of symbols found in Romanian prints and manuscripts, necessitating complex configurations, specialized fonts, or hybrid character sets. Even under such conditions, recognition quality may remain inadequate for effective proofreading, underscoring the need to evaluate transliteration not in isolation but in relation to the entire technical workflow and the required level of human intervention (Benţa et al. 2020).

In the case of manuscripts, increasing attention has been directed toward Handwritten Text Recognition (HTR), particularly through the use of the Transkribus platform. Unlike OCR, HTR operates at the line and word levels and incorporates contextual information, which can significantly improve recognition accuracy for cursive handwriting (Malahov et al. 2017).

The performance of Handwritten Text Recognition (HTR) systems is highly dependent on both the volume and the specificity of the training data. Models trained on printed texts or on other languages or variants of Cyrillic have proven ineffective for historical Romanian manuscripts. Although reported Character Error Rate (CER) values may appear acceptable in controlled experimental settings, they remain highly sensitive to variations in script, period, and writing style, which significantly limits model transferability (Burlacu and Rabus 2021). Some studies report CER values of approximately 5–10% for carefully selected Cyrillic manuscript corpora, provided that sufficiently large and appropriate training datasets are available. Performance improves markedly when the corpus is paleographically homogeneous (e.g., produced by a single scribe or representing a single writing type) and declines substantially for heterogeneous materials (Tikhonov et al. 2022). Moreover, the manual creation of ground truth data is resource-intensive, requiring up to approximately 30 minutes per page in some cases, which poses a serious constraint on scalability (Frîncu et al. 2023).

Within this research landscape, several methodological approaches to transliteration and transcription have been explored. Deterministic, rule-based methods define explicit grapheme-to-grapheme correspondences, prioritizing reversibility and formal consistency. While such approaches are suitable for diplomatic transliteration, they fail to resolve graphemic polysemy without human intervention and tend to yield texts that are formally accurate but difficult to read (Cristea et al. 2023). Interpretative transcription, whether phonetic or morpho-lexical, relies on contextual analysis (including position within the word, graphemic proximity, morphological structure, and sometimes

semantic considerations). Although indispensable for editions aimed at readability and linguistic analysis, this approach is difficult to automate and, due to its lack of standardization, must be explicitly documented to ensure transparency and reproducibility (Gîfu and Onofrei 2014; Frîncu et al. 2023).
Hybrid approaches, widely adopted in philological research, attempt to reconcile these perspectives by producing parallel layers: a diplomatic, reversible transliteration alongside an interpretative, reader-oriented transcription. When aligned in a digital environment, these layers reduce information loss and allow users to navigate between formal fidelity and linguistic interpretation (Petic and Gîfu 2014, Frîncu et al. 2023; Burlacu and Rabus 2021).

For manuscript materials in particular, HTR, most notably through platforms such as Transkribus, has become the dominant transcription method. HTR systems primarily generate raw transcriptions based on sequence recognition, which subsequently require additional stages of transliteration and interpretative processing (Malahov et al. 2017, Burlacu and Rabus 2021). Recent approaches increasingly combine deep learning techniques with symbolic resources, such as historical dictionaries and morphological rules, within pipeline architectures designed to support validation and disambiguation (Cristea et al. 2023). Where parallel Cyrillic–Latin texts are available, through modern editions or contemporary variants, their alignment enables the identification of graphic conventions, the detection of editorial interventions, and the refinement of transliteration rules. In such cases, diachronic lexical resources play a crucial role in validating automatic outputs and improving system performance (Petic and Gîfu 2014, Malahov et al. 2017).

Regarding the results and limitations of existing transcription methods, recent studies report steady progress. For printed texts, the adaptation of OCR systems, such as ABBYY FineReader and Tesseract, through training sets tailored to specific historical periods and typographic conventions can, in controlled corpora, achieve Character Error Rate (CER) values below 1–5% and Word Error Rate (WER) values of approximately 3%. For manuscripts, HTR approaches can yield CER values of around 5–10%, provided that sufficiently large training datasets (on the order of tens of thousands of words) are available and that historical dictionaries and iterative correction procedures are integrated into the workflow (Malahov et al. 2017; Burlacu and Rabus 2021).

To further improve accuracy, several researchers emphasize the need to develop annotated corpora and modular infrastructures that clearly separate processing stages, such as preprocessing, recognition, classification, linearization, transliteration, and lexical validation, while enabling alignment across levels (image–text–transliteration–interpretative transcription) (Cristea et al. 2023). Nevertheless, the limitations of these approaches remain structural: graphemic polysemy, diachronic variation, and the absence of a standardized orthography continue to preclude fully automatic and unambiguous transliteration solutions (Cristea et al. 2023, Petic and Gîfu 2014).

