How to (Auto) Collate Big Manuscript Data with Minimal HTR Training

Elpida Perdiki

How to (Auto) Collate Big Manuscript Data with Minimal HTR Training

Elpida Perdiki (eperdiki@helit.duth.gr)

Department of Greek Philology, Democritus University of Thrace, Greece

***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 Despite the state-of-the-art algorithms and software, historical documents (especially those written in Greek) remain a real-world challenge for researchers.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 (transkribus.eu) 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. This study also tests the usability of these transcriptions, which are automatically produced by the HTR models, through several text collation tools. The aim is to distinguish each manuscript's position in the textual tradition of Chrysostom's works, i.e., the grouping of manuscripts according to the text variations they transmit. For large manuscript traditions, manually processed table alignment of text variations is unattainable. Automated collation is achieved through various methods, e.g., a) heat map histograms of text variants or b) multiple sequence analysis in a tree visualisation, forming a series of branching points that connect ancestors (manuscripts serving as prototypes for copies). Less could be more if we can correctly evaluate HTR learning and results. 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 is, therefore, unfeasible to be evaluated by humans.

***Keywords*—**Big data, Byzantine manuscripts, deep learning, HTR models, text collation.

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 ~880 manuscripts of Homer, ~3,991 of the New Testament, or ~21,482 of John Chrysostom's opera (as currently listed in [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 a diplomatic 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, HTR algorithms –assisted by AI neural networks and can therefore train highly accurate models– require a quantum of training data to be effective. 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 millstone 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 the second phase of this research, the HTR-produced transcriptions are tested further as data input in experiments of manuscripts auto-collation. 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 [2], [3].

For all the following methodological experiments, a set of 11 manuscripts with homilies of John Chrysostom served as a case study on HTR and manuscript collation testing. This author was chosen for two main reasons: a) his opera are numbered at ~21,482 manuscripts, which is equal to almost half a million words, and, thus, unfeasible 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 main manuscript families, known as recensions, from which one is the revision of the other [4], [5]. 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. Manuscripts' collation was tested in three applications: a) Juxta, b) CollateX, and c) Orange data mining.

Methods of Manuscripts Auto-Transcription

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) and HTR. HTR is furtherly divided into offline (meaning recognition from a scanned document image) or online (text recognition while the text is being written) [6]. 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 [7], [8], TensorFlow [9], Kraken [10], eScriptorium [11], or Transkribus [12]–[14]. Already conducted experiments [15]–[17] have demonstrated that from the HTR mentioned above tools, Transkribus, Kraken and e-Scriptorium (which implements Kraken) are the most successful in producing low CER (Character Error Rate) text recognitions. However, since only some Humanities scholars are tech savvy, it was decided to exploit the Transkribus system for experimenting with Greek manuscripts HTR. The reason behind this decision is that while all other systems are executed via CLI (Command Line Interface) assuming coding fluency, Transkribus is offered as a GUI (Graphical User Interface) and Web-based application,[1] making it accessible to most researchers [18], [19].

Methodology on HTR

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*. Homilies 1 and 5 were used as data sets in all experiments conducted based on the availability of digital images.

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. Early experiments on Transkribus indicated that most erroneous outputs involved misrecognition of accents, punctuation or word tokens splitting (due to *scripta continua* form of the writing style), see Fig. 1. Such errors can be easily cleaned to significantly lower CER. Moreover, despite successful OCR demands of 90% accuracy, the complex peculiarities of HTR allow for a lower, close to 80%, level of accuracy [20], [21]. As a result, a minimum 20% threshold was decided for a model to be deemed adequately accurate. 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 in order to avoid overfitting [5], [16], [22].

[1] Currently, eScriptorium offers a Web-based platform upon registration and further contact with the eScriptorium team.

