HTREC 2022: "One Rule Based System to rule them all" Improving the HTR output of Greek papyri and Byzantine manuscripts in simple way

Konstantina Liagkou
Athens University of
Economics and Business
konstantinalia4@gmail.com

Emmanouil Papadatos
Athens University of
Economics and Business
manospapadatos@gmail.com

Abstract

An abundance of ancient Greek and Byzantine transcripts are preserved online in the form of images. In order to extract the underlying transcripts and record them in machine-encoded text, Optical Character Recognition (OCR) systems are used to extract text, albeit with little accuracy. In order to reduce the errors, Natural Language Processing (NLP) techniques are employed. The proposed system is based on simple heuristic rules that correct the text. Our rule based system (RBS) utilizes a corpus of ancient Greek and Byzantine words and does not require training or human annotations. Moreover, we provide a collection of ancient Greek and Byzantine texts, either to be converted into a lexicon for the RBS or to pretrain deep learning models. We also provide a lexicon of words separate from the collection of books. We release our datasets and code for public use: https://github.com/Connalia/htrec.

1 Introduction

The digitization of ancient texts is essential for analyzing ancient corpora and preserving cultural heritage. However, the transcription of ancient handwritten text using optical character recognition (OCR) methods remains challenging. Handwritten text recognition (HTR) concerns the conversion of scanned images of handwritten text into machine-encoded text. In contrast with OCR, where the text to be transcribed is printed, HTR is more challenging and can lead to transcribed text that includes many more errors or even to no transcription at all when training data on the specific script (e.g., medieval) are not available.

Existing work on HTR combine OCR models and Natural language processing (NLP) methods from fields such as grammatical error correction (GEC), which can assist with the task of post-correcting transcription errors. The post-correction task has been reported as expensive, time-consuming, and challenging for the human expert, especially for OCRed text of historical newspapers, where the error rate is as low as 10%. The HTREC focus of this challenge was the post-correction of HTR transcription errors, attempting to build on recent NLP advances such as the successful applications of Transformers and transfer learning. The ground truth of the evaluation set was used to score participating systems in terms of character error rate (CERR). (See Figure 5)

	Human annotations	***OCR system***
# of unique char	131	77
# of strings	12797	11042
# of unique strings	6210	7206

Table 1: **Training set statistics:** Table showcasing statistics regarding the number of characters and tokens both for the OCR extracted texts as well as the human annotated ones.

2 Exploratory Data Analysis

Two datasets were provided separately, training and a test set. The train contained 1,875 and the test 338 texts from ancient Greek Byzantine papyrus that were extracted from an OCR system and human annotators. The unique characters that were identified in the human-annotated text were 131, in contrast to the much fewer found in the system texts. The OCR system's unique strings were comparable in number to the human-annotated ones, with 183 being the difference in unique strings.

Table 1 summarizes all the statistical information.

The OCR system performed overall well in understanding all the characters, but it struggled to recognize the space in the correct position ("ἀλλατῆε κλησει"→"ἀλλὰ τῆ εκλύσει") and in other cases, some characters, thus outputting a space ("β ου"→"βίου").

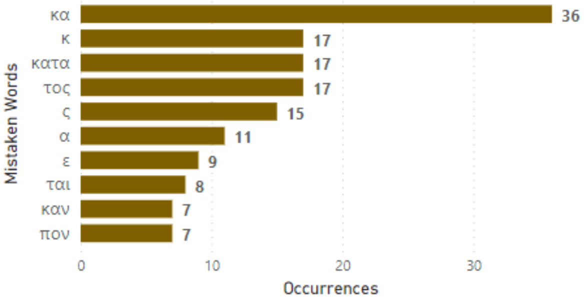


Figure 1: **Distribution of most frequent mistake:** The blue bars depict the frequency

Another key shortcoming of the system was the word stress ("εμπυριζου"→"ἐμπυρίζουσι"). The ancient Greek language had a number of different intonations, which the OCR system had trouble capturing. Moreover, the system often confused characters that were visually similar ("δ"→"ὸ"). Complex words ("ἐγγενομεναπαδημησμεννωτες"→"ἐγγινομένα πάθη μὴ σβεννύντε") were also not clearly understood by the OCR system which often concatenated them into one big word. Figure 1 emphasized the most common mistakes made by the OCR system with "κα" being the most frequent misrepresentation of "και" with over 35 occurrences.

Iterating through the train set, we identified that words could be split at the end of each line and continue to the next. ([id3]"τῶν ἀλλ" [id4]"οτρίων"→"ἀλλοτρίων")

More challenges arose, as the annotators made mistakes as well. A significant challenge was the lack of word stress, which was missing many words. To give an example, the "and" in Greek ("και") was the most common word in the train set with 599 appearances, where "καὶ" was 300 and "και" with 299 (see Table 2). The Figure 2, represents the removal of word stress from "καὶ", where the annotator corrected "μος, καὶ τοπρο ποτω απτω χω ηκατίςχων ερ" to "μου και το προσωπον των πτωχων καταισχυνετε". The OCR system outputted the correct word, but the annotator wrongly replaced it. Thus, there are inconsistencies between the same words. In addition, the annotators tried to correct the OCR system without considering the actual text; indeed, the system was at times closer to the real one. For instance, the papyri wrote, "ίδίας", which is very similar to "ἐδίας" that the OCR system outputted, in contrast to "ἡδεῖας" that the human annotator wrote, which is an entirely different word in Greek, as shown in Figure 3.

	OCR System				
Word	και	καὶ	του	τὴν	την
Frequency	168	157	100	98	88
	Human Annotations				
Word	καὶ	και	το	τὴν	δὲ
Frequency	300	299	110	105	95

Table 2: **Distribution of most common words:**5 most common words in the dataset

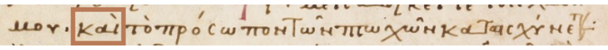


Figure 2: **Image showcasing part of the real transcript (και):** While the real transcript wrote "μος, καὶ τοπρο ποτω απτω χω ηκατίςχων ερ", the annotator wrote "μου και το προσωπον των πτωχων καταισχυνετε", text line 21, Bodleian-Library-MS-Barocci-102_00157_fol-75r.jpg

3 Methods

We first employed the three baselines offered, namely Edit_distance, LM_bases and LMing. Since those could not be scaled up in performance, the next step was to utilize deep learning models. We initially utilized a pretrained transformer with both the encoder and the decoder based on BERT architecture (Bert-to-Bert), as was introduced in the bibliography (Kaneko et al., 2020) for spell correction. The Bert-to-Bert[1] model utilized both a Greek Bert [2] and an ancient Greek Bert[3], each pretrained in texts from their respective languages. The next deep learning model was an EncoderDecoder with LSTM layers (Seq-to-Seq), based on the character representation of the text and outputted whole sentences. Neither of the deep learning models performed better than the baselines due to a lack

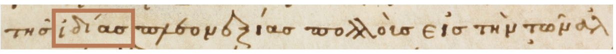


Figure 3: **Image showcasing part of the real transcript ("ίδίας"):**OCR"της ἐδίας πλσον ἐξιας πολλους ἐις τὴν τῶν ἀλ" Human: "τῆς ἡδεῖας πλεονεξίας πολλοὺς εἰς τὴν τῶν ἀλλ",, text line 4, Bodleian-Library-MS-Barocci-102_00157_fol-75r.jpg

[1] https://huggingface.co/docs/transformers/model_doc/encoder-decoder

[2] https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1

[3] https://huggingface.co/pranaydeeps/Ancient-Greek-BERT

of resources for training and hyperparameter tuning and thus were not included in the Experimental Results.

Finally, we developed a rule-based system (RBS) that was based on a small corpus of ancient Greek words. First and foremost, the dataset was split into tokens between whites-spaces. If a token in the test set contained a word in corpus A, it would be split. To give an example, the token "παιξτεεχηρευ", which contained the word "παιξτε". Applying the rule, the system returned "παιξτε εχηρευ". The same rule but with stricter parameters was applied with an additional corpus B that included pronouns, like "αυτοῦ", and conjunctions, like "εαν" and "και", because these words were often found or were part of other words, like "καινούργιος" meaning something is new. The rule joined all possible combinations of words from corpora A and B, and if the word was found in the test set, it was split. For instance, the words "και" from corpus A and "παιξτε" from corpus B. If the new word "καιπαιξτε" was found in the dataset, it was split like "και παιξτε". Contrary to that would be the word "καινούργιος", where "νούργιος" is not a real word, thus the join would never happen.

The next rule checked if the token with edit distance equal to one character was in corpora A and C, where C contained all the possible alterations of "και", like ["και", "καὶ", "καί"] and replaced it, else it left the token unchanged. The same rule was separately applied for words with more than 8 characters, like "γυναικας", but with edit distance equal to two characters.

Another challenging concept was that the model produced white spaces in the wrong positions, messing up words like "δικαι ονπεριτους", which stems from "δικαιον περιτους". To counter this issue, the system took bigrams and merged them, like "δικαιονπεριτους". Then it would split the words into known words found in the corpora and create a new correct bigram like "δικαιον περιτους".

The rule based model detected single characters tokens and applied two rules. The first rule checked whether the character was an article like "ή", in "ἀπλῶς ή πῶς". In the case where the character was not an article, then the second rule merged the token to the left with the character on its' end and the token to the right with the character on its' start. If any of those combinations matched with a word in the corpora, it would be replaced, like "φοῦ ἐ δίαι", which was corrected to "φοῦ δίαι".

Wrong Token	*Corrected Tokens*
ειντοις	ἐν τοῖς
εντοις	ἐν τοῖς
ἐντῷ	ἐν τῷ
ηκτοις	εκ τοις
εκτῆς	εκ τῆς
εκτης	εκ τῆς

Table 3: Transformations of tokens starting with "ἐν" , "εκ" and ending with one of the above substrings

Duplicate characters found in tokens were replaced with single ones, unless the token was found in the corpora as it was, for example, "ἐεστιν", which was converted to "ἐστιν".

For pronouns like "τὴν" and "τῶν", edit distance was not viable because too many words contained those pronouns with an edit distance of two characters. Thus, should the system recognize them in any wrong order, it would replace them with the correct version, like ("τνω", "των") or ("τνῶ", "τῶν"). Moreover, if those pronouns were located at the beginning of a token, the system would split them into two new tokens with a white space in between, like "τηνκαρδιαν", which was altered to "την καρδιαν".

For commonplace words that begin with "ἐν" and "εκ", little could be achieved with the above rules. Thus when the system recognized tokens like "ειντοις", it replaced them with two new tokens like "ἐν τοῖς". Find indicative replacements on Table 3.

4 Experimental Results

The main metrics that were used were The Character Error Rate (CERR) and The Word Error Rate (WERR)(Platanou et al., 2022). Our rule based model did not perform well on the subset of real data, as shown in Table 5. Nevertheless, it achieved first place on the whole test dataset and separately on synthetic data with CERR 0.278 and 0.096, respectively.

There is inconsistency in intonation on the text corrected by the annotators, with 50 sentences having intonations and 130 not having intonations in the test set. Thus, an experimental version of our best rule based system (RBS) where the intonation was removed was tested. The experimental RBS managed to produce far better results (4.632 CERR) than the winning submission (2.525 CERR) on real data and on the ranking of real synthetic (5.947

CERR			
Models \ Data	*RuleBase with intonation*	*RuleBase without intonation*	*1st Place*
real	0.439	**4.632**	2.525
synthetic	**0.096**	-7.826	-7.719
real & synthetic	0.278	**5.947**	-2.264

Table 4: **Scores Table:** The CERR score for our model (with and without intonation) versus the winner.

CERR), Table 4. It did not perform as well on synthetic data, as those data preserved intonations on their text as they were produced. To sum up, we observe that our RBS, should the intonations be removed, would supersede all other models in the competition.

5 Error Analysis

The Rule Based system (RBS) sought to correct words with the correct intonations. Our observation is that the annotators were not consistent with adding or removing intonations because while the intonations were present in the actual text, and the OCR model detected them, they were still removed in the corrected text (see Figure 4). An example is the sentence "μολλειν αὗτη διδειγματα περιτῆς ἐν τοῖς νο", as it was detected from the OCR system. The annotators corrected the sentence to "μελλειν αυτη διδαγματα περι της εν τοις νοη", having removed intonations, while the RBS had added intonations to parts of that sentence. The RBS also struggled with small words that were very similar. Based on the above example, is the word "μολλειν", which the RBS corrected to "μολεῖν" (an infinitive of "βλώσκω"), instead of the correct word "μελλειν". Another shortcoming of the RBS is its small corpora, which limits its correcting capabilities. Based on the same example sentence, the system could not correct the word "διδειγματα" to "διδαγματα" because the specific word was not present in the corpus.

The RBS did not perform well on the original dataset since the corrected text was often completely different from the OCR"s output. The OCR system often identified real words that needed minimum correction, while in reality, the correct text was far from that representation. For instance the OCR identified the sentence "καξεύδὢςτα ξήμής", while the actual was "κἀξηῦρες ὥστ ἀζημίους". Such correcting capabilities are out of scope for the RBS since it utilizes basic rules and is not trained to identify grammatical and syntactical structures.

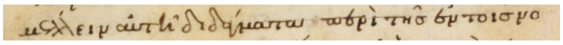


Figure 4: **Image showcasing part of the real transcript:** The text in Figure is "μελλειν αυτη διδαγματα περι της εν τοις νοη", that depicted in line 4, on Bodleian-Library-MS-Barocci-102-80r

6 Discussion and Future Work

A rule based system (RBS) is only as good as the corpus size it has access to. We hypothesize that the system's performance would improve with a bigger corpus. To that end, we provide over 100 books of text in ancient Greek [4] and Byzantine [5], scraped from various online sources. Unfortunately, due to time constraints, we could not utilize them, and thus we provide them for future work. The biggest collection, titled 'Σύνοψις Ιστοριών' from I.Skylitzis totals 5 books, 153,709 words and 885,259 characters [6]. Furthermore, we provide a lexicon [7] of over 42,107 ancient Greek words independent of the collection of books, which was also not utilized by the RBS.

The OCR system recognized complementary text in the ancient transcripts, which had a different colour from the main texts. The annotators handled this case by removing the supplementary text altogether. However, since the RBS only sought to correct the system transcripts, it failed to capture the correct text and thus corrected words that did not belong in the actual text. One way to handle the supplementary different color coded text (shades of red) in the ancient transcripts would be to map those pixels to the colour of the papyri to be filtered out by the system altogether.

The ancient Greek language had several different intonations so as to give different meanings depending on the use of each word. A bidirectional LSTM could understand the OCRed sentences' syntactical and grammatical order and produce the correct intonations for each word, and another sys-

[4] http://users.uoa.gr/~nektar/history/tributes/ancient_authors/index.htm

[5] https://byzantium.gr/keimena/keimena.php

[6] https://wordcounter.tools/

[7] https://www.greek-language.gr/digitalResources/ancient_greek/tools/liddel-scott/search.html?start=20&lq=

			CERR				
Data / Models	***Edit_distance***	***LM_based***	***LMing***	***RuleBase***	***1st Place***	***2nd Place***	***3rd Place***
real	-0.187	-6.057	0.022	0.439	**2.525**	1.034	0.567
synthetic	-0.543	-5.629	0.000	**0.096**	-7.719	-7.784	-0.629
real & synthetic	-0.353	-5.857	0.012	**0.278**	-2.264	0.049	0.247
			WERR				
Data / Models	***Edit_distance***	***LM_based***	***LMing***	***RuleBase***	***1st Place***	***2nd Place***	***3rd Place***
real	-0.286	-0.802	0.062	1.822	**14.967**	4.629	3.398
synthetic	-2.477	-3.133	0.000	**1.292**	-23.140	-31.859	-3.112
real & synthetic	-1.310	-1.891	0.033	1.575	-2.846	-0.101	**1.836**

Table 5: **Scores Table:** The CERR and WERR score for our model, baselines and the top 3 contestants.

tem could produce a better representation of the sentence based on historical knowledge. An ideal solution to solve both intonations and grammatical and syntactical errors would be to develop a char-to-word model with CNN and RNN layers (Ghosh and Kristensson, 2017).

An interesting approach to the intonation problem would be to remove all word stress from vowels and apply a machine learning algorithm to specify which word stress is appropriate for each n-gram of words. As a final remark, language evolves over time. The texts used belonged to different centuries; thus, different grammar or syntactic rules might have been used and should ideally be accounted for.

7 Conclusion

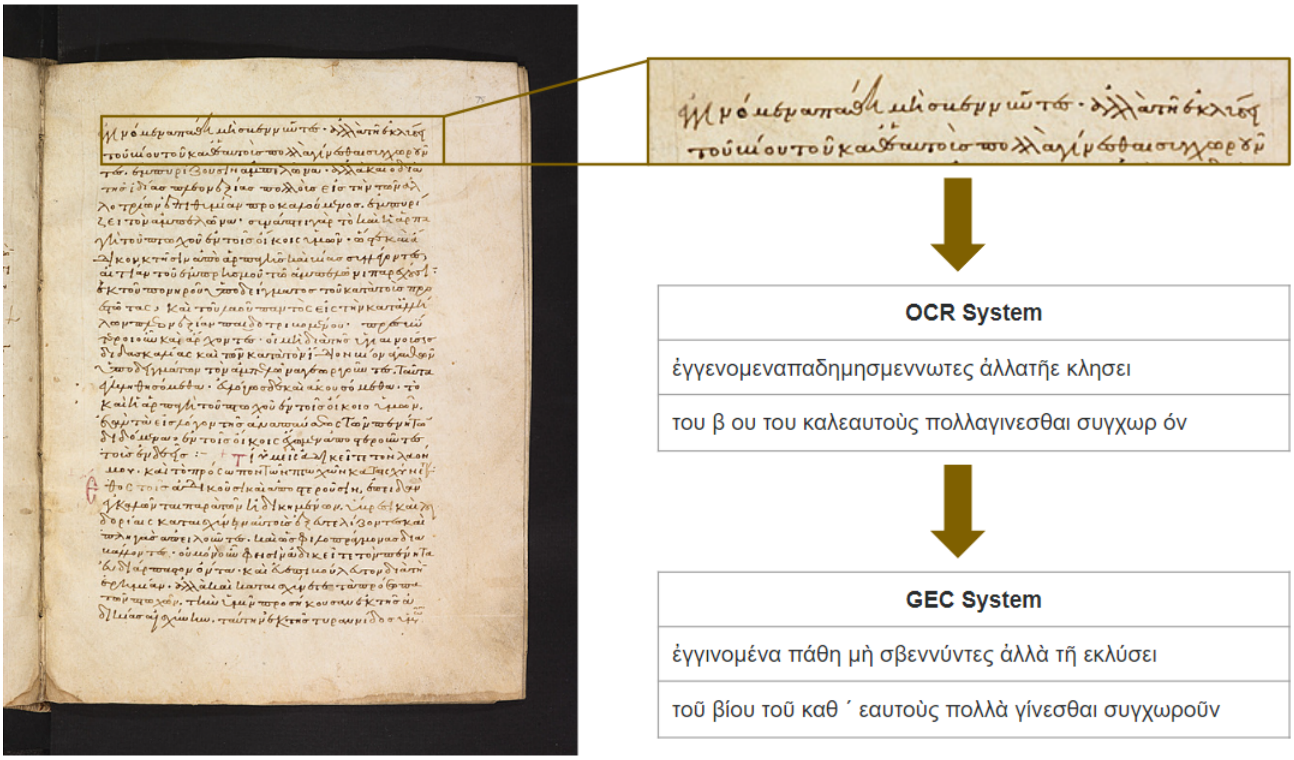
In this paper, we propose a rule based system (RBS) for text correction and the improvement of the HTR output of Greek papyri and Byzantine manuscripts. We show that our model achieves a CERR of 0.439 on real data, 0.096 on synthetic data and 0.278 on real and synthetic data. Traditional text correction relies on the use of machine learning models. In contrast, our model applies rules to correct and includes a small corpus with correct words without any training. Further, we propose an alternative experimental rule based system (eRBS) to address the inconsistency of intonations in the human-annotated text. Since more than half of those sentences in the test set did not include intonations, the eRBS outperformed all other models achieving a CERR of 4.632 on real data, 5.947 on real and synthetic data and -7.826 on synthetic data. Finally, we also provide over 100 books of ancient Greek and Byzantine texts to supplement the existing corpora.

References

Shaona Ghosh and Per Ola Kristensson. 2017. Neural networks for text correction and completion in keyboard decoding. *arXiv preprint arXiv:1709.06429*.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*.

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022. Handwritten paleographic greek text recognition: A century-based approach. pages 6585–6589.



OCR System
ἐγγενομεναπαδημησμεννωτες ἀλλατῆε κλησει
του β ου του καλεαυτοὺς πολλαγινεσθαι συγχωρ όν

GEC System
ἐγγινομένα πάθη μὴ σβεννύντες ἀλλὰ τῆ εκλύσει
τοῦ βίου τοῦ καθ ΄ εαυτοὺς πολλὰ γίνεσθαι συγχωροῦν

Figure 5: **Description of Approach:** In this image we present the pipeline followed from extraction to correction of the ancient transcripts. The OCR system extracts the text from the image, which is then processed by the GEC system and corrected. Image url: https://digital.bodleian.ox.ac.uk/objects/8102c257-80c2-4d52-97e8-f23b38b5ab0e/surfaces/8ffbeea4-630a-4580-a2df-c2103abbe554/