Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6585–6589
Marseille, 20-25 June 2022

Handwritten Paleographic Greek Text Recognition: A Century-Based Approach

Paraskevi Platanou, John Pavlopoulos, Georgios Papaioannou
Athens University of Economics and Business, Greece
{platanou20, annis, gepap}@aueb.gr

Abstract

Today classicists are provided with a great number of digital tools which, in turn, offer possibilities for further study and new research goals. In this paper we explore the idea that old Greek handwriting can be machine-readable and consequently, researchers can study the target material fast and efficiently. Previous studies have shown that Handwritten Text Recognition (HTR) models are capable of attaining high accuracy rates. However, achieving high accuracy HTR results for Greek manuscripts is still considered to be a major challenge. The overall aim of this paper is to assess HTR for old Greek manuscripts. To address this statement, we study and use digitized images of the Oxford University Bodleian Library Greek manuscripts. By manually transcribing 77 images, we created and present here a new dataset for Handwritten Paleographic Greek Text Recognition. The dataset instances were organized by establishing as a leading factor the century to which the manuscript and hence the image belongs. Experimenting then with an HTR model we show that the error rate depends on the century of the image.

Keywords: handwritten text recognition, Greek handwriting, digital paleography

1. Introduction

Recently, there is an increased interest for the improvement of Optical Character Recognition (OCR) software performance in terms of recognition accuracy(Jang, 2020). OCR targeting printed text material can indeed offer impressive results, but OCR is significantly hindered, when switching to Handwritten Text Recognition (HTR) (Bathla et al., 2016; Nair et al., 2017). The HTR task for handwritten text in Greek, among other languages, is notoriously difficult due to various traits, such as the existence of multiple representations for the same character, character visual similarity, significant differences in writing style, as well as the impact of writing implements and writing practice.

This paper investigates HTR of Greek paleographic manuscripts, by introducing a new dataset and fine-tuning an AI-powered HTR model to improve the current state of the art, in terms of recognition rate. Since style, the language and writing tools change in the span of the target era, we study the performance as a function of the century the manuscript was written. This model can be further used to assist the text recognition of many more Greek paleographic manuscript images. More specifically, the contributions of this work are the following two:

- The transcription of 77 Greek paleographic text images, dating from the tenth to the sixteenth century, developing a dataset that we release for public use.[1]
- An investigation of the performance of AI-powered HTR by century, which reveals the challenges in paleographic Greek HTR and concludes that texts belonging to the fifteenth and sixteenth centuries are especially challenging.

The rest of the article first presents the new dataset, then describes an empirical evaluation that was performed. The paper concludes with our remarks and suggestions for future work.

2. The new HPGT dataset

The dataset that was developed and presented in this work consists of images of Handwritten Paleographic Greek Text. This section presents information concerning the source, the content, the selection procedure, and the characteristics of this dataset.

2.1. Source

The images of folios used in our study can be freely browsed from the website of a wide digitization project generously supported by The Polonsky Foundation.[2] The Polonsky Foundation Digitization Project was carried out between 2012 and 2017 and its digitized collections include early printed books, Greek manuscripts, Hebrew manuscripts and Latin manuscripts. The digitized collection used in our study is the Greek manuscripts one. The Barocci collection consists of 244 volumes and it is the largest acquisition of the Bodleian collection. The dates of these manuscripts range from the 8th century AD to the 17th century AD. In our study we decided to categorize manuscripts into seven groups based on the century to which they date. This means that we are not interested in manuscripts belonging to more than one group. To

[1] https://github.com/vivianpl/hpgtr

[2] Retrieved January 15, 2021, from http://bav.bodleian.ox.ac.uk/greek-manuscripts.

