Preparing Big Manuscript Data for Hierarchical Clustering with Minimal HTR Training

Elpida Perdiki[1] [*]

1 Department of Greek Philology, Democritus University of Thrace, Greece

*Corresponding author: eperdiki@helit.duth.gr

Abstract

HTR (Handwritten Text Recognition) technologies have progressed enough to offer high-accuracy results in recognising handwritten documents, even on a synchronous level. Despite the state-of-the-art algorithms and software, historical documents (especially those written in Greek) remain a real-world challenge for researchers. A large number of unedited or under-edited works of Greek Literature (ancient or Byzantine, especially the latter) exist to this day due to the complexity of producing critical editions. To critically edit a literary text, scholars need to pinpoint text variations on several manuscripts, which requires fully (or at least partially) transcribed manuscripts. For a large manuscript tradition (i.e., a large number of manuscripts transmitting the same work), such a process can be a painstaking and time-consuming project. To that end, HTR algorithms that train AI models can significantly assist, even when not resulting in entirely accurate transcriptions. Deep learning models, though, require a quantum of data to be effective. This, in turn, intensifies the same problem: big (transcribed) data require heavy loads of manual transcriptions as training sets. In the absence of such transcriptions, this study experiments with training sets of various sizes to determine the minimum amount of manual transcription needed to produce usable results. HTR models are trained through the Transkribus platform on manuscripts from multiple works of a single Byzantine author, John Chrysostom. By gradually reducing the number of manually transcribed texts and by training mixed models from multiple manuscripts, economic transcriptions of large bodies of manuscripts (in the hundreds) can be achieved. Results of these experiments show that if the right combination of manuscripts is selected, and with the transfer-learning tools provided by Transkribus, the required training sets can be reduced by up to 80%. Certain peculiarities of Greek manuscripts, which lead to easy automated cleaning of resulting transcriptions, could further improve these results. The ultimate goal of these experiments is to produce a transcription with the minimum required accuracy (and therefore the minimum manual input) for text clustering. If we can accurately assess HTR learning and outcomes, we may find that less data could be enough. This case study proposes a solution for researching/editing authors and works that were popular enough to survive in hundreds (if not thousands) of manuscripts and are, therefore, unfeasible to be evaluated by humans.

keywords

big data; byzantine manuscripts; deep learning; HTR models; Transkribus;

I INTRODUCTION

The humanitarian spirit of Antiquity and Byzantium has passed down to younger generations a multitude of manuscripts that preserve ancient and Byzantine Greek literary texts. Many of these manuscripts remain unedited or under-edited due to the complexity of producing critical editions. This process requires heavy loads of manuscript research until all disparate text variations (instances of manuscripts that contain the same opus transmitting different text) are collected and collated to the last detail. Especially in cases of rich manuscript traditions (i.e., a significant number of manuscripts transmitting the same work), this process is not only tedious, but it could also take years to complete. Therefore, some otherwise well-known authors remain archived in libraries or under poorly edited publications. Such is the case of the ~800 manuscripts of Homer, ~3,900 of the New Testament, or ~21,400 of John Chrysostom's opera (as currently listed in *Pinakes* database,[1] yet those numbers might be even higher).

HTR technology could greatly assist the monumental task of massively and accurately collating hundreds, if not thousands, of manuscripts. After all, no collation can be done without (at least) a diplomatic or (preferably) a regularised transcription of the sources. Although HTR systems have evolved significantly in the last decades, and despite the several state-of-the-art systems readily available to the scholars' community, the peculiarities of handwritten historical documents remain a real challenge. This phenomenon is especially true for documents written in Ancient Greek, in which special characters, such as accents (at least five unique characters for accents and six combinations of those) and ligatures or abbreviations of letters, perplex character recognition even more. Apart from these factors, AI neural networks can assist HTR algorithms in training highly accurate models, but a substantial amount of training data is necessary for effectiveness. In order to produce these input data, one should return to the same old process: manual transcription of a bulk of manuscripts.

Due to the scarcity of these transcriptions and the human efforts milestone of producing them ex nihilo, this paper examines the limits of HTR technology by defining the optimum amount of data needed to train an AI model successfully. In an extended version of this research (currently ongoing by the author), the HTR-produced transcriptions are tested further as data input in experiments of manuscripts' hierarchical clustering. The aim is to produce a classification system by which all instances of a text can be traced back to their ancestors through a series of branching points –much like the phylogenetics method in Biology, but with DNA sequences replaced by manuscripts [Macé and Baret, 2004; Spencer *et al.*, 2004]. Considering a big data scenario, hierarchical clustering was preferred over a stemmatical analysis for its speed and simplicity.

For all the following methodological experiments, a set of 11 manuscripts with homilies of John Chrysostom served as a case study on HTR.[2] This author was chosen for two main reasons: a) his opera are numbered at ~21,400 manuscripts, which is equal to almost half a billion words –unfeasible, thus, for humans to transcribe–, and b) almost 3,000 of these manuscripts are known for the *double recension* phenomenon, simply meaning that there are at least two prominent manuscript families, known as recensions, from which one is the revision of the other [Konstantinidou, 2021; Perdiki and Konstantinidou, 2021]. Thus, to classify these thousands of manuscripts into the relevant recension, one should first extract the raw text data from the manuscripts. For both tasks, exploitation of pertinent technology seems necessary to rapidly and massively handle the bulk of data.

HTR experiments were conducted on the Transkribus platform.[3]

[1] https://pinakes.irht.cnrs.fr/ (Accessed: 8 October 2020). Cf. [Augustin, Binggeli and Cassin, 2009].

[2] Not all 11 manuscripts were used in every one of the following recorded experiments. Some of them proved to be insufficient due to image noise (which would result in inaccurate HTR results) or data availability. Many Byzantine manuscripts are not available in digital copies or are distributed under copyright restrictions. The copyright constraints and data availability were a significant impediment to this research.

[3] https://readcoop.eu/transkribus/?sc=Transkribus.

II LITERATURE REVIEW

Text recognition systems are well-researched and continuously developing. Currently, there are two main systems for image text extraction: OCR (Optical Character Recognition), targeting printed text, and HTR for handwritten documents, where character recognition is not straightforward [Firmani *et al.*, 2018; Ströbel *et al.*, 2022]. HTR is further divided into offline (meaning recognition from a scanned document image) or online (text recognition while the text is being written) [Ingle *et al.*, 2019]. Furthermore, the ever-increasing need for transcription of historical documents, currently archived in libraries and collections worldwide, has led to the development of HTR systems, mainly focused on ancient or medieval handwriting.

