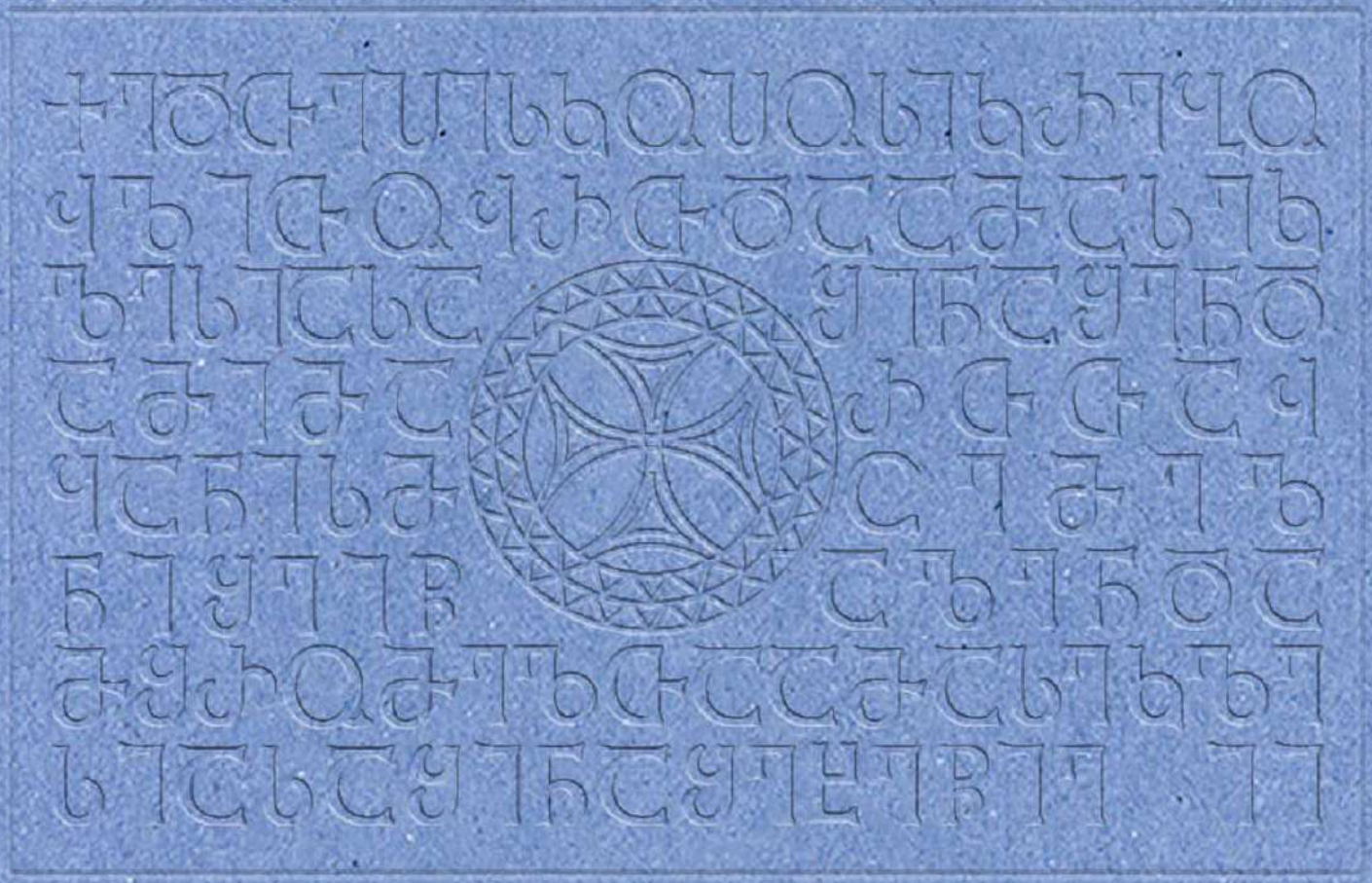
ACADEMY FOR DIGITAL HUMANITIES – GEORGIA

ციფრული ქართველოლოგია



Digital Kartvelology

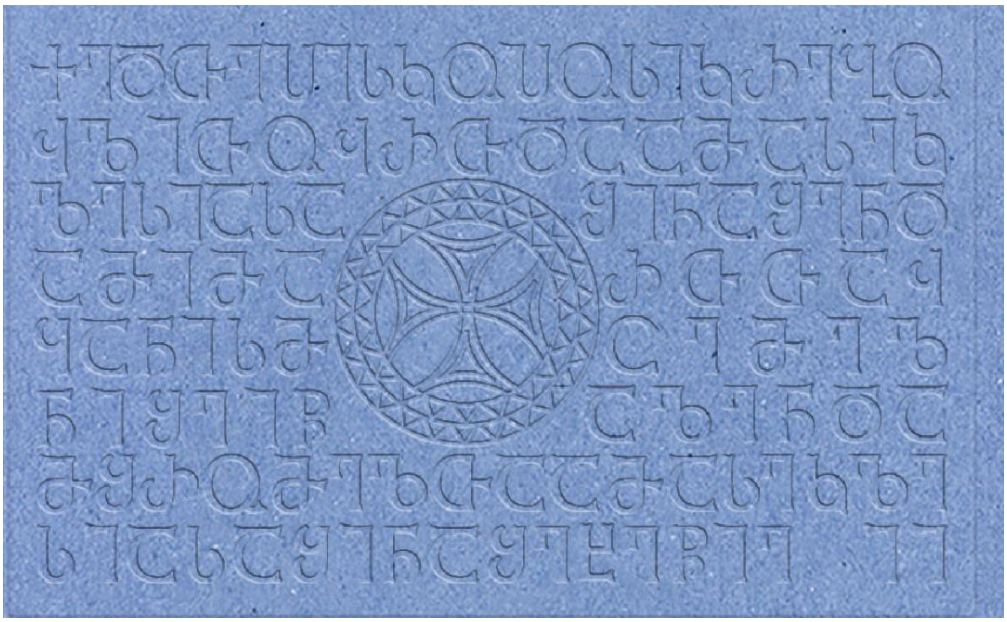
INTERNATIONAL BILINGUAL JOURNAL

VOLUME 3

The Bilingual Scientific Online Journal of the **Academy for Digital Humanities**

დიგიტალური ჰუმანიტარიის აკადემია - საქართველოს
ორენოვანი საერთაშორისო სამეცნიერო ონლაინჟურნალი

დიგიტალური ქართველოლოგია



Digital Kartvelology

Volume 3

Dedicated to the Memory of Iza Chantladze (1.12.1939 – 6.6.2024)
ეძღვნება იზა ჩანტლაძის (1.12.1939 – 6.6.2024) ხსოვნას

Tbilisi / თბილისი

2024

The journal **Digital Kartvelology** aims to promote the development of Digital Kartvelology and to create an academic platform for specialists working in Digital Humanities to publish scientific papers. The journal covers the following fields: Digital Humanities, Digital Kartvelology, Digital Caucasiology, Digital Lexicography, Translation Studies, Georgian Language Technologies, Corpus linguistics, Documentary Linguistics.

ჟურნალი **დიგიტალური ქართველოლოგია** მიზნად ისახავს დიგიტალური ქართველოლოგიის განვითარების ხელშეწყობას და წარმოადგენს აკადემიურ პლატფორმას აღნიშნული მიმართულების სპეციალისტებისთვის სამეცნიერო ნაშრომების გამოსაქვეყნებლად. ჟურნალში წარმოდგენილია შემდეგი დარგები: დიგიტალური ჰუმანიტარია, დიგიტალური ქართველოლოგია, დიგიტალური კავკასიოლოგია, დიგიტალური ლექსიკოგრაფია, თარგმანმცოდნეობა, ქართული ენის ტექნოლოგიები, კორპუსლინგვისტიკა, დოკულინგვისტიკა.

ISSN: 2720-8427 (online)

Digital Kartvelology, Vol. 3

დიგიტალური ქართველოლოგია, ტომი 3

Content/სარჩევი

I. Digital Manuscript Studies / ხელნაწერთა დიგიტალური შესწავლა

Jost Gippert (Hamburg), **Daniel Stökl Ben Ezra** (Paris) 5–24
Reading Georgian Manuscripts Automatically on the eScriptorium Platform
იოსტ გიპერტი (ჰამბურგი), **დანიელ შტეკლ ბენ ეზრა** (პარიზი)
ქართული ხელნაწერების ავტომატური ამოკითხვა eScriptorium-ის პლატფორმაზე
DOI: https://doi.org/10.62235/dk.3.2024.8508

Maia Matchavariani (Tbilisi) 25–40
Digital Humanities and the K. Kekelidze Georgian National Center of Manuscripts
მაია მაჭავარიანი (თბილისი)
დიგიტალური ჰუმანიტარია და კ. კეკელიძის სახელობის
საქართველოს ხელნაწერთა ეროვნული ცენტრი
DOI: https://doi.org/10.62235/dk.3.2024.8509

II. Digital Kartvelology / დიგიტალური ქართველოლოგია

Hélène Gérardin (Paris) 41–52
Building a Dialectological Lexicological Database of Georgian Cognates for Digital Analysis
ელენე ჟერარდინი (პარიზი)
დიალექტოლოგიური ლექსიკოლოგიური მონაცემთა ბაზის აგება
ქართული ენის ზიარი კოგნატების ანალიზისათვის
DOI: https://doi.org/10.62235/dk.3.2024.8511

Ketevan Datukishvili, Nana Loladze, Merab Zakalashvili (Tbilisi) 53–76
The Linguistic Model of the Georgian Language (Synthesis of Nominal Word forms)
ქეთევან დათუკიშვილი, ნანა ლოლაძე, მერაბ ზაკალაშვილი (თბილისი)
ქართული ენის ლინგვისტური მოდელი (სახელური სიტყვაფორმების
სინთეზი)
DOI: https://doi.org/10.62235/dk.3.2024.8512

III. Digital Rustvelology / დიგიტალური რუსთველოლოგია

Manana Tandaschwili (Frankfurt), **Mariam Kamarauli** (Hamburg) 77–108
Research into Equivalence in Multilingual Parallel Corpora
მანანა თანდაშვილი (ფრანკფურტი), **მარიამ ყამარაული** (ჰამბურგი)
ეკვივალენტობის კვლევისათვის მრავალენოვან პარალელურ კორპუსებში
DOI: https://doi.org/10.62235/dk.3.2024.8514

IV. Linguistics / ენათმეცნიერება

George Hewitt (London) 109–124
Paul's 1st Epistle to the Corinthians ch. 13. - A Comparison of Georgian and Abkhaz Translations (Taking into account the Greek Original)
ჯორჯ ჰიუიტი (ლონდონი)
პავლე მოციქულის პირველი ეპისტოლე კორინთელთა მიმართ - თავი 13 ქართული და აფხაზური თარგმანების შედარება (ბერძნული ორიგინალის გათვალისწინებით)
DOI: https://doi.org/10.62235/dk.3.2024.8516

Maia Lomia, Ketevan Margiani (Tbilisi) 125–136
The Theoretical Grounds and the Practical Value of Distinguishing the Markers of Evidentiality in the Verb Paradigms of the Kartvelian Languages
მაია ლომია, ქეთევან მარგიანი (თბილისი)
ქართველურ ენათა ზმნურ პარადიგმებში ევიდენციალობის მარკერის გამოყოფის თეორიული საფუძვლები და პრაქტიკული ღირებულება
DOI: https://doi.org/10.62235/dk.3.2024.8517

V. Lexicography / ლექსიკოგრაფია

Tinatin Margalitadze, Salome Tchighladze (Tbilisi) 137–150
Unknown Pages of English-Georgian Lexicography (George Ellis and his Comparative Lexicon of Kartvelian Languages)
თინათინ მარგალიტაძე, სალომე ჭიღლაძე (თბილისი)
ინგლისურ-ქართული ლექსიკოგრაფიის უცნობი ფურცლები (ჯორჯ ელისის ქართველური ენების შედარებითი ლექსიკონი)
DOI: https://doi.org/10.62235/dk.3.2024.8518

Reading Georgian Manuscripts Automatically on the *eScriptorium* Platform

Jost Gippert (Hamburg), Daniel Stökl Ben Ezra (Paris)

DOI: https://doi.org/10.62235/dk.3.2024.8508
jost.gippert@uni-hamburg.de || ORCID: 0000-0002-2954-340X
Daniel.Stoekl@ephe.psl.eu || ORCID: 0000-0001-5668-493X

Abstract: The article outlines the development of means for an automatic reading of Georgian manuscripts on the *eScriptorium* platform and the first results achieved with them. After an overview of the efforts undertaken in applying Optical Character Recognition (OCR) to Georgian printed books since the late 1980's and a short introduction into the basics of the *eScriptorium* approach to Handwritten Text Recognition (HCR) and its functionalities, it exemplifies the application of the three core procedures of *eScriptorium*, which consist in the automatic segmentation of text-covered regions and lines, the automatic transcription of the detected lines based on manual input and the training of appropriate models, and the alignment with existing electronic texts in order to provide reliable ground truth for further training. With a total of 292 manually transcribed pages and 7488 pages with aligned (but not yet always corrected) text that have been processed so far, there is a strong material basis for further improvement of the models and the reading results depending on them.

Keywords: Optical Character Recognition, Handwritten Text Recognition, Text corpora, Old Georgian, manuscripts

1. From OCR to HTR

Nothing has changed the work of linguists, philologists and other researchers dealing with textual data as drastically as the availability of large corpora that can be searched and analysed digitally.[1] In the compilation of such corpora, two types of text data must be distinguished: those that exist in electronic form right from the beginning (i.e. that are "born digital"), and those that must be transferred into digital form (i.e. "digitised") from other media, i.e. usually printed or handwritten sources.[2] The National Corpus of the Georgian language that has been compiled over the past 13 years (GNC, http://gnc.gov.ge) is a typical witness of this: the biggest of its subcorpora (the Georgian Reference Corpus [GRC], with 202,728,329 tokens) mostly consists of materials that were harvested in the World Wide Web, whereas the other subcorpora, especially those dealing with older stages of the language (GNC Old Georgian, with 7,101,021 tokens; GNC Middle Georgian, with 1,432,262 tokens; GNC Law texts, with 1,495,985 tokens), are more or less based on digitisations of printed matter that have been undertaken in the framework of the TITUS and ARMAZI projects.[3] Only in a few cases was the digitisation done manually, i.e. by typing a given text with the keyboard; in most cases the printed model was entered page by page via an optical scanner and then "read" into a digital text format via special software that was able to isolate the individual letters on the page, identify them with the

[1] On the general issue of text digitisation, see Stökl Ben Ezra (forthcoming). For an early programmatic approach concerning Georgian and other languages, see Gippert (1990).
[2] This includes text data from audiovisual media (e.g. recordings of spoken language) which must be transcribed before they can be integrated into corpora.
[3] See https://titus.uni-frankfurt.de/texte/texte2.htm#georgant and https://armazi.fkidg1.uni-frankfurt.de.

corresponding characters of the alphabet, and storing the sequences of letters as a coherent electronic text.

When the endeavours to compile a diachronic corpus of Georgian began in the late 1980s, this method, usually styled Optical Character Recognition (OCR), was by far not as well developed as it is today, and scanners were much less effective. The biggest problem consisted in the fact that personal computers of that time were not yet designed to deal with languages written in scripts other than Latin – they were first based on a 7-bit encoding system (ASCII) that comprised just the characters used in English, and this was only gradually substituted by an 8-bit system (ANSI) which also covered the "extra characters" necessary for "Western" languages like German ("umlaut characters") or French ("accented characters"). With the introduction of the international standard named ISO/IEC-8859, Cyrillic and Greek became applicable alongside "North", "Central" and "South European" Latin-based alphabets; however, the so-called "code pages" representing them did not imply a unique encoding, given that different characters had to be "mapped" across the set of 256 available code points (e.g., Greek *θ* shared its code [232] with Cyrillic *ш* and, depending on the code page, Latin *è* and *č*).[4] For scripts like Georgian, there was no support at all, which means that in the attempt of dealing with Georgian texts, a mapping with Latin characters had to be devised at the beginning.

OCR, too, was strongly dependent on Latin at that time, which means that most software that came (or could be used) with the available scanners was preconditioned to recognise Latin characters, no others, and only those pertaining to the ASCII and, later, ANSI standards. It goes without saying that such software could not reasonably be applied for reading Georgian.[5] Even in 1988, however, there was already some OCR software available that could be "trained" to distinguish Georgian characters and to "map" them onto encodable Latin characters. One such software package which proved usable for this purpose and which was applied (in connection with a Xerox Datacopy scanner) for the digitisation of the first hundreds of pages of Georgian printed texts now stored in the GNC, was the "SPOT" program developed by a company named Flagstaff Engineering. Fig. 1 shows the functionality of the program, which was still based on the DOS operating system, in a set of screenshots including its "training" function.[6]

During the 1990s, another commercial product came on the market which provided a similar functionality but with higher efficiency; this was the "FineReader" program, now designed for Windows systems, which was developed by the Russian company ABBYY. Georgian did not belong to the many languages that the "FineReader" supported, nor was the Georgian script; however, the "mapping" principle could still be applied in both "training" and reading here, too, and so the bulk of the Old and Middle Georgian materials digitised in the course of the TITUS and ARMAZI projects were based on its application. It goes without saying that for the development of a consistent electronic corpus of Georgian, the "mapped" Latin-script renderings of Georgian characters (see Table I for an example) had to be corrected and then

[4] For a comprehensive survey see https://en.wikipedia.org/wiki/ISO/IEC_8859.
[5] This is true, e.g., for the widespread *OmniPage* software (first developed by Caere Corporation, see https://en.wikipedia.org/wiki/OmniPage) as well as the scanning systems of Makrolog (see https://www.makrolog.de/home/) and Kurzweil (https://en.wikipedia.org/wiki/Ray_Kurzweil).
[6] Screenshots on the basis of the demo version in German.

converted into adequate codes; this became possible in the late 1990s with the development of the (32-bit) Unicode standard (ISO/IEC-10646), which meanwhile offers unique code points for all three Georgian scripts (*mkhedruli*, *nuskha-khutsuri*, *asomtavruli*/*mrglovani*).

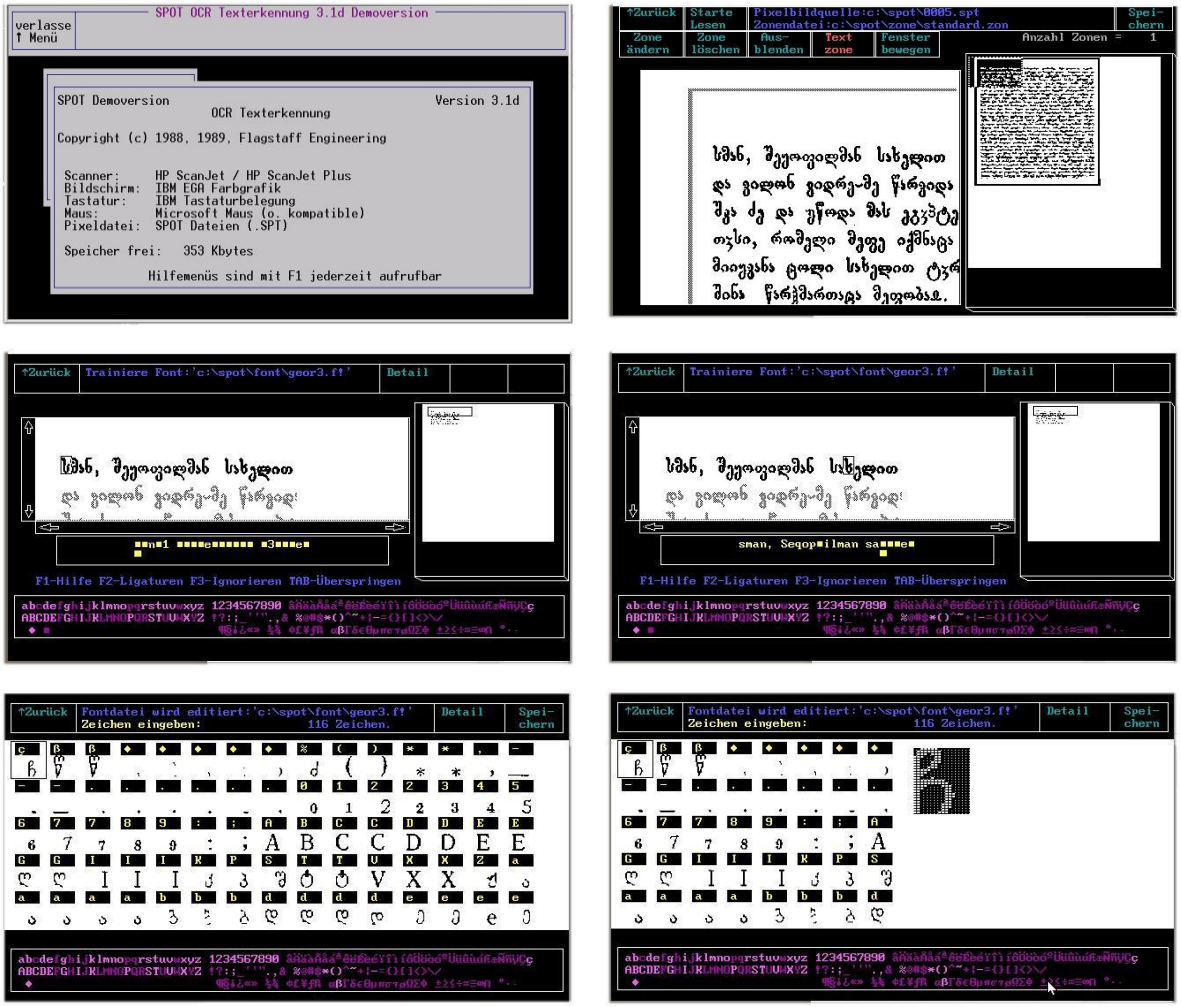


Fig. 1: DOS-based OCR Software (SPOT) with "training" function (screenshots)

Table I: OCR result with "mapping" onto Latin characters

's~ da ubr&ana didebulta: amas Kacsa saKvirveli nadiri una-
xavs, aC misi unaxaoba Çvengan ar egebis da, vinattgan esre~
~aKvirveli aris, mnebavs, ratam~a mteli SeviPQrato.
gaiQara nadimi da SeeKazmes mepe da Qovelni didebulni
misni, Cina Cari&Gvanes igi Kaci da &ivides mas mi~-
dorsa Sina. naxes igi kurciKi mit saxit~ masve adgilsa. br&a-
na mepeman: aha male, aba, Qovelman Kacman cxenebi SeuTevet da
siKvdilsa eKr&alenito. SeuTeves cxenebi da Cavida kurc~,Ki
igi. arbives dGisa savali da mosCQda sxva Kaci Qvela, mepe da
samni sxvani didebulni SerÇes da arbives vidre mCuxramd`,s
~vidisa dGisa savali: verca mieCives da arca mosCQdes.
adgilsa ertsa mivides Kldtvansa. ÇaexveCa kurciki da
uÇino ikmna. odes daxednes kveQana~a, verasada cnes da ver~a
naxes kalaki da verca sopeli, verca igi nadiri. kveQana
ucxod auÇndat; ara uCQodes, sadamca Cavides. mixed~ me~
peman da naxa saxli erti kvitKirisa. gauKvirda da tkva: ra-

In this way, the Georgian National Corpus has incorporated the textual heritage of Old and Middle Georgian to a large extent, namely, as far as it has been available via printed editions. However, a great bulk of highly interesting materials are still awaiting their inclusion, namely, everything written in unpublished manuscripts. Attempts to automatically read text from images of manuscript pages with commercial OCR software yielded no acceptable results; the reason is that the variation of letter shapes is much higher in handwritten texts than in printed ones. For printed texts, it was necessary to undertake "training" for every single type face used in them; in the case of manuscripts, it would not even have been enough to undertake the same effort for every single scribe's hand since the variation in character shapes is enormous even where scribes intended to be consistent in their handwriting style.

This problem has now been overcome by the development of AI-based trainable software for Handwritten Text Recognition (HTR), and the first endeavours to apply such a system to Georgian manuscripts undertaken in the framework of the DeLiCaTe project[7] during the past six months have yielded remarkable results. The system used is the *eScriptorium* platform, developed at the Paris Sciences et Lettres University as part of the projects *Scripta* and RESILIENCE with contributions from other institutions, partly funded by the EU's Horizon 2020 funding program and a grant from the Andrew W. Mellon Foundation.

2. *eScriptorium*, its background and its functionality

eScriptorium is currently the only *cutting* edge open source HTR programme with an ergonomic interface.[8] It has been developed by the Ecole pratique des hautes études, Paris Sciences et Lettres University (PSL), in Paris since 2019 in collaboration with other teams, in particular the ALMAnaCH team of the Institut national de recherche en sciences et technologies du numérique (INRIA, National Institute for Research in Digital Science and Technology, Paris) and the openITI consortium of University of Maryland (College Park, MD), North-Eastern University (Boston/London), and Aga-Khan University (London).[9] Its central functionalities allow users to automatically analyse the layout of uploaded pictures of handwritten or printed documents and then to automatically transcribe them. This requires the existence of trained AI models for layout segmentation and text recognition of such documents, their script and/or language types. If there are no already trained adequate AI "models" available, users can manually create the necessary training material and train new models and/or fine-tune existing ones, even switching from one language or one script to another. The key to *eScriptorium*'s success is that this can be done relatively quickly and with less human effort than in the past.

With "models" we refer here to multilayered neural networks composed of convolutional neural networks (CNN) and bidirectional long-short-term memory neural networks (bi-LSTM). They are like mathematical formulae with millions of parameters that the computer optimises itself during the training process by being shown "ground truth". For layout-segmentation, this

[7] *The Development of Literacy in the Caucasian Territories*, see 3.1 below.

[8] Its user-interface stack is available at https://gitlab.com/scripta/escriptorium/. The AI part is available at https://github.com/mittagessen/kraken. There are other open source HTR and OCR programs but their interface is by far not as ergonomic. And there are other HTR programs with an ergonomic interface but they are not open source or not cutting edge.

[9] See Kiessling et al. 2019: 19; Stokes et al. 2021: 18 (1).

ground truth consists of document pages with annotations that describe the polygons of regions, line-contours and baselines as well as their types (e.g. main text, running header, etc). For recognition or transcription, the ground truth is Unicode-encoded strings of characters corresponding (and bound) to specific lines.

For the creation of ground truth and the correction of automatic layout segmentation or transcription, ergonomics are paramount. An unergonomic interface not only makes the correction process more time-consuming but also more tedious and more error-prone.

Data in *eScriptorium* is structured into projects which consist of documents, in their turn consisting of images (of pages, bifolios, or cutouts – as the user defines). Users can collaborate in teams and share the material among them. Each document image can have only one layout segmentation linked to it, but each segmentation can be linked with as many transcriptions as one would like, either automatically or manually, so that one can contrast e.g. several independent transcribers, keep abbreviations both unresolved and resolved, or have a normalised rendering alongside a diplomatic one.

Almost everything is user-definable. This means that users are free to define their own way to handle the layout segmentation and the typology of region- and line-types used. They can decide for their own transcription conventions, i.e. whether to expand abbreviations or leave them unresolved, whether to use a graphematic, a diplomatic or a normalising transcription and which conventions to follow, and which metadata to add on the document or image level. Transcriptions can be annotated and enriched with a versatile annotation interface where users can define what they want to annotate (e.g. named entities referring to places or people; dates; resolutions of abbreviations; connections between marginal or interlinear annotations and their insertion spots; etc).

In the edit mode, the user can currently choose to visualise one to five parallel panels: (1) metadata, (2) facsimile, (3) segmentation, (4) transcription and (5) text-annotation (see Fig. 2 for an example).[10] The facsimile (2), segmentation (3) and transcription (4) panels are twinned

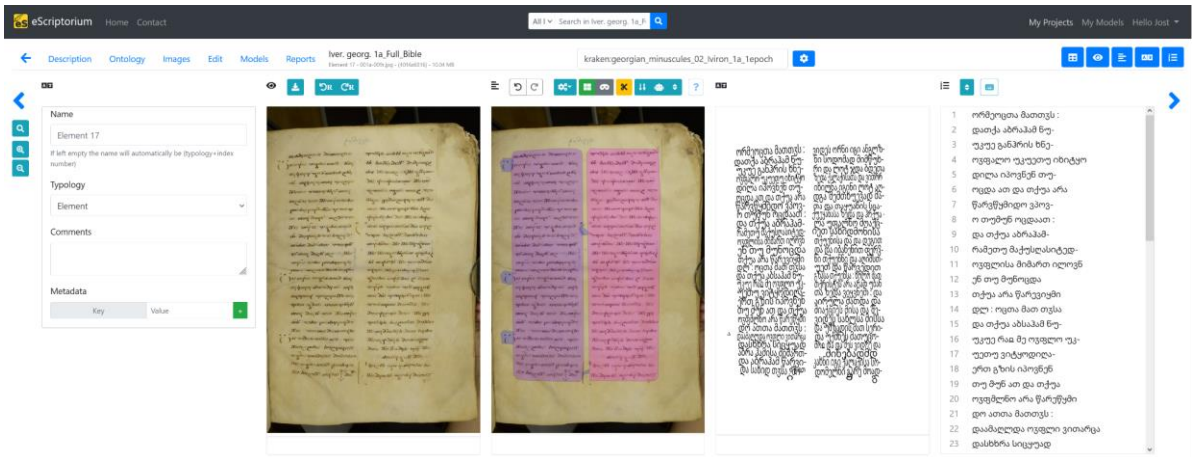


Fig. 2: The five panels of *eScriptorium* in edit mode

[10] In the upcoming new interface created with the massive support of the openITI team, the metadata panel will become a modal window and the facsimile panel will be deprecated.

for zooming in/out or panning to visualise different document parts in the required enlargement. This enables the users to always see the specific region of the document image, its annotation and transcription in a what-you-see-is-what-you-get fashion. The segmentation panel serves the complex procedure of creating and correcting layout annotations, similar to a multilayer drawing programme (cf. Fig. 3).

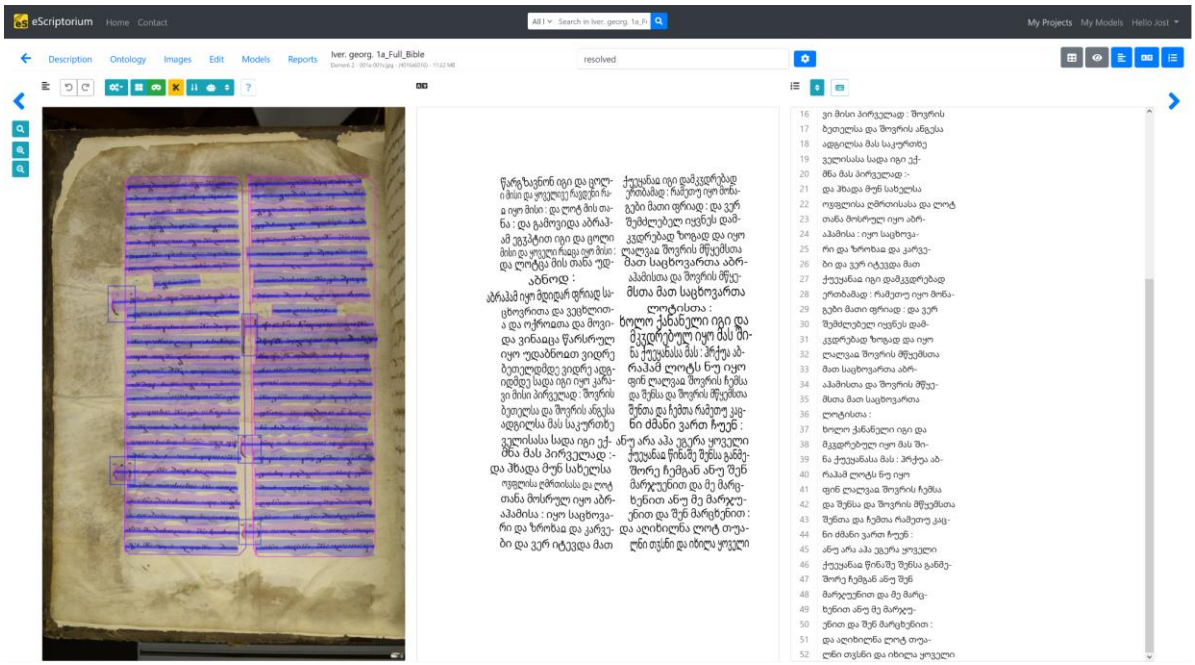


Fig. 3: Segmentation, transcription and text-annotation panels

In the transcription panel, clicking on a line opens it up in a modal window that displays the image of this line and its transcription directly beneath it so that the eye has the shortest distance between the image and the transcription (cf. Fig. 4). The text-annotation panel permits the users

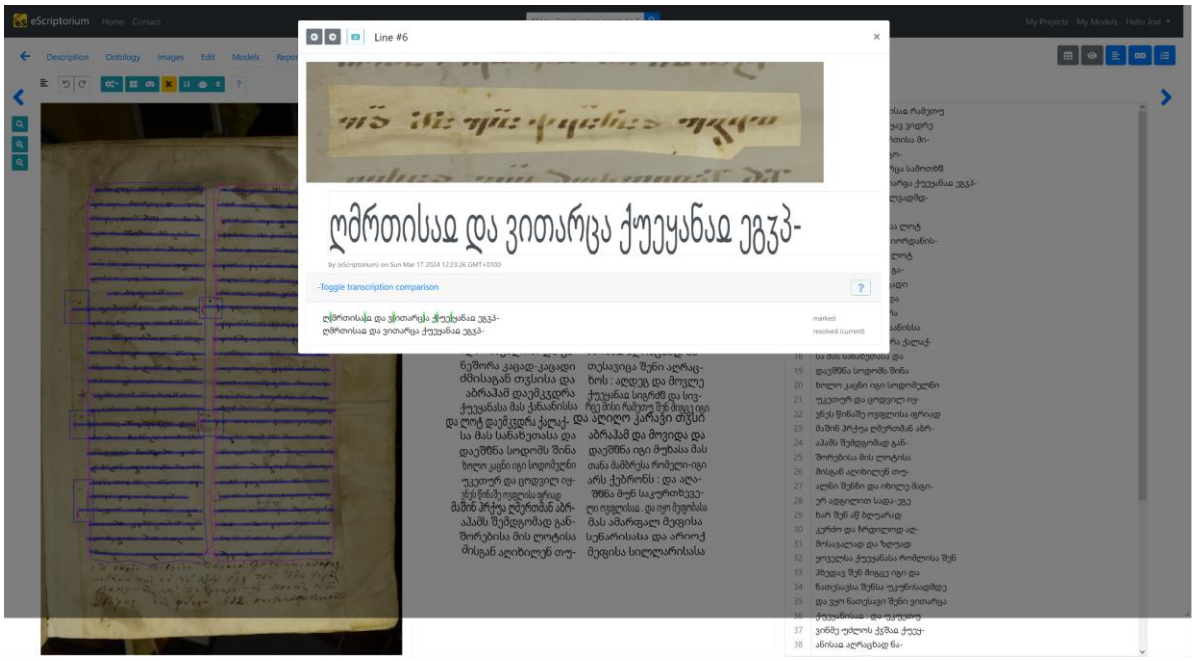


Fig. 4: Modal window for transcription

to create their own annotation buttons above it and then to add e.g. IDs for people, grammatical information, etc. It also allows users to correct the line-ordering.

A most important additional automatic feature is the text-to-text alignment based on the *passim* project by David A. Smith.[11] It allows e.g. to semi-automatically provide training ground truth from existing e-texts by first creating a rough automatic transcription and then aligning it with an existing gold standard transcription of a similar or the same text or manuscript.[12]

A powerful automated programming interface (API) permits to deal with almost all data points and functions automatically. Of course, images, annotations, and trained models can be exported, published and transferred between different instances of *eScriptorium*. The available formats for transcription export are (plain) text, Alto-XML[13] and Page-XML.[14] The latter two XML schemas include the layout coordinates and typology. In the near future, the text-annotation will also be directly exportable via METS.[15] Currently, one needs to use the API for this.

At present, excellent recognition models exist for languages written in Latin, Hebrew, Hindi, Chinese, Syriac, and Palaeo-Slavonic characters for certain periods. Various teams and projects are working on Greek, Avestan, Arabic, Japanese, Coptic, Armenian, Sanskrit, etc. In May 2021 we trained a first recognition model for printed Georgian (*mkhedruli*) on a small dataset consisting of 3668 lines. The model reached an accuracy of 99.2% on the test data.

eScriptorium can be installed on Linux and MAC-OS for individual users. A graphics processing unit (GPU) is needed if one wants to train layout-segmentation models. For the training of recognition models, a GPU is highly recommended as well. For teams, it is preferable to work on a more powerful server with a high speed internet connection. There are many configurations for different use cases and budget constraints, extending from small teams working on a single server with one GPU to very large multi-server high performance computing systems (HPCs) with multiple GPUs. If no GPU is available, we have also seen users exporting their data to a Google Colab(oratory),[16] installing *kraken* (the OCR system underlying *eScriptorium*)[17] and training the model there and then reimporting it. For large models created on data from multiple documents, we recommend exporting the data and training a model on an HPC cluster with a GPU because one has more control over hyper-parameters then. An active user-community on Gitter[18] helps new users and system-administrators to overcome eventual difficulties during the installation or initial use.

[11] See https://github.com/dasmiq/passim.
[12] Several parameters can be set for the alignment; for our Georgian manuscripts, the best results were achieved with a line length match threshold of 0.5, an N-gram value of 8 and a beam size of 100.
[13] "Analyzed Layout and Text Object", see https://en.wikipedia.org/wiki/Analyzed_Layout_and_Text_Object.
[14] "Page Analysis and Ground Truth Elements", see https://en.wikipedia.org/wiki/Page_Analysis_and_Ground_Truth_Elements.
[15] "Metadata Encoding and Transmission Standard", see https://en.wikipedia.org/wiki/Metadata_Encoding_and_Transmission_Standard.
[16] See https://colab.research.google.com/.
[17] See https://kraken.re/main/index.html. We started using version 4.1.2 (see Kiessling 2022). The latest version used for this article is 5.2.9 (see Kiessling 2024).
[18] See https://gitlab.com/scripta/escriptorium.

In the following section, we will describe ongoing efforts undertaken within the DeLiCaTe project to use *eScriptorium* for the automatic reading of Georgian manuscripts from the first and second millennia of our era.

3. Using *eScriptorium* for Georgian Manuscripts: Achievements and Future Tasks

3.1 The context

The DeLiCaTe project, an ERC Advanced Grant devoted to investigate the "Development of Literacy in the Caucasian Territories",[19] has been running since April 2022 in the Centre for the Study of Manuscript Cultures (CSMC) of the University of Hamburg.[20] A decisive part of its objects consists of palimpsest manuscripts in Georgian, Armenian and Caucasian Albanian, which represent the oldest written sources available for the three languages of the Caucasus that have developed their own literacy in the course of the Christianisation of the region in the 4th–5th centuries CE. The decipherment of palimpsests, i.e. handwritten texts that were erased so that the writing support could be reused for writing down other text materials at a later time,[21] is usually extremely difficult and requires peculiar imaging methods, among which multispectral imaging (MSI) has meanwhile yielded the best results.[22] Even with these methods, however, many overwritten texts cannot be identified easily, especially when only a few words per page can be made out. Whenever the overwritten text is available from other manuscripts in electronic form and retrievable via search engines of corpora, a text passage can often be identified even if only a few characters are discernible; in the case of Georgian and Armenian, this is typically true of biblical matter.[23] With other relevant text genres such as hagiography, homiletics, and hymnography, the amount of digitally available text materials that can be used for searching is much more restricted, given that important parts of the written tradition of the Caucasus have never been edited and are only accessible in manuscript form to the present day. This is the reason why we decided in our project to apply *eScriptorium* in order to digitise and transcribe manuscripts that may contain relevant textual materials.

The manuscripts we are working on with *eScriptorium*[24] are usually parchment codices of the 9th–11th centuries of which colour images of sufficient quality are available, provided by (or procured from) the major repositories (for Georgian, the Korneli Kekelidze National Centre of Manuscripts, Tbilisi; the Iviron Monastery on Mount Athos; the University Libraries of Graz and Leipzig; the Bibliothèque nationale de France, Paris; the Bodleian Libraries, Oxford; and others). In some cases, we have to rely on greyscale images that were produced from microfilms; this is true, first of all, of the large collections kept in St Catherine's monastery on

[19] European Research Council, grant agreement no. 101019006.
[20] See https://www.csmc.uni-hamburg.de/delicate.html.
[21] As to palimpsests in general and the DeLiCaTe approach, cf. Gippert (2025a and 2025b).
[22] Cf. Gippert (2025b), Kamarauli (2025), Kvirkvelia (2025), Sargsyan (2025) and Bonfiglio (2025) as to palimpsests that are investigated with multispectral imaging in the DeLiCaTe project and Mohammed, Jampour and Gippert (2025) as to a new method of reconstructing the erased layers of palimpsests.
[23] The decipherment of the Caucasian Albanian palimpsests, too, was only possible because the content could be identified as being biblical; see Gippert (2023: 99–141) for details.
[24] First attempts to use *eScriptorium* for reading the erased lower layers of palimpsests have failed – the system is optimised to ignore the traces of the overwritten text, treating them as "noise".

Mount Sinai and the Greek Patriarchate in Jerusalem, which are accessible thanks to the efforts of the Library of Congress (Washington DC).[25] The images may show either one page each (usually in the case of recently produced colour images, see Figures 2–4 above) or two pages side by side; this is typically true of the digitised microfilms (see Fig. 5). The text on the manuscript pages can be arranged in one or two columns so that a maximum of four columns may have to be dealt with in one image if this covers two pages.

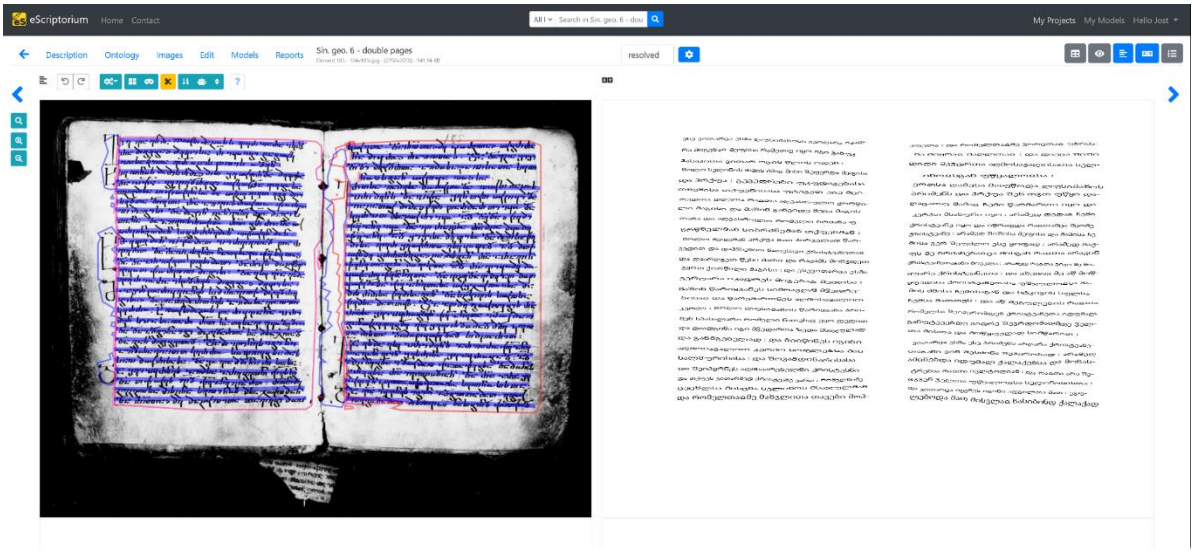


Fig. 5: Double-page microfilm images in *eScriptorium* (fols 184v–185r of Sin. georg. 6)

Old Georgian manuscripts of the first millennium are mostly written in majuscules (*mrglovani*), often with large initial letters (also in majuscules) indicating the beginning of chapters, paragraphs or sections (*asomtavruli*); later manuscripts usually appear in (*nuskha-khutsuri*) minuscules with (*asomtavruli*) majuscules applied to initials, sometimes also titles and other special elements. Whereas the shape of the majuscules is comparatively uniform throughout the given period (roughly the 8th–10th centuries), the minuscules vary to a considerable extent, in both (relative) size and shape and with different degrees of slanting; in addition, the use of abbreviations (usually elisions of one or several characters within a word, indicated by the so-called *karagma*, a tilde-shaped diacritic written above) increases remarkably from century to century. All this means that the development of a unique (one-for-all) solution for a model covering the Old Georgian manuscript production seemed unfeasible right from the beginning; instead, two independent models (one for majuscules only, one for minuscules in combination with majuscules) had to be planned and realised.

3.2 A spiral bootstrapping procedure

Since 17 March 2024, four of our project team members[26] have been busy training the two models for Old Georgian.[27] In order to minimise human effort, we have applied a bootstrapping

[25] See https://www.loc.gov/manuscripts/?q=georgian+jerusalem and https://www.loc.gov/manuscripts/?q=georgian+sinai.

[26] Jost Gippert, Mariam Kamarauli, Eka Kvirkvelia and Sandro Tskhvedadze.

[27] From March to October 2024 we used the msIA server in Paris (https://msia.escriptorium.fr/); meanwhile an instance has been set up in Hamburg by Magnus Bender of the Institute for Humanities-Centred Artificial Intelligence (CHAI; see https://escriptorium.chai.uni-hamburg.de/).

procedure based on four principles: (1) dividing the manuscripts into two lots according to whether they are written in minuscule or majuscule script, (2) applying "transfer learning" or "fine-tuning" existing segmentation and recognition models for other scripts to Georgian, (3) identifying low-hanging fruit where transcriptions of Georgian manuscripts already exist, and (4) exploiting *eScriptorium*'s automatic alignment feature (text-to-text). Most of this is methodology of Daniel Stökl ben Ezra's ERC Synergy grant on Hebrew manuscripts, MiDRASH, whose task among others is a transcription of millions of images of manuscripts in Hebrew characters.[28]

We started with three pages of an important minuscule manuscript, the so-called Oshki Bible of 978 CE (manuscript Iviron georg. 1, fols 1r–2r, written in two columns, see Figures 2–4 above), by applying a standard Latin layout segmentation model and correcting it, and then entering the transcription (taken from Akaki Shanidze's edition of 1947) manually line-by-line. We then fine-tuned a first transcription model optimised for the Oshki Bible on top of the *Biblia* model trained for Medieval Hebrew.[29] On the basis of this tiny dataset, the character accuracy only reached 76.4%, and the word accuracy even only 44.6% after 43 epochs;[30] nevertheless, the model could be used to transcribe subsequent pages automatically and to align existing transcriptions for a first small dataset of five manuscripts (88 pages).[31] For each of them we trained specific models on the same day. On the second day of our "transcribathon" in March, we added text from three further manuscripts[32] and trained a first mixed model for early minuscules, which had a more sizable textual foundation of 116 pages and has since served as a suitable basis for all other minuscule manuscripts we have been dealing with so far.[33] With a similar approach, we also developed the first model for majuscule manuscripts, on the basis of samples from three codices.[34]

We further devised three essential tools for the project to proceed: (1) a tracking sheet (on Google Sheets) where all team members could simultaneously update the status as to which steps had been performed on which part of which manuscript with which model; (2) conventions for the segmentation, and (3) transcription conventions. Right from the beginning we decided not to render the Old Georgian scripts as such in the transcriptions but to transcribe

[28] On an even more automatised methodology applied to Syriac corpora see Bambaci et al. (2024).

[29] Daniel Stökl Ben Ezra (2021): Medieval Hebrew manuscripts version 1.0. (see https://doi.org/10.5281/zenodo.5468286).

[30] An "epoch" is a training round in which the computer has seen each element of the training material once.

[31] Single pages, colour, two columns: Oxford, Bodleian Libraries, georg. b1, fols 181r–185v; Tbilisi, Korneli Kekelidze Georgian National Centre of Manuscripts (hereafter: KKNCM), A-95 (the so-called *Parkhali Mravaltavi*), fols 305v–319v; one column: Athos, Iviron Monastery, georg. 45 (autograph by George the Athonite), fols 2v–8v; double pages, black and white, one column: Sinai, St Catherine's Monastery, georg. 6, fols 184r–200v.

[32] Single pages, colour, two columns: Athos, Iviron Monastery, georg. 10, fols 331v–337v, and georg. 62, fol. 9r; single page, colour, one column: Vienna, Austrian National Library, georg. 2, fol. 1r.

[33] The model is named georgian_minuscules_02_Iviron_1a_best.

[34] Single pages, colour, one column: Graz, University Library, ms. 2058/1 (the so-called *Khanmeti Lectionary*), fols 1r–27v; two columns: Athos, Iviron Monastery, georg. 9, fols 1r–3v; Tbilisi, KKNCM, A-1109 (the *Udabno Mravaltavi*), fols 100r–104v. The processing for both models was undertaken by Mariam Kamarauli, Eka Kvirkvelia, and Jost Gippert, with support by Daniel Stökl Ben Ezra.

them into the modern script (*mkhedruli*), in accordance with the practice of scholarly editions of today.

A major challenge consisted in the treatment of the abbreviations (see above). In modern editions of Old Georgian handwritten texts, they are normally expanded. We therefore experimented with three different transcription types: (1) a “diplomatic” rendering of the letters that actually appear in the manuscript, including the *karagma* diacritic; (2) a “resolved” rendering as used in the modern editions; and (3) a “marked” rendering, which is a mixture of both (1) and (2) with the expansion of the abbreviation indicated by parentheses. After some testing, we decided to apply the “resolved” rendering (2) throughout because the system learned those expansions that are not too rare relatively well and because almost all of the existing transcriptions with which we could align contain resolved abbreviations, too.

Another problem we encountered is the division of words between lines. Currently, *passim*, the text-to-text alignment algorithm inside *eScriptorium*, silently splits words between lines if the letters recognised by the rough HTR in each line make this solution more probable. It does, however, not add a hyphen that would indicate the word split at the end of a line as we would find it in modern editions of Old Georgian texts. This makes it more difficult to restitute a running text from the automatic transcription. Even if such hyphens, which are practically never graphically represented in any form in the manuscripts, are eventually inserted (manually), the system has only the length and sequence of letters at the end of the given line as data on which to later base its estimation whether to add a hyphen or not. In the current line-based OCR/HTR approach of *kraken*, the contents of a subsequent line are unknown when a given line is processed.

3.3 The first results

The two models that resulted from our “transcribathon” of 17 and 20 March 2024 have since then been applied to more than 40 Old Georgian manuscripts, of very different types and sizes, among them the complete Georgian collections of the University Libraries of Graz (manuscripts 2058/1–6c) and Leipzig (manuscripts Vollers 1094–1098), as well as several further manuscripts of the Iviron Monastery on Mount Athos (codices 8, 11, 20, 42 and 89).

The versatility of the models we have developed so far can be illustrated on the example of Leipzig, University Library, Vollers 1095, a bundle comprising, among other fragments, seven pages filled with commemorative notes (აღაპები) from the Georgian community of Jerusalem, written by at least 30 different hands in various inks and styles. After correcting the automatic segmentation, which had yielded an astonishingly correct result in distinguishing lines of extremely unequal shape (see Fig. 6 showing fols 13v–14r), the automatic transcription left no element unread (see Fig. 7), and the alignment with the edition of the notes by Elene Metreveli[35] (see Fig. 8) shows that the reading accuracy was high enough to cover most of the two pages. The few unaligned items could then easily be corrected manually (see Fig. 9).

[35] Metreveli (1962: 72–78); cf. the electronic edition in https://titus.uni-frankfurt.de/texte/etcg/cauc/ageo/liturg/masjer/masje.htm. A few corrections and additions were provided in Assfalg (1963: 60–72).

Fig. 6: Leipzig, University Library, Vollers 1095, fols 13v–14r, automatically segmented

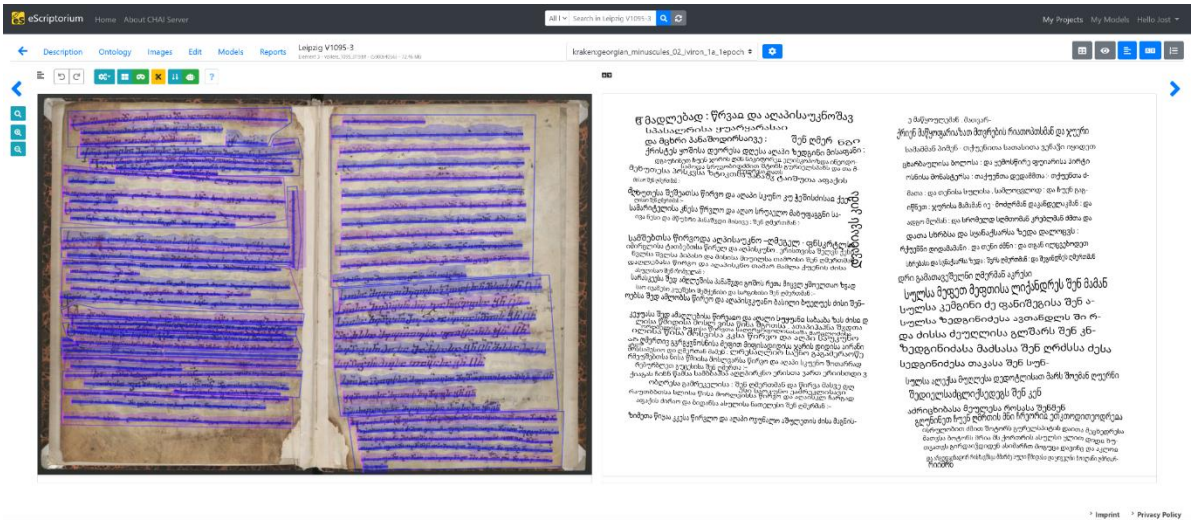


Fig. 7: Same, automatically transcribed after correction of segmentation

As in this case, it has turned out that the automatic segmentation sometimes leads to less satisfying results, mostly due to damages or stains of the writing support, fainted inks,[36] erasures, or simply insufficient quality of the images available; this is why we decided to manually correct the segmentation wherever necessary.[37]

[36] Red ink may also pose problems because it appears less dark in the greyscale images into which colour images are converted for the use of the segmentation and transcription algorithms. Another persisting problem is caused by large initials that are outdented and extend over several lines as visible in Figures 11–13 below.

[37] In the manual correction we have been supported by Yorrick Stute and Fahimeh Rahravan.

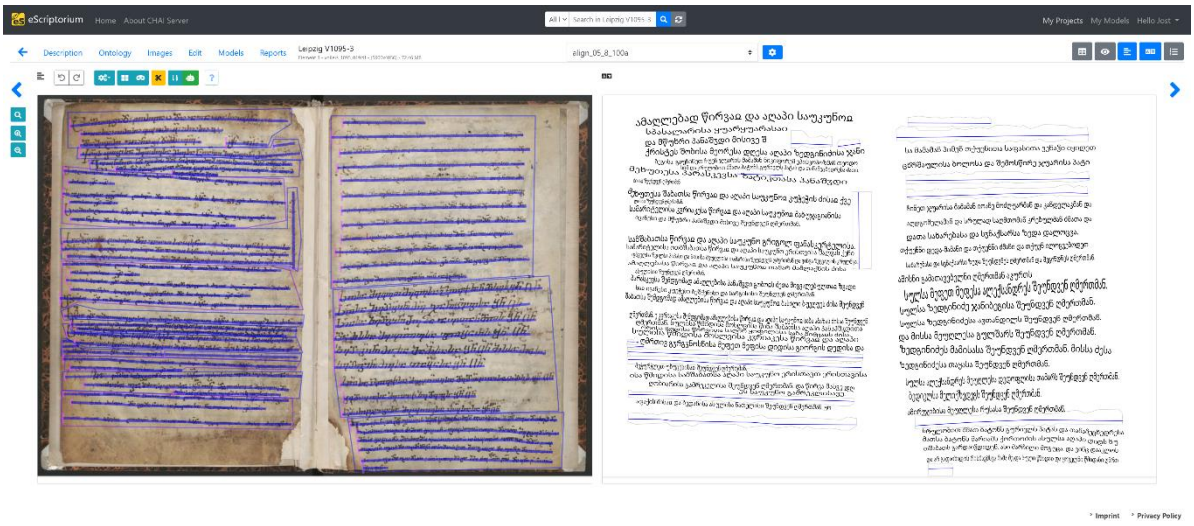


Fig. 8: Same, automatically aligned with existing e-text

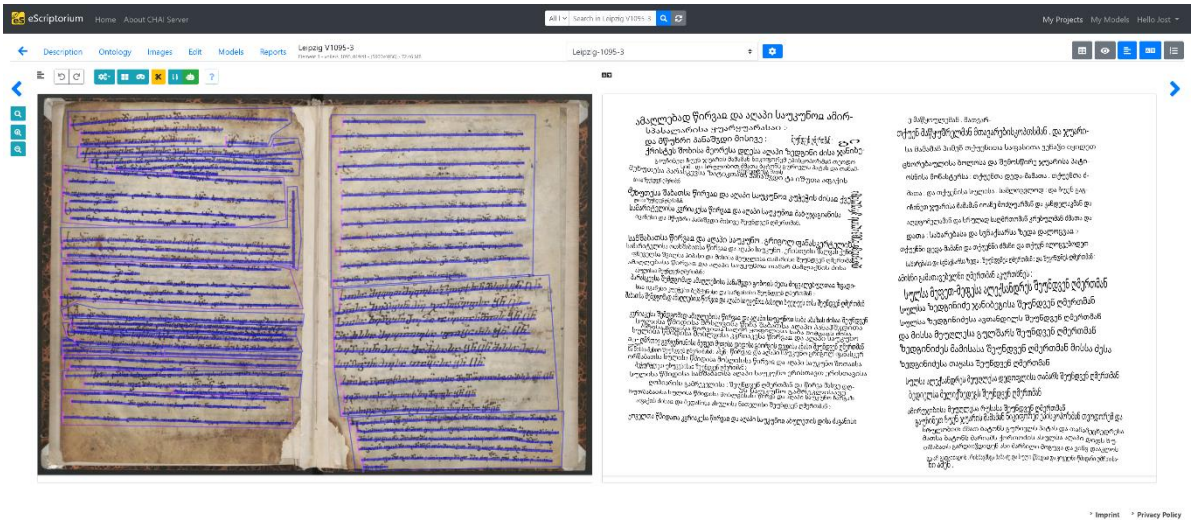


Fig. 9: Same, after manual correction

One of the most problematic cases was manuscript 2058/4 of the Graz collection, a codex written all in majuscules (except for the colophons on fols 94v–95r and 110v, which are in minuscules);[38] Fig. 10 shows fol. 51v after the automatic segmentation and its manual correction. In the given case, the reason for the partial failure of the automatic segmentation seems to be that the image size is smaller than the input size of the neural network for segmentation. After the manual correction, the complete codex was transcribed automatically using the model that was trained for majuscules. The resulting transcription was then aligned with the edition of the text (the liturgy by James and the *Missa Praesanctificorum*), which had been provided by Vakhtang Imnaishvili in electronic form for the TITUS project in 2004.[39] As a result, the complete manuscript is now available with near-to full alignment, with only a few lines (less than 6%) needing further manual correction, mostly because they are damaged,

[38] According to the colophon on fol. 95r, the first part of the codex was written by Ioane Zosime on Mount Sinai in 985; the scribe of the second part (fols 96r–110r) was another Iovane (according to his colophon on fol. 110r) who wrote in a totally different hand.

[39] See https://titus.uni-frankfurt.de/texte/etca/cauc/ageo/liturg/litjak/litja.htm; the electronic text corresponds to that published in print by Imnaishvili (2004: 266–294).

contain corrections, or abound in abbreviations; cf. Figures 11–13 showing fol. 104v after automatic transcription, automatic alignment, and with the modal window usable for manual correction. Once the aligned text has been corrected in the way indicated and the problem of the insertion of hyphens at the line break has been solved, the transcript thus produced can be used for the development of extended models.

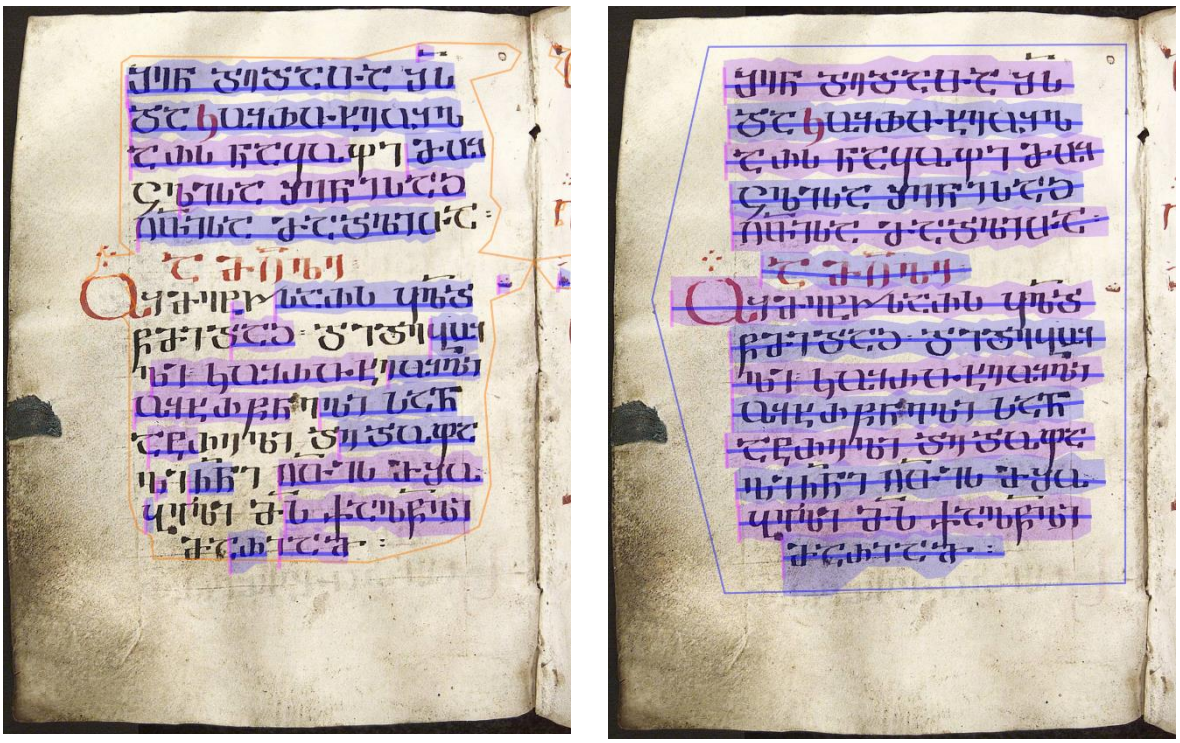


Fig. 10: Segmentation of Graz, University Library, MS 2058/4, fol. 51v (left: automatic; right: manually corrected)

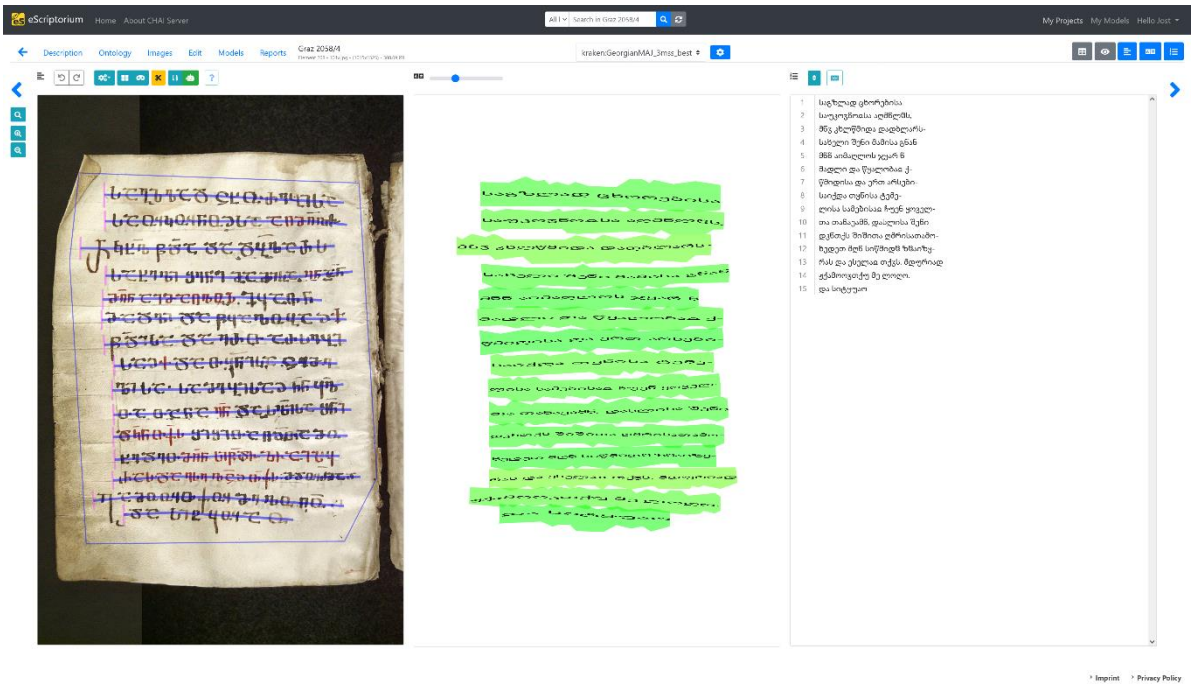


Fig. 11: Graz, University Library, MS 2058/4, fol. 104v, automatically transcribed

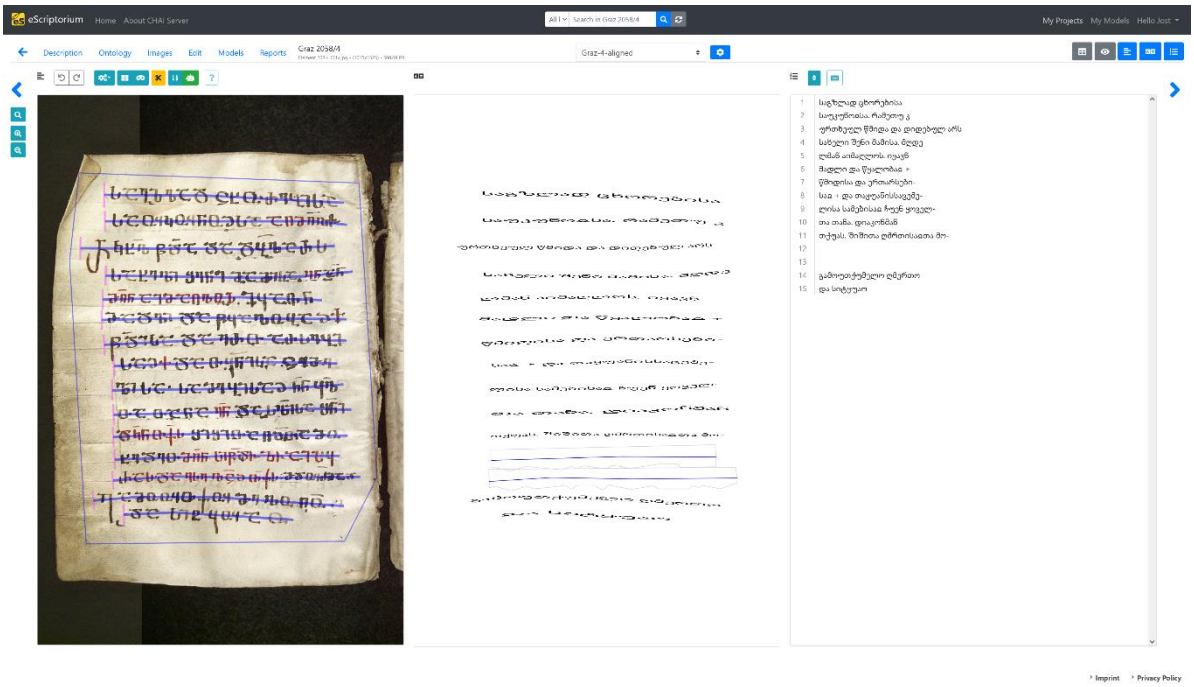


Fig. 12: Same, autmatically aligned

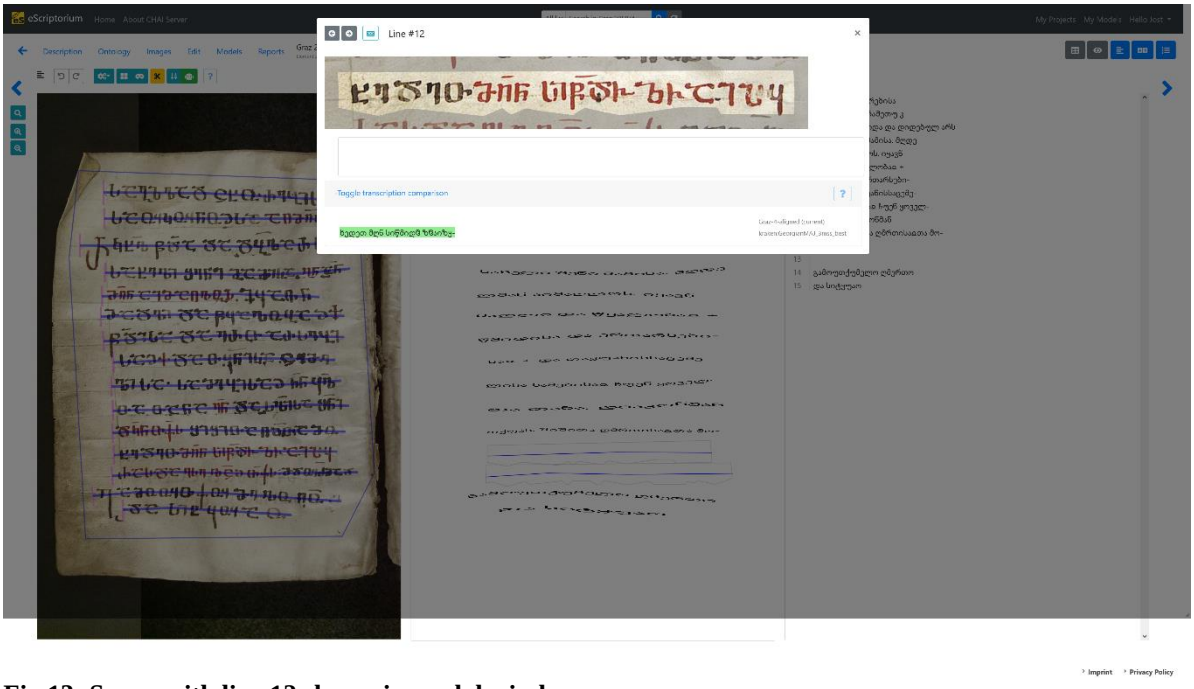


Fig 13: Same, with line 12 shown in modal window

Table II summarises the number of images (= pages) of Georgian manuscripts that have been processed on the *eScriptorium* platform so far (as of 10 November, 2024).

Table II: Statistics			
	manually transcribed	auto-aligned	total
majuscules	76	1280	1356
minuscules	216	6208	6424
total	292	7488	7780

3.4 Future Tasks

On the basis of the 292 manually transcribed pages and 7488 pages with aligned (but not yet always corrected) text, we are now about to turn towards the next generation of models. We expect these models soon to reach an accuracy reading rate higher than 90%, which is far beyond what we could arrive at with OCR of printed Georgian books 20 years ago. It goes without saying that we intend to make our results openly available; this is not only true of the models but also of the finalised transcripts.[40]

Acknowledgements

The present research was funded by the European Union's research and innovation programmes Horizon 2020 (ERC Advanced Grant "DeLiCaTe", grant agreement no. 101019006) and Horizon Europe (ERC Synergy Grant "MiDRASH", grant agreement no. 101071829) and by the German Research Foundation (DFG, Cluster of Excellence 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796); the work was partly carried out at the Centre for the Study of Manuscript Cultures, University of Hamburg. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union, the European Research Council Executive Agency or the German Research Foundation. Neither the European Union nor the other granting authorities can be held responsible for them.

References

Assfalg (1963): Julius A., *Georgische Handschriften* (Verzeichnis der orientalischen Handschriften in Deutschland, 3), Wiesbaden: Steiner. https://digital.staatsbibliothek-berlin.de/werkansicht/?PPN=PPN1878674927.

Bambaci et al. (2024): Luigi B., George Kiraz, Christine Roughan, Matthieu Freyder, Daniel Stökl Ben Ezra, Daniel, "Steps Towards Mining Manuscript Images for Untranscribed Texts: A Case Study From the Syriac Collection at the Vatican Library", CHR 2024: *Computational Humanities Research Conference*, December 4–6, 2024, Aarhus, Denmark.

Bonfiglio (2025): Emilio B., "Uncovering Lost Armenian Texts: Schøyen Collection MS 575 and the Armenian Translation of John Chrysostom's Commentary on the Psalms", in Gippert, Maksimczuk & Sargsyan (2025), 351–380. https://doi.org/10.1515/9783111552170-013.

Gippert (1990): Jost G., "Perspektiven des Computereinsatzes in der Orientalistik [Perspectives of the application of computers in Near Eastern studies]", *Forschungsforum* [University of Bamberg] 2, 133–136. https://titus.fkidg1.uni-frankfurt.de/personal/jg/pdf/jg1990c.pdf.

— (2023): Jost G., "The Textual heritage of Caucasian Albanian". In Jost Gippert, Jasmine Dum-Tragut (eds.), *Caucasian Albania. An International Handbook*. Berlin / Boston: de Gruyter, 95–166. https://doi.org/10.1515/9783110794687-003.

— (2025a): Jost. G., "Removed and Rewritten: Palimpsests and related phenomena from a cross-cultural perspective", in Gippert, Maksimczuk & Sargsyan (2025), 1–19. https://doi.org/10.1515/9783111552170-001.

— (2025b): Jost G., "Palimpsests from the Caucasus: Two Case Studies", in Gippert, Maksimczuk & Sargsyan (2025), 253–281. https://doi.org/10.1515/9783111552170-009.

[40] Whether or not the images can be made available along with the transcripts depends on the institutions owning the manuscripts. We are confident that we will come to the necessary agreements in most cases.

Gippert, Maksimczuk & Sargsyan (2025): Jost G., José M. and Hasmik S. (eds.), *Palimpsests and Related Phenomena across Languages and Cultures*. Berlin / Boston: de Gruyter (Studies in Manuscript Cultures, 42). https://doi.org/10.1515/9783111552170.

Imnaishvili (2004): ვახტანგ იმნაიშვილი, უძველესი ქართული ხელნაწერები ავსტრიაში. ხანმეტი ლექციონარი, ფსალმუნი, სვიმეონ სალოსის ცხორება, იოვანე ოქროპირისა და იაკობ მოციქულის ჟამისწირვები. თბილისი: ტრიადა.

Kamarauli (2025): Mariam K., "The Oldest Georgian Witness of the Martyrdom of St Febronia", in Gippert, Maksimczuk & Sargsyan (2025), 283–308. https://doi.org/10.1515/9783111552170-010.

Kiessling (2022): Benjamin K., *The Kraken OCR system* (Version 4.1.2) [Computer software]. https://kraken.re; https://github.com/mittagessen/kraken.

— (2024): Benjamin K., *The Kraken OCR system* (Version 5.2.9) [Computer software]. https://kraken.re; https://github.com/mittagessen/kraken.

Kiessling et al. (2019): Benjamin K., Robin Tissot, Peter A. Stokes, Daniel Stökl Ben Ezra, "eScriptorium: An Open Source Platform for Historical Document Analysis", *2019 International Conference on Document Analysis and Recognition Workshops* (ICDARW), Sydney, NSW, Australia, 2019: 19. https://doi.org/10.1109/ICDARW.2019.10032.

Kvirkvelia (2025): Eka K., "New Witnesses of the Jerusalem-Rite Lectionary: Georgian Palimpsests Ivir. georg. 47 and Ivir. georg. 59", in Gippert, Maksimczuk & Sargsyan (2025), 309–329. https://doi.org/10.1515/9783111552170-011.

Metreveli (1962): ელენე მეტრეველი, მასალები იერუსალიმის ქართული კოლონიის ისტორიისათვის (XI-XVII სს.), თბილისი: საქართველოს სსრ მეცნიერებათა აკადემია, ხელნაწერთა ინსტიტუტი. https://dspace.nplg.gov.ge/handle/1234/406705.

Mohammed, Jampour & Gippert (2025): Hussein M., Mahdi J. and Jost G., "Inpainting with Generative AI: A Significant Step towards Automatically Deciphering Palimpsests", in Gippert, Maksimczuk & Sargsyan (2025), 535–545. https://doi.org/10.1515/9783111552170-019.

Sargsyan (2025): Hasmik S., "Linguistic Divergence in Armenian Bible and Lectionary Palimpsests", in Gippert, Maksimczuk & Sargsyan (2025), 331–350. https://doi.org/10.1515/9783111552170-012.

Shanidze (1947): წიგნი ძოჯელისა აღთქოჯმისანი 978 წლის ხელნაწერის მიხედვით. ტ. I. ნაკვეთი 1. დაბადაბისაჲ. გამოსლვათაჲ. გამოსცა აკაკი შანიძემ (ძველი ქართული ენის ძეგლები, 4). თბილისი: საქართველოს სსრ მეცნიერებათა აკადემიის გამომცემლობა.

Stökl Ben Ezra (2021): Daniel S., Medieval Hebrew manuscripts version 1.0. *Zenodo*. https://doi.org/10.5281/zenodo.5468286.

— (forthcoming): Daniel S., "Text Digitization". To appear in Christopher A. Nunn and Frederike van Oorschot (eds.), *Compendium of Computational Theology 1. Introducing Digital Humanities to Theology*, Heidelberg: heiBOOKS (probably 2024), 99–115.

Stökl Ben Ezra et al. (2021): Daniel S., Bronson Brown-DeVost, Pawel Jablonski, Hayim Lapin, Benjamin Kiessling, Elena Lolli, "BiblIA – a General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset". HIP@ICDAR 2021: 61–66. https://doi.org/10.1145/3476887.3476896.

Stokes et al. (2021): Peter Anthony S., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, El Hassane Gargem, "The eScriptorium VRE for Manuscript Cultures", *Classics@*, 2021, 18 (1). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.

ქართული ხელნაწერების ავტომატური ამოკითხვა *eScriptorium*-ის პლატფორმაზე

იოსტ გიპერტი (ჰამბურგი), დანიელ შტეკლ ბენ ეზრა (პარიზი)

DOI: https://doi.org/10.62235/dk.3.2024.8508
jost.gippert@uni-hamburg.de || ORCID: 0000-0002-2954-340X
Daniel.Stoekl@ephe.psl.eu || ORCID: 0000-0001-5668-493X

ჰუმანიტარიის არსებობის ისტორიის განმავლობაში არაფერს არ შეუცვლია კვლევის პროცესი ენათმეცნიერებაში, ფილოლოგიასა და მომიჯნავე დარგებში ისე მკვეთრად, როგორც ეს ციფრულ ეპოქაში მოხდა: დიდ მონაცემთა ტექსტური ბაზების შექმნამ, კორპუსული კვლევების განვითარებამ და ავტომატური ძიების მრავალფეროვანმა მეთოდებმა სწრაფი და ეფექტიანი კვლევის საშუალება მისცა მკვლევრებს. ტექსტური კორპუსების განხილვისას უნდა განვასხვავოთ ორი ტიპის ტექსტური რესურსები: **„ციფრულად დაბადებული“** (ანუ ის რესურსები, რომლებიც თავიდანვე ელექტრონული ფორმით შეიქმნა) და **რეტროდიგიტალური** რესურსები (ხელნაწერი ან ბეჭდური სახით არსებული რესურსები, რომლებიც გაციფრებას საჭიროებს). ქართული ენის ეროვნული კორპუსი, რომელიც ბოლო 13 წლის განმავლობაში შეიქმნა (GNC, http://gnc.gov.ge), ორივე ტიპის რესურსებს მოიცავს: მისი ყველაზე დიდი ქვეკორპუსი (ქართული ენის რეფერენციული კორპუსი [GRC], რომელიც 202 728 329 ტოკენის მოცულობისაა) ძირითადად შედგება იმ მასალებისგან, რომლებიც შეგროვდა WWW (World Wide Web) ქსელში; ხოლო ქვეკორპუსები, განსაკუთრებით ისინი, რომლებიც ძველი (GNC-ის ძველი ქართულის ქვეკორპუსი, 7 101 021 ტოკენი) და საშუალი ქართული ენის (GNC-ის საშუალი ქართულის ქვეკორპუსი, 1 432 262 ტოკენი) რესურსებს მოიცავს, ასევე თემატური კორპუსები, როგორებიცაა ქართული სამართლის ტექსტები (GNC-ის იურიდიული ტექსტების ქვეკორპუსი, 1 495 985 ტოკენი) ან პოლიტიკური ტექსტები (GNC-ის პოლიტიკური ტექსტების ქვეკორპუსი, 1 436 075 ტოკენი) მეტ-ნაკლებად ეფუძნება ბეჭდური მასალის გაციფრებულ რესურსებს. ისინი TITUS-ისა და ARMAZI-ის პროექტების ფარგლებში შეიქმნა. რესურსების მხოლოდ მცირე ნაწილი გაციფრულდა ხელით, ანუ მოცემული ტექსტის კლავიატურაზე აკრეფით; უმეტეს შემთხვევაში დაბეჭდილი ტექსტები მუშავდებოდა ოპტიკური სკანერის მეშვეობით და შემდეგ „იკითხებოდა“ ციფრული ტექსტის ფორმატში სპეციალური პროგრამული უზრუნველყოფის საშუალებით, რომელსაც შეეძლო გრაფემების განცალკევება, მათი ამოცნობა ანბანის შესაბამის სიმბოლოებთან შედარების გზით და ამოცნობილი ასოების თანმიმდევრულად შენახვა ელექტრონული ტექსტის ფორმატში.

1980-იანი წლების ბოლოს, როდესაც საფუძველი ეყრებოდა ქართული ენის დიაქრონიული კორპუსის შექმნას, სიმბოლოების ოპტიკური ამოცნობის (OCR) პროგრამები ჯერ კიდევ არ იყო ისე კარგად განვითარებული, როგორც დღეს არის, და სკანერებიც ტექნიკურად ბევრად ჩამორჩებოდნენ თანამედროვე სკანერებს. თუმცა ყველაზე დიდი პრობლემა მაინც იმაში მდგომარეობდა, რომ იმდროინდელი პერსონალური კომპიუტერები არ იყო გათვლილი იმ ენებზე სამუშაოდ, რომელთაც ლათინურისაგან

განსხვავებული ანბანები აქვთ. ისინი 7-ბიტიანი კოდირების სისტემაზე იყო დაფუძნებული (ASCII) და მხოლოდ ინგლისურ ენაში გამოყენებულ სიმბოლოებს მოიცავდა. თანდათანობით 7-ბიტიანი სისტემა შეიცვალა 8-ბიტიანი სისტემით (ANSI), რომელიც უკვე ფარავდა „დამატებით სიმბოლოებს“, რომლებიც აუცილებელი იყო „დასავლური“ ენებისთვის, მაგალითად, გერმანულისა („უმლაუტიანი სიმბოლოები“) ან ფრანგულისთვის („მახვილებიანი სიმბოლოები“). საერთაშორისო სტანდარტის ISO/IEC-8859 დანერგვით ლათინურზე დაფუძნებული „ჩრდილო“, „ცენტრალური“ და „სამხრეთ ევროპული“ ანბანების გვერდით შესაძლებელი გახდა კირილიცასა და ბერძნული ანბანის გამოყენება. მათგან განსხვავებით ქართულ ანბანს იმ პერიოდში საერთოდ არ ჰქონდა კომპიუტერული მხარდაჭერა, რაც საგრძნობლად აფერხებდა ქართული ენის გაციფრულებას. მიუხედავად მეცნიერთა და ინფორმატიკის დარგის სპეციალისტთა ძალისხმევისა, რომლებმაც რამდენიმე შუალედური ოპტიკური ხელსაწყო შექმნეს, არსებითად არ შეცვლილა ქართული ენის გაციფრულებასთან დაკავშირებული ტექნიკური სირთულეები.

მოგვიანებით, 1990-იან წლებში, ბაზარზე გამოჩნდა ახალი ოპტიკური პროგრამა „FineReader“, რომელმაც გადამწყვეტი როლი შეასრულა ქართული ტექსტების მასშტაბურ გაციფრულებაში. მიუხედავად იმისა, რომ ქართული არ იყო იმ მრავალ ენათა რიცხვში, რომელთაც „FineReader“ უჭერდა მხარს, მაინც მოხერხდა მისი „გავარჯიშება“ ქართული ენის ტექსტების წასაკითხად. TITUS-ისა და ARMAZI-ის პროექტების ფარგლებში გაციფრულებული ძველი და საშუალი ქართულის რესურსების უმეტესი ნაწილი სწორედ მისი გამოყენებით შეიქმნა. 1990-იანი წლების ბოლოს კი დადგა გადამწყვეტი ეტაპი – შემუშავდა უნიკოდის სტანდარტი (ISO/IEC-10646), რომელიც მოიცავდა უნიკალური კოდირების სისტემას ქართული ანბანის სამივე სახეობისათვის (მხედრული, ნუსხა-ხუცური და ასომთავრული/მრგლოვანი).

თანამედროვე, ციფრული საუკუნის პირობებში მიმდინარე ტექნიკურმა რევოლუციამ ოპტიკური ამოცნობის სრულიად ახალი პერსპექტივები შექმნა, რამაც შესაძლებელი გახადა აქამდე გადაუჭრელი არაერთი პრობლემის დაძლევა, განსაკუთრებით ხელნაწერების გაციფრულების თვალსაზრისით. დღეისათვის შექმნილია ხელოვნურ ინტელექტზე (AI) დაფუძნებული ხელნაწერი ტექსტის ამოცნობის (HTR) წვრთნადი პროგრამული უზრუნველყოფა. პროექტის „მწიგნობრობის განვითარება კავკასიის ტერიტორიებზე (DeLiCaTe)“ ფარგლებში ამ სისტემის ქართულ ხელნაწერებზე გამოყენების პირველმა ცდამ ბოლო ექვსი თვის განმავლობაში შესანიშნავი შედეგი გამოიღო. ხელნაწერი ტექსტების ამოცნობის პლატფორმა *eScriptorium* შემუშავებულია პარიზის უნივერსიტეტში (Paris Sciences et Lettres University) სამეცნიერო პროექტების *Scripta* და RESILIENCE ფარგლებში, რომელთა განხორციელებაში ჩართული იყო სხვა ინსტიტუტებიც. პროექტები ნაწილობრივ დაფინანსდა ევროკავშირის პროგრამის „ჰორიზონტი 2020“-ისა და ენდრიუ ვ. მელონის ფონდის (Andrew W. Mellon Foundation) მიერ.

1980-იანი წლების ბოლოდან ქართული ბეჭდური წიგნების გაციფრულებისთვის სიმბოლოების ოპტიკური ამოცნობის (OCR) ხელსაწყოების დასახვეწად გაწეული ძალისხმევის მიმოხილვის, აგრეთვე, ხელნაწერი ტექსტის ამოცნობისა (HTR) და *eScriptorium*-ის შესახებ მოკლე შესავლის შემდეგ სტატიაში მოცემულია ამ უკანასკნელის ფუნქციონირების ძირითადი ასპექტები. ქართული ხელნაწერების ავტომატურად გადმოწერის ხარისხის გაუმჯობესების მიზნით *eScriptorium*-ის წვრთნა მოიცავს რამდენიმე

ძირითად ეტაპს: პლატფორმაზე ხელნაწერის ციფრული ფოტოების ატვირთვას, თითოეულ გვერდზე ტექსტით დაფარული სვეტ(ებ)ისა და სტრიქონების ავტომატურ სეგმენტაციას (სეგმენტაციის კორექტირება ზოგჯერ ხელით ხდება), მონიშნული გვერდების ავტომატურ ტრანსკრიბირებასა და ტრანსკრიბირებული ტექსტის ავტომატურ გასწორებას („alignment“) ხელნაწერის უკვე არსებული გამოცემის ტექსტთან. ზოგიერთი ხელნაწერის შემთხვევაში ტრანსკრიბირება მოხდა ხელით (ასეთი ამ ეტაპისთვის 292 გვერდია). რაც შეეხება ავტომატურად ტრანსკრიბირებულსა და გასწორებულს არსებული გამოცემის ტექსტთან, მათი ოდენობა 7488 გვერდს შეადგენს. დაგროვილი მასალა საშუალებას გვაძლევს, გავაუმჯობესოთ ტრანსკრიბირების აქამდე არსებული მოდელი. ველით, რომ ხელნაწერის ამოკითხვის სიზუსტის მაჩვენებელი 90%-ს გადააჭარბებს. აღსანიშნავია ისიც, რომ დაინტერესებული საზოგადოებისათვის ჩვენი შედეგები სრულად ხელმისაწვდომი იქნება. ეს ეხება როგორც უშუალოდ ტრანსკრიბირების მოდელს, ასევე, გამოუცემელ ხელნაწერთაგან გადმოწერილ და გასწორებულ ტექსტებსაც.