OCR for Greek polytonic (multi accent) historical printed documents: development, optimization and quality control

Anna-Maria Sichani
Sussex Humanities Lab, University of Sussex
East Sussex BN1 9RG, United Kingdom

Panagiotis Kaddas
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research Demokritos
Agia Paraskevi GR-153 10, Greece

Georgios K. Mikros
Department of Italian Language and Literature School of Philosophy, National and Kapodistrian University of Athens
Athens GR-157 84, Greece

Basilis Gatos
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research Demokritos
Agia Paraskevi GR-153 10, Greece

ABSTRACT

This paper presents the development and implementation of a robust OCR tool and a related comprehensive workflow for the recognition of Greek printed polytonic scripts. This project is initiated and developed by an interdisciplinary team with expertise in the areas of document image processing, character segmentation and recognition, machine learning, corpus creation and digital humanities. Our paper aims to describe the design and development of the workflow around this project, including data gathering and structuring, OCR tool development, user interface development, experiments on the training procedure of the tool, evaluation, post-correction and quality control of the results.

KEYWORDS

Optical Character Recognition, Greek polytonic scripts, historical printed documents, image processing, page segmentation, machine learning, post correction workflow

ACM Reference Format:
Anna-Maria Sichani, Panagiotis Kaddas, Georgios K. Mikros, and Basilis Gatos. 2019. OCR for Greek polytonic (multi accent) historical printed documents: development, optimization and quality control. In *3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019), May 8–10, 2019, Brussels, Belgium.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3322905.3322926

DATeCH2019, May 8–10, 2019, Brussels, Belgium
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7194-0/19/05...$15.00
https://doi.org/10.1145/3322905.3322926

1 INTRODUCTION

Although the accurate recognition of Latin machine-printed text is now considered largely a solved problem, recognition of non-Latin minority languages or scripts having a large number of character classes is still subject of active research. Greek polytonic (multi accent) scripts have a large variety of diacritic marks and as a result a large number of character classes (more than 270). Due to that, Greek polytonic documents (both handwritten and printed) cannot be successfully processed by the currently available Optical Character Recognition (OCR) technologies. Furthermore, as the Greek polytonic scripts were actively used from around 200 BC to modern times until 1982, we can easily observe that a huge amount of digitised (handwritten and printed) Greek documents still remains without full text processing and analysis capabilities.

Drawing expertise from the areas of document image processing, page segmentation, pattern recognition, machine learning and digital humanities, our paper aims to describe the development and implementation of a robust OCR tool and workflow for the recognition of printed Greek polytonic scripts. This interdisciplinary work is partially supported by and initiated within the COST action [1] Distant Reading for European Literary History(CA16204), aiming to create the first Modern Greek Literary Textual Collection as part of a larger European Literary Textual Collection (ELTeC) and to apply Distant Reading methods in order to revisit the European Literary History. The interdisciplinary team has been working on this project for almost a year now on a semi-funded basis and is currently formed by 7 members from the areas of Digital Humanities, Computational Linguistics and Corpus creation, Modern Greek Literature and Computer Science.

In what follows, we are briefly discussing the state-of-the-art and previous work in the field of OCR for Greek polytonic scripts (Section 2). Then we are describing our workflow, currently consisted of four (4) stages. Firstly, we will discuss the data acquisition process and data quality specifications, while

[1] https://www.distant-reading.net

discussing obstacles and common problems in data gathering and corpus creation (Section 3). Furthermore, we will present the development of a tool for the recognition of old Greek polytonic scripts, by using very simple machine learning algorithms as well as the development of a ground-truthing Graphical User Interface (GUI) application for the manual annotation of old Greek polytonic documents (Section 4). We will then focus on the application and the evaluation of the trained model for the recognition of the data corpus (Section 5). Lastly, we will discuss the development and implementation of a quality control and post-correction framework for the resulted OCR-ed documents (Section 6). We will conclude with some plans for future development of this research (Section 7).

2 PREVIOUS WORK

The automated recognition of Greek printed polytonic scripts was for many years an open question/challenge for researchers in the Humanities and Computer Science.

Initiated by the community of (Digital) Classics, OCR for Greek polytonic (mainly Ancient Greek) is still a claim and an open challenge for Humanities and Cultural Heritage researchers working with textual resources using Greek polytonic. Several previous attempts developing an OCR engine [2], focusing on line recognition and character patterns, alongside a GUI interface [3], are offering limited accuracy and poor results for non-Unicode (old) typefaces [4].

Related work within the Perseus project in [3] on progressive multiple alignment and constrained spell checking to different OCR outputs of printed critical editions of Ancient Greek texts is a promising development in Greek polytonic OCR.

In [6], an OCR framework for the recognition of machine-printed Greek polytonic documents is proposed that is based on combining different recognition modules. One module is used for accent recognition while four modules are used for recognizing characters that belong to different horizontal text zones. Moreover, pre-processing stages like text dewarping and text line detection are included in order to assist the recognition modules.

In [5], adaptive zoning features are proposed as an improvement to the already robust zoning feature descriptors. The main idea is to adjust the position of every zone based on local pattern information. Movements are based on the maximization of the local pixel density around each zone.

In [7], a recognition system for ancient Greek polytonic scripts is based on Hidden Markov Models (HMM) [1] is used. This system is character segmentation free and can be applied both on character and word level. The HMM-based method uses language models of bigrams at character level and yields more accurate results when compared to previous techniques which require a more sophisticated character segmentation technique.

3 DATA ACQUISITION

In order to train the OCR tool, a corpus of Greek texts using polytonic scripts was created as a test bed. As this project was related to and partially funded by the Cost action Distant Reading, the Modern Greek corpus was mainly compiled by hundred (100) 19th century Greek novels and their length ranges between 7-400 pages. There have been also several other criteria for the corpus sampling, in order to create a text corpus that would be balanced in terms of authors'gender representation, date of publication, length, et.c.