The most recent bibliography suggests applications such as Tesseract [Patel *et al.*, 2012; White, 2012],[4] Kraken [Schoen and Saretto, 2022; Kiessling, 2019],[5] eScriptorium [Kiessling *et al.*, 2019], Transkribus [Kahle *et al.*, 2017; Muehlberger *et al.*, 2019], or μDoc [Tsochatzidis *et al.*, 2021]. Despite being listed in line here, it should be noted that the above methods cannot be measured reliably against each other, as they are diverse in architecture and function (i.e., Kraken and Tesseract are OCR engines, while eScriptorium and Transkribus are interface platforms for HTR). Regardless, already conducted experiments [Ströbel and Clematide, 2019; Ströbel, Clematide and Volk, 2020; Clérice, 2022b] have demonstrated that from the HTR mentioned above tools, Transkribus and e-Scriptorium (which implements Kraken) are the most successful in producing low CER (Character Error Rate) text recognitions. Part of this success is due to the different and more efficient layout analysis performed by both Transkribus and eScriptorium, an analysis that does not restrict segmentation in rectangular regions, as handwritten text can expand in many forms and directions [Stokes *et al.*, 2021]. It was decided to exploit the Transkribus system for experimenting with Greek manuscripts HTR and the reason behind this decision was twofold. First of all, while most other systems are executed via CLI (Command Line Interface), assuming coding fluency, Transkribus is offered as a GUI (Graphical User Interface)[6] and Web-based application,[7] making it accessible to most researchers. Currently, eScriptorium also offers a Web-based platform upon registration and further contact with their team [Kiessling *et al.*, 2019; Stokes *et al.*, 2021]. However, their interface was partially developed when this research began experiments. That being said, we intend to expand our training methods with the exploitation of eScriptorium. Secondly, unless one uses its official servers, eScriptorium/Kraken needs high computational power to train models [Stokes *et al.*, 2021]. Unfortunately, not many scholars have access to high-performing hardware. On the contrary, since Transkribus is connected to the Innsbruck server, all computations and training are performed there [Kahle *et al.*, 2017; Muehlberger *et al.*, 2019, pp. 959, 962]. So, each user can train models even from a low-cost laptop. It should also be noted that, to our best knowledge, currently there are no published and readily available HTR models for ancient Greek or Byzantine textual data.[8]

[4] https://web.archive.org/web/20220125061256/https:/github.com/tesseract-ocr/tesseract (Accessed: 20 October 2022). Nevertheless, there are also some voices of opposition regarding the results of Tesseract compared to similar systems, as stated in [Smith, 2007].

[5] http://kraken.re/ (Accessed: 18 September 2020).

[6] http://web.archive.org/web/20211113063459/https:/readcoop.eu/transkribus/download/ (Accessed: 19 July 2022).

[7] http://web.archive.org/web/20220119164148/https:/transkribus.eu/lite/ (Accessed: 19 July 2022).

[8] An HTR application for Byzantine manuscripts was described by [Tsochatzidis et al., 2021], but it is currently unavailable to the public.

III METHODS OF MANUSCRIPTS AUTOMATIC TRANSCRIPTION

3.1 Data Availability

As described previously, the 11 case study manuscripts were used as training data. The manuscripts are dated to the 10th-14th century and transmit John Chrysostom's *Homilies on St. Paul's Epistles to Titus*.[9] Homilies 1 and 5 were used as data sets in all conducted experiments based on the availability of digital images.

Most digital reproductions of manuscripts are under some degree of copyright protection. As a result, data gathering is not always a straightforward process. Quite the opposite, the author produced ground truth data from scratch for the following experiments, which are fully provided to the research community. Instead of the digital files of the manuscripts, links to the libraries' digital archives are given, if applicable. The detailed dataset can be found in [Perdiki, 2023]. However, the best-performing general model from our training is currently not available in Transkribus. It will be released upon my thesis publication in the coming months, subject to Transkribus approval [Perdiki, forthcoming].

3.2 Methodology

Transkribus documentation denotes that for a successful HTR model training, at least 15,000 words of diplomatic transcription input are required. However, since such transcriptions are unavailable, producing them from scratch would demand heavy economic and human resources, for a data set consisting of million words. Early experiments on Transkribus indicated that most erroneous outputs involved misrecognition of accents, punctuation or word tokens splitting (due to the *scripta continua* form of the writing style – which, however, is normalised in the transcription, while abbreviations are marked up and later expanded in the metadata section),[10] see Figure 1. Most of the time, the last character of a word token (usually pronouns or conjunctions with an average of three characters in length) was erroneously connected with the following or the previous word token. In addition, probably for the same reason,[11] when tokenisation fails, so does accent recognition. See, i.e., line 5 of Figure 1, where instead of *μὴ νεκροὺς* (transl.: not the dead) HTR recognises *μὴν ἐκροὺς* (meaningless) and so adds the smooth breathing diacritical mark above the letter *ε*. This diacritics addition is an interesting mistake; according to ancient Greek grammar, when a word begins with the letter *ε*, it commonly has the smooth breathing mark. Other times, accent recognition rightfully fails because accents are already misplaced in the manuscripts. For instance, cf. Figure 2 and line 3 of Figure 1, where the grave accent of the adjective *πολὺς* is not recognised since it is misplaced above the last character, the consonant *ς*. Such instances, should not be considered as HTR errors. It should also be mentioned that accent misrecognition might be negatively affected by the non-expansion of the line region to the upper margin (see Figure 2; the line region is depicted with the blue rectangle and baseline with the purple underline). Following Transkribus guidelines regarding the higher importance of baseline region (i.e., the imaginary horizontal line on which text characters rest) in HTR results [Muehlberger *et al.*, 2019, p. 959],[12] adjustments were made only on the relevant