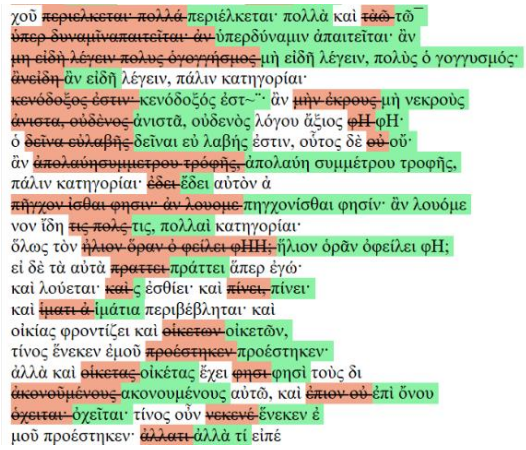

Figure 1 HTR errors due to scripta continua

Upon these three criteria, experiments were conducted under four main methods. 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 Fig. 2, 24 models were trained (via CITlab HTR+ method) from 8 different manuscripts (three models per manuscript), 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 set aside 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. The breaking point of the model training was usually around 5-10 epochs.

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. 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 Fig. 3, 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. Nevertheless, since these nine successful applications were not apparent in advance as similar writing styles, clustering algorithms that would predict script similarity seem necessary.

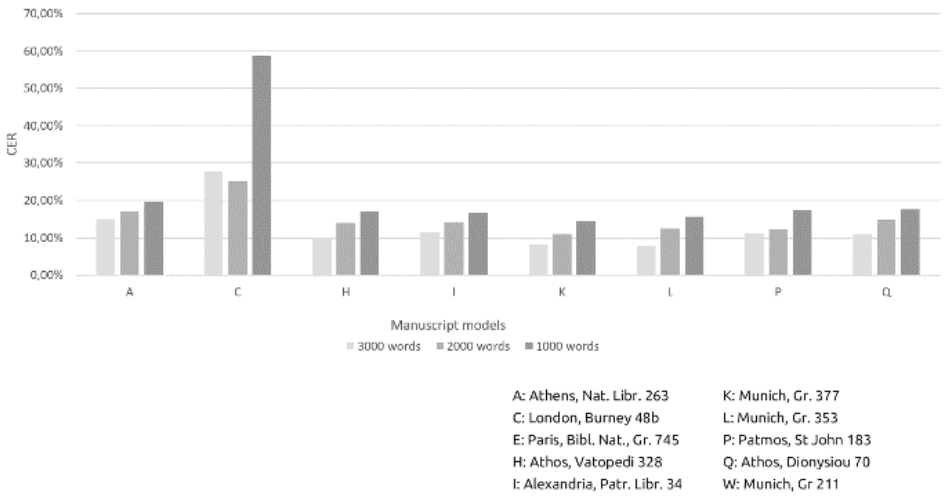

Figure 2 CER results on a decreasing number of training data (number of words of manually generated training data), as in [5]

A: Athens, Nat. Libr. 263
C: London, Burney 48b
E: Paris, Bibl. Nat., Gr. 745
H: Athos, Vatopedi 328
I: Alexandria, Patr. Libr. 34
K: Munich, Gr. 377
L: Munich, Gr. 353
P: Patmos, St John 183
Q: Athos, Dionysiou 70
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 3 Cyclic application of each model to all manuscripts, as in [5]

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 data augmentation 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 Fig. 4 (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.[2] 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 Fig. 4).

A: Athens, Nat. Libr. 263
C: London, Burney 48b
E: Paris, Bibl. Nat., Gr. 745
H: Athos, Vatopedi 328
I: Alexandria, Patr. Libr. 34
K: Munich, Gr. 377
L: Munich, Gr. 353
P: Patmos, St John 183
Q: Athos, Dionysiou 70
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 4 CER of models with mixed training sets, as in [5]

The fourth and last methodology on HTR experiments via the Transkribus platform came in the form of 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, that is) was performed on a different training set (the 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 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 [22]. 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. In an attempt to further augment the data of the experiments and improve, thus, the recognition results, the model with the best performance –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 0.60% on the training set, which translates to 3.90% on

[2] 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 digitisation of microfilm. However, that was also the case with the K manuscript, yet the relevant CERs returned under 20%.

the validation set (at 3-10 epochs range breaking point), the minimum CER of all conducted experiments (see Table 1).