The majority of the texts were initially hold and digitised by various Greek institutions, including the National Library of Greece, using different digitisation standards, thus, resulting to variations in terms of quality of digitisation. In addition, as the texts were derived mainly from original printed editions of the 19th and 20th century using various typefaces and diacritics, there were also major challenges in terms of image quality and variations in the digitisation results.

4 OCR TOOL

4.1 Proposed OCR Method

The proposed system for the recognition of old Greek printed polytonic scripts consists of four very simple modules and can be seen in Figure 1.

The first one is the initial character segmentation and recognition module, where all input document images are fed into an existing OCR tool [2]. The acquired information is used as an initial stage for our system. Next, these results are post-processed manually (second module) in order to improve the quality of the initial character recognition and create an accurate training set.

The third module is a feature extractor algorithm, were each already segmented character is normalized into a fixed size and then splitted into 3x3 zones. Then, from each sub-image pixel intensity is extracted as image feature and the final descriptor for each character is the zone feature concatenation. This module is used in order to create a train set containing samples from all unique character classes. Also and during test phase, the same features are extracted from each new character and used as input for the classification module.

The last module uses the well-known machine learning algorithm, k-NN [4], in order to classify characters to the nearest class of the train set. The decision relies upon the feature values extracted from the previous module. Experiments with different values of the k parameter (k=1, 3, 5, 7) were conducted and k-NN classification algorithm with k=5 gave was the most accurate results (Section 5).

4.2 User Interface

A Graphical User Interface (GUI) was created in order to assist the manual annotation of old Greek polytonic documents. This GUI application is easy-to-use and provides a very fast and accurate way of marking each character of a document

[2] http://tesseract.projectnaptha.com
[3] https://dcthree.github.io/antigrapheus/
[4] https://unicode.org

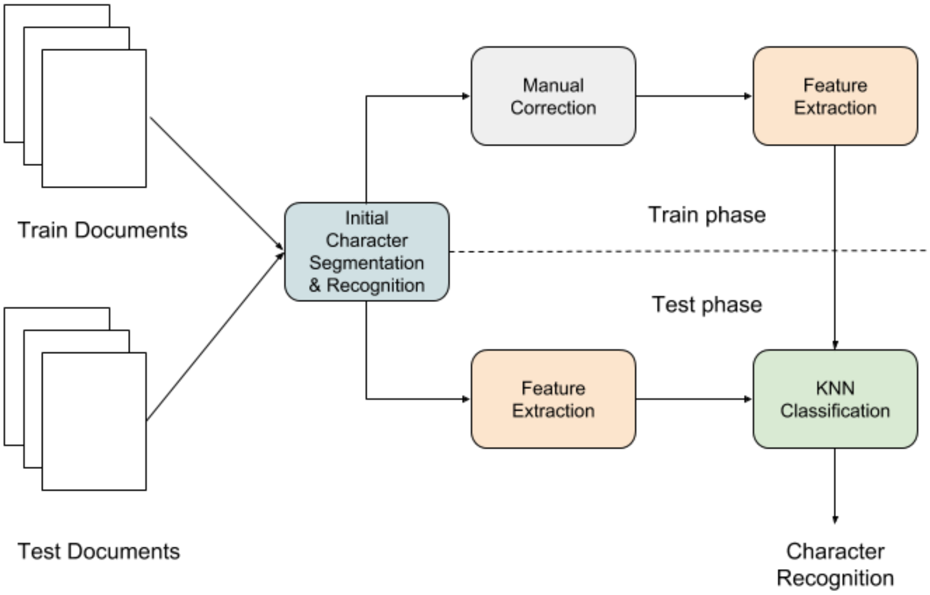
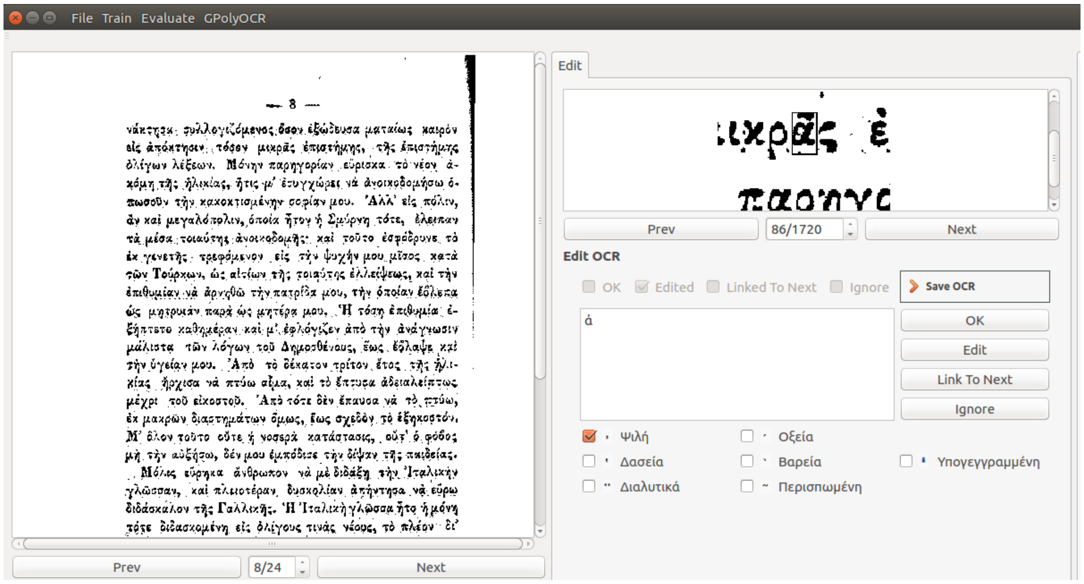


Figure 1: Overview of the proposed OCR system. Blue: Initial Segmentation & Recognition module used as input to the system. Gray: Manual correction module for input enhancement. Orange: Feature Extraction using pixel intensity for different zones of each character. Green: Classification module using the k-NN algorithm.