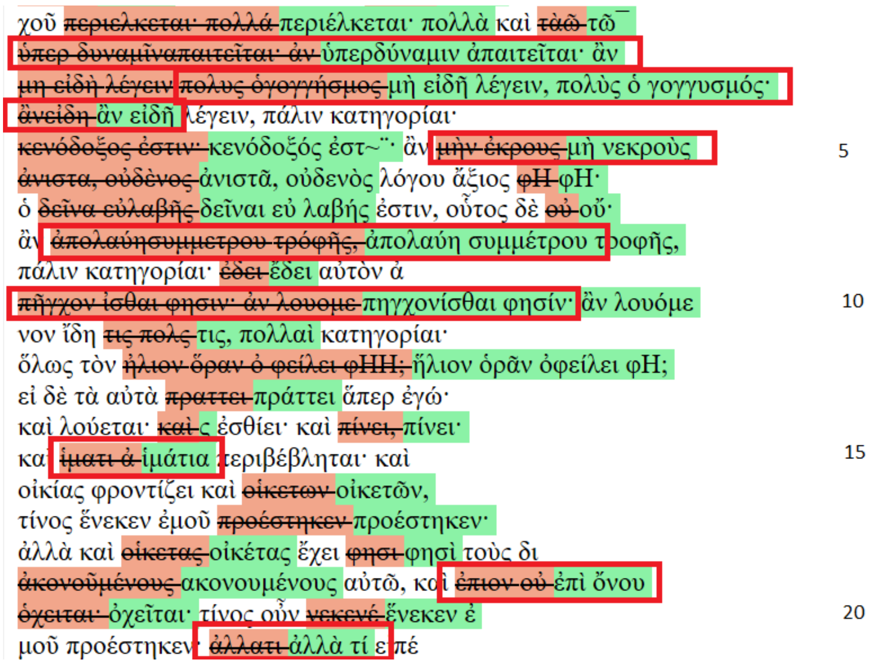
[9] Namely, the manuscripts are NLG Athens 263, BL Burney 48B, ONB Vindob. theol. gr. 14, BNF Par. gr. 745, Athos Vatopedi 328, Patriarchal Library Alexandria 34, BSB Monac. gr. 377, BSB Monac. gr. 353, Patmos St. John 183, Athos Dionysiou 70, BSB Monac. gr. 211.

[10] Segmentation was normalised in order to test the HTR limits – i.e., how successful can a model be with such a condensed writing style? On the other hand, abbreviations were not normalised, as they can be more complex both structurally and semantically. Cf. Figure 1.

[11] Although, extensive experiments should be conducted before a conclusion is made on the matter.

[12] Cf. https://readcoop.eu/glossary/line-region/ (Accessed: 13 February 2023).

baselines in all mentioned experiments. This inaction on line-region adjustment proves to be a limitation and should be corrected in future experiments. Moreover, not all manuscripts consist of a baseline. In rare cases, letters are written from a hanging line instead, a feature that can be challenging for the HTR system – cf. Figure 3 and [Perdiki and Konstantinidou, 2021].



χοῦ ~~περιελκεται· πολλά~~ περιέλκεται· πολλὰ καὶ ~~τὰῶ~~ τῶ¯
~~ὑπερ δυναμῖναπαιτεῖται· ἀν~~ ὑπερδύναμιν ἀπαιτεῖται· ἂν
~~μη εἰδὴ λέγειν~~ ~~πολυς ὁγογγήσμος~~ μὴ εἰδῆ λέγειν, πολὺς ὁ γογγυσμός·
~~ἂνειδη~~ ἂν εἰδῆ λέγειν, πάλιν κατηγορίαι·
~~κενόδοξος ἐστιν·~~ κενόδοξός ἐστ~¨· ἂν ~~μὴν ἐκρους~~ μὴ νεκροὺς
~~ἀνιστα, οὐδὲνος~~ ἀνιστᾶ, οὐδενὸς λόγου ἄξιος ~~φΗ~~ φΗ·
ὁ ~~δεῖνα εὐλαβῆς~~ δεῖναι εὐ λαβής ἐστιν, οὗτος δὲ ~~οὐ~~ οὔ·
ἂν ~~ἀπολαύησυμμετρου τρόφῆς,~~ ἀπολαύη συμμέτρου τροφῆς,
πάλιν κατηγορίαι· ~~ἐδει~~ ἔδει αὐτὸν ἀ
~~πῆγχον ἰσθαι φησιν· ἀν λουομε~~ πηγχονίσθαι φησίν· ἂν λουόμε
νον ἴδη ~~τις πολς~~ τις, πολλαὶ κατηγορίαι·
ὅλως τὸν ~~ἡλιον ὅραν ὀ φείλει φΗΗ;~~ ἥλιον ὁρᾶν ὀφείλει φΗ;
εἰ δὲ τὰ αὐτὰ ~~πραττει~~ πράττει ἅπερ ἐγώ·
καὶ λούεται· ~~καὶ~~ ς ἐσθίει· καὶ ~~πίνει,~~ πίνει·
καὶ ~~ἱματι ἀ~~ ἱμάτια περιβέβληται· καὶ
οἰκίας φροντίζει καὶ ~~οἰκετων~~ οἰκετῶν,
τίνος ἕνεκεν ἐμοῦ ~~προέστηκεν~~ προέστηκεν·
ἀλλὰ καὶ ~~οἰκετας~~ οἰκέτας ἔχει ~~φησι~~ φησὶ τοὺς δι
~~ἀκονοῦμένους~~ ακονουμένους αὐτῶ, καὶ ~~ἐπιον οὐ~~ ἐπὶ ὄνου
~~ὀχειται·~~ ὀχεῖται· τίνος οὖν ~~νεκενέ~~ ἕνεκεν ἐ
μοῦ προέστηκεν· ~~ἀλλατι~~ ἀλλὰ τί εἰπέ

Figure 1. Highlighted in red rectangles are instances of HTR errors (at a normalised segmentation) due to scripta continua. I.e., line 21 depicts how tokens *ἀλλὰ τί* (transl.: but what) are falsely recognised as the single (meaningless) token *ἀλλατι*. Manuscript: NLG Athens 263, f. 158v.