HTR systems for Cyrillic manuscripts operate on continuous sequences (lines or words) using recurrent and convolutional neural networks, an approach that is particularly well suited to manuscripts characterized by calligraphic variation, ligatures, and abbreviations. The Cyrillic script is often treated as a special case due to its graphemic instability and the high density of paleographic variation. A significant innovation in recent years has been the development of multilingual HTR models (e.g., trained on German, Russian, Serbian, or Ottoman corpora), which learn from heterogeneous datasets in which identical graphemes may carry different values. These models can achieve performance levels comparable to monolingual systems, while offering improved generalization for mixed, transitional, or poorly standardized documents.

The typical workflow for such approaches includes digitization and image preprocessing (e.g., skew correction and contrast enhancement), segmentation, recognition (often implemented via Transkribus), lexical or statistical post-correction, and subsequent conversion through transliteration or interpretative transcription. HTR systems generally produce a raw diplomatic transcription, with conversion into the Latin alphabet handled as a separate, context-dependent step. Some studies report CER values in the range of 5–10% for Russian Cyrillic and Church Slavonic manuscripts under conditions of adequate training data, with higher accuracy observed in paleographically coherent corpora and an unavoidable reliance on human expertise for correction and validation (Tikhonov et al. 2022).

Overall, existing research agrees that automatic Cyrillic–Latin transliteration for historical documents is feasible only as a semi-automatic process, in which OCR or HTR, transliteration rules, lexical validation, and human linguistic intervention are integrated within a single workflow. Consequently, it is not possible to develop a universal or definitive solution; rather, transliteration can only be achieved through contextualized and inherently hybrid approaches.

2.1. Transkribus: overview and functional principles

Handwritten Text Recognition (HTR) has rapidly evolved from an experimental technique into a research infrastructure, largely due to advances in deep learning. Its principal value for archival work lies in scalability: collections that once required years of manual transcription can now be processed far more quickly, even if the resulting transcriptions still require human correction. At the same time, HTR functions as a collaborative practice grounded in the creation of "ground truth": users train models, the models accelerate transcription, and subsequent human corrections further refine model performance. This iterative cycle shifts scholarly effort away from exhaustive manual transcription toward validation, interpretation, and analysis.

From an access perspective, HTR-generated transcriptions enable full-text search and contribute to the democratization of archival materials, provided that digital literacy and methodological transparency are ensured. At the same time, several risks must be acknowledged, including strong dependence on training data, the potential reproduction of biases, performance variability across authors, periods, and document types, and the concentration of errors in rare forms, abbreviations, and paleographic variants. These risks underscore the need for auditability and thorough documentation of both data and models.

Moreover, while access to large volumes of automatically transcribed texts facilitates quantitative and comparative analyses, it may also marginalize close philological reading if not accompanied by critical reflection. For this reason, the literature recommends transparent and well-documented workflows, the involvement of specialist communities in the definition and evaluation of models, and the careful maintenance of distinctions between automatic transcription, critical edition, and historical interpretation (Terras, 2022).

Transkribus[4] is one of the applications that plays a central role in contemporary transcription efforts. The platform emerged in response to a pressing need: while the large-scale digitization of manuscript collections in libraries and archives has dramatically increased access to digital images of historical documents, it has not resolved the problem of effective access to their textual content. In this context, Handwritten Text Recognition (HTR) has become an essential technology for converting manuscript images into machine-processable text, and Transkribus currently represents the principal platform through which HTR is operationalized in the field of cultural heritage. As such, it has become the dominant tool used in humanities and archival

[4] https://app.transkribus.org/

research. Transkribus was initially developed within the European research projects tranScriptorium (FP7) and READ – Recognition and Enrichment of Archival Documents (Horizon 2020), with the goal of creating a public, scalable infrastructure for the automatic recognition of historical handwritten texts. The platform was later taken over and further developed by READ-COOP, operating as a community-supported service model (Muehlberger 2019). From a conceptual standpoint, Transkribus provides an integrated environment that brings together image processing, machine learning, and human intervention, thereby supporting the transcription, full-text search, and analysis of large-scale manuscript corpora (Caers 2024).

The HTR technology implemented in Transkribus is based on deep neural networks that model text recognition as a sequential process. Unlike traditional OCR, which focuses on identifying isolated characters, HTR operates at the level of entire text lines, extracting complex visual features and estimating probabilities for character and word sequences. This approach is particularly well suited to historical manuscripts, which are characterized by graphic variation, abbreviations, ligatures, and a lack of orthographic standardization.