— 8 —

νάκτησα· συλλογιζόμενος ὅσον ἐξώδευσα ματαίως καιρὸν εἰς ἀπόκτησιν τόσον μικρᾶς ἐπιστήμης, τῆς ἐπιστήμης ὀλίγων λέξεων. Μόνην παρηγορίαν εὕρισκα τὸ νέον ἀκόμη τῆς ἡλικίας, ἥτις μ' ἐσυγχώρει νὰ ἀνοικοδομήσω ὁπωσοῦν τὴν κακοκτισμένην σοφίαν μου. Ἀλλ' εἰς πόλιν, ἂν καὶ μεγαλόπολιν, ὁποία ἦτον ἡ Σμύρνη τότε, ἔλειπαν τὰ μέσα τοιαύτης ἀνοικοδομῆς· καὶ τοῦτο ἐσφάδρυνε τὸ ἐκ γενετῆς τρεφόμενον εἰς τὴν ψυχήν μου μῖσος κατὰ τῶν Τούρκων, ὡς αἰτίων τῆς τοιαύτης ἐλλείψεως, καὶ τὴν ἐπιθυμίαν νὰ ἀρνηθῶ τὴν πατρίδα μου, τὴν ὁποίαν ἔβλεπα ὡς μητρυιὰν παρὰ ὡς μητέρα μου. Ἡ τόση ἐπιθυμία ἐξήπτετο καθημέραν καὶ μ' ἐφλόγιζεν ἀπὸ τὴν ἀνάγνωσιν μάλιστα τῶν λόγων τοῦ Δημοσθένους, ἕως ἔβλαψε καὶ τὴν ὑγείαν μου. Ἀπὸ τὸ δέκατον τρίτον ἔτος τῆς ἡλικίας ἤρχισα νὰ πτύω αἷμα, καὶ τὸ ἔπτυσα ἀδιαλείπτως μέχρι τοῦ εἰκοστοῦ. Ἀπὸ τότε δὲν ἔπαυσα νὰ τὸ πτύω, ἐκ μακρῶν διαστημάτων ὅμως, ἕως σχεδὸν τὸ ἑξηκοστόν. Μ' ὅλον τοῦτο οὔτε ἡ νοσερὰ κατάστασις, οὔτ' ὁ φόβος μὴ τὴν αὐξήσω, δέν μου ἐμπόδιζε τὴν δίψαν τῆς παιδείας.

Μόλις εὕρηκα ἄνθρωπον νὰ μὲ διδάξῃ τὴν Ἰταλικὴν γλῶσσαν, καὶ πλειοτέραν δυσκολίαν ἀπήντησα νὰ εὕρω διδάσκαλον τῆς Γαλλικῆς. Ἡ Ἰταλικὴ γλῶσσα ἦτο ἡ μόνη τότε διδασκομένη εἰς ὀλίγους τινὰς νέους, τὸ πλέον δι'

□ OK ☑ Edited □ Linked To Next □ Ignore

ά

☑ ᾿ Ψιλή
□ ῾ Δασεία
□ ¨ Διαλυτικά
□ ΄ Οξεία
□ ` Βαρεία
□ ~ Περισπωμένη
□ ͺ Υπογεγραμμένη

Figure 2: Example of the developed Graphical User Interface

page. An existing OCR tool [2] was used in order to acquire an initial character recognition, but recognition results are poor for old Greek polytonic fonts (accuracy is below 80% in most cases) and need further optimization. User feedback can be embedded by complementing possible polytonic accents to a character just by clicking buttons and without the need of typing special Unicode characters, usually hard to type and impossible to remember. Furthermore, the user is able to discard or modify erroneous character segmentation and recognition results and finally save all changes into a database file. An example of the developed interface is shown in Figure 2.

5 EXPERIMENTS

Having developed all the required tools, the training procedure of the model took place by annotating a 1-2 pages per document (depending on the document degradation) using the already described ground-truthing framework. This led into a database system containing information from more than 100 document pages (50k of character entries belonging to 143 unique classes). After completing the training of the model, the OCR tool was applied in batch mode for all the 100 old Greek polytonic novels.

Evaluation of the results using the described framework and random pages as test set yielded accuracy of 94.2% for the best parameter of the k-NN algorithm (k=5) Table 1. This set contained 50k characters belonging to 143 unique classes. The accuracy metric that was used for evaluation was the Levenshtein (or Edit) distance between the predicted transcription of each recognized page and the manually annotated transcription. A maximum accuracy threshold for a page was 98.6% and was usually achieved in document pages of low degradations and noise-free fonts. Respectively, the lowest accuracy was 82.3%.

Moreover, the results are somehow bound to the character segmentation results acquired from the existing OCR tool [2] used as initial stage of the developed system. Despite this, the proposed system results into more accurate results (25% improvement) when compared to the initial recognition results.

Table 1: Character recognition results for the proposed system using the Levenshtein/Edit Distance.

Method	Accuracy (%)
ABBYY FineReader [2]	68.7
k-NN(k=1)	89.3
k-NN(k=3)	92.1
KNN(k=5)	**94.2**
k-NN(k=7)	93.8

6 POST-CORRECTION AND QUALITY CONTROL

In order to further assess the resulted OCRed files, a post correction and quality control strategy was designed and put in place. The team working on the post-correction was formed by five Modern Greek Literary scholars with digital humanities skills and experience.

Firstly, the resulted OCRed files (in txt format) were manually compared with the original digitised files and corrected accordingly (Figure 3). When checking a text, the aim was to ensure that the text is accurately compared to the original source, with alterations and corrections made where considered necessary. At this point, an auxiliary resource was also developed as a look-up vocabulary to assist with and speed up the post-correction phase. The corrected texts were then used as ground truth files in txt format: a comparison with the resulted OCR file could be easily performed, resulting to an evaluation of the entire procedure. The texts are further proceeded and converted in XML TEI format, following the guidelines and the schema of the Distant Reading project working group.

In addition, and in order to document both the quality of the tool and the post-correction resources needed, we are collecting data, mainly concerning three variables: the time needed for the correction, the type and the amount of errors found. With this type of data we are able, alongside qualitative comments, to accurately estimate the resources (both human and financial) needed for post-correction as well as to locate the type of errors and further train the tool for future use.

