DHQ: Digital Humanities Quarterly
Preview
2024
Volume 18 Number 1

Recognition and Analysis of the Proceedings of the Greek Parliament after WWII

Epameinondas-Konstantinos Barmpounis <konbarbou_at_aueb_dot_gr>, Athens University of Economics and Business https://orcid.org/0009-0002-2322-1175

John Pavlopoulos <annis_at_aueb_dot_gr>, Athens University of Economics and Business https://orcid.org/0000-0001-9188-7425

Panos Louridas <louridas_at_aueb_dot_gr>, Athens University of Economics and Business https://orcid.org/0000-0002-3971-4612

Dritsa Konstantina <dritsakon_at_aueb_dot_gr>, Athens University of Economics and Business https://orcid.org/0000-0003-3395-2182

Abstract

The first post-WWII years in Greece were devastating. After a brutal Nazi occupation, the Greek Civil War (1946–1949) erupted. It wrecked the economy and the country's infrastructure and altered politics and the social fabric for decades to come. A study of the issues discussed in the Greek Parliament during the tense and unstable first years of the conflict (1946-1947) could facilitate our understanding of the society at the time. An obstacle is that parliament proceedings are publicly available in a machine-readable form beginning in 1989; before that only scanned images of the original records exist. We show that text recognition followed by natural language processing can unlock this corpus for historical research. Using Transkribus, we trained a text recogniser (1.5% CER) that we applied to 3,156 images from 1946 and 1947. As low-quality recognition is inevitable, we trained a language model on the transcribed text and applied it to recognised text, discarding records with high average cross-entropy. Using information extraction techniques, we sampled speeches that were applauded and we introduced the first quantification of issues that were thus received. All our resources are made available at https://zenodo.org/record/8302990.

1 Introduction

Parliamentary proceedings are a valuable resource for the historian, facilitating the investigation of the political and social history of a country. The history of post-WWII Greece is full of political instability, ideological conflicts, and social discontinuity, with the proceedings of the Greek Parliament being one of the most valuable sources for the period. Their preservation in digital form is vital, opening access to more researchers and potentially allowing them to work more efficiently with a machine-readable form. In Greece, only the proceedings of the last three decades exist in a machine-readable form. Earlier proceedings are digitised only as scanned images of the original printed records. This hinders the computational exploration of the material and the application of Natural Language Processing (NLP) techniques. 1

In this study, we introduce a corpus comprising the proceedings from 1946–1947, the first years of the Greek Civil War of 1946–1949. The corpus was developed by recognising scanned files found online on the website of the parliament's library. We trained Transkribus on 16 double-column book pages, achieving a Character Error Rate (CER) of 1.5%. We then used the trained model to recognise the text from 1,589 images of 3,178 pages (each image contains two pages). That is a large-scale, real-world recognition experiment that contains noisy text, for example from pages that include tables or distorted (bent) pages. Borrowing from authorship analysis [Pavlopoulos and Konstantinidou 2023], to discard low-quality recognised text, we trained a language model on the transcriptions and applied it to the recognised text, 2 discarding 144 recognised texts with considerably high Bits Per Character (BPC) [Graves 2013]. The resulting corpus is publicly released.

Our Contribution:

- A method for recognising and exploiting text from a challenging corpus. The method is based on text recognition, language modeling-based filtering, and NLP-assisted quality assurance and information extraction, shown in Figure 1 and explained in the rest of this work. The method allowed us to focus on a sample of high-quality recognised pages, which we could study qualitatively, something that would have been infeasible otherwise. 3
- An application of our method in a real-world corpus:
  - We trained a Transkribus text recognition model on Greek parliamentary proceedings from 1946–1947, which we release to the public along with our training resources.
  - We developed and presented the first corpus of Greek parliamentary proceedings from 1946–1947, the first years of the Greek civil war, comprising machine-readable texts.
  - We used NLP to assess the quality of the recognition and to filter out low-quality texts, a method which can be applied to other domains and datasets.