Over time, Transkribus has successively integrated different generations of HTR engines, ranging from convolutional architectures combined with LSTM networks to more recent Transformer-based models pre-trained on large volumes of historical data. Model performance is primarily evaluated using the Character Error Rate (CER), the standard metric in automatic text recognition. A crucial feature of HTR is its relative independence from language: recognition accuracy depends less on linguistic factors than on the availability of sufficiently large and representative training data for a given type of handwriting or document (Nockels et al. 2022).

The use of Transkribus follows a standardized workflow typical of supervised machine learning. First, users upload digital images of documents, which are then segmented into text regions and lines. While segmentation can be performed automatically, it is often manually corrected to improve recognition accuracy. The next step involves creating a reference dataset ("ground truth") consisting of manual transcriptions aligned with the corresponding image segments; this dataset is used to train an HTR model tailored to a specific corpus, author, period, or document type.

Once trained, the model can be reused and shared with other users, contributing to the cumulative development of community resources. It can then be applied to similar documents to generate automatic transcriptions, which may be further corrected and refined. The resulting texts become fully

searchable and can be exported for large-scale linguistic, statistical, or historical analysis (Leifert et al. 2024).

3. Training an HTR model for the transcription of Romanian parish registers from Transylvania written in Cyrillic script

The difficulties encountered in training a language model for the automatic transliteration of the Cyrillic alphabet into the Latin alphabet for Romanian are comparable to those identified in earlier attempts discussed above. Although still at an early stage, this process has already brought to light the full range of challenges inherent in such an undertaking.

For model training, a corpus of 50 pages comprising a total of 7,546 words was used. The material was drawn from two parish registers: a death register from the Orthodox parish of the Holy Trinity in Cluj, covering the period 1813–1858, and a baptism register from the Orthodox parish of St. Nicholas in the Şcheii Braşovului district of Braşov, dating from 1812 to 1845. The Cluj register is available at the Archives, where it was photographed, while the Braşov register is available online on the National Archives portal[5] . These registers were selected for two main reasons. First, they constitute key sources for research on population dynamics in major urban centers of Transylvania conducted at the Center for Population Studies. Second, their internal complexity, characterized by multiple handwriting styles, makes them particularly suitable as training data. Although the registers are structured in tabular format, this feature is not relevant for training a linguistic model. Overall, the training corpus includes six distinct handwriting styles, and all entries are in Romanian written in Cyrillic characters.

Transkribus was used to transcribe the selected register pages. To date, seven training cycles have been completed, with Character Error Rate (CER) values ranging from 28.95% to 52.35% in the most recent iteration. The model was trained to perform transliteration from Cyrillic to Latin. These high CER values indicate that the model is still far from achieving functional performance.

Transkribus includes two public models for transcribing documents written in Romanian using Cyrillic characters. The first, *19th-century Romanian Transitional Script - GT corrected*[6] , was trained on the transitional alphabet (used during the transition from the Cyrillic to the Latin alphabet and containing a mixture of Cyrillic and Latin letters); in addition, it used printed documents rather than handwritten ones. The second model, *RTA2 (Romanian Transition*

[5]https://descopera.arhivelenationale.ro/cota/?cid=10254797
[6]https://app.transkribus.org/models/text/117113

Alphabet)[7], also refers to the transitional alphabet. These two models were tested and found to be unsuitable for training a model for parish registers, even as a starting point.

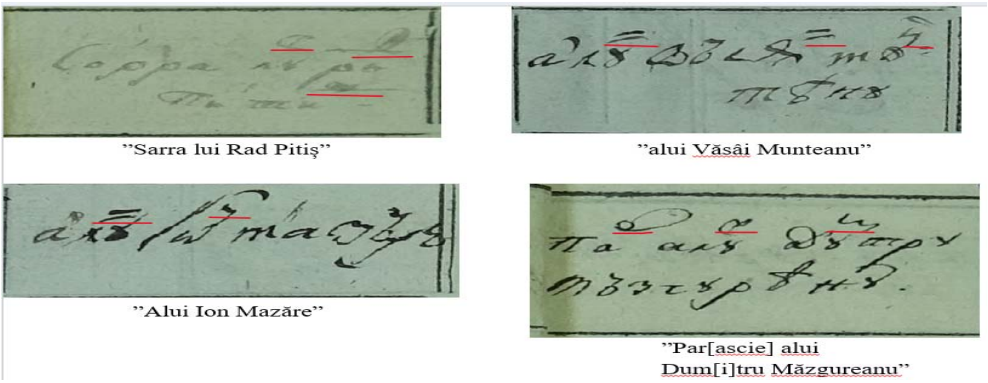
To date, there is no effective public solution for the automatic transcription of parish registers written in Romanian using Cyrillic characters. The difficulties involved in training a model for this type of source stem from multiple factors. Some of these are common to many archival documents, such as the poor physical condition of the registers, blurred or skewed images, and illegible handwriting. Others are specific to the relationship between the two alphabets.