Such errors, as mentioned above, can be easily cleaned to lower CER significantly. Moreover, despite successful OCR demands of above 90% accuracy [Holley, 2009; Clérice 2022a], the complex peculiarities of HTR allow for a lower level of accuracy, especially if we consider that keyword spotting techniques return accurate results even with a 30% CER [Muehlberger *et al.*, 2019, p. 963; Tomoiaga *et al.*, 2019; Stokes *et al.*, 2021; Ströbel *et al.*, 2022]. Consequently, aiming for better HTR results, a maximum 20% CER threshold was initially decided – as a hypothesis – for a model to be deemed adequately accurate (i.e., to serve as data for text clustering).[13] That being said, it should be pointed out that this threshold must be set upon firm grounds, so the author's current benchmarking experiments with HTR results (unpublished for the moment) are adapting a different methodology to test the limits of such a metric accurately. Tests include HTR-produced text normalization by eliminating accents, brake lines (i.e., in hyphenated words), and even data contamination (i.e., combining HTR and manual transcriptions' data to reduce/increase CER percentage). Although important in evaluation [Sánchez *et al.*, 2019, p. 124; Kang *et al.*, 2022], using WER (Word Error Rate)

[13] Of course, this percentage of error rate is not an ideal aim for HTR performance. Yet, it is a cost-efficient choice in order to balance low input data and high output accuracy.

instead of CER was ruled out due to the errors described in the previous section (i.e., despite being a single erroneous character, a falsely recognised accent can lead to a WER, but such inaccuracies seem irrelevant as they can be normalised). Lastly, previous research concluded that, although AI machine learning algorithms require a quantum of data to be effective, there is certainly a limit to the data set volume or the training epochs number to avoid overfitting [Rabus, 2019; Ströbel, Clematide and Volk, 2020; Perdiki and Konstantinidou, 2021].

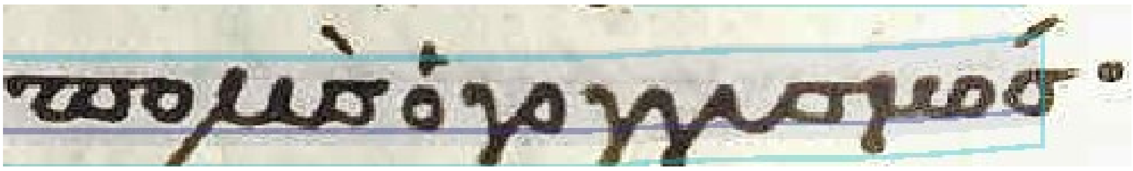


Figure 2. An instance of erroneous accent recognition. The depicted text should be transcribed as *πολὺς ὁ γογγυσμός*. For the HTR prediction cf. Figure 1, line 3. Manuscript: NLG Athens 263, f. 158v.

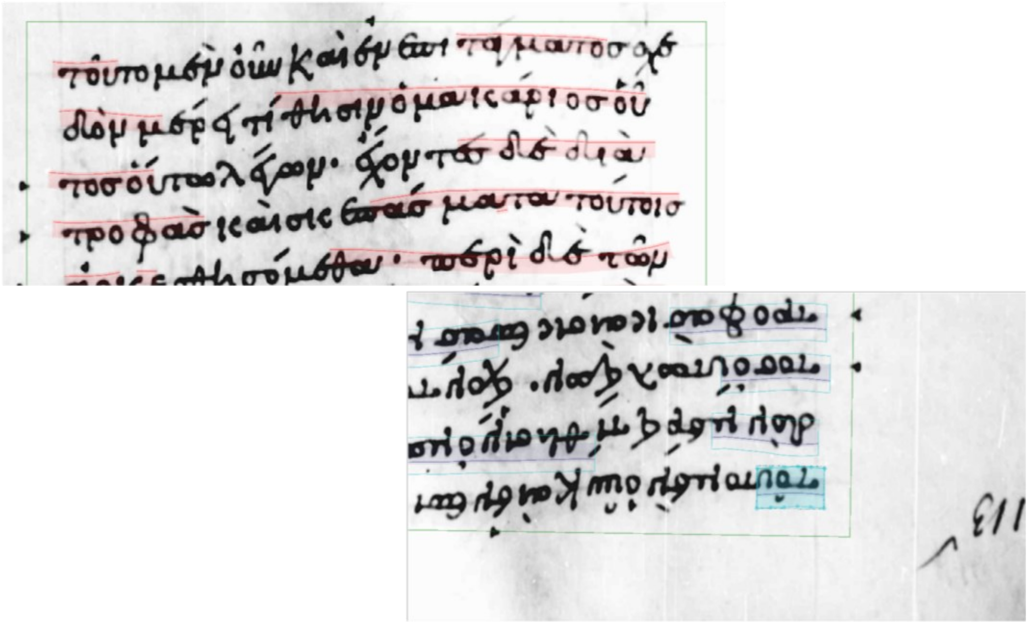


Figure 3. An instance of a (HTR-challenging) hanging line, as recognised in Transkribus automatic segmentation. The HTR correctly recognises the line's dependent traces but, expecting a baseline, it rotates the page upside-down.

IV RESULTS

Upon these three criteria, experiments were conducted under four main methods and mainly under the same system configurations (reported in Table 1). The four methods did not share the same objective. The first's methods objective was to determine the minimum amount of required training data (with a hypothesis of a 20% CER threshold). The second method aimed both to validate the HTR model with unseen data and test the possibility of revealing scripts' similarity. On the other hand, to achieve better performance with less data, the third method's main objective was the contamination of training data. In addition to that, in the fourth experiment stage, aiming at the optimum performance, comparing the two HTR training methods (HTR+ and PyLaia) provided by Transkribus, a different training set of the same manuscripts was used to train HTR models via PyLaia.

Particularly, the first method was HTR model training with gradual data set reduction to define the minimum amount of data needed to produce usable results. As depicted in Figure 4, 24 models were trained (via CITlab HTR+ method) from 8 different manuscripts (three models per manuscript),[14] with a decreasing number of words: transcription input of ~3,000, ~2,000 and ~1,000 words from John Chrysostom's *1st Homily*, with a minimum of 50 epochs of each training set. A 10% portion of the data input was reserved in each training set as validation data. Most models performed below the 20% CER threshold, even under the low 1,000-word input test. The few exceptions of poor recognition results overlapped with some low-quality manuscript digitisations. Usually, the breaking point of the model training was around 5-10 epochs.