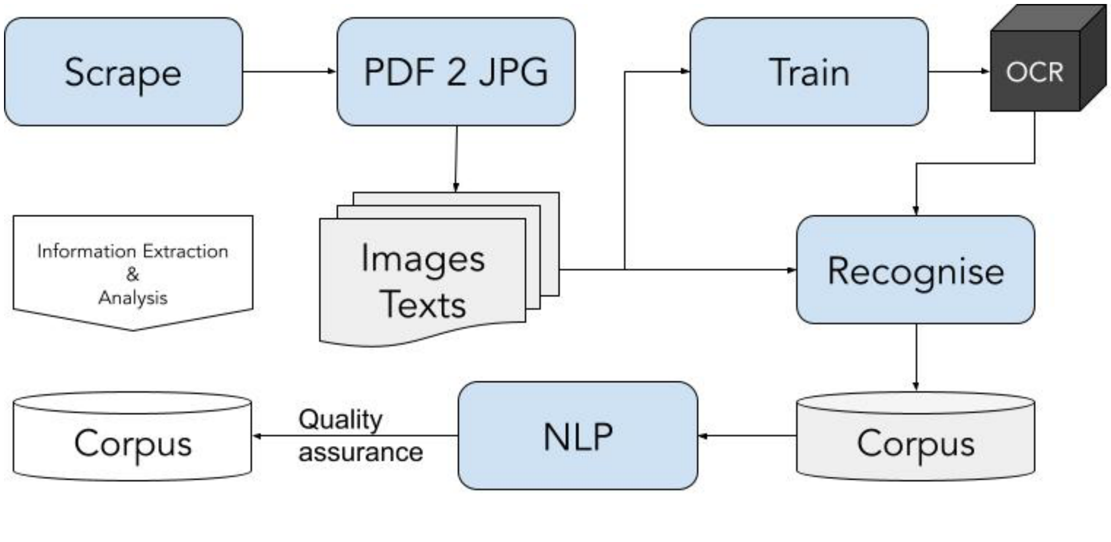

Figure 1. Diagram of workflow for the creation of the corpus.

Our Findings:

4 The proposed method unlocks research capabilities for the historian that were not available before. We used NLP to extract useful information, including dates and speaker names. Furthermore, we extracted indicators of applause, revealing a positive reception of the issue being discussed and agreement with the speaker. We compiled a set of 800 texts with an applause indicator and the corresponding speaker's name. From this set, we sampled 50 texts, which were qualitatively studied further by a historian (part of our team). The historical analysis revealed an estimation of the distribution of the topics being discussed at the time, with the most frequent one (23.56%) being anticommunist speeches, followed by comments on government policies (16.67%) and issues of territorial integrity (7.78%).

5 Although statistical analysis is not the main tool of modern approaches to historical research, its results will always provide a solid foundation for the acceptance or rejection of existing conclusions or advancing new ones. For instance, our method can help historians approach a large corpus of historical evidence to review or construct a narrative, in this case the dominant narrative of Greece's political history after WWII. Also, although broad estimates, our findings provide the first quantification of the issues supported by the political system of that era in Greece.

6 In what follows, we outline the related work and the historical context of our corpus in Section 2. The challenges are

presented in Section 3. We then proceed to the method we devised regarding text recognition in Section 4, and regarding information extraction in Section 5. We present our historical analysis of the extracted material in Section 6 and round up with conclusions in Section 7.

2 Related Work and Context

The analysis of parliamentary proceedings as historical evidence with the help of NLP is not novel. Related work has drawn conclusions using text analysis [Walter 2021] and has provided tools to serve the work of researchers and historians [Puren et al. 2022]. The development of language technology helps historians who are faced with an increasing amount of information. For instance, in a study of the Slovenian parliamentary data, Named Entity Recognition (NER) was suggested as a tool to extract information automatically from large collections [Pančur and Šorn 2016]. The authors, however, noted that precision for the recognition of persons was not perfect (85%), calling for awareness of reliability. In this work, we suggest that NLP can be used to “screen” parliamentary data, allowing the historian to search and focus on targeted records. Other work attempts to access and analyze the potential of Parliamentary Markup Language (PML) for enabling contextualization of linguistic data in parliamentary proceedings and facilitating large-scale studies in this area [Gartner 2018]. 7

The computational study of the Greek Parliament proceedings dates to 2008 when document classification was applied to detect the topics being discussed [Nikolaos and George 2008]. Such studies, however, use records after 1989 because only these exist in a machine-readable form on the website of the Greek Parliament's library. Even in machine-readable data, extensive preprocessing, followed by NLP, is required to create a structured dataset [Konstantina et al. 2022], which can then be readily used for political analysis without burdening the researcher with the considerable effort required for text processing [Gkoumas et al. 2018]. 8

The development of a machine-readable corpus of the proceedings of the Greek Parliament for the years before 1989 has not been attempted, to the best of our knowledge. The potential of unlocking this material for computational studies, however, is great not only for historians but also for political and social scientists. The period following WWII is exceptional, starting with the Greek Civil War that broke out after the end of the Nazi occupation, and then encompassing decades in which the state evolved and gradually created a very particular identity. The main political characteristic of the state that emerged in the post-WWII period was its “sickly” democracy [Νικολακόπουλος [Nikolakopoulos] 2009]. 9