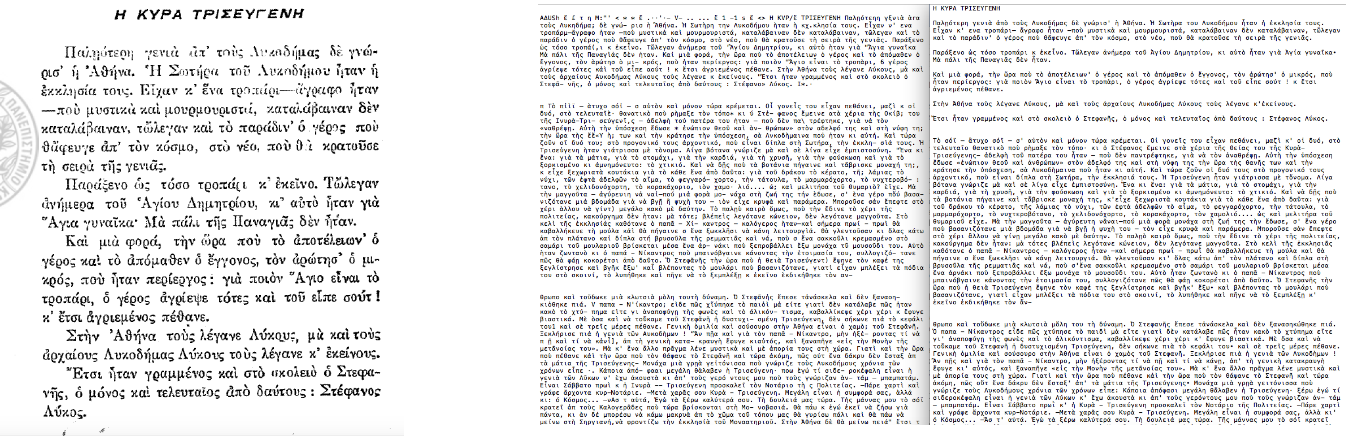
7 CONCLUSION & FUTURE PLANS

This research project aims to further contribute to and advance the Optical Character Recognition (OCR) methods for Greek printed polytonic scripts. While working towards the development of the OCR tool, we are also able to point out existing obstacles and gaps in areas such as digitisation standards, ground truth as well as (manual and automated) post-correction. This new OCR framework will facilitate accessing and further processing a huge amount of historical printed polytonic documents, currently digitised and held by numerous cultural institutions but with limited usability in terms of full text access and analysis. By developing a user-friendly and robust OCR framework for Greek polytonic documents, scholars in areas such as Literary Studies, Computational Linguistics, Archival and Information Science will be able to perform previously unavailable advanced computational processes on Greek printed polytonic textual resources.

As for the current phase of the tool, development of a corpus of 19th century printed texts was used as training set; we are willing to enlarge the test set with earlier textual resources from the 18th century as well as latest ones from the 20th century. From a technical point of view, there are a couple of aspects that we are willing to work on in future iterations, such as a more sophisticated classification of characters and perhaps an update on the feature extraction mechanisms.

8 ACKNOWLEDGMENTS

This work was partially funded by the COST Action "Distant Reading for European Literary History" (CA16204). We are also grateful to the National Library of Greece for



Η ΚΥΡΑ ΤΡΙΣΕΥΓΕΝΗ

Παληότερη γενιὰ ἀπ' τοὺς Λυκοδήμας δὲ γνώρισ' ἡ Ἀθήνα. Ἡ Σωτήρα τοῦ Λυκοδήμου ἦταν ἡ ἐκκλησία τους. Εἶχαν κ' ἕνα τροπάρι—ἄγραφο ἦταν —ποὺ μυστικὰ καὶ μουρμουριστά, καταλάβαιναν δὲν καταλάβαιναν, τὤλεγαν καὶ τὸ παράδιν' ὁ γέρος ποὺ θἄφευγε ἀπ' τὸν κόσμο, στὸ νέο, ποὺ θὰ κρατοῦσε τὴ σειρὰ τῆς γενιᾶς.

Παράξενο ὡς τόσο τροπάρι κ' ἐκεῖνο. Τὤλεγαν ἀνήμερα τοῦ Ἁγίου Δημητρίου, κι' αὐτὸ ἦταν γιὰ Ἅγια γυναῖκα· Μὰ πάλι τῆς Παναγιᾶς δὲν ἦταν.

Καὶ μιὰ φορά, τὴν ὥρα ποὺ τὸ ἀποτέλειων' ὁ γέρος καὶ τὸ ἀπόμαθεν ὁ ἔγγονος, τὸν ἀρώτησ' ὁ μικρός, ποὺ ἦταν περίεργος: γιὰ ποιὸν Ἅγιο εἶναι τὸ τροπάρι, ὁ γέρος ἀγρίεψε τότες καὶ τοῦ εἶπε σούτ! κ' ἔτσι ἀγριεμένος πέθανε.

Στὴν Ἀθήνα τοὺς λέγανε Λύκους, μὰ καὶ τοὺς ἀρχαίους Λυκοδήμας Λύκους τοὺς λέγανε κ' ἐκείνους.

Ἔτσι ἦταν γραμμένος καὶ στὸ σκολειὸ ὁ Στεφανῆς, ὁ μόνος καὶ τελευταῖος ἀπὸ δαύτους: Στέφανος Λύκος.