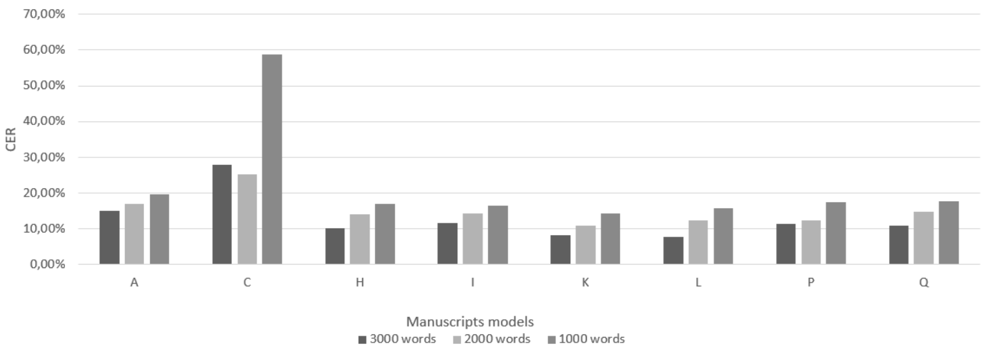
	Training Set (words)	**Validation Set**	**Epochs**[15]	**Early Stopping**[16]	**Base model**	**Training system**
Single models I	3,000	10%	50	N/A	No	HTR+
	2,000	10%	50	N/A	No	HTR+
	1,000	10%	50	N/A	No	HTR+
Mixed models	2,000	10%	50	N/A	No	HTR+
General model I	9,000	10%	50	N/A	Yes	HTR+
Single models II	3,000	10%	50	20	No	PyLaia
	3,000	10%	250	20	No	PyLaia
General model II	25,621	10%	50	20	Yes	PyLaia
	25,621	10%	250	20	Yes	PyLaia

Table 1. System configurations for all experiments

[14] Two manuscripts of the dataset, E and W, were excluded from this testing due to their insufficient number of words.

[15] In most cases, the number of epochs was set up to 50. This fixed choice was an attempt to standardise configurations so as to clarify which factors affect the HTR performance, i.e., the number of training data, the number of epochs, the quality of the image data or otherwise.

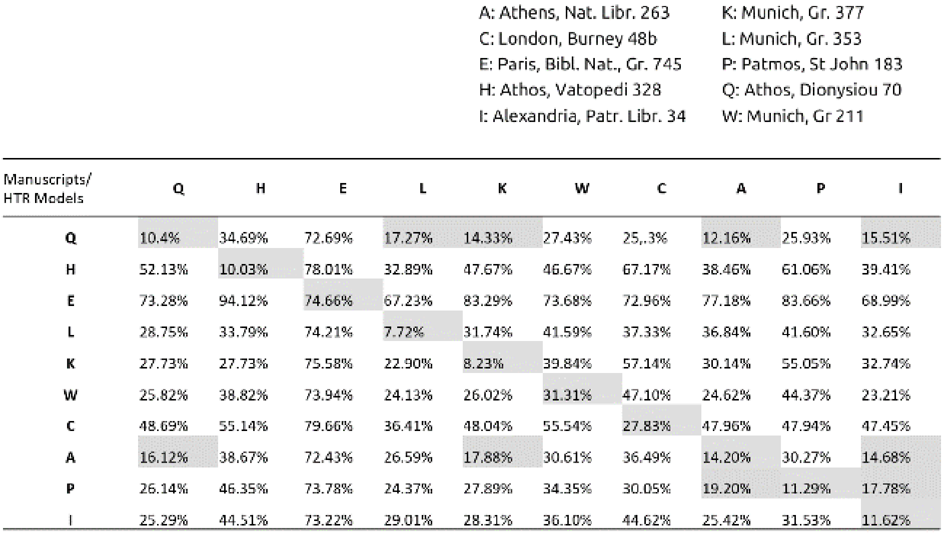
[16] The early stopping technique is not applicable in the HTR+ method (the system of choice for the majority of the experiments). PyLaia, on the contrary, does provide that option, so early stopping was set to the predefined number of 20. Since Transkribus had announced that HTR+ would not be supported any further than November 2022, the exploitation of both systems was decided as a form of systems performance comparison. See, https://readcoop.eu/glossary/htr-plus/ (Accessed: 24 February 2023).



A: Athens, Nat. Libr. 263	K: Munich, Gr. 377
C: London, Burney 48b	L: Munich, Gr. 353
H: Athos, Vatopedi 328	P: Patmos, St John 183
I: Alexandria, Patr. Libr. 34	Q: Athos, Dionysiou 70

Figure 4. CER results of decreasing training data (number of words of manually generated training data), as in [Perdiki and Konstantinidou, 2021].

With the aim of testing script similarity out of text recognition and limiting the manual production of data input even further, a cyclical application of each trained model to 10 manuscripts was performed (some of these manuscripts were palaeographically similar in style). The experiment hypothesis of this method was whether an already trained model could accurately recognise the text of a different but similar writing style manuscript. This process would also serve as a manuscript clustering method if proven successful. However, as seen in Figure 5, the resulting 90 text recognitions were mainly inaccurate. Only 9 out of 90 combinations recovered text with a lower than 20% CER, despite exploiting the 3,000-word input training sets. Prior to the results, it was uncertain whether the given manuscripts' combinations would reflect the models' optimum performance. As a result, clustering algorithms that would predict script similarity seem necessary (cf. [Stutzmann, 2016]).

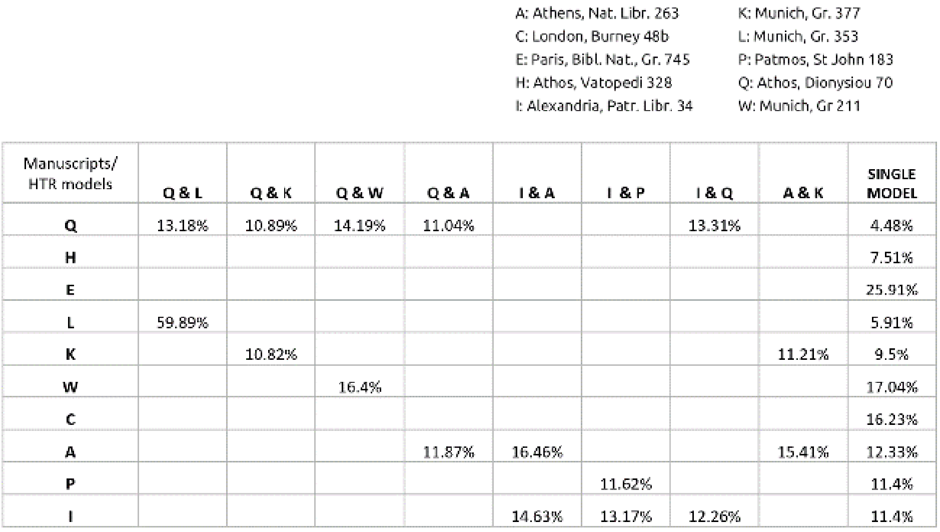


A: Athens, Nat. Libr. 263	K: Munich, Gr. 377
C: London, Burney 48b	L: Munich, Gr. 353
E: Paris, Bibl. Nat., Gr. 745	P: Patmos, St John 183
H: Athos, Vatopedi 328	Q: Athos, Dionysiou 70
I: Alexandria, Patr. Libr. 34	W: Munich, Gr 211