One feature of this sickly democracy was its anticommunist nature. Anticommunism was espoused not only by politicians from the right (and influenced their political decisions) [Χατζηβασιλείου [Chatzivasiliou] 2011], but was also embraced by a wider part of the political spectrum and of society at large. It can be traced to the persecution of communists and suspected fellow travellers whose acts were designated “crimes against the Nation” [Μπουρνάζος [Bournazos] 2009]. The communist party was illegal and suspected members or supporters were sent to prison or exile. The blending of communism into antinationalist ideologies was part of the jingoistic climate that also characterised the period. This leads to a second feature of the Greek polity of the time, irredentism regarding regions near the borders of Greece, like Northern Epirus and Cyprus [Stefanidis 2016]. 10

The exploration of these two issues, anticommunism and irredentism, cannot be complete without the investigation of the parliamentary proceedings of that period. The proceedings provide critical information about the expressed views of the parties on the topics being discussed. The attentive studying of this evidence can reveal both quantitatively and qualitatively, based on actual data, the ways that anticommunist and irredentist discourse gradually evolved and became a feature of the state's identity. 11

3 Data Challenges

A challenge in the corpus of the Greek parliamentary proceedings is the language itself. In Greece, until 1976 there were two versions of the Greek language: Katharevousa and Demotic. Katharevousa was an artificial form of Greek, created in the 18th century by combining features of Ancient Greek and the modern vernacular, the idea being to have a linguistic vehicle free from impurities that had accumulated in Greek over the centuries (Katharevousa means 12 “purifying”). It was used in formal occasions and official documents, and less frequently in literature. Demotic was the vernacular, spoken by the people in everyday life, and the primary means of artistic expression. In the texts of the parliamentary proceedings those two versions coexist. Not only do the Members of Parliament (MPs) speak both Katharevousa and Demotic, but also in many cases a text (of a bill, or a letter, which can be written in either of them) is read out aloud in the house. Apart from Greek, in its two guises, in several cases, pages in French or English can be found in the corpus.

13 Another challenge concerns the script. The text is written in Greek polytonic, which was the way of writing Greek until 1982. This writing system includes a variety of diacritics (accents, aspirants, and subscripts) and thus is more similar to late antiquity's Ancient or Byzantine Greek than Modern Greek writing. It was the official Greek spelling of the era and, consequently, the spelling used in the parliamentary proceedings.

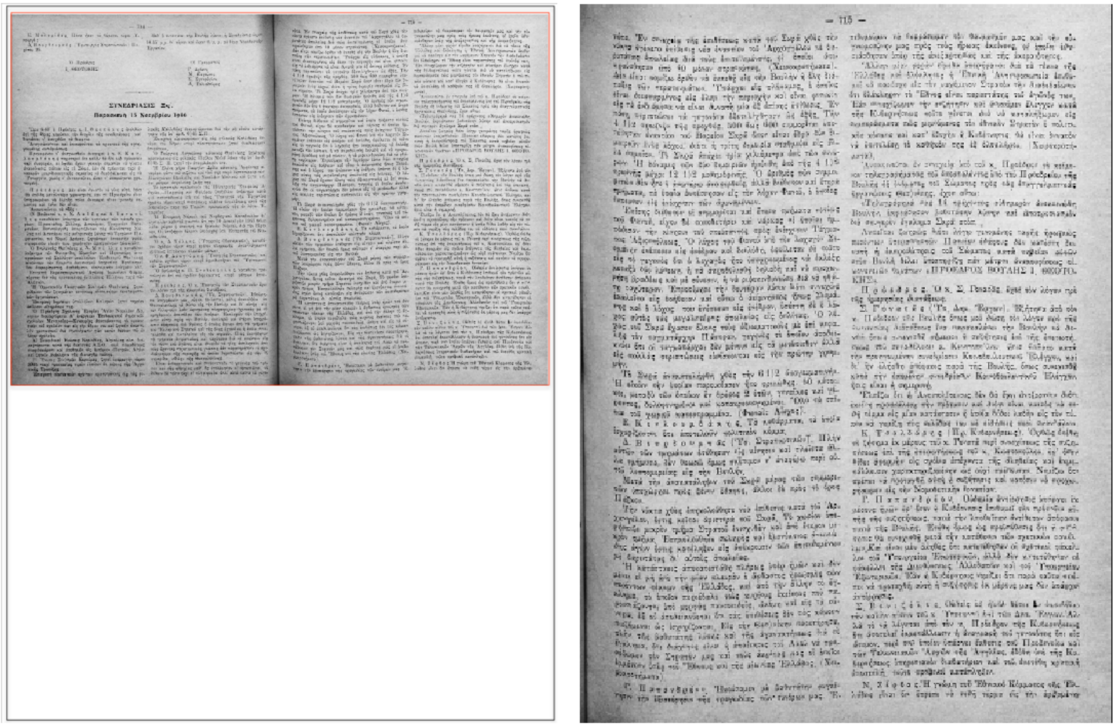

Figure 2. On the left (a) is a scanned image. On the right (b) is the right curved 2-column page. Images often comprised white space (a; black plate) that was cropped and led to the two double-column pages (a; red plate). The pages resulting after splitting may include curved areas (b; left column).

4 Text Recognition

14 We downloaded 1,589 images (3,178 book pages) of parliamentary proceedings from the website of the Greek Parliament's library.[1] The images are in PDF format. Each PDF file contains a scanned picture of two book pages (recto and verso, see Figure 2). Each scanned image is an A4 size page scanned vertically; the image is in JPEG format, stored inside the PDF document. Images came in three resolutions: 850 x 1170 (2038 images), 1213 x 1750 (570 images), 1288 x 1750 (570 images). For each image, we automatically cropped the part of the image that included text, such as the upper half of Figure 2(a). No other manual processing methods (e.g., de-skewing or contrast adjusting) were employed. Images were then split into two parts so that each instance in our dataset is a two-column page in JPEG format, as is shown in Figure 2(b). [2]

15 To recognise the text, we experimented with eScriptorium [Kiessling et al. 2019] and Transkribus [Kahle et al. 2017], and selected the latter as we found it superior in preliminary experiments. Specifically, when trained on the same six pages, Transkribus achieved a CER of 9% (Table 1, first row) while the reported error rate of eScriptorium was 26.9%.[3] A manual evaluation of the recognised text by the two platforms verified the superiority of Transkribus, which was better at capturing the two-column layout and was overall more accurate in recognition. Transkribus is the most commonly

used text recognition tool in the cultural heritage space [Nockels et al. 2022] and provides free credits until a paid subscription is required. Its architecture, as far as we know, has not been published, but we could infer from information on the tool's website that it uses a neural network with three convolutional layers (combined with two max pooling layers), followed by three stacked bidirectional Long Short-Term Memory (LSTM) layers and one convolutional layer on top to produce the output.[4] In our experiments, using text transcribed from 16 pages, we achieved a CER of 1.5% (Table 1, third row). The addition of more training pages (up to 12 more, 4th row in Table 1), brought no improvement, which might be related to the curvature of the added pages. Therefore, we used the well-performing model trained on the 16 pages to recognise the text from 3,178 pages.[5]

Pages #	Training CER (%)	Validation CER (%)
6	1.3	9.0
8	1.9	2.3
16	1.8	1.5
28	2.4	1.5

Table 1. The CER achieved on the training and on the validation dataset by using a different number of pages to train the recognition model.

Quality Assessment

We found that 24 out of 3,178 recognised texts were empty and were therefore discarded. Some of the remaining 3,154 recognised texts, however, will likely comprise noisy output, due to images of bent pages, as in the left column of Figure 2(b), or due to content that is hard to transcribe (e.g., pages with tables or with text in a foreign language). In these cases, the recognised text consists of scrambled meaningless text, characters one next to the other as in "Ἐάξμξζ,λ,βθλΗξζλΕζ (ΕΥ ΙΚΚῳ ν8Υ ψπυ". This noisy text, however, hinders search and information extraction for the historian. Therefore, in order to discard low-quality texts from our corpus, we built a language model to assess the quality of each of our recognised texts. Language models are suitable for this task, as they can assess the quality of the recognised text and can be used for the evaluation of the output of text recognition [Ströbel et al. 2022]. 16

We used a statistical language model of character trigrams, following work on authorship analysis [Pavlopoulos and Konstantinidou 2023], using text from our transcriptions to train the model. The transcribed text comprises 1,707 lines that add up to 81,965 character tokens and 135 unique characters. Then, we used the language model M that was trained on transcribed text to compute the average cross-entropy of the N starting characters from each of our recognised texts, where $N = 500$. Also known as Bits Per Character (BPC), this is the average negative base two log probability that M will generate character c_{t+1} if the preceding characters were $c_1 ... c_t$. More formally, BPC can be defined as: 17

$$\mathrm{BPC} = -\frac{1}{N}\sum_{t=1}^{N}\log_2 P(c_{t+1}; M; c_1 ... c_t)$$