Table 1 CER of 50 and 250 Epochs Models

Training Data	Model Name	50 epochs CER	250 epochs CER
Q: Athos, Dionysiou 70	Q-3000[3]	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	(noisy data)	14.20%
E: Paris, Bibl. Nat., Gr. 745	E-3000	(noisy data)	14.90%
K: Munich, Gr. 377	K-3000	8.93%	12.30%
L: Munich, Gr. 353	L-3000	8.12%	13.00%
General Model	GM	17.18%	3.90%

Methods of Manuscripts Auto-Collation

Literature Review

One of the most demanding stages of philological research is determining a manuscript tradition for producing a critical edition. Several manuscripts can transmit the same text, despite variations ranging from a single word to a complete paragraph alteration. Such text variations include not only text insertions but also deletions or substitutions. Some have happened accidentally, whereas others are intentional interventions (i.e., to improve the copied text) [23]. To critically edit a literary text, scholars need to pinpoint those variations on every manuscript and then try to define the manuscripts' origins (as in which source was copied from which). Following that path, scholars can argue which manuscript is the ancestor. That last step is usually depicted in the form of a hierarchical tree, known as *stemma codicum*.

The above-described process already brings to mind the phylogenetics method and the alignment algorithms, which are highly used in Biology for attributing DNA sequences. The resemblance between the two processes became apparent early to researchers so that from 1968, scholars exploited computing algorithms to automate and enhance manuscripts collation [2], [3], [24], [25]. Most of the existing auto-collation solutions emphasise visualising the differences among manuscripts, thus assisting with the much-needed pattern matching. In the last decades, several collation applications have been developed (one can find a somewhat detailed overview of most of them in [24]). The most applicable and widely used seem to be Juxta [26],[4] CollateX [27] and TRAViz [28].

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

[4] Until May 2022, Juxta was accessible to everyone from the software's website. Since then, though, Juxta has been available only via a login process after contact with the developers.

Apart from purpose-built collation applications, research has also focused on data mining techniques and algorithms for text classification. Thus, texts that are considered to be relevant to each other can be classified into correlating groups that would, by extension, lead to the manuscripts' ancestors and, finally, the *stemma codicum* formation. Comparative research has been conducted by Wahbeh et al. [29], where data mining tools are tested. According to this study, the WEKA toolkit [30] and Orange [31] seem the most efficient.

Methodology of Manuscript Collation and Classification

The data set for that cycle of experiments was based on the 11 manuscripts of John Chrysostom's *Homilies* (*Homily 1 on St. Paul's Epistles to Titus*), as outlined above (Section II *B*). Only 6 out of 11 were used for the experiments, while the other 5 were excluded due to the transcription quality and the amount of transcribed text. Data consisted of manually produced and not entirely accurate transcriptions of the digitised manuscripts, with a minimum of 98% CER.[5] Due to some noise in the text data (spelling, accents, case sensitivity, or punctuation mistakes), a simple cleaning process was first applied to all data. Punctuation was eliminated, and all uppercase letters were converted to lowercase via Python functions. No conversion was applied to the accented characters because, in ancient Greek, sometimes similarly written words are distinctive in meaning only by the accent character, i.e., *εἶναι* and *εἶναι*, however alike, are two different infinitive forms with different meanings: the first one means *to exist* while the second *to set in motion*. The only optical difference is the special *breathing* character, *smooth* and *rough,* respectively, that distinguishes one from the other.

After cleaning, the transcriptions were uploaded to Juxta, CollateX, and Orange data mining. Konstantinidou [4] and Goodall [32] have already identified in their research the manuscripts' recensions to which each was classified. As a result, the following experiments served equally as a methodological technique and a validation of the human-produced collation. The results of Konstantinidou's [4] research regarding the recensions are given in Table 2.

Table 2 Classification of Manuscripts in Tradition. The data in this table are drawn from Konstantinidou's stemma codicum [4]

Manuscript Name	Manuscript Family
Q: Athos, Dionysiou 70	(intermediary)
H: Athos, Vatopedi 328	α
A: Athens, Nat. Libr. 263	α
I: Alexandria, Patr. Libr. 34	α
D: Venice, ONB theol. gr.14	γ
E: Paris, Bibl. Nat., Gr. 745	β
P: Patmos, St John 183	γ
K: Munich, Gr. 377	α
L: Munich, Gr. 353	α

[5] This choice was made as a form of early testing of auto-collation methods in order to minimise potential errors. In the later stages of this research, the same methodology will be tested again with HTR-produced data.