From a theoretical standpoint, the transliteration of the Cyrillic alphabet into the Latin alphabet does not in itself pose major difficulties, as well-established manuals and clear rules exist for this process. A human operator with appropriate training can transcribe such documents without significant obstacles (Bianu and Cartojan 1940, Boroianu 1971, Dragnev and Gumenâi 2003, Vîrtosu 1968). The main challenges arise when attempting to train an application to automatically convert Romanian written in Cyrillic characters into Romanian written in Latin characters.

The principal challenges involved in the transcription of these registers can be summarized as follows:

1. Writing peculiarities, meaning the use of abbreviations and superscriptions specific to the Cyrillic alphabet, as can be seen below:

Figure 1. Examples of abbreviations and superscriptions specific to the Cyrillic alphabet



Source: Parish Baptism Register no BV-F-00259-1-00039, Parish Registers Collection, SJAN Braşov.

[7] https://app.transkribus.org/models/text/51515.

In cases of superscription, the established rule is that letters written above the line are treated, for transliteration purposes, as integral parts of the word, a convention consistently applied in the manual transcription of the registers. However, training an automatic model to recognize such cases correctly requires substantially more training data and iterations than are needed for regular transcriptions. Superscribed letters are not only positioned above the baseline, but often differ markedly in shape from their standard forms. For instance, in the example "Alui Ion Mazăre", the superscript forms of the letters *i* and *n* differ so significantly from their normal forms that they can effectively be considered distinct characters, requiring the model to be explicitly trained to identify and interpret them correctly.

Moreover, there are no standardized rules governing the use of superscription. Analysis of the registers shows that the decision to place certain letters above the line depends both on individual scribal preferences and on spatial constraints within the tabular layout, some letters are superscribed simply because the word would not otherwise fit within a cell. No consistent pattern of superscription could be identified across the two registers, further complicating the training and generalization of the model.

In addition, the existence of abbreviations in the records necessitates subsequent manual intervention in the transcribed text to complete the words.

2. *Poor spacing*, meaning the absence of spaces between words, as shown in the example below:

Figure 2. Examples of absence of spaces between words

"Dezmireanvasili din Cluj 43" "TăuțanCostandin = acum născut"

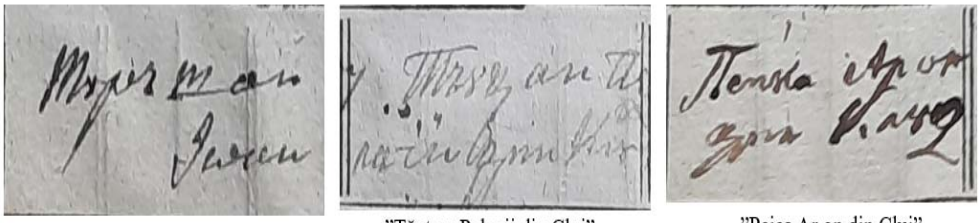
Source: Deaths Parish Register no 71/22, Parish Registers Collection, SJAN Cluj

In these examples, the names *Dezmirean* and *Vasili* in the first image should be separated by a space, as should *Tăuțan* and *Costandin* in the second. In such contexts, these forms typically correspond to given names and surnames. Given that the pool of first and last names within a historical community is

relatively limited, it is plausible that a linguistic model, if trained on a sufficiently large number of spelling variants, could learn to recognize these elements as distinct entities even when they are written without explicit spacing.

Conversely, the opposite phenomenon is also frequently observed: words are sometimes split by spaces that are large enough for Transkribus to interpret them as separate tokens, even though they belong to a single lexical unit, as illustrated below.

Figure 3. Examples of spaces within names

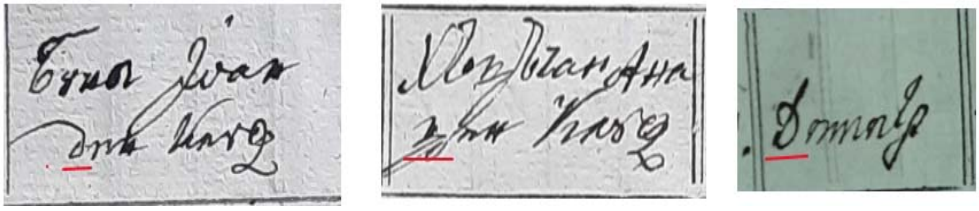


Source: Deaths Parish Register no 71/22, Parish Registers Collection, SJAN Cluj

These examples illustrate the presence of spaces within personal names such as *Murăşan*, *Tăuţan*, and *Aron*. At present, the Transkribus model interprets these segments as separate words, which necessitates additional manual post-processing after transcription. It remains an open question whether future training, incorporating a sufficiently large number of spelling and spacing variants of these names, will enable the model to correctly recognize them as single lexical units despite the internal spacing.