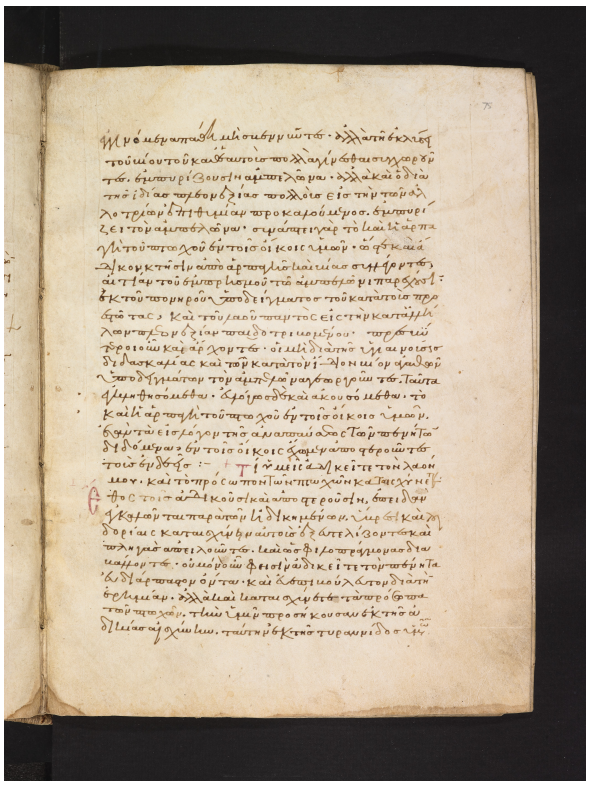


Figure 1: HPGTR.N full-page sample.

our knowledge, there is not an online available published study concerning HTR for the Barocci collection. Recognition studies with other Bodleian collections involve handwritten text recognition in historical manuscripts (Edwards III, 2007), the Bodleian Library's Book of Curiosities Project (Krätli and Lydon, 2011) and Letter identification in tremulous medieval handwriting with the aid of an ensemble of evolutionary algorithms (da Silva et al., 2021).

2.2. Dataset contents

Our dataset consists of two incrementally built editions of data. HPGTR.N involves no segmentation, whereas HPGTR.S is the result of careful line segmentation. That said, both editions comprise folio images processed, and thus, of different size. HPGTR.N consists of an image dataset of 77 items. Each image displays the folio and hence text, while it may also display other items necessary for manuscript digitization. These items may include a ruler as well as tools enabling the manuscripts to stay open. Apart from external items, characteristics other than the text itself may also be demonstrated on the image. These characteristics concern the manuscript and may be associated with either its production or its condition. Page numbers, dust and dirt are some of the usual characteristics. The seventy HPGTR.N images are grouped into seven categories according to the date of the manuscripts; ten images for each century between the 10th and the 16th. The ten images of the 16th century group belong to two different manuscripts. That said, images of eight digitized manuscripts in total, such as the one presented in Figure 1, are used in our study.

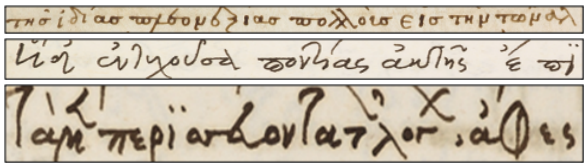


Figure 2: Three HPGTR.S samples.

HPGTR.S is the second edition for the development of which we examined carefully the first edition's images and we selected the images that follow specific criteria. We are particularly interested in images which demonstrate dirt- and dust-free pages as well as characters set in horizontal upright position. This is because we wish to provide Machine Learning models with training images, which demonstrate easy-to-segment text. After image selection, we segment five text lines in each image. Text line segmentation is associated with document analysis, while automatic line segmentation is still an open research issue (Ratheash and Sathik, 2019). The goal of our line segmentation task is to provide training images, which can be used as input to machine learning models. Two images from each century image group are used for this task. After establishing a list of guidelines for segmenting our lines, which includes cropping and keeping the area of interest only in line shape and thus, letting aside characters violating the rule, while ignoring breaths, accents, stresses and punctuation marks in case they are not part of our line shape, we succeeded to segment 1,906 lines in total. Figure 2 provide three samples of this edition. The new edition consists of images of cursive script as well, as shown in Figure 2 - middle. In this work, we only experimented with HPGTR.N , leaving experimentation with HPGTR.S for future work.

2.3. Image dataset selection criteria

The specific image dataset is used due to the fact that the respective folios serve as characteristic examples of the writing style they represent. They include both Greek minuscule script and the cursive style of the minuscule script. In this way, our work involves examination of different styles and is not limited to one style of script. This enables us to draw conclusions on machine reading ability demonstrated in different styles.

2.4. Image dataset characteristics

The digitized manuscript texts share a significant number of common characteristics. At first, there is the distinction between minuscule clear writing style (Figure 2 - top) and cursive style (Figure 2 - middle) (Thompson, 2013) which one can see in the sixteenth century's image group. In our cursive style samples, the characters tend to form connections with one another in such a way that one cannot easily determine