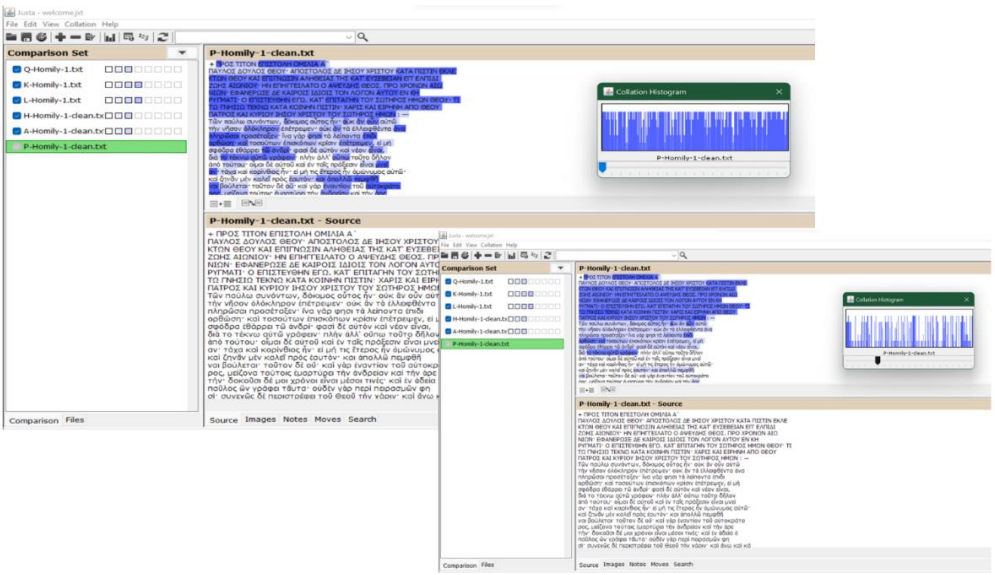

Figure 5 Juxta visualisation of collation process

CollateX is executed via Python scripts but accepts txt files as input. The same files were processed via a python script in which the collatex library was imported. Collation was performed with the Dekker algorithm [33], and output was exported in the forms of a) alignment table, b) html2 format (with coloured table visualisation), and c) SVG graph. Results on revision classification are not apparent due to the amount of data; visualisation always emphasises differentiation to the token point, which proves inefficient when researching a rich manuscript tradition. Nevertheless, as seen in Fig. 6, manuscripts P and Q are usually more related, whereas A, K, H, and L belong to another branch. However, since there is no general classification result, further research or data editing is needed for more accurate manuscript taxonomies.