3. The same phoneme represented by several distinct graphemes, as illustrated by the letter *d*, which appears in three different graphic forms. These are not mere variants of the same sign but distinct graphemes, as shown below in the words *din*, *din*, and *Dumitr*, respectively.

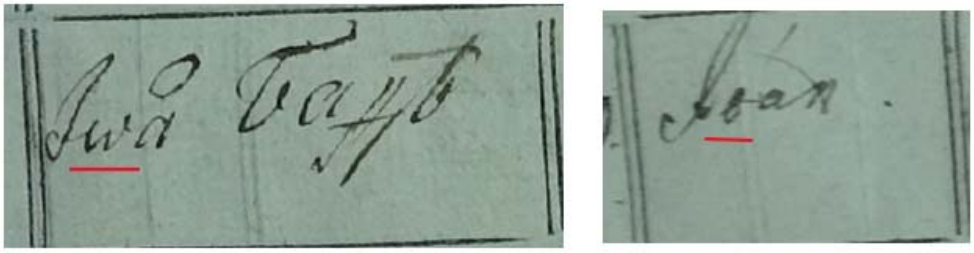
Figure 4. Examples of distinct graphemes for the same phoneme



Source: Deaths Parish Register no 71/22, Parish Registers Collection, SJAN Cluj, Parish Baptism Register no BV-F-00259-1-00039, Parish Registers Collection, SJAN Braşov.

In such cases, the Transkribus model must be trained to recognize these distinct signs as representing the letter *d* in transliteration.
A comparable situation arises with the letter *o*, which can likewise be rendered by two different graphemes, both of which appear in the present example in the name *Ioan*.

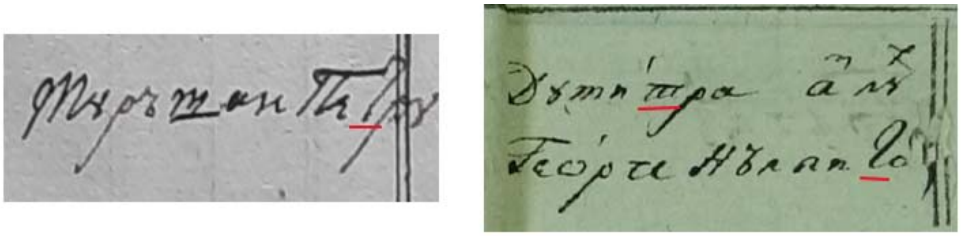
Figure 5. Examples of distinct graphemes for the same phoneme



Source: Parish Baptism Register no BV-F-00259-1-00039, Parish Registers Collection, SJAN Braşov.

A similar issue arises with the letter *t*, which is sometimes written in different graphic forms within the same table cell, as illustrated by the names *Murăşan Petru* and *Dumitra ai lui George Nălbitor*:

Figure 6. Examples of distinct graphemes for the same phoneme



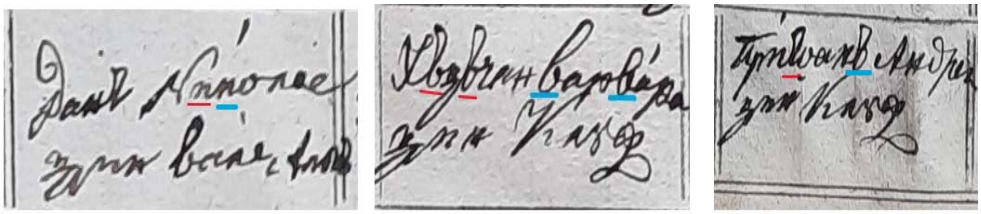
Source: Deaths Parish Register no 71/22, Parish Registers Collection, SJAN Cluj, Parish Baptism Register no BV-F-00259-1-00039, Parish Registers Collection, SJAN Braşov

In such cases, the variability occurs even at a very local level, further complicating automatic recognition and requiring the model to associate multiple distinct graphemic forms with a single Latin equivalent.

The existence of multiple graphic forms for the same letter reflects the historical evolution and internal variation of the Cyrillic alphabet. In the registers examined here, however, the choice between these forms does not appear to follow any consistent rule and seems to depend largely on individual scribal preference. As a result, no stable patterns emerge that would facilitate training the model to systematically recognize two or even three distinct graphemes as corresponding to a single letter in transliteration.

4) The use of identical or nearly identical signs to represent different letters. In some of the handwritings found in these registers, certain graphemes are visually very similar, or effectively indistinguishable, even though they correspond to different phonemes. This phenomenon, illustrated below, introduces an additional level of ambiguity that further complicates automatic recognition and transliteration.