an empty space between them. The characters are of various sizes, since some of them can extend beyond the line, while it is possible that there is size inconsistency in the same character group as one can see in the last word of the lowermost sample (ἄφες) in Figure 2, where the character "φ" differs in size from the others. One can also notice that the characters are not always grouped together, having both different kerning and base line offset, which is often the case of the fourteenth century's image group. Figure 2 - bottom illustrates this fact. When in a group, they may join each other and form ligatures. Ligatures, which are actually characters containing two or more united characters, are more or less frequent and appear in manuscripts of all centuries investigated. Another characteristic of the dataset is that, although the script is lowercase, in many cases the text includes both uppercase and lowercase characters. For instance, one can see in Figure 2 - top that the capital letter "H" is used instead of the small letter "η". Moreover, since we work with Greek scripts, we often encounter other symbols apart from letters, such as the Greek breaths and accents. These marks may be placed over the associated syllable or even further in the text. Going back to the character group ἄφες in Figure 2 - bottom, we can see that the stress is placed well beyond the affected character "α".The fact that they are not often aligned with the corresponding letter, renders the marks difficult to read for the machine, in addition to other factors. For these reasons we decided that we will not transcribe them although we admit that their inevitable presence in our many cases of our editions may come at a cost; i.e., the recognised text might not include the expected marks and accents.

3. Methodology

The HTR tool we opted for is Transkribus 1.15.1, which is designed to accommodate Artificial Intelligence-powered text recognition and transcription of historical documents (Kahle et al., 2017).

3.1. Training and test data

We randomly selected 63 out of the 77 HPGTR.N images for training and we used 7 images for validation and 7 for testing. The selection was stratified across centuries, so that one image per century was used for validation and testing while nine images per century were used for training. We used the training and the validation data to train Transkribus and we report its accuracy on the seven test images.

3.2. Experimental results

Training was performed for 25 epochs. The Character Error Rate (CER) on the training set was 14.96% and on the validation set it was 17.16%. Plotting the CER against the individual centuries the texts come from, as shown in Figure 3, we observe an imbalance in CER across centuries. Although the model performs well

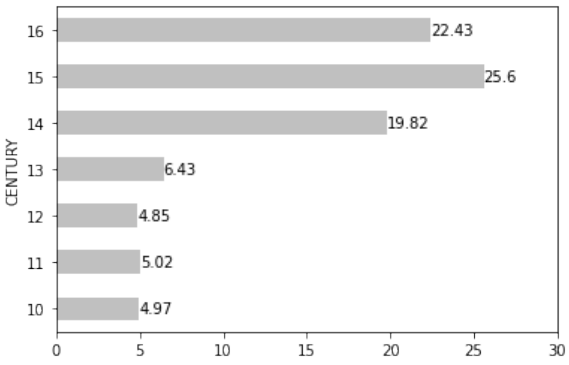


Figure 3: Character error rate (percent; shown horizontally) in HPGTR.N per century.

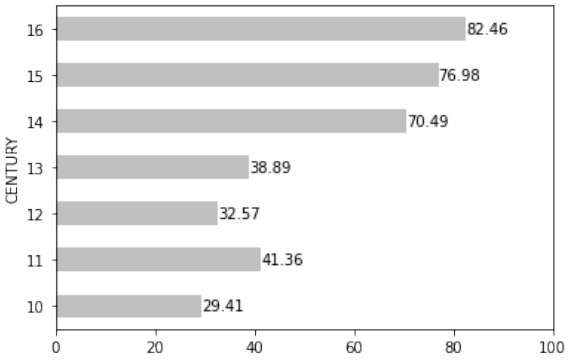


Figure 4: Word error rate (percent; shown horizontally) in HPGTR.N per century.

in the first four century image groups, poor character recognition performance characterizes the century image groups after the thirteenth one.

Figure 4 presents the Word Error Rate (WER) scores, which are higher than the character-based scores. A similar imbalance is shown, with WER being much higher for images of the last three centuries, from the 14th to the 16th, but with the highest error rate now corresponding to the 16th century. This may be due to the presence of cursive style in the 16th manuscript pages.

3.3. Error analysis

The dataset organization by century enabled us to find that Character Error Rates vary across the centuries of interest. An error analysis discussed in this section, allowed us then to study the reasons behind character misrecognition. Words are not easy for the model to recognize because of the often missing spaces between them; the model cannot detect spaces, where there is no clear character separation. Although there are reasons which can account for false word recognition (e.g., unexpected empty space between characters of the same word or the size inconsistency discussed in Section 2.4), the cause of the misinterpretation of individual characters is not that obvious, given that the text used is clear enough for the machine to read. However, error analysis combined with a better look at the

characters reveal that a considerable number of characters in the case of the fourteenth, fifteenth and sixteenth century data groups present at least one of the difficulties discussed below.