(a) Original Document Image

```
AΔUSh ἔ £ τ η M:"' < * * ἔ .··'·- V- .. ... ἔ 1 -1 s ἔ <> Η KVP/ἔ ΤΡΙΣΕΥΓΕΝΗ Παλησότεηη γξνιὰ ἀτα
τοὺς Λυκηδήμα; δὲ γνώ- ρισ ἡ Ἀθήνα. Ἡ Σωτὴρη την Λυκοδήμου ἡταν ἡ κχ.κλησία τους. Εἶχαν ν' ενα
τροπάρμ—ἄγραφο ἡταν —ποὺ μυστικά καὶ μουρμουριστά, καταλάβαιναν δὲν καταλάβαιναν, τὤλεγαν καὶ τὸ
παράδιν ὁ γέρος ποὺ θἄφευγε ἀπ' τὸν κόσμο, στὸ νέο, ποὺ θὰ κρατοῦσε τὴ σειρὰ τῆς γενιᾶς. Παράξενο
ὡς τόσο τροπά(,ι κ ἐκεῖνο. Τὤλεγαν ἀνήμερα τοῦ "Ἁγίου Δημητρίου, κι αὐτὸ ἡταν γιὰ "Ἅγια γυναῖκα
Μὰ πάλι τῆς Παναγὶάς δὲν ἡταν. Καὶ μιὰ φορά, τὴν ὣρα ποὺ τὸ ἀποτέλειων ὁ γέρος καὶ τὸ ἀπόμαθεν ὁ
ἔγγονος, τὸν ἀρώτησ ὁ μι- κρός, ποὺ ἡταν περίεργος: γιὰ ποιὸν "Ἅγιο εἶναι τὸ τροπὰρι, 6 γέρος
ἀγρίεψε τότες κἀὶ τοῦ εἶπε αούτ ! κ ἔτσι ἀγριεμένος πέθανε. Στὴν Ἀθήνα τοὺς λέγανε Λύκους, μὰ καὶ
τοὺς ἀρχαίους Αυκοδήμας Λύκους τοὺς λέγανε κ ἐκείνους. "Ἔτσι ἡταν γραμμένος καὶ στὸ σκολειὸ ὁ
Στεφᾶ- νῆς, ὁ μόνος καὶ τελευταῖος ἀπὸ δαύτους : Στέφανο» Λύκος. I*.·

π Τὸ πὶὶϊ — ἀτυχο σόί — σ αὐτὸν καὶ μόνον τώρα κρέμεται. Οἷ γονεἷς του εἶχαν πεθάνει, μαζὶ κ οἱ
δυό, στὸ τελευταῖὲ· θανατικὸ ποὺ ρὴμαξε τὸν τόπο* κι ύ Στέ- φανος ἔμεινε ατὰ χέρια τὴς Οκίβ; του
τῆς Ιυυρὰ-Τρι- σεύγενῖ,ς — ἀδελφὴ τοὐ πατέρα του ἡταν — ποὺ δὲν πα\ τρέφτηκε, γιὰ νά τὸν
«ναθρέψη. Αὐτὴ τὴν ὑπόσχεση ἔδωσε * ἐνώπιον θεοῦ καὶ ἀν- θρώπων» στὸν αδελφό της καὶ στὴ νύφη τη;
τὴν ὥρα τὴς ἔἕ«Y ἠ; των καὶ τὴν κράτησε τὴν ὑπόσχεση, σὰ Λνκοδὴμαινα πού ἡταν κι αὐτή. Καὶ τώρα
ζοῦν οἷ δυό του; στὸ προγονικό τους ἀρχοντικό, ποὺ εἶναι δίπλα στὴ Σωτήρα, τὴν ἐκκλη- σιά τους. Ἡ
Τρισεύγενη ἡταν γιάτρισσα μὲ τὸνομα. Αίγα βότανα γνώριζε μὰ καὶ σὲ λίγα εἶχε ἐμπιστοσύνη. "Ἕνα κι
ἕνα: γιὰ τὰ μάτια, γιὰ τὸ στομάχι, γιὰ τὴν καρδιά, γιὰ τὴ χρυσῆ, γιὰ τὴν φούσκωση καὶ γιὰ τὸ
ξορκισμένο κι ἀμνημόνευτο: τὸ χτικιό. Καὶ νὰ δῆς ποὺ τὰ βοτάνια πήγαινε καὶ τἄβρισκε μοναχή τη;,
κ εἶχε ξεχωριατὰ κουτάκια γιὰ τὸ κάθε ἕνα ἀπὸ δαῦτα: γιὰ τοῦ δράκου τὸ κέρατο, τῆ; λάμιας τὸ
νύχι, τῶν ἐφτὰ ἀδελφῶν τὸ αἷμα, τὸ φεγγαρό- χορτο, τὴν τάτουλα, τὸ μαρμαρόχορτο, τὸ νυχτεροβό- :
τανο, τὺ χελιδονόχορτη, τὸ κορακάχοριο, ιὸν χαμο· λιό.... ὡ; καὶ μελιτήοα τοῦ θυμαριὸ? ὲἷχε. Μὰ
τὴν μαγγοῦτα — ἀγύρευιη νά ναί—πού μιά φορὰ μο- νάχα στὴ ζωή της τήν ὲδωσε, σ' ὲνα γέρο πὄύ βασα-
νιζότανε μιά βδομάδα γιὰ νὰ βγῆ ῆ ψυχὴ του — ιὸν εἶχε κρυφὰ καὶ παράμερα. Μποροῦσε σάν ἔπεφτε στὸ
χέρι ἀλλου νὰ γίντ) μεγάλο κακὸ μὲ δαύτην. Τὸ παληὺ καιρὸ ὅμως, ποὺ τὴν ἔδινε τὸ χέρι τῆς
πολιτείας, κακούργημα δὲν ἡταν: μὰ τότε; βλέπεὶς λεγότανε κώνειον, δὲν λεγότανε μαγγοῦτα. Στὸ
κελὶ τῆς ἐκκλησίἀς καθότανε ὀ παπᾶ — Χί- καντρος — καλόγερος ἡταν—καὶ σήμερα πρωί — πρωὶ θὰ
καβαλλήκευε τὴ μούλα κἀὶ θὰ πήγαινε σ ἕνα ξωκκλῆσι νὰ κάνη λειτουργὶά. Θὰ γλεντοῦσαν κι ὅλας κάτω
ἀπ τὸν πλάτανο καὶ δίπλα στή βρυσοῦλα τῆς ρεμματιᾶς καὶ νά, ποὺ σ ἕνα σακκοῦλι κρεμασμένο στὸ
σαμάρι τοῦ μουλαριοῦ βρίσκεται μέσα ἕνα ἀρ- νάκι ποὺ ξεπροβάλλει ἔξω μονάχα τῦ μουσοῦδι του. Αὐτὸ
ἡταν ζωντανὸ κι ό παπά — Νίκαντρος ποὺ μπαινόβγαινε κάνοντας τὴν ἑτοιμασία του, συλλογιζό- τανε
πῶς θὰ φάη κοκορέτσι ἀπὸ δαῦτο. Ὁ Στεφᾶνῆς τὴν ώρα ποὺ ή θειὰ Τρισεύγεντ) ἔψηνε τὸν καφέ της
ξεγλίστρησε καὶ βγῆκ ἔξω' καἶ βλέποντας τὸ μουλάρι ποὺ βασανιζότανε, γιατὶ εἶχαν μπλέξει τὰ πόδια