Manuscripts/ HTR Models	Q	H	E	L	K	W	C	A	P	I
Q	10.4%	34.69%	72.69%	17.27%	14.33%	27.43%	25,.3%	12.16%	25.93%	15.51%
H	52.13%	10.03%	78.01%	32.89%	47.67%	46.67%	67.17%	38.46%	61.06%	39.41%
E	73.28%	94.12%	74.66%	67.23%	83.29%	73.68%	72.96%	77.18%	83.66%	68.99%
L	28.75%	33.79%	74.21%	7.72%	31.74%	41.59%	37.33%	36.84%	41.60%	32.65%
K	27.73%	27.73%	75.58%	22.90%	8.23%	39.84%	57.14%	30.14%	55.05%	32.74%
W	25.82%	38.82%	73.94%	24.13%	26.02%	31.31%	47.10%	24.62%	44.37%	23.21%
C	48.69%	55.14%	79.66%	36.41%	48.04%	55.54%	27.83%	47.96%	47.94%	47.45%
A	16.12%	38.67%	72.43%	26.59%	17.88%	30.61%	36.49%	14.20%	30.27%	14.68%
P	26.14%	46.35%	73.78%	24.37%	27.89%	34.35%	30.05%	19.20%	11.29%	17.78%
I	25.29%	44.51%	73.22%	29.01%	28.31%	36.10%	44.62%	25.42%	31.53%	11.62%

Figure 5. Cyclic application of each model to all manuscripts, as in [Perdiki and Konstantinidou, 2021]. The combination of identical letters (e.g., Q & Q) reflects model performance on the training data manuscript.

The third method of experimenting with HTR extended the second method's hypothesis. If one mixes training data from more than one manuscript, as a corpus expansion process, the data set will be enlarged without demanding extra manual input. So, a hypothesis was made to test whether this would result in mixed models of high accuracy. Furthermore, combining all transcriptions in a single training set would test the possibility of building an optimum model capable of accurately transcribing any of our Greek manuscripts. The nine best matches, produced under the second method experiments, served as the data for the manuscripts' combinations. The data set was formed out of randomly selected pages from each manuscript. These combined data consisted of ~2,000-word input transcription per combination. Each model was trained with 50 epochs via the CITlab HTR+ method. The validation set was formed out of a random 10% of the training data. Afterwards, the trained mixed models were applied to each of the training set's manuscripts, as in Figure 6 (i.e., the Q&L model was trained from Q and L manuscripts' combined data and then applied to each for text recognition). These models performed at 80% accuracy, within threshold limits, with a breaking point around 5-10 epochs.[17] Lastly, a 9,000-word input from all 10 joined manuscripts, trained on 50 epochs and CITlab HTR+ method, validated from a random 10% of the data, and applied to every single manuscript, resulted in top-end CER performance (down to 4,48%, see Figure 6).



A: Athens, Nat. Libr. 263	K: Munich, Gr. 377
C: London, Burney 48b	L: Munich, Gr. 353
E: Paris, Bibl. Nat., Gr. 745	P: Patmos, St John 183
H: Athos, Vatopedi 328	Q: Athos, Dionysiou 70
I: Alexandria, Patr. Libr. 34	W: Munich, Gr 211

Manuscripts/ HTR models	Q & L	Q & K	Q & W	Q & A	I & A	I & P	I & Q	A & K	SINGLE MODEL
Q	13.18%	10.89%	14.19%	11.04%			13.31%		4.48%
H									7.51%
E									25.91%
L	59.89%								5.91%
K		10.82%						11.21%	9.5%
W			16.4%						17.04%
C									16.23%
A				11.87%	16.46%			15.41%	12.33%
P						11.62%			11.4%
I					14.63%	13.17%	12.26%		11.4%

Figure 6. CER of models with mixed training sets, as in [Perdiki and Konstantinidou, 2021].

[17] The poor performance of the Q & L model, when applied to the L manuscript, has yet to be fully explained. Apart from the writing style, the only difference between the two manuscripts was that Q's data set was coloured digitisation, whereas L's was grey-scaled microfilm digitisation. However, that was also the case with the K manuscript, yet the relevant CERs returned under 20%. On the other hand, E's bad score is probably due to the insufficient amount of data and the poor image quality; this particular manuscript had a missing folio which drastically affected the HTR results (because of the missing word tokens). Its inclusion in the data set was deliberate to test the degree of effect such manuscript damages might have on our methodology.

The fourth and last methodology on HTR experiments via the Transkribus platform came as validation. With the same data set of the 11 manuscripts mentioned above and under the same methodology, the most successful of the above experiments (the first and third method) was performed on a different training set (John Chrysostom's *5th Homily* transcription) in two testing phases. Firstly, the training set consisted of ~3,000-word input and the HTR method was altered to PyLaia with 250 epochs. The validation set was a random 10% portion of the training data. The resulting CER was lower than 10% (breaking point on 20-30 epochs), yet higher compared to previous experiments with the HTR+ method and 50 epochs of training. In addition to the CITlab HTR+ method's better performance, it appears that, with insufficient data, a higher epoch number returns the worst results, as already shown by Rabus [Rabus, 2019]. Secondly, another attempt was made to build a general model, as all manuscripts are characterised by a certain (perceived) script uniformity and clarity (none of the manuscripts was heavy in ligatures, abbreviations, or damaged areas), however unique in writing style. By joining up the *5th Homily's* transcriptions (from 9 out of the 11 manuscripts), a 25,621-word input trained the general model. This training results in a 4.60% CER. In an attempt to further expand the data of the experiments, fine-tune this general model, and improve, thus, the recognition results, the model with the best performance[18] –during phase 1 of the fourth method experiments– was added to the training process as a base model. The final CER on that last experiment was 3.90% on the validation set (at 3-10 epochs range breaking point), the minimum CER of all conducted experiments (see Table 2).