Table 1 provides a list of some of these problematic characters and their respective transcriptions. We present representative images of characters found in the fifteenth century data group. Most of these characters appear in the fourteenth and sixteenth century data groups as well. However, they are very frequent in the particular century data group we examine. These characters are the result of a particular procedure which is usually merging, grouping or symbolization. There is a tendency towards cursive writing style in the particular pages, which leads to character union; the scribe did not lift up the hand, when the next character was to be written. Nevertheless, according to Table 1, it seems that such union does not take place everywhere in the text but, on the contrary, there are specific characters which tend to form groups. Another way of drawing characters, which is quite interesting, is matching characters with symbols. In this case, lines and curved lines stand for specific character combinations and they usually appear at the end of the word or the line.

What we wish to show here is the fact that when the model is to read such complicated characters, it is extremely possible that it will make false predictions, leading to higher CER. Our hypothesis is verified by the frequency numbers, which show that problematic characters in total can account for misclassification.

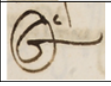
Apart from the character list provided above, there are other difficulties as well. There is a tendency in the sixteenth century image group texts to writing characters nested inside other characters. This is usually the case with the character "o" which is often magnified and specific characters are written inside it. Table 2 includes representative images of this character combination.

Character shape and legibility go hand in hand but other difficulties arise in the data groups of interest as well. The text images of the fourteenth and sixteenth century data groups do not demonstrate only the main text but also characters between the lines of interest, which are either "glosses" or text analysis. This may also account to some extent for high character error rates in these data groups.

Table 1: Problematic 15th c. AD characters.

MANUSCRIPT ID	CHARACTER	INTENDED MEANING	FREQUENCY
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Το	6
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Ει	9
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Στ	19
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Ου	14
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Σθ	5
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Ευ	8
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Υν	5
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Με	21
Bodleian-Library-MS-Barocci-59_00076_fol-42v		Εν	7

Table 2: Problematic 16th c. AD characters.

MANUSCRIPT ID	CHARACTER	INTENDED MEANING	FREQUENCY
Bodleian-Library-MS-Barocci-66_00322_fol-155v		Οι	13
Bodleian-Library-MS-Barocci-66_00322_fol-155v		Ον	3
Bodleian-Library-MS-Barocci-66_00323_fol-156r		δσ	3

4. Conclusion

In this work, we focused on the effort to automate transcription of Greek paleographic manuscripts dating from the tenth to the sixteenth century. To this end, we created two datasets with a parallel corpus of transcriptions and we experimented on HPGTR.N with an AI-powered handwritten text recognition tool. The data organization by century shows that paleographic text belonging to the 15th and 16th centuries may be challenging, when it comes to handwritten text recognition, as demonstrated by our error analysis.

In future work, we plan to extend the dataset with more images. Also, we will experiment with post-correction techniques that could potentially decrease the high word error rate. Finally, we will investigate the generalisation of our findings in other languages, where the error may also depend on the century the image belongs to (Derolez, 2003).

5. Acknowledgements

All digitized content comes from the Bodleian Libraries ©2013, used here under a Creative Commons Attribution-Noncommercial-ShareAlike licence. This work was partially funded by the Department of Informatics, Athens University of Economics and Business.

6. Bibliographical References

Bathla, A. K., Gupta, S. K., and Jindal, M. K. (2016). Challenges in recognition of devanagari scripts due to segmentation of handwritten text. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 2711–2715. IEEE.

da Silva, R. S., Costa-Abreu, D., Smith, S., et al. (2021). Investigating the use of an ensemble of evolutionary algorithms for letter identification in tremulous medieval handwriting. *Evolutionary Intelligence*, 14(4):1657–1669.

Derolez, A. (2003). *The palaeography of Gothic manuscript books: From the twelfth to the early sixteenth century*, volume 9. Cambridge University Press.

Edwards III, J. A. (2007). *Easily adaptable handwriting recognition in historical manuscripts*. University of California, Berkeley.

Jang, S. J. (2020). Ocr related technology trends. *European Journal of Engineering and Technology Vol*, 8(1).

Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017). Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE.

Krätli, G. and Lydon, G. (2011). *The Trans-Saharan Book Trade: Manuscript Culture, Arabic Literacy and Intellectual History in Muslim Africa*, volume 3. Brill.

Nair, P. P., James, A., and Saravanan, C. (2017). Malayalam handwritten character recognition using convolutional neural network. In *2017 International conference on inventive communication and computational technologies (ICICCT)*, pages 278–281. IEEE.

Ratheash, R. S. and Sathik, M. M. (2019). A detailed survey of text line segmentation methods in handwritten historical documents and palm leaf manuscripts. *International Journal of Computer Sciences and Engineering*, 7(8):99–103.

Thompson, E. M. (2013). *An introduction to Greek and Latin palaeography*. Cambridge University Press.