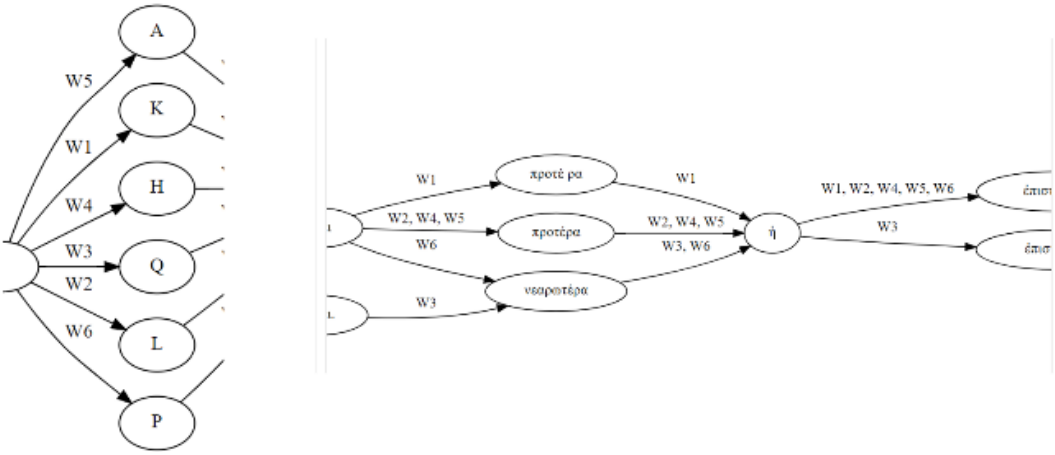

Figure 6 CollateX graph visualisation of the manuscripts' collation

On the other hand, experiments with data mining produced more accurate results. The Orange toolkit can be executed either as a Python library or as a GUI application. For the purposes of this research, the GUI platform was preferred. The workflow of the experiments is graphically depicted in Fig. 7. Firstly, the same documents were imported as txt files. Due to data noise, a text pre-process pipeline was performed. Text transformation included (as in previous steps) lowercase conversion and tokenisation via Regexp, which by default keeps only words. In addition, a Chrysostom lexicon was added to the pipeline to enhance tokenisation by filtering out non-present in the lexicon words. Lexicon was built from the vocabulary of all of John Chrysostom's *Homilies* via Python. Afterwards, a Bag of Words model was applied for feature extraction with count frequency, and then the cosine distance algorithm calculated the documents' similarity. Finally, a hierarchical clustering algorithm produced a dendrogram that visualised the manuscripts' classification. Orange's workflow resulted in the same taxonomies as Konstantinidou [4] (also partially shown in the CollateX experiment). A, K, L, and H manuscripts originate from the same revisionist family. P is from another and older manuscript family. Furthermore, Orange's algorithm seems to validate Konstantinidou's [4] hypothesis that Q is somehow connected to P as an intermediary source to the γ family (see Fig. 8).