του στὸ σκοινί, τὸ λυπήθηκε καὶ πῆγε νὰ τὸ ξεμπλέξη κ ἐκεῖνο ὲκδικήθηκε τὸν αν–

θρωπο καὶ τοῦδωκε μιὰ κλωτσιὰ μὸλη τουτὴ δύναμη. Ὁ Στεφᾶνής ἔπεσε τάνάσκελα καὶ δὲν ξαναοη-
κιόθηκε πιά. V παπα — Ν'ίκαντρο; εἰδε πῶς χἴύπηαε τὸ παιδὶ μά εἴτε γιατὶ δὲν κατάλαβε πῶς ἡταν
κακὸ τὸ χτύ- πημα εἴτε γι ἀναποφύγη τὴς φωνὲς καὶ τὸ άλικόν- τισμα, καβαλλίκεψε χέρι χέρι κ ἔφυγε
βιαστικά. Μὲ ὀσα καὶ νὰ τοῦκαμε τοῦ Στεφᾶνῆ ἡ δυστυχι- σμένη Τρισεύγενη, δὲν σήκωνε πιά τὸ κεφάλι
του1 καὶ σὲ τρεῖς μέρες πέθανε. Γενικὴ ὀμιλία καὶ σούσουρο στὴν Ἀθήνα εἶναι ὁ χαμὸ; τοῦ Στεφᾶνῆ.
Ξεκλήρισε πιά ή γενιὰ τῶν Αυκοδὴμων ! "Ἂν πῆα καὶ γιὰ τὸν παπᾶ — Νίκαντρο, μὴν ἡξέ- ροντας τί νὰ
π ῆ καὶ τί νὰ κάνἶ], ἀπ τὴ γενικὴ κατα- κραυγὴ ἔφυγε κιαύτός, καὶ ξαναπῆγε «είς τὴν Μονὴν τῆς
μετᾶνοίας του». Μὰ κ' ἕνα ἄλλο πρᾶγμα λένε μυστικὰ καὶ μὲ ἀπορία τους στὴ χώρα. Γιατὶ καὶ τὴν ὣρα
ποὺ πέθανε κἀὶ τὴν ὣρα ποὺ τὸν θάψανε τὸ Στεφᾶνῆ καὶ τώρα ἀκόμη, πῶς ούτ ἕνα δάκρυ δὲν ἔσταξ ἀπ
τὰ μάτια τῆς Τρισεύγενης— Μονάχα μιὰ γρηὰ γειτόνισσα ποὺ γνώριζε τοὺς Λυκοδήμους χρόνια τῶν
χρόνων εἶπε ·. Κάποια ἀπό- φααι μεγάλη θὰλαβεν ἡ Τρισεύγενη· ποω ὲγώ τί σιδε- ροκέφαλη εἶναι ἡ
γενιὰ τῶν Λΰκων ν' ὲχω ἀκουστὰ κι ἀπ' τοὺς γερό ντους μου ποὺ τοὺς γνώριζαν ἀν- τάμ — μπαμπατάμ.
Εἶναι Σάββατο πρωὶ κ ή Ιυυρὰ -- Τρισεύγενη προσκαλεῖ τὸν Νοτάριο τὴ ς Πολιτείας. —Πάρε χαρτὶ καὶ
γράφε ἄρχοντα κυρ-Νοτάριε. —Μετὰ χαρᾶς σου Κυρὰ — Τρισεύγενη. Μεγάλη εἶναι ἡ συμφορά σας, ὰλλά
κι: ό Κόσμος... —νΑσ τ αὐτά, Ἐγώ τὰ ξέρω καλύτερά σου. Τὴ δουλειά μας τώρα. Τής μάννας μου τὸ σόϊ
κρατεῖ ἀπ τοὺς Καλογεράδες ποὺ τώρα βρίσκονται στὴ Μο- νοβασιά. Θὰ πάω κ ἐγώ ἐκεῖ νὰ ζήσω γιὰ
πάντα, κι ὰν δέ μπορέσω νὰ κάμω μακρυὰ ἀπ τὸ χῶμα τοῦ τόπου μας θὰ γυρίσω πάλι καὶ θὰ πάω νὰ
μείνω στὴ Σηργιανή,νὰ φροντίζω τὴν ἐκκλησία τοῦ Μοναατηριοῦ. Στὴν Ἀθήνα δὲ θὰ μείνω πειά" ἔτσι τ
```

```
Η ΚΥΡΑ ΤΡΙΣΕΥΓΕΝΗ

Παληότερη γενιὰ ἀπὸ τοὺς Λυκοδήμας δὲ γνώρισ' ἡ Ἀθήνα. Ἡ Σωτὴρα του Λυκοδήμου ἦταν ἡ ἐκκλησία τους.
Εἶχαν κ' ενα τροπάρι— ἄγραφο ἦταν —ποὺ μυστικὰ καὶ μουρμουριστά, καταλάβαιναν δὲν καταλάβαιναν, τὤλεγαν
καὶ τὸ παράδιν' ὁ γέρος ποὺ θἄφευγε ἀπ' τὸν κόσμο, στὸ νέο, ποὺ θὰ κρατοῦσε τὴ σειρὰ τῆς γενιᾶς.

Παράξενο ὡς τόσο τροπάρι κ ἐκεῖνο. Τὤλεγαν ἀνήμερα τοῦ Ἁγίου Δημητρίου, κι αὐτὸ ἦταν γιὰ Ἁγία γυναῖκα·
Μὰ πάλι τῆς Παναγιᾶς δὲν ἦταν.

Καὶ μιὰ φορά, τὴν ὥρα ποὺ τὸ ἀποτέλειων' ὁ γέρος καὶ τὸ ἀπόμαθεν ὁ ἔγγονος, τὸν ἀρώτησ' ὁ μικρός, ποὺ
ἦταν περίεργος: γιὰ ποιὸν Ἅγιο εἶναι τὸ τροπὰρι, ὁ γέρος ἀγρίεψε τότες καὶ τοῦ εἶπε σούτ ! κ ἔτσι
ἀγριεμένος πέθανε.

Στὴν Ἀθήνα τοὺς λέγανε Λύκους, μὰ καὶ τοὺς ἀρχαίους Λυκοδήμας Λύκους τοὺς λέγανε κ'ἐκείνους.

Ἔτσι ἦταν γραμμένος καὶ στὸ σκολειὸ ὁ Στεφανῆς, ὁ μόνος καὶ τελευταῖος ἀπὸ δαύτους : Στέφανος Λύκος.


Τὸ σόϊ — ἄτυχο σόϊ — σ' αὐτὸν καὶ μόνον τώρα κρέμεται. Οἱ γονεῖς του εἶχαν πεθάνει, μαζὶ κ' οἱ δυό, στὸ
τελευταῖο θανατικὸ ποὺ ρήμαξε τὸν τόπο· κι ὁ Στέφανος ἔμεινε στὰ χέρια τῆς θείας του τῆς Κυρὰ–
Τρισεύγενης– ἀδελφὴ τοῦ πατέρα του ἦταν — ποὺ δὲν παντρέφτηκε, γιὰ νὰ τὸν ἀναθρέψη. Αὐτὴ τὴν ὑπόσχεση
ἔδωσε «ἐνώπιον θεοῦ καὶ ἀνθρώπων» στὸν ἀδελφό της καὶ στὴ νύφη της τὴν ὥρα τὴς θανῆς των καὶ τὴν
κράτησε τὴν ὑπόσχεση, σὰ Λυκοδήμαινα ποὺ ἦταν κι αὐτή. Καὶ τώρα ζοῦν οἱ δυό τους στὸ προγονικό τους