Figure 7. Examples of identical looking graphemes for distinct phonemes



Source: Deaths Parish Register no 71/22, Parish Registers Collection, SJAN Cluj

Although the signs highlighted in different colors appear very similar, or even identical, they in fact represent distinct graphemes. In the first example, the sign underlined in red corresponds to the letter *n*, while the one in blue represents *c*. In the second example, the graphemes marked in red represent *ă*, whereas those in blue correspond to *v*. In the final example, the grapheme highlighted in red represents *g*, while the one in blue corresponds to *ă*.

A human reader with solid knowledge of Cyrillic paleography and of the Romanian language can usually infer the correct value of these signs intuitively, despite their graphic similarity. Training an automatic system to make such distinctions is considerably more challenging, and it remains an open question whether incorporating a sufficiently large number of such ambiguous cases into the training data will be sufficient to resolve this issue.

5) Difficulties caused by structural differences between the Cyrillic and Romanian alphabets. Although most Cyrillic signs can be transliterated unambiguously into the Latin alphabet, there are several cases in which a direct one-to-one correspondence does not apply. One such situation arises when a single Cyrillic grapheme corresponds to two Latin letters that also exist as independent units. For instance, the grapheme *ЮО* maps to the Latin sequence *iu*, while the individual letters *i* and *u* are represented in Cyrillic by distinct signs such as *ŭ*, *ĭ*, and *ѵ*, respectively.

While normative rules governing the use of these graphemes did exist, they were rarely observed in parish registers. Instead, scribes appear to have used these signs in an arbitrary manner, as illustrated in the example below.

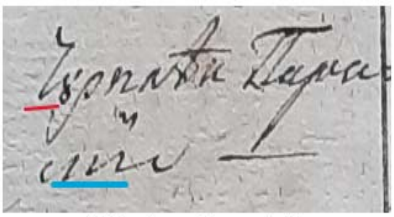
Figure 8. Examples of lack of correspondence grapheme-grapheme

Source: *Parish Baptism Register no BV-F-00259-1-00039, Parish Registers Collection, SJAN Braşov*

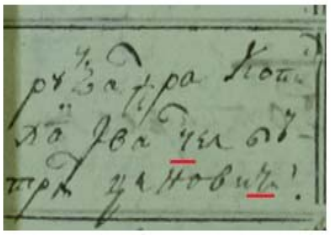
In such cases, the model must be trained to recognize not only the graphemes *ŭ*, *ĭ*, and *ѵ* as corresponding to *i* and *u*, but also the grapheme *ЮО* as representing the sequence *iu*. The task becomes even more complex when *ЮО*

is written with its two constituent elements spatially separated, as in the first example, *Iulie*. In this instance, the first component resembles the letter *i*, while the second resembles *o*. A human reader can readily infer the intended grapheme from context, whereas an automatic model is likely to transliterate these shapes separately as *i* and *o*, rather than correctly interpreting them as *iu*.
A similarly challenging case is presented by the grapheme *ч*, which corresponds to the sound *ci*, given that Cyrillic also includes distinct graphemes for the individual letters *c* and *i*. This overlap, illustrated in the following example, further complicates automatic recognition and transliteration.

Figure 9. Examples of lack of correspondence grapheme-grapheme



"Ciurilean Parascie"



"Ruxandra copila Ioan ciel bătrân Ţenovicii"

Source: Deaths Parish Register no 71/22, Parish Registers Collection, SJAN Cluj, Parish Baptism Register no BV-F-00259-1-00039, Parish Registers Collection, SJAN Braşov

In the first example, the sound *ci* is rendered in two different ways: once by the grapheme *ч* and once by the separate signs corresponding to the letters *c* and *i*. This situation highlights a further difficulty, namely the rendering of the sound *chi*, for which the Cyrillic alphabet provides no dedicated grapheme. Thus, the form transliterated here as *Parascie* corresponds in fact to the correct name *Paraschie*. According to established rules, the grapheme corresponding to the letter *c* may also be read as *ch* in appropriate contexts, a distinction that a human operator can usually infer without difficulty.

In the context of training a model in Transkribus, however, it is not feasible to systematically assign the value *ch* to the grapheme corresponding to *c*, since in many instances this grapheme represents only *c* and not the digraph *ch*. A comparable situation arises with the grapheme *z*, which corresponds to the letter *g* but can also be read as *gh*, as in the name *Gheorghe*. Within the application, it is likewise impractical to assign both values (*g* and *gh*) to this

grapheme, because in most cases it carries only the value *g* (for example, in the name *Gligor*). As a result, such cases inevitably require additional manual intervention after automatic transliteration in order to correct words in which *ꙁ* must be interpreted as *gh*.