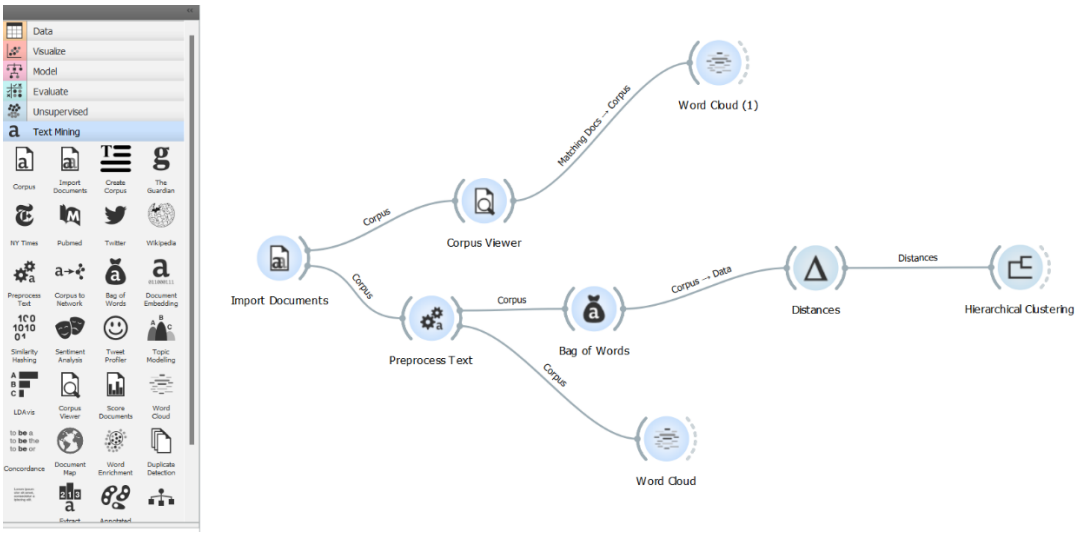

Figure 7 Experiment's workflow on Orange canvas

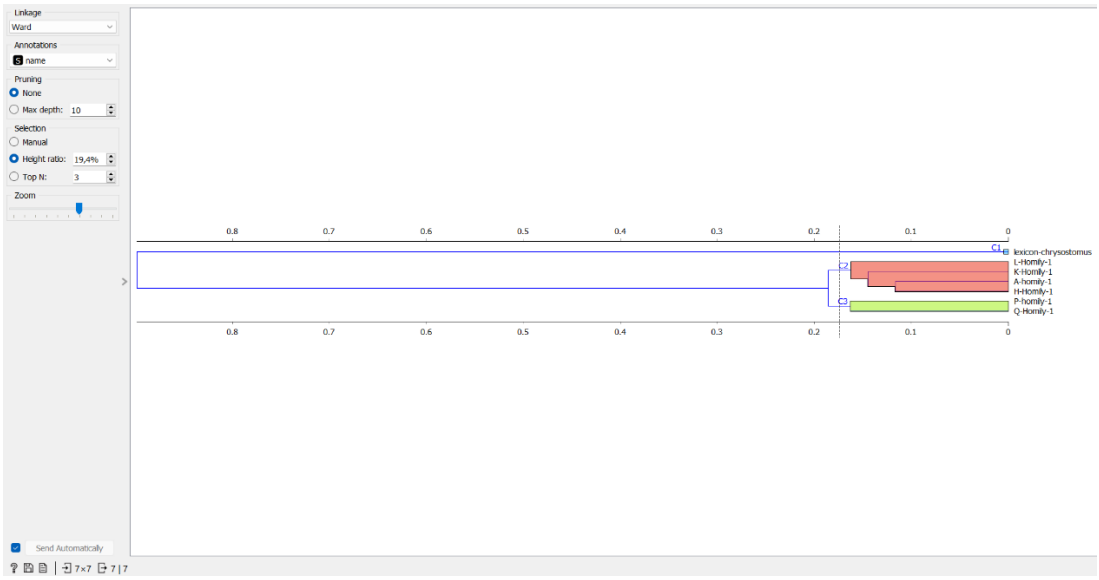

Figure 8 Revisions classification via Orange

Conclusion Remarks

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 on how to exploit specific tools in order 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. With further testing and fine-tuning, developing general models that could transcribe a good portion of Greek manuscripts is more than possible. Mass transcription from historical documents can fuel the research with much-needed data. On the other hand, even not entirely accurate data can produce a valuable outcome when researching manuscripts tradition as demonstrating, collating algorithms that mimic DNA sequencing extracted particular text features that assisted in visualising the manuscripts' interrelations. Especially data mining algorithms, such as the Orange toolkit, include NLP (Natural Language Processing) algorithms that can predict and illustrate all complex relationships between the sources.

Machine learning, indeed, benefits from data plethora, but sometimes data augmentation and manipulation can produce functioning results. According to the research questions, humans can then 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.

References

[1] ‘Pinakes | Πίνακες Institut de Recherche et d’histoire des textes’. https://pinakes.irht.cnrs.fr/ (accessed Oct. 08, 2020).

[2] C. Macé and P. V. Baret, 'Why Phylogenetic Methods Work : The Theory of Evolution and Textual Criticism', *Linguist. Comput.*, no. 24/25, 2004, doi: 10.1400/54380.

[3] M. Spencer, E. A. Davidson, A. C. Barbrook, and C. J. Howe, 'Phylogenetics of artificial manuscripts', *J. Theor. Biol.*, vol. 227, no. 4, pp. 503–511, Apr. 2004, doi: 10.1016/j.jtbi.2003.11.022.

[4] M. Konstantinidou, 'St John Chrysostom's Homilies on the Letters of St Paul to Titus. A Critical Edition with Introduction and Notes on Selected Passages', Thesis submitted to the Faculty of Classics in partial fulfilment of the requirements for the degree of Doctor of Philosophy, University of Oxford, Oxford, 2006.

[5] E. Perdiki and M. Konstantinidou, 'Handling Big Manuscript Data', *Classics@*, vol. 18, no. Ancient Manuscripts and Virtual Research Environments, special issue, 2021, Accessed: Jan. 21, 2022. [Online]. Available: https://classics-at.chs.harvard.edu/classics18-perdiki-and-konstantinidou/

[6] R. R. Ingle, Y. Fujii, T. Deselaers, J. Baccash, and A. C. Popat, 'A Scalable Handwritten Text Recognition System', in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sep. 2019, pp. 17–24. doi: 10.1109/ICDAR.2019.00013.

[7] 'GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)'. https://web.archive.org/web/20220125061256/https:/github.com/tesseract-ocr/tesseract (accessed Oct. 20, 2022).

[8] N. White, 'Training Tesseract for Ancient Greek OCR', *Εὔτυπον*, vol. 28–29, pp. 1–11, Oct. 2012.

[9] 'GitHub - githubharald/SimpleHTR: Handwritten Text Recognition (HTR) system implemented with TensorFlow.' https://web.archive.org/web/20220124180144/https:/github.com/githubharald/SimpleHTR (accessed Oct. 20, 2022).

[10] ‘kraken — OCR system’. http://kraken.re/ (accessed Sep. 18, 2020).

[11] 'eScriptorium: A Digital Text Production Pipeline for Print and Handwritten Texts using machine learning techniques.' https://web.archive.org/web/20220719175947/https://escriptorium.fr/ (accessed Jul. 19, 2022).