ἀρχοντικό, ποὺ εἶναι δίπλα στὴ Σωτήρα, τὴν ἐκκλησιά τους. Ἡ Τρισεύγενη ἦταν γιάτρισσα μὲ τὄνομα. Λίγα
βότανα γνώριζε μὰ καὶ σὲ λίγα εἶχε ἐμπιστοσύνη. Ἕνα κι ἕνα: γιὰ τὰ μάτια, γιὰ τὸ στομάχι, γιὰ τὴν
καρδιά, γιὰ τὴ χρυσῆ, γιὰ τὴν φούσκωση καὶ γιὰ τὸ ξορκισμένο κι ἀμνημόνευτο: τὸ χτικιό. Καὶ νὰ δῆς ποὺ
τὰ βοτάνια πήγαινε καὶ τἄβρισκε μοναχή της, κ'εἶχε ξεχωριστὰ κουτάκια γιὰ τὸ κάθε ἕνα ἀπὸ δαῦτα: γιὰ
τοῦ δράκου τὸ κέρατο, τῆς λάμιας τὸ νύχι, τῶν ἐφτὰ ἀδελφῶν τὸ αἷμα, τὸ φεγγαρόχορτο, τὴν τάτουλα, τὸ
μαρμαρόχορτο, τὸ νυχτεροβότανο, τὸ χελιδονόχορτο, τὸ κορακάχορτο, τὸν χαμολιό.... ὡς καὶ μελιτήρα τοῦ
θυμαριοῦ εἶχε. Μὰ τὴν μαγγοῦτα — ἀγύρευτη νᾶναι—πού μιὰ φορὰ μονάχα στὴ ζωή της τὴν ἔδωσε, σ' ἕνα γέρο
ποὺ βασανιζότανε μιὰ βδομάδα γιὰ νὰ βγῆ ἡ ψυχὴ του — τὸν εἶχε κρυφὰ καὶ παράμερα. Μποροῦσε σὰν ἔπεφτε
στὸ χέρι ἄλλου νὰ γίνη μεγάλο κακὸ μὲ δαύτην. Τὸ παληὸ καιρὸ ὅμως, ποὺ τὴν ἔδινε τὸ χέρι τῆς πολιτείας,
κακούργημα δὲν ἦταν: μὰ τότες βλέπεὶς λεγότανε κώνειον, δὲν λεγότανε μαγγοῦτα. Στὸ κελὶ τῆς ἐκκλησιᾶς
καθότανε ὁ παπᾶ — Νίκαντρος — καλόγερος ἦταν —καὶ σήμερα πρωΐ — πρωῒ θὰ καβαλλήκευε τὴ μούλα καὶ θὰ
πήγαινε σ ἕνα ξωκκλῆσι νὰ κάνη λειτουργιά. Θὰ γλεντοῦσαν κι' ὅλας κάτω ἀπ' τὸν πλάτανο καὶ δίπλα στὴ
βρυσοῦλα τῆς ρεμματιᾶς καὶ νά, ποὺ σ'ἕνα σακκοῦλι κρεμασμένο στὸ σαμάρι τοῦ μουλαριοῦ βρίσκεται μέσα
ἕνα ἀρνάκι ποὺ ξεπροβάλλει ἔξω μονάχα τὸ μουσοῦδι του. Αὐτὸ ἦταν ζωντανὸ κι ὁ παπᾶ — Νίκαντρος ποὺ
μπαινόβγαινε κάνοντας τὴν ἑτοιμασία του, συλλογιζότανε πῶς θὰ φάη κοκορέτσι ἀπὸ δαῦτο. Ὁ Στεφανῆς τὴν
ὥρα ποὺ ἡ θειὰ Τρισεύγενη ἔψηνε τὸν καφέ της ξεγλίστρησε καὶ βγῆκ' ἔξω· καὶ βλέποντας τὸ μουλάρι ποὺ
βασανιζότανε, γιατὶ εἶχαν μπλέξει τὰ πόδια του στὸ σκοινί, τὸ λυπήθηκε καὶ πῆγε νὰ τὸ ξεμπλέξη κ'
ἐκεῖνο ἐκδικήθηκε τὸν ἄν–


θρωπο καὶ τοὖδωκε μιὰ κλωτσιὰ μὄλη του τὴ δύναμη. Ὁ Στεφανῆς ἔπεσε τἀνάσκελα καὶ δὲν ξανασηκώθηκε πιά.
Ὁ παπα — Νίκαντρος εἶδε πῶς χτύπησε τὸ παιδὶ μὰ εἴτε γιατὶ δὲν κατάλαβε πῶς ἦταν κακὸ τὸ χτύπημα εἴτε
γι' ἀναποφύγη τὴς φωνὲς καὶ τὸ ἀλικόντισμα, καβαλλίκεψε χέρι χέρι κ' ἔφυγε βιαστικά. Μὲ ὅσα καὶ νὰ
τοὖκαμε τοῦ Στεφανῆ ἡ δυστυχισμένη Τρισεύγενη, δὲν σήκωνε πιὰ τὸ κεφάλι του· καὶ σὲ τρεῖς μέρες πέθανε.
Γενικὴ ὁμιλία καὶ σούσουρο στὴν Ἀθήνα εἶναι ὁ χαμὸς τοῦ Στεφανῆ. Ξεκλήρισε πιὰ ἡ γενιὰ τῶν Λυκοδὴμων !
Ἂν πῆς καὶ γιὰ τὸν παπᾶ — Νίκαντρο, μὴν ἠξέροντας τί νὰ πῆ καὶ τί νὰ κάνη, ἀπ' τὴ γενικὴ κατακραυγὴ
ἔφυγε κι' αὐτός, καὶ ξαναπῆγε «εἰς τὴν Μονὴν τῆς μετᾶνοίας του». Μὰ κ' ἕνα ἄλλο πρᾶγμα λένε μυστικὰ καὶ
μὲ ἀπορία τους στὴ χώρα. Γιατὶ καὶ τὴν ὥρα ποὺ πέθανε κἀὶ τὴν ὥρα ποὺ τὸν θάψανε τὸ Στεφανῆ καὶ τώρα
ἀκόμη, πῶς οὔτ ἕνα δάκρυ δὲν ἔσταξ' ἀπ' τὰ μάτια τῆς Τρισεύγενης• Μονάχα μιὰ γρηὰ γειτόνισσα ποὺ
γνώριζε τοὺς Λυκοδήμους χρόνια τῶν χρόνων εἶπε: Κάποια ἀπόφασι μεγάλη θἄλαβεν ἡ Τρισεύγενη· ξέρω ἐγώ τί
σιδεροκέφαλη εἶναι ἡ γενιὰ τῶν Λύκων κ' ἔχω ἀκουστὰ κι ἀπ' τοὺς γερόντους μου ποὺ τοὺς γνώριζαν ἀν– τάμ
— μπαμπατάμ. Εἶναι Σάββατο πρωῒ κ' ἡ Κυρὰ — Τρισεύγενη προσκαλεῖ τὸν Νοτάριο τῆς Πολιτείας. —Πάρε χαρτὶ
καὶ γράφε ἄρχοντα κυρ–Νοτάριε. —Μετὰ χαρᾶς σου Κυρὰ — Τρισεύγενη. Μεγάλη εἶναι ἡ συμφορά σας, ἀλλὰ κι'
ὁ Κόσμος... —Ἄσ τ' αὐτά. Ἐγὼ τὰ ξέρω καλύτερά σου. Τὴ δουλειά μας τώρα. Τῆς μάννας μου τὸ σόϊ κρατεῖ
```