Training Data	Model Name	50 epochs CER (HTR+)	250 epochs CER (PyLaia)
Q: Athos, Dionysiou 70	Q-3000[19]	11.62%	14.41%
H: Athos, Vatopedi 328	H-3000	10.03%	13.70%
A: Athens, Nat. Libr. 263	A-3000	10.03%	12.20%
I: Alexandria, Patr. Libr. 34	I-3000	13.00%	14.60%
D: Venice, ONB theol. gr.14	D-3000	(n/a)[20]	14.20%
E: Paris, Bibl. Nat., Gr. 745	E-3000	(n/a)	14.90%
K: Munich, Gr. 377	K-3000	8.93%	12.30%
L: Munich, Gr. 353	L-3000	8.12%	13.00%
General Model (without a base model)	GM	17.18%	4.60%
General Model (with a base model)	GMbm	(n/a)[21]	3.90%

Table 2. CER of 50 and 250 epochs models.

Expanding upon the last two experiments (the last part of method 3 and method 4), it is important to emphasise the significance of general models. The aforementioned experiments have shown that even with a minimal data set one could easily provide the initial impulse for the training of a larger and more accurate HTR model. The 20% CER threshold needs yet to be proven as a sufficient amount of HTR-produced errors. If not, however, general models (despite their limits) would be an effective technique for fast and clean text extraction, regardless of the writing style.

[18] In terms of CER percentage – i.e., the best model is considered the one with the lowest CER in the same textual data, but in a different manuscript (which means a different writing style).

[19] The "3000" tag indicates the amount of word input on the training data set.

[20] Manuscripts D and E were excluded from the 50 epochs training due to insufficient training data.

[21] No base model was used during the general model's training via the HTR+ method.

4.1 Limitations

The end-all of the conducted experiments was to determine the minimum amount of data needed to produce an accurate transcription and to deliver data usable in text clustering. The hierarchical clustering ought not to be a detailed collation as stemmatological analysis is out of the scope of this research. Currently, experiments conducted by the author with data mining techniques are examining promising preliminary results for the task. However, data quality evaluation[22] for text clustering needs to be further fine-tuned, by eliminating text noise. More often than not, HTR erroneous results include forced-hyphenated words in brake lines (without a hyphen symbol though), false accents (which do not result in semantic shift), or tokens concatenation due to the scripta continua. Such instances need to be automatically normalised,[23] before proceeding to clustering algorithms.

As pointed out previously in this paper, the metrics used in our experiments should be further adjusted. While the 20% CER threshold remains to be proved for ensuring successful text clustering of automated transcriptions, benchmarking experiments have the potential to refine the required accuracy. Furthermore, the uneven distribution of errors on each page probably affects the erroneous HTR results, which should also be accounted for in future testing. More than that, the low number of input data equals insufficient data on the validation set. These experiments produce seemingly unexpected results. In other words, since our case study was a limit test, traditional metrics cannot always be applied. We seek to find more suitable evaluation techniques as we continue our experiments.

The aforementioned dual approach to early stopping (none for HTR+ and 20 for the PyLaia method) might prove to be a limitation on our results. As already explained, both systems were used to compare performance fairly. Nevertheless, since early stopping is not available for the HTR+ training, perhaps the reasonable choice would be to nullify PyLaia early stopping. The system's configurations need further improvement on the matter.

V CONCLUSION

Computational processes can highly assist philological research when dealing with a bulk of data. Time-consuming and painstaking tasks, often leading to errors due to their complexity, produce fruitful results when conducted via special algorithms. This paper presented methodologies for exploiting a specific HTR tool to enhance manuscript tradition research.

The Transkribus platform proved highly efficient in training HTR models and recognising text from digitised manuscripts. Even with minimal training data input, the accuracy of the produced models was high (cf. first experiment). With further testing and fine-tuning, developing general models that could transcribe a good portion of Greek manuscripts is more than possible. The two main conclusions of the previous experiments were: a) HTR models perform worse on unseen data, so including training data from more than one writing style is crucial (cf. experiments two and three), and b) one can produce successful HTR models with low CER percentage simply by combining small datasets of many handwriting instances (cf.

[22] A method proposed by Thibault Clérice [Clérice, 2022a] suggests using language models on sentences instead of n-grams to assess the quality of Character Error Rate (CER) in line-level classification systems. While this method has been employed for Kraken-trained models (an OCR system), it shows promise for efficiently processing large HTR datasets, particularly with previously unseen textual material.

[23] Regarding the automatic correction of HTR-produced errors, see the work of [Pavlopoulos *et al.*, 2023].

experiments three and four). Mass transcription from historical documents can fuel the research with much-needed data. Ongoing author's experiments are testing whether it is possible to perform algorithmic hierarchical clustering, through data mining techniques, even with some inaccuracy in HTR results.

Indeed, machine learning benefits from data plethora, but sometimes data can be expanded – rather than produced from zero – and so training can produce functioning results. According to the research questions, humans can evaluate the process and fine-tune algorithms to high performance. By outsourcing tedious and prone to errors tasks to computing power and accuracy, researchers can concentrate on more analytical quests and lead the way forward.

Datasets and Models

Perdiki, E. (2023). List of manuscripts containing John Chrysostom's Homilies and the relevant manual transcriptions (1.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8102662

Perdiki, E. (forthcoming). HTR model 'Chrysostomicus I' (ID 44872). Transkribus.

References

Augustin, P., Binggeli, A. and Cassin, M. La base de données Pinakes: textes et manuscrits grecs. *Scriptorium*, 2009;63(1):148-149.

Clérice, T. Ground-Truth Free Evaluation of HTR on Old French and Latin Medieval Literary Manuscripts. *Proceedings of the Computational Humanities Research Conference 2022*, edited by Folgert Karsdorp, Alie Lassche, and Kristoffer Nielbo, 2022a;3290:1–24. CEUR Workshop Proceedings. Antwerp, Belgium: CEUR. https://ceur-ws.org/Vol-3290/#long_paper2081.