[12] 'Transkribus | AI powered Handwritten Text Recognition'. https://web.archive.org/web/20211108183341/https:/readcoop.eu/transkribus/ (accessed Jul. 19, 2022).

[13] P. Kahle, S. Colutto, G. Hackl, and G. Muhlberger, 'Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents', in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Nov. 2017, pp. 19–24. doi: 10.1109/ICDAR.2017.307.

[14] 'About us - READ-COOP'. https://web.archive.org/web/20211213175714/https:/read-coop.eu/about/ (accessed Oct. 20, 2022).

[15] P. Ströbel and S. Clematide, 'Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images', *Digit. Humanit. 2019*, 2019, doi: 10.5167/UZH-177164.

[16] 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', in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 3551–3559. Accessed: Oct. 14, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.436

[17] D. Firmani, M. Maiorino, P. Merialdo, and E. Nieddu, 'Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio -- Episode 1: Machine Transcription of

the Manuscripts', in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2018, pp. 263–272. doi: 10.1145/3219819.3219879.
[18] 'Download Transkribus | Download the Expert Client', Nov. 13, 2021. http://web.archive.org/web/20211113063459/https:/readcoop.eu/transkribus/download/ (accessed Jul. 19, 2022).
[19] 'Transkribus Lite'. http://web.archive.org/web/20220119164148/https:/transkribus.eu/lite/ (accessed Jul. 19, 2022).
[20] R. Holley, 'How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs', *-Lib Mag.*, vol. 15, no. 3/4, Mar. 2009, doi: 10.1045/march2009-holley.
[21] C. Tomoiaga, P. Feng, M. Salzmann, and P. Jayet, 'Field typing for improved recognition on heterogeneous handwritten forms'. arXiv, Sep. 22, 2019. Accessed: Oct. 27, 2022. [Online]. Available: http://arxiv.org/abs/1909.10120
[22] A. Rabus, 'Recognising handwritten text in Slavic manuscripts: A neural-network approach using Transkribus', *Scr. E-Scr.*, vol. 19, pp. 9–32, 2019.
[23] M. L. West, *Textual Criticism and Editorial Technique: Applicable to Greek and Latin texts*. B. G. Teubner, 1973. Accessed: Oct. 23, 2022. [Online]. Available: https://www.degruyter.com/document/isbn/9783598774010/html
[24] E. Nury, 'Visualising Collation Results', *Variants*, no. 14, pp. 75–94, Mar. 2019, doi: 10.4000/variants.950.
[25] F. Boschetti, 'Methods to extend Greek and Latin corpora with variants and conjectures: mapping critical apparatuses onto reference text', *Proc. Corpus Linguist. Conf. Birm. 2007*, p. 9, 2007.
[26] 'Juxta | Collation Software for Scholars', Apr. 19, 2022. https://web.archive.org/web/20220419175634/https:/www.juxtasoftware.org/ (accessed May 27, 2022).
[27] 'CollateX'. https://collatex.net/ (accessed May 27, 2022).
[28] 'TRAViz'. http://www.traviz.vizcovery.org/index.html (accessed Oct. 23, 2022).
[29] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. Al-Shawakfa, 'A Comparison Study between Data Mining Tools over some Classification Methods', *Int. J. Adv. Comput. Sci. Appl.*, vol. 1, no. 3, Art. no. 3, 21 2012, doi: 10.14569/SpecialIssue.2011.010304.
[30] 'Weka 3 - Data Mining with Open Source Machine Learning Software in Java'. https://www.cs.waikato.ac.nz/ml/weka/index.html (accessed Oct. 23, 2022).
[31] B. L. University of Ljubljana, 'Orange Data Mining'. https://orangedatamining.com/ (accessed May 16, 2022).
[32] B. Goodall, *The homilies of St. John Chrysostom on the letters of St. Paul to Titus and Philemon: prolegomena to an edition*, vol. 20. University of California Press, 1979.
[33] R. Haentjens Dekker, D. van Hulle, G. Middell, V. Neyt, and J. van Zundert, 'Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project', *Digit. Scholarsh. Humanit.*, vol. 30, no. 3, pp. 452–470, Sep. 2015, doi: 10.1093/llc/fqu007.