Other cases in which a single Cyrillic grapheme corresponds to two Latin letters include *ѫ* (rendered as *ţi*), *ѣ* (*ea*), and *ѩ* (*ia*). These cases introduce the same transliteration difficulties, since the corresponding Latin letters (*t, e, a, i*) are also represented individually by distinct Cyrillic graphemes. While such situations are readily resolved by a human operator through linguistic intuition and contextual knowledge, they become highly problematic when formalized for training a linguistic model, where each grapheme must be assigned a stable and unambiguous value.

The results obtained so far suggest that the standard approach to model training in Transkribus may not, on its own, be sufficient to achieve a fully functional solution. Under normal circumstances, supplying the application with a sufficiently large volume of representative handwriting samples leads to progressively improved recognition accuracy. In the present case, however, the unstable correspondence between certain Cyrillic and Latin graphemes, together with the wide range of irregularities discussed above, significantly complicates this process.

Initial training attempts produced Character Error Rate (CER) values that could be considered optimistic given these constraints, with a best result of 28.95%. However, the subsequent inclusion of additional handwriting styles, each characterized by its own internal variability, overlapping letter forms, and distinct graphemic conventions, led to a marked increase in CER, rendering some model versions unusable. This outcome suggests, on the one hand, that achieving a model with acceptable accuracy will require a prolonged and iterative training process, and, on the other hand, that additional tools and strategies will need to be explored in order to support the recognition and disambiguation of particularly problematic characters.

4. Conclusions

This paper has explored both the challenges and the initial results of training an automatic transcription model for Romanian parish registers written in Cyrillic script, situating this work within the broader fields of historical population research and digital humanities. The experience gained so far confirms what recent international studies have already shown: the automatic transcription and transliteration of historical Cyrillic sources is a complex,

historically grounded process that sits at the crossroads of paleography, historical linguistics, and machine learning.

Romanian texts written in Cyrillic pose particularly serious difficulties. The absence of a standardized orthography, the imperfect fit between the Cyrillic alphabet and Romanian phonology, the frequent use of letters with multiple values, and the widespread presence of abbreviations, superscripts, and individual writing habits all create layers of ambiguity. While these are relatively easy for trained human readers to interpret, they are extremely hard to formalize for an automated system. Experiments conducted using Transkribus show that, despite the platform's advanced HTR capabilities, the standard training workflow is currently insufficient to produce a reliable model for automatically transliterating Romanian Cyrillic parish registers into the Latin alphabet. The relatively high Character Error Rates observed reflect not only the limited size of the training data, but also the structural constraints inherent in the sources themselves.

These results, however, provide empirical confirmation of the limits of full automation in this domain. The findings support the view that any effective approach must be hybrid and iterative, combining HTR with rule-based transliteration, lexical and onomastic checks, and sustained human involvement. In the case of parish registers, where formulaic language coexists with various personal names and diverse handwriting styles, the use of historical dictionaries, name lists, and context-sensitive post-correction tools appears indispensable. For historical demography and population research, even imperfect automatic transcriptions can be highly valuable, as long as their limitations are clearly acknowledged and they are treated as research aids rather than authoritative editions.

In the longer term, developing an efficient transcription tool for Romanian Cyrillic sources will require much larger and more consistent training corpora. Despite the challenges encountered, the potential gains are substantial. Such a tool would greatly improve access to a vast body of pre-modern Romanian sources, reduce reliance on specialized paleographic expertise, and facilitate the integration of these materials into large-scale historical population databases. This paper represents an early but necessary step in that direction, clarifying both the possibilities and the limitations of applying HTR technologies to one of the most demanding categories of Romanian historical documents.

Acknowledgement
This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project number PN-IV-P1-PCE-2023-0339, within PNCDI IV.

References

Babayev, J. (2025). "Orthographic Challenges in the Transliteration of Proper Names between the Languages with Different Spelling". *Acta Globalis Humanitatis et Linguarum* 2(4): 345-356.

Benţa, D., Bud, P., Platon, E., Paşca-Tuşa, S., Oneţiu, E., Mihăilă, A., & Floca, F. (2020). "Challenges in proofing the Cyrillic MCVRO resources – Equability between the technical component and the role of the researcher". *Philobiblon* 25(2): 337-353. https://doi.org/10.26424/philobib.2020.25.2.09

Bianu, I., & Cartojan, N. (1940). *Album de paleografie românească (scrierea chirilică)* (Ed. a III-a). Bucureşti: Cartea Românească.

Boroianu, C. (1971). *Texte vechi româneşti. Album de paleografie româno-chirilică*. Bucureşti: Universitatea din Bucureşti.

Burlacu, C., & Rabus, A. (2021). "Digitalizarea scrierilor cu alfabet chirilic (românesc) prin utilizarea platformei Transkribus: noi perspective". *Diacronia* 14 (A196): 1-10. https://doi.org/10.17684/i14A196ro.

Caers, B. (2024). "Teaching handwritten text recognition: Can new technologies save old skills?" *Quærendo* 54: 198–209. https://doi.org/10.1163/15700690-bja10024.

Cristea, D., Pădurariu, C., Rebeja, P., & Onofrei, M. (2020). "From scan to text: Methodology, solutions and perspectives of deciphering old Cyrillic Romanian documents into the Latin script". In *Knowledge, Language, Models,* pp. 38–56.

Cristea, D., Cleju, N., Rebeja, P., Haja, G., Coman, E., Vasilescu, A., Marinescu, C., & Dascălu, A. (2023). "Bringing the old writings closer to us: Deep learning and symbolic methods in deciphering old Cyrillic Romanian documents". *Memoirs of the Scientific Sections of the Romanian Academy* 46: 87–125.

Dragnev, D., & Gumenâi, I. (2003). *Paleografia slavo-română şi româno-chirilică*. Chişinău: CIVITAS.

Frincu, M., Frincu, S., & Penteliuc, M. E. (2023). "Challenges and solutions in transliterating 19th century Romanian texts from the transitional to the Latin script". In *Proceedings of the 4th Conference on Language, Data and Knowledge,*Vienna, Austria. NOVA CLUNL, Portugal, pp. 226–231.

Petic, M., & Gîfu, D. (2014). "Transliteration and alignment of parallel texts from Cyrillic to Latin". In *Proceedings of the International Conference on Language Resources and Evaluation (LREC),* pp. 1819–1823.

Grama, A. (1989). "Mesajul scrisului românesc din documente săteşti transilvănene (1810–1860)". *Revista Arhivelor* 51(4): 349–356.

Iliev, I. (2013). "Short history of the Cyrillic alphabet". *International Journal of Russian Studies* 2(2): 221-285.

Leifert, G., Romein, C. A., Rabus, A., Ströbel, P. B., & Hodel, T. (2024). *Transkribus and beyond: Pioneering the future of transcription technology*. Royal Netherlands Academy of Arts and Sciences.

Malahov, L., Colesnicov, A., Cojocaru, S., & Bumbu, T. (2017). "On recognition of manuscripts in the Romanian Cyrillic script". In *Proceedings of the Conference on Mathematical Foundations of Informatics (MFOI 2017)*. Chişinău, Republic of Moldova.

Muehlberger, G., et al. (2019). "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study". *Journal of Documentation* 75(5): 954–976. https://doi.org/10.1108/JD-07-2018-0114.

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). "Understanding the application of handwritten text recognition technology in heritage contexts". *Archival Science* 22: 367–392. https://doi.org/10.1007/s10502-022-09397-0.

Vasilescu, V. & Boiangiu, A. (1982). *Scrierea chirilică românească. Album de paleografie.* Bucureşti : Universitatea din Bucureşti.

Rebeja, P. (2023, November 6). *Digital analysis of old Romanian texts* (Extended abstract). Universitatea „Alexandru Ioan Cuza" din Iaşi. https://scdoc.info.uaic.ro/wp-content/uploads/2023/11/Rezumat-teza-EN_Rebeja-Petru.pdf

Terras, M. (2022). "Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription". In L. Jaillant (Ed.). *Archives, access and artificial intelligence: Working with born-digital and digitized archival collections*. Bielefeld University Press, pp. 179–200. https://doi.org/10.14361/9783839455845.

Tikhonov, A., Loew, L., Matić-Chalkitis, M., Meindl, M., & Rabus, A. (2023). "Multilingual handwritten text recognition (MultiHTR): Reading your grandma's old letters in German, Russian, Serbian, and Ottoman Turkish with artificial intelligence". In A. Schwan & T. Thomson (Eds.). *The Palgrave handbook of digital and public humanities*. Palgrave Macmillan, pp. 1–18. https://doi.org/10.1007/978-3-031-11886-9.

Vakulenko, M. (2024). "Transliteration of non-Latin texts: From everyday practice to linguistic technologies". *Proceedings of the World Conference on Foreign Language Education* 1(1): 1–11. https://doi.org/10.33422/worldfle.v1i1.545.

Vîrtosu, E. (1968). *Paleografia româno-chirilică*. Bucureşti: Editura Ştiinţifică.

Zagórski, B. R. (2015). *Difficult historical problems of transliteration and transcription in South-Eastern European toponomastic practice*. Paper presented at the UNGEGN–ECSEED Meeting, Ljubljana, Slovenia.