Clérice, T. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. arXiv, 2022b. Available at: https://doi.org/10.48550/arXiv.2207.11230.

Firmani, D., Maiorino, M., Merialdo, P., & Nieddu, E., Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio -- Episode 1: Machine Transcription of the Manuscripts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018:263–272. Available at: https://doi.org/10.1145/3219819.3219879.

Holley, R., How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 2009;15(3/4).

Ingle, R. R., Fujii, Y., Deselaers, T., Baccash, J., & Popat, A. C., A Scalable Handwritten Text Recognition System. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019:17–24. Available at: https://doi.org/10.1109/ICDAR.2019.00013.

Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G., Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto: IEEE, 2017:19–24. Available at: https://doi.org/10.1109/ICDAR.2017.307.

Kang, L., Riba, P., Rusiñol, M., Fornés, A., & Villegas, M., Pay attention to what you read: Non-recurrent handwritten text-Line recognition. *Pattern Recognition*, 2022;129:108766. Available at: https://doi.org/10.1016/j.patcog.2022.108766.

Kiessling, B., Kraken – an Universal Text Recognizer for the Humanities. In *ADHO, Éd., Actes de Digital Humanities Conference*, 2019.

Kiessling, B., Tissot, R., Stokes, P., & Ezra, D. S. B., eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019:19–19. Available at: https://doi.org/10.1109/ICDARW.2019.10032.

Konstantinidou, M. The Double Tradition of John Chrysostom's Exegetical Works: Revisions Revisited. M. Vinzent, G. Bady, and C. Broc-Schmezer (eds) *Studia Patristica. Vol. CXIV - Papers presented at the Eighteenth International Conference on Patristic Studies held in Oxford 2019*. Peeters Publishers (Volume 11: John Chrysostom through Manuscripts, Editions and History), 2021:5–26. Available at: https://doi.org/10.2307/j.ctv27vt5gb.4.

Macé, C. and Baret, P.V. Why Phylogenetic Methods Work : The Theory of Evolution and Textual Criticism. *Linguistica computazionale* [Preprint], 2004:(24/25). Available at: https://doi.org/10.1400/54380.

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., ... & Zagoris, K., Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 2019;75(5):954–976. Available at: https://doi.org/10.1108/JD-07-2018-0114.

Pang, B., Nijkamp, E. and Wu, Y.N. Deep Learning With TensorFlow: A Review. *Journal of Educational and Behavioral Statistics*, 2020;45(2):227–248. Available at: https://doi.org/10.3102/1076998619872761.

Patel, C., Patel, A. and Patel, D. Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications*, 2012;55(10):50–56. Available at: https://doi.org/10.5120/8794-2784.

Pavlopoulos, J., Kougia, V., Platanou, P., Shabalin, S., Liagkou, K., Papadatos E., Essler H., Camps J.B., Fischer, F., Error Correcting HTR'ed Byzantine Text, 15 May 2023, PREPRINT (Version 1) available at Research Square: https://doi.org/10.21203/rs.3.rs-2921088/v1

Perdiki, E. and Konstantinidou, M. Handling Big Manuscript Data. *Classics@*. Edited by C. Clivaz and V.A. Garrick, 2021;18(Ancient Manuscripts and Virtual Research Environments, special issue). Available at: https://classics-at.chs.harvard.edu/classics18-perdiki-and-konstantinidou/ (Accessed: 21 January 2022).

Rabus, A. Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus. *Scripta & e-Scripta*, 2019;19:9–32.

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., & Vidal, E. A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition*, 2019;94:122–134. Available at: https://doi.org/10.1016/j.patcog.2019.05.025.

Schoen, J. and Saretto, G.E. Optical Character Recognition (OCR) and Medieval Manuscripts: Reconsidering Transcriptions in the Digital Age. *Digital Philology: A Journal of Medieval Cultures*, 2022;11(1):174–206. Available at: https://doi.org/10.1353/dph.2022.0010.

Smith, R. An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007:629–633. Available at: https://doi.org/10.1109/ICDAR.2007.4376991.

Spencer, M., Davidson, E. A., Barbrook, A. C., & Howe, C. J., Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, 2004;227(4):503–511. Available at: https://doi.org/10.1016/j.jtbi.2003.11.022.

Stokes, P. A., Kiessling, B., Ezra, D. S. B., & Tissot, R., The eScriptorium VRE for Manuscript Cultures. *Classics@*. Edited by C. Clivaz and V.A. Garrick, 2021;18(Ancient Manuscripts and Virtual Research Environments, special issue). Available at: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (Accessed: 16 February 2023).

Ströbel, P. and Clematide, S. Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images. *Digital Humanities 2019* [Preprint], 2019. Available at: https://doi.org/10.5167/UZH-177164.

Ströbel, P. B., Clematide, S., Volk, M., Schwitter, R., Hodel, T., & Schoch, D., Evaluation of HTR models without Ground Truth Material. *Proceedings of the Thirteenth Language Resources and Evaluation Conference. LREC 2022*, Marseille, France: European Language Resources Association, 2022:4395–4404. Available at: https://aclanthology.org/2022.lrec-1.467 (Accessed: 17 January 2023).

Ströbel, P.B., Clematide, S. and Volk, M. 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. LREC 2020*, Marseille, France: European Language Resources Association, 2020:3551–3559. Available at: https://www.aclweb.org/anthology/2020.lrec-1.436 (Accessed: 14 October 2020).

Stutzmann, D., Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol. *Digital Medievalist*, 2016;10. DOI: http://doi.org/10.16995/dm.61

Tomoiaga, C., Feng, P., Salzmann, M., & Jayet, P., Field typing for improved recognition on heterogeneous handwritten forms. arXiv, 2019. Available at: http://arxiv.org/abs/1909.10120 (Accessed: 27 October 2022).

Tsochatzidis, L., Symeonidis, S., Papazoglou, A., & Pratikakis, I., HTR for Greek Historical Handwritten Documents. *Journal of Imaging*, 2021;7(12):260. Available at: https://doi.org/10.3390/jimaging7120260.

White, N. Training Tesseract for Ancient Greek OCR. *Εὔτυπον*, 2012;28–29:1–11.