(b) OCR-ed (left) and corrected (right) images

Figure 3: Post-correction of the OCR result through comparison with the digitised original source.

their collaboration in providing access to digitised textual resources.

REFERENCES

[1] 2008. *Markovian Models for Sequential Data*. Springer London, London, 265–303. https://doi.org/10.1007/978-1-84800-007-0_10

[2] 2019 (accessed January 15, 2019). *ABBYY FineReader*. https://www.abbyy.com/en-us/finereader/

[3] Federico Boschetti, Matteo Romanello, Alison Babeu, David Bamman, and Gregory Crane. 2009. Improving OCR Accuracy for Classical Critical Editions. In *Research and Advanced Technology for Digital Libraries*, Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 156–167.

[4] Thomas M. Cover and Peter E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Trans. Information Theory* 13 (1967), 21–27.

[5] B. Gatos, A. L. Kesidis, and A. Papandreou. 2011. Adaptive Zoning Features for Character and Word Recognition. In *2011 International Conference on Document Analysis and Recognition*. 1160–1164. https://doi.org/10.1109/ICDAR.2011.234

[6] B Gatos, Georgios Louloudis, and Nikolaos Stamatopoulos. 2011. Greek Polytonic OCR Based on Efficient Character Class Number Reduction. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 1155 – 1159. https://doi.org/10.1109/ICDAR.2011.233

[7] V. Katsouros, V. Papavassiliou, F. Simistira, and B. Gatos. 2016. Recognition of Greek Polytonic on Historical Degraded Texts Using HMMs. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. 346–351. https://doi.org/10.1109/DAS.2016.60