Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 37–44
DOI: 10.15439/2024F1772
ISSN 2300-5963 ACSIS, Vol. 41

Semi-automatic annotation of Greek majuscule manuscripts: Steps towards integrated transcription and annotation

Carina Geldhauser
0000-0002-9997-6710
Munich Centre for Machine Learning
Technical University Munich
Boltzmannstr. 3,85748 Garching bei München, Germany
and
ETH Zürich, Department of Mathematics
Rämistrasse 101, 8092 Zürich, Switzerland
Email: carina.geldhauser@ma.tum.de

Konstantin A. Malyshev
0009-0009-6338-5941
Saint Petersburg Theological Academy
nab. Obvodnogo kanala 17
Saint Petersburg, 191167, Russian Federation
Email: konstantin.a.malyshev@gmail.com

***Abstract*—We present a prototype for the integration of HTR transcription and semi-automated markup of textual features in the eScriptorium GUI.**

The prototype is designed for scholars working with ancient texts, who desire to perform standardized markup for a larger research project or digital edition. Motivated by research questions in Classics and Theology, we simultaneously investigate upcoming specific transcription challenges arising when working with ancient Greek manuscripts of majuscule type.

***Index Terms*—handwritten text recognition, named entity recognition, annotation, majuscule script**

I. Introduction

TEXTS play a central role in the work of a humanities scholar. New “distant reading” methods, where huge text corpora are analyzed through queries or statistical methods, open a whole new world of scholarly research questions.

However, not all objects of scholarly inquiry in the humanities are available as digital texts, and non-digital textual data must be, as a first step, extracted from the paper sources through Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR), and then processed further. Many research questions are sensitive to OCR quality, see section II-A1 for some examples.

As a second step, the digital plain text is *annotated*: Annotation[1] or markup, used interchangeably in this work, is the enrichment of a digital (plain) text with tags, categories, or standardized encoding for textual features.

This work is concerned with possibilities to integrate these two steps, transcription and annotation. We explore two possibilities on the example of a critical edition of manuscripts from the graeco-roman antique.

[1]If the text is already available digitally, then annotation may be the only step to prepare for "distant reading" research questions.

a) Example setting: The critical editions of works from the graeco-roman antique, used frequently by classicists and theologians, have to take into account that it is often not clear what is the actual text: Before the print age, works of popular authors had to be copied manually to be distributed. Both mistakes and deliberate alterations in the texts of such *manuscripts* happened, and *critical editions* have to display the variations between the available manuscripts.

As the variants of the text are a crucial part of the dataset, it is not advisable to use the same OCR post-processing methods as for printed material such as newspapers, see e.g. [1]. Those are often based on the comparison of words with available dictionaries, the algorithms remove hyphens etc, which means a potentially significant piece of information is lost through the post-processing.

Instead, at the current moment, scholars creating such digital editions are either manually correcting and annotating HTR or OCR- digitized manuscripts, or, in cases when the OCR software cannot deal with the used font/handwriting, the desired digital text needs to be established by manual typing, often employing a four-eyes principle to minimize transcription mistakes.

A separate research question is to draw conclusions by comparing the different variants. To automatize that comparison, a flawless *digital edition* (cf section II-B) with relevant annotation is crucial.

b) Motivation: Our work originates in the motivation to ease and facilitate the following task:

Group a set of available manuscripts into different tradition lines, i.e. order them by "similarity". For very significant ancient texts, e.g. Homer’s works or biblical texts (see the discussion in II-B), the aim may be to potentially establishing hypotheses of a common "parent manuscript". This is a very common, but challenging research question, see e.g. [2] for an example in medieval history. Depending on the precise formu-

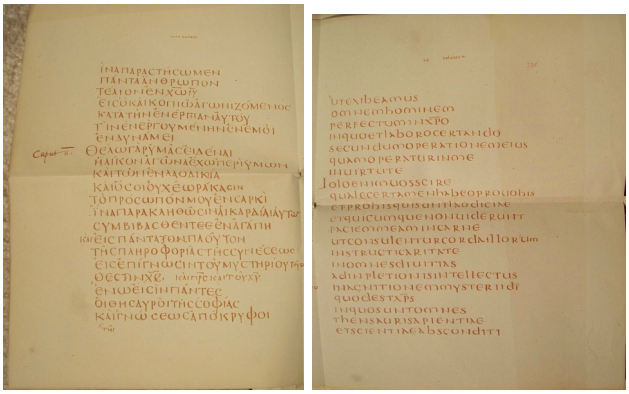

Fig. 1. Sample pages from Codex Claromontanus, taken from [3]. Both pages contain a textual variant of Colossians 1:28–2:3, nomina sacra with their characteristic overline bar, such as visible corrections by later scribes, most notably the insertion of καὶ in line 15 and diacritics in the Greek manuscript page.

lation, which might be cluster/group identification, similarity measures or the establishment of a proper "family tree" of manuscripts, the complexity of the task may vary, and usually involves non-trivial mathematical algorithms.

Particularly interesting for a comparison task as we described it above are so-called bilingual codices, see figure I-0b: these are manuscripts which contain the same content in two languages, in our case Greek and Latin. Here, a careful annotation of corresponding words and sentences may enable scholars in Classics or Theology to perform more diverse quantitative research tasks on the text at hand, and hence to potentially reach an improved hypotheses of manuscript relationships than if they were to look only at the Greek text.

To this aim, a very accurate transcription of the texts of each manuscript needs to be available, and a fine-grained annotation, with additional elements than what is commonly done, is necessary.

c) Main contribution of this work: The goal of our work was twofold: First, to address certain peculiarities in the ancient manuscripts at hand, listed in section III-B, by targeted HTR training to increase output quality. Second, we provide a prototype for relevant feature annotation, that may contribute to ease or speed up the preparatory work of classicists and theologians, so that they have more time to dedicate on their actual research questions.

A. Paper layout

The paper is structured as follows: In section II, we introduce the necessary background and state of the art. In section III, we describe our "dataset", i.e. Greek manuscripts of majuscule type, their special characteristics, and how the treatment of these characteristics are reflected in our work.

II. Background

A. OCR/HTR for the humanities

A lot of effort has been put into the digitization of old texts. The most prominent are OCR methods for printed texts, e.g. old newspapers, collections of letters, or early print editions of major literary works. For input available as text printed on old paper, the output quality of an OCR method may be negatively affected by fading ink or poor paper quality, which could be further worsened by a suboptimal scanning process, leading to additional distortions in the image. Furthermore, the OCR quality is also negatively affected by heterogeneity within the printed text: these may be layout features like changes in fonts or colors of letters within a page, but also text-inherent features like spelling changes (which enlarge the diversity of possible character sequences) or low distinctiveness of characters in se, e.g. the long "s" and "f" in older German texts[2]. Still, for printed texts, significant progress was made, leading to very satisfactory output, in terms of character error rates (CER) lower than 2%, see Ströbel and Clematide for results using Transkribus[3], Wick et al (2018) using OCRopy[4] and Calamari[5] or Martinek et al. [5] using a convolutional and recurrent neural network, combined with suitable preprocessing (e.g. binarization) and data augmentation.

For handwritten text recognition (HTR), the situation is different, and depends very much on the concrete case at hand, therefore, we do not even attempt to give a comprehensive overview here. In general, as text recognition tasks fall in the category of supervised machine learning, their performance depends on the available data, in particular on the number of samples and their variability, and reflects potential biases that are present in the data, see e.g. [6].

Broadly speaking, most progress has been made for contemporary handwriting in widely-used languages and writing systems, especially in the Latin alphabet[6]. Line recognition seems to be essential: most off-the shelf products have issues with recognizing rotated text.

The quality of HTR results depend, among other factors, on the amount of training data with ground truth available. Very good results were achieved e.g. for Manu McFrench [7], a model trained on almost 78.000 lines (more than 4 million characters) of French handwriting[7] from the 17th to 20th century, which reached a character recognition accuracy of 90.56% in version 3.

The situation is different for so-called rare scripts or historical writing styles. As the wording intends, we might have a significantly lower amount of training data available. Other obstacles are built-in assumptions on the nature of the script, i.e. the writing being from left to right, on horizontal lines, etc. We refer to [8] for extensive reflections on the matter.

[2]See Ströbel et al [4] for a detailed analysis.

[3]https://transkribus.eu/Transkribus/

[4]https://github.com/tmbdev/ocropy

[5]https://github.com/Calamari-OCR/calamari

[6]We did perform tests of Tesseract, kraken (command-line) and Google's Cloud Vision API for several languages using Cyrillic alphabet, with satisfactory results, however, this is rather anecdotal evidence; we did not aim at a systematic evaluation of OCR tools in this work.

[7]According to the authors [7], the overwhelming majority of training data is in French language, mainly handwriting, but with some percentages of print, and a few thousand lines of Spanish and English handwriting were added in the training dataset.

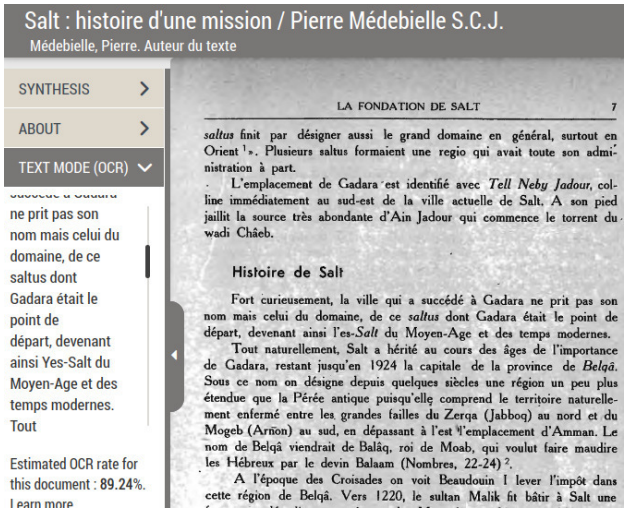

Fig. 2. Screenshot of a page in the Gallica collection [10], with estimated OCR accuracy given.

Furthermore, the current restrictions within the Transkribus HTR platform, which does not allow users to export models, even those which they trained themselves on their data, is a great disadvantage to the progress of HTR for rare scripts and historical writing styles.

1) Quick status assessment of DH for OCR: To summarize, Digital Humanities heavily profit from technical developments in OCR, but there is still space for improvement: the accuracy obtained so far may not be enough for reliable results on certain types of research question. For example, Chiron et al. [9] showed that OCR errors lead to significant missing relevant documents, as output of user queries in the OCRed Gallica collection: We see in Figure II-A1 a sample page from the Gallica collection, with an estimated OCR accuracy of below 90%. While this may sound fair, there is a drastic variance, and some infrequent, but highly relevant words may be wrongly ORC-red up to two thirds of their occurrences[8]. This may lead to biased query results, and therefore inaccurate answers to research questions.

Some scholars in the humanities, e.g. Smith and Cordell [11], even argue that the remaining errors in the digitized text is still "impeding advances in Digital Scholarship".

One of the applications of OCR/HTR where an accuracy of 90% is not enough is the establishment of a scholarly edition. As the research questions that motivated our prototype are closely related to the work on critical editions of ancient texts, we describe them briefly in the next paragraph.

B. Digital editions of ancient manuscripts

Digital scholarly editions are said to be the "crown jewels of Digital Humanities" [12], offering a plethora of ways of representing texts[9] and their transmission histories.

[8]See [9], page 5.

[9]Sahle has argued in [13] for the usage of the word "document" instead of "text", but due to the dominance of "text" also in scholarly literature, we use this intuitive notion also for the purpose of this work.

Roughly, we may distinguish three steps in digital editorial work of an ancient work: First, the provision of a digital (main) text[10] from the available sources, which usually involves the transcription/OCR/HTR of the raw material. Second, a markup/annotation[11] step, and third, an appropriate visualization[12], which includes tools for scholarly work with the edition. Scholarly editions of ancient manuscripts have certain peculiarities, among else:

a) Complex transmission histories: The transmission histories of centuries-old texts, such as Homer's works or biblical texts, is rich and highly complex, due to partial losses or damage, scribal errors, editorial decisions, and in general the huge spread and impact of these works. Hence, there may be many variants for the text to be presented in a new edition.

To make an illustrative example: there exist about 1000 manuscripts[13] of Homer's works, written on papyrus or parchment, and later on paper. These were copied by scribes multiple times over the course of two millennia, resulting in numerous losses and the introduction of many variations in the texts along the way. Which text is closest to the "real" Homer is an ongoing question of scholarly debate.

In the case of biblical manuscripts, the INTF Münster collects, curates and transcribes all available manuscripts of biblical texts. It prepares a scholarly critical edition of the New Testament, called the *Editio Critica Maior*, based on roughly 5800 manuscripts available to them at the moment.

b) Transparency about textual variations: A critical editions needs to be transparent about the different readings that are present in the different manuscripts. A "reading" in this case means an occurrence of a specific string on the paragraph, sentence, word or character level, in one manuscript, which is different from the string occurring in another manuscript. The two strings are then called "variants".

In order for this to be possible, the correctness of the transcription of each manuscript's text is extremely important, and human post-processing and error correction strictly necessary.

c) Less hierarchy of readings: A great opportunity of a digital edition is the possibility, through clever digital presentation and visual effects, to present different readings as equally plausible to the user, so that the user can decide for themselves which reading they want to adopt. This is a huge advantage w.r.t. the classical paper-based editions, which could only display one reading as the main text, and kept all variants in the apparatus.

[10]To establish the "correct text" in presence of variants is a huge field, which we do not want to enter here. To remain neutral in the debate, we use the uncommon wording "main text".

[11]As in the literature, we use "annotation" and "markup" as synonyms.

[12]With visualization, we mean the presentation of the TEI-XML code in a user-friendly environment, which helps scholars to answer their research questions. Available tools are e.g. The Versioning Machine or EVT https://visualizationtechnology.wordpress.com , but many larger projects build their own, customized tools.

[13]Estimate taken from https://www.lib.uchicago.edu/collex/exhibits/homer-print-transmission-and-reception-homers-works/homer-print/, last accessed: 30.09.2023.

C. Annotation

In this section, we describe the usage of annotation or markup of texts in a Digital Humanities context.

The aim of the annotation process is to provide a sufficient data enrichment as to allow adequate tools to answer a specific research question. Very popular is the markup of named entities, e.g. persons or places. Several stand-alone GUI tools[14] designed for scholars already exist, and they allow humanities scholars to find relevant attestations of named entities relevant to them. However, it is not possible to add further texts to the database of these tools, so that a scholar working on a different text has no advantage of them. Machine-Learning based NER tools such as *spaCy*[15] and *greCy*[16] have emerged lately, with accuracy strongly depending on context and the training data used by the developer. For example, as the models were trained on classical Greek, they do not lead to satisfactory results with late ancient Greek texts such as patristics. An overview of Named Entity Recognition models and challenges can be found in [14].

However, for some circumstances, the research question posed implies a need for manual annotation, as no appropriate tool exists. We illustrate our claim here on the example of creating visualizations on the usage of certain word categories, used in [15]: On the one hand, defining tags or categories is an independent research step that requires individual case decisions and therefore must be carried out by qualified personnel. On the other hand, most annotation tools are specialized, require significant time to learn and are limited in their distinctive features [16].

On the positive side, once mastered and provided with suitable annotated data, these tools not only make it possible to conduct detailed text-scientific research, but also to create visual forms of presentation of the text such as graphs, heatmaps and network graphs. Semantic markup of texts has been used for various purposes including categorizing handwritten annotations of an author [17], visualizing collaboration networks [18] and analyzing the lexical variance that occurs in the transmission of a medieval text [19].

As far as the creation of a digital edition is concerned, annotation of textual features in mark-up languages plays an integral role, as to provide functionalities to the user that they could not enjoy in a paper edition. Such annotation can be used to mark a variety of stylistic (such as text breaks and re-inkings) and semantic (such as place names, proper names and lemmatization) features, and we will describe this in detail in section III-B.

a) The text encoding initiative: Ideally, annotation follows a common standard, which allows for a group of scholars to build up upon each other's work. The text encoding initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. It develops and maintains a set of guidelines, the TEI Guidelines, which specify encoding methods, designed for the digital humanities community. According to the initiative[17], the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. The guidelines specify the semantics and interpretation of tags and attributes for basically all different textual components and concepts, from words to glyphs, persons, named entities etc.

In this work, we will follow the TEI guidelines for our automated markup, as it seems to be widely used within the digital classicist and digital theologian community.

III. Greek Majuscule texts and their HTR

In this section, we introduce the reader to the main points to consider when dealing Greek Majuscule manuscripts, and how it shaped our work. To keep this paragraph short and readable to a diverse audience, we consciously simplify and may use wordings which are intuitively understandable, yet suboptimal for the experts. We apologize in advance to the well-versed classicist or theologian among our readers.

A. Classical text features to consider

In the following, we emphasize those textual features in Greek majuscule manuscripts that may be unknown to the reader, but which need careful treatment or algorithm adjustments in the HTR step.

a) Scriptio continua: scriptio continua is an inherent feature of majuscule manuscripts. It means that there is no visual gap between two words, but the letters of the text are aligned with uniform distance to each other until a line break. See Figure [3] for an example. There are usually also no punctuation signs, diacritics, or distinguished letter cases used.

A line break does not have to coincide with the end of a word or syllable, but the scribe may decide to break the line at any point. This may be simply, for aesthetic reasons, at a certain, specified distance to the margin, irrespective of the ending of a word. An example is the fourth century Codex Sinaiticus.[18].

Scriptio continua is one of the factors that deteriorate the HTR quality significantly. Indeed, as reported also by Perdiki [20], who used the commercial software Transcribus, most erroneous output was caused by scriptio continua, such as misrecognition of accents, wrong punctuation or wrong word token splitting. In our case, we implemented a separate word-split functionality, described in section IV-C to get a transcription in our modern way of writing ancient Greek.

b) Nomina Sacra: "nomina sacra" are specific abbreviations for frequent words such as "God", "Christ" or "Jerusalem", used in biblical codices. In biblical manuscripts written on papyri or parchment, such as in figure III-A0b or figure I-0b, a nomen sacrum is marked with an overline bar, it is usually two or three letters long, with letters taken from the

[14]E.g. Recogito https://recogito.pelagios.org/, or Kima https://data.geo-kima.org/, which is specifically for places in Hebrew script already exist that find the occurrences (attestations) of these named entities in digital texts.

[15]https://spacy.io/universe/project/grecy

[16]https://github.com/jmyerston/greCy

[17]https://tei-c.org/

[18]High-quality photographs of Codex Sinaiticus are openly available on https://codexsinaiticus.org.

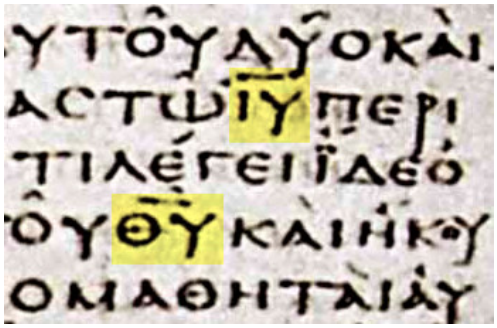

Fig. 3. Nomina Sacra (highlighted in yellow) in Codex Vaticanus, from [21], both are in genitive form, $\iota\upsilon$ represents the declined form of the word "Jesus", $\theta\upsilon$ represents the declined form of the word "God".

word it stems from. The last letter indicates the grammatical form of the abbreviated word. These specific characteristics make nomina sacra differ from generic abbreviations, e.g. on stone inscriptions.

Current research of New Testament textual scholars shows that the abbreviations are not unique, i.e. there might be a multitude of possible abbreviated forms used to encapsulate the same word. However, every nomen sacrum corresponds to a unique word.

c) Multiple hands and corrections: The major biblical codices all underwent changes by later scribes, to different extend or in different forms - diacritica were added, corrections were made, and more. The above-mentioned Codex Sinaiticus has undergone a particularly complex manual editing process over the centuries, and the investigation of the number of scribal hands, see e.g. [22], [23], is still ongoing research.

To our best knowledge, current machine-learning based results on scribe distinction for ancient or medieval handwritten texts are rather scarce, limited to binary classification ("is it scribe A or not?"), need a full page consisting of only one writing hand, and have, up to now, unsatisfactory accuracy. The automatic identification of different scribes in one manuscript remains a highly desired, but challenging feature.

B. Relevant textual features

The following text features are relevant towards a possible quantitative analysis, already within one text, before comparison takes place:

1) line breaks and other breaks in the original manuscript
2) multiple hands, especially corrections by a later scribe
3) re-inking (redrawing of letters)
4) highlighting, e.g. initial letters and flared letters
5) paratextual elements such as titles, marginal glosses, etc.

IV. First results in semi-automatic annotation of textual features

In this section, we first give an overview of our work, followed by subsections on each specific step.

A. Outline and Assumptions

In this work, we focus on the situation of scholars working with ancient manuscripts. Our **test case** and primary example are biblical manuscripts in majuscule style and scriptio continua, see Figure I-0b for an example. We assume that high quality images of the manuscripts are available, and the goal of the digital edition is to provide both a transcription of the text contained in these images, enriched with annotations in TEI-XML, and, in a later step, additional features that allow scholarly work with the text at hand, e.g. an apparatus, a way to compare the texts[19] displayed in different witnesses, or named entity recognition [24]–[26].

B. Overview

In our work, we developed a very first tool that may allow scholars working on digital editions to carry out annotations or other textual enrichment with less effort and without knowledge of a programming language. To be of real use for the humanities community, we aimed to make these tools available to the public in an easy-to-use form, i.e. inside a tool that is already in use, and with a visual frontend to avoid usage of scripts, codes or the opening of the command-line.

Taking into account the needs of our humanities colleagues, we decided to contribute to eScriptorium, a digital text production pipeline for print and handwritten texts using machine learning techniques [27]. The advantages of eScriptorium, from our perspective, are described in section IV-F.

The fork we created includes a couple of extra functions for annotation in TEI-XML standard (see section II-C) in a semi-automated manner, i.e. the user uploads a text file which contains the words, names or places they wish to be annotated. Our fork allows to export this skeleton annotation as “custom XML”, see figure IV-E, which can be further enriched manually in any other tool, e.g. Oxygen XML editor.

C. Word splitting

For word-splitting tasks we decided to use the *SymSpell* library[20]. It is based on so called "Symmetric Delete" spelling correction algorithm[21]. The crucial advantage of this library is that it can not only split the continuous non-space string into words, but also can correct errors to some degree. This is an important case, because the OCR/HTR step never delivers 100% recognition accuracy. For example, the string "cnebigeIefant" has two errors: "c" instead of "o", and "I" instead of "l", and if they are fixed and the words are splitted, we get "one big elefant". The SymSpell library must be provided with the words dictionary, sorted by their assumed frequency in the given text. In addition, it can also use a bigram or trigram dictionary, where occurrences of two or three words

[19] Here, we take a viewpoint similar to Sahle’s [13] “Text als Fassung”, where we define the “text” as the “reading”, i.e. the ensemble of words presented in a particular physical witness.

[20] https://github.com/wolfgarbe/SymSpell

[21] We refer the interested reader to the introductory notes in https://seekstorm.com/blog/fast-word-segmentation-noisy-text/ and https://seekstorm.com/blog/fast-approximate-string-matching/.

are sorted according to their frequency in the language. The most probable words candidates are defined according to the frequency dictionaries and the Damerau-Levenshtein distance between the given string and the candidate. Note that this dictionary should also include nomina sacra (see III-A) or other abbreviations, in order to increase accuracy.

At the moment, two unsolved problems remain: First, each text has its own distribution of words frequencies, and this can vary a lot. This means that the standard language dictionary should not be used, but an additional custom-dictionary-generation step is required, which takes into account the peculiarities of the given text. The second unsolved question is: how to adjust the maximum Damerau-Levenshtein spelling correction distance to the optimal value? As far as we oversee the issue, this step is dependent on the accuracy of the HTR step.

D. Semi-automated annotation

Though machine-learning based NER tools are available, see the discussion in section II-C, we decided for a rule-based implementation, as also other recent work, e.g. the *Opera Graeca Adnotata* [28]. Our decision relied on the following thoughts:

First, there are currently no good greCy models for majuscule Greek on which we could rely on. This means also that annotation will not be available to a user with datasets of different type than what an external tool can handle. A rule-based implementation is independent on external tools.

Second, despite necessary post-processing might always be done for delicate tasks related to digital editions, our users prefer reliable output, instead of having to deal with "false-positives": Therefore, we prefer to annotate less, but to do that with the highest precision possible, avoiding ML biases.

Third, a direct linkage with external tools is both delicate to implement and vulnerable to break down, due to software changes from version to version.

Therefore, we decided for the following approach: the user creates a list of relevant words for annotation, using a tool of their choice. This list, in plain text format, can then be uploaded into our eScriptorium fork, where the annotation algorithm creates TEI-standard markup of all words in the list with one click. This means that all grammatical forms of a word have to be provided in the word list.

E. Annotated features

We coded and incorporated into our fork of eScriptorium a semi-automated TEI mark-up functionality for Hebrew and Greek personal names and place names, numerals, nomina sacra,(see section III-A), punctuation signs and line and page breaks in the original manuscript.

F. Why eScriptorium?

After preliminary tests using eScriptorium, Transkribus, but also tools without GUI, such as Tesseract, kraken and Google Cloud Vision API's Document OCR tool, we realized that very few transcription tools are able to deal with Greek majuscule manuscripts: some were simply not designed for ancient texts, and therefore gave terrible results, probably due to the uncommon font and the scriptio continua.

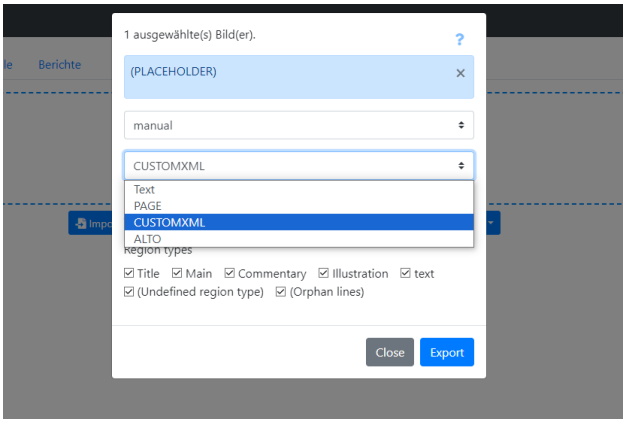

Fig. 4. Screenshot of our prototype front-end: exporting the annotated text XML file.

As a GUI was important to us due to its friendliness to the relevant scholarly community, the decision stood between eScriptorium and Transkribus.

As our tests found no significantly better performance of Transkribus w.r.t. eScriptorium we decided to take advantage of the open source nature of eScriptorium, which made us more flexible, saved resources, and hopefully allows for an easy adaptation of our fork in the relevant community.

During the work with eScriptorium, we came to enjoy its additional advantages: The core code base appeared to be clean, understandable and well-designed. The internal architecture of the tool is modular, well-structured and easy-to-extent. This makes it ideal for an expandable open-source project. It was not complicated to integrate our custom code into the eScriptorium core.

G. Annotation of nomina sacra

The current fork of our project[22] contains a function that expands abbreviated nomina sacra from the transcribed text. For this, the user needs to provide a list of abbreviations used, or adapt our default, provided for Greek and Latin letters.

The advantage of this approach is its versatility towards different writing styles in Greek, Latin, or any other language the user works with inside eScriptorium: This way of expanding a nomen sacrum works as soon as the transcription obtained by eScriptorium's transcription step is accurate, it is independent of the model parameters used or trained. As such, it can be used also to expand and annotate nomina sacra in texts written in minuscule style, or even printed editions, e.g. by 19th century scholars like K. von Tischendorf [29].

Another method, dependent on the used model, is to link transcription and annotation of nomina sacra and other "relevant" features directly. With "relevant" we intend annotation that can be used to answer research questions or to serve the

[22] Available at https://gitlab.com/archtype/escriptorium/-/branches

reader of a critical edition in their exploration and understanding of the text. With "direct linking" we intend an integrated HTR and annotation pipeline, namely to recognize a nomen sacrum from the overline bar in the handwritten document, and then to suggest the correct expansion directly in an annotation, without a human-created list.

This implies training of the underlying machine learning model to an extremely high accuracy, in order to recognize nomina sacra by the overline bar used by the scribe on the abbreviated version (recall Figure III-A0b). We discuss this in section V.

H. Annotation of numerals

The annotation of numerals is less common, as only relevant for very specific research questions. We included it both as a 'placeholder annotation', i.e. it may be replaced by the user, to instead annotate something else, and in order to show the limitations of the algorithmic side: Our algorithm technically checks all words in their order of appearance in the text, linearly going through the text word-by-word. This linear processing will recognize a numeral and annotate it immediately. However, if this numeral is part of a compound number word, the linear processing will not be successful.

I. Towards an integrated HTR and annotation pipeline

As discussed above, an integrated HTR and annotation pipeline could use certain image features directly for "relevant" annotation. To achieve the necessary high accuracies and density of appearance of nomina sacra, we used a version of "data augmentation": we created 50 pages of artificial digital manuscripts [30] containing all possible grammatical forms of the nomina sacra, available on Zenodo[23]. See figure IV-I for an example. Our set of augmented images uses genuine biblical unicial fonts[24] and incorporates a variety of visual characteristics that deteriorates the quality of a scanned manuscript page, such as distortions, heterogeneities in the background color, damaged or partially degraded paper ("dark spots") etc.

V. Summary and Outlook: Integrating HTR and annotation

The presented prototype provides a semi-automated markup functionality to the HTR transcription step of eScriptorium. We started with a couple of exemplary features to annotate through a rule-based algorithm. While a rule-based approach provides maximum accuracy, it needs well-prepared input of the user. Also, the annotation of multi-word numerals turned out to be difficult.

a) Current work: At the moment, we work on the improvement of the performance of the new transcription model in terms of the recognition of nomina sacra. While our team developed a bounding box model (via Kraken) achieving high accuracy levels (>90%), the training of a baseline model (done directly in eScriptorium) resulted in lower accuracy rates. The

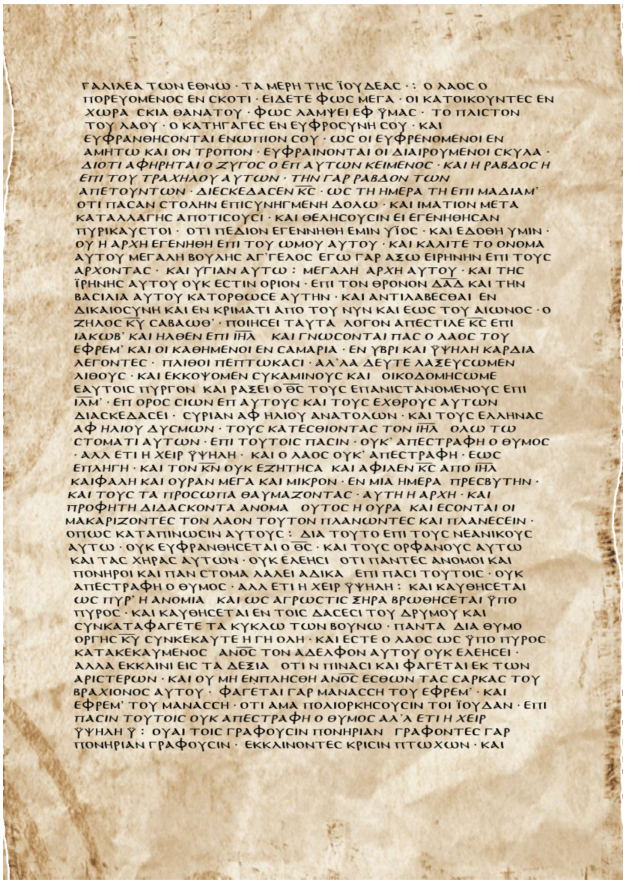

Fig. 5. Sample page from our data augmentation set. The dataset is available on Zenodo [30].

improvement of our transcription model in a baseline format will continue through the feeding of additional training data. We envision that the complete automated annotation of nomina sacra will be possible through these improvements.

b) Discussion: This work was motivated by the work of our colleagues on digital editions and manuscript comparison.

We are happy that our prototype saves their time in the annotation step of their work: the most common or basic annotations are already taken care of algorithmically, and the researcher can correct both OCR errors and missing annotations in the same round of manual interventions.

Note, however, that this is only a small part of the time that our colleagues invest in building a good dataset / a good critical edition: For these aims, an OCR accuracy of 90 % is barely enough to make OCR-transcription and subsequent manual correction as fast as they were when they transcribed completely manually.

Hence, in order for a digital tool to be truly useful for them, a much higher OCR accuracy is needed.

c) Further Steps: A few open questions and improvement points have been pointed out in the various sections above, e.g. in IV-C we pointed out the need of custom-dictionary-generation and other open issues in the word split-

[23]URL: https://zenodo.org/records/12755706

[24]Available at http://individual.utoronto.ca/atloder/uncialfonts.html

ting step. With regard to further developments in the project, we aim to explore other annotation options for our prototype, to accommodate a larger variety of research questions. One idea is adding GPS information for annotated place names. In view of scholarly research questions on manuscript transmission history, scribal habits and cultural heritage questions, the annotation of a variety of more subtle visual features, such as different scribal hands, deletions and re-inkings, are also envisioned.

ACKNOWLEDGMENT

This work was partially supported by the German Federal and Bavarian Ministries of Education and Research, through its financing of the Munich Center for Machine Learning. We thank J. Heilmann and N. Müller for supplying us ground truth data, and our students J. Ehness, N. Dominko Kobilica, P. Kumar and I. Tuncel for their help in the code development and testing phase.

REFERENCES

[1] B. Alex, C. Grover, E. Klein, and R. Tobin, "Digitised historical text: Does it have to be mediocre?." in *KONVENS*, 2012, pp. 401–409.

[2] P. Roelli and D. Bachmann, "Towards generating a stemma of complicated manuscript traditions: Petrus alfonsi's dialogus," *Revue d'histoire des textes*, vol. 5, pp. 307–321, 2010.

[3] Wikimedia, "Codex claromontanus, the greek text of colossians 1:28-2:3," 2024, [Online; accessed July 20, 2024]. [Online]. Available: https://en.wikipedia.org/wiki/Codex_Claromontanus#/media/File:Claromontanus_2_greek.jpg

[4] P. B. Ströbel, S. Clematide, and M. Volk, "How much data do you need? about the creation of a ground truth for black letter and the effectiveness of neural ocr," *Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, 2020*, 2020.

[5] J. Martínek, L. Lenc, and P. Král, "Training strategies for ocr systems for historical documents," in *Artificial Intelligence Applications and Innovations: 15th IFIP WG 12.5 International Conference, AIAI 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings 15*. Springer, 2019, pp. 362–373.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] A. Chagué, T. Clérice, J. Norindr, M. Humeau, B. Davoury, E. Van Kote, A. Mazoue, M. Faure, and S. Doat, "Manu mcfrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets," in *Digital Humanities 2023: Collaboration as Opportunity*, 2023.

[8] P. Stokes and B. Kiessling, "Sharing data for handwritten text recognition (htr)," *Digital Humanities in Practice*, 2024.

[9] G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J.-P. Moreux, "Impact of ocr errors on the use of digital libraries: towards a better access to information," in *2017 ACM/IEEE joint conference on digital libraries (JCDL)*. IEEE, 2017, pp. 1–4.

[10] T. G. Collection, "Pierre médebielle s.c.j. gallica (auteur); salt: Histoire d'une mission (texte)," 2024, [Online; accessed July 20, 2024]. [Online]. Available: https://gallica.bnf.fr/ark:/12148/bpt6k91248315/f7.item#

[11] D. A. Smith and R. Cordell, "A research agenda for historical and multilingual optical character recognition," *NUlab, Northeastern University.@ https://ocr. northeastern. edu/report*, p. 36, 2018.

[12] E. Pierazzo, "A rationale of digital documentary editions," *Literary and linguistic computing*, vol. 26, no. 4, pp. 463–477, 2011.

[13] P. Sahle, "What is a scholarly digital edition?" *Digital scholarly editing: Theories and practices*, vol. 1, pp. 19–39, 2016.

[14] K. Pakhale, "Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges," *arXiv preprint arXiv:2309.14084*, 2023.

[15] W. Riess, "Prolegomena zu einer digitalen althistorischen Gewaltforschung: Gewaltmuster bei Solon, Alkibiades und Arat im Vergleich," *Klio*, vol. 102, no. 2, pp. 445–473, 2020.

[16] A. Przepiórkowski, "Tei p5 as an xml standard for treebank encoding," in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*, 2009, pp. 149–160.

[17] S. A. and A. M. Del Grosso, "Giorgio bassani's notes between tradition and innovation." *Digital Humanities 2023: Book of Abstracts*, 2023.

[18] M. A. Cipolla, A. Cappellotto, M. Rospocher *et al.*, "Collaboration practices between people and tools: the case of" snorra edda. a collaborative bibliography (snecb)"," in *Digital Humanities 2023: Book of Abstracts*, 2023, pp. 93–94.

[19] S. Moors, "Constrained. a computational study of the influence of formal characteristics on the transmission of the middle dutch martijn trilogy by jacob van maerlant." *Digital Humanities 2023: Book of Abstracts*, 2023.

[20] E. Perdiki, "Preparing big manuscript data for hierarchical clustering with minimal htr training," *Journal of Data Mining & Digital Humanities*, no. Sciences of Antiquity and digital humanities, 2023.

[21] Wikimedia, "Nomina sacra in codex vaticanus john 1," 2024, [Online; accessed July 20, 2024]. [Online]. Available: https://commons.wikimedia.org/wiki/File:Nomina_Sacra_in_Codex_Vaticanus_John_1.jpg

[22] D. Jongkind, *Scribal Habits of Codex Sinaiticus*. Gorgias Press, 2013.

[23] A. Wilson, "Scribal habits in greek new testament manuscripts," *Filología neotestamentaria*, vol. 24, pp. 95–126, 2011.

[24] R. Hanslo, "Deep learning transformer architecture for named-entity recognition on low-resourced languages: State of the art results," in *PROCEEDINGS OF THE 2022 17TH CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS (FEDCSIS)*, ser. Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., 2022, pp. 53–60.

[25] R. Sharma, D. Chauhan, and R. Sharma, "Named entity recognition system for the biomedical domain," in *PROCEEDINGS OF THE 2022 17TH CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS (FEDCSIS)*, ser. Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., 2022, pp. 837–840, 17th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, BULGARIA, SEP 04-07, 2022.

[26] R. Hanslo, "Evaluation of neural network transformer models for named-entity recognition on low-resourced languages," in *PROCEEDINGS OF THE 2021 16TH CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS (FEDCSIS)*, ser. Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Slezak, Eds., 2021, pp. 115–119, 16th Conference on Computer Science and Intelligence Systems (FedCSIS), ELECTR NETWORK, SEP 02-05, 2021.

[27] B. Kiessling, "Kraken-an universal text recognizer for the humanities," in *Proceedings of the DH2019 Conference*, 2019.

[28] G. G. Celano, "Opera graeca adnotata: Building a 34m+ token multilayer corpus for ancient greek," *arXiv preprint arXiv:2404.00739*, 2024.

[29] C. v. Tischendorf, *Novum Testamentum graece*. Leipzig, 1841.

[30] C. Geldhauser, "Artificially created image files resembling ancient greek manuscripts in majuscule script (version v1) [data set], zenodo." 2024, [Online; accessed July 23, 2024]. [Online]. Available: https://doi.org/10.5281/zenodo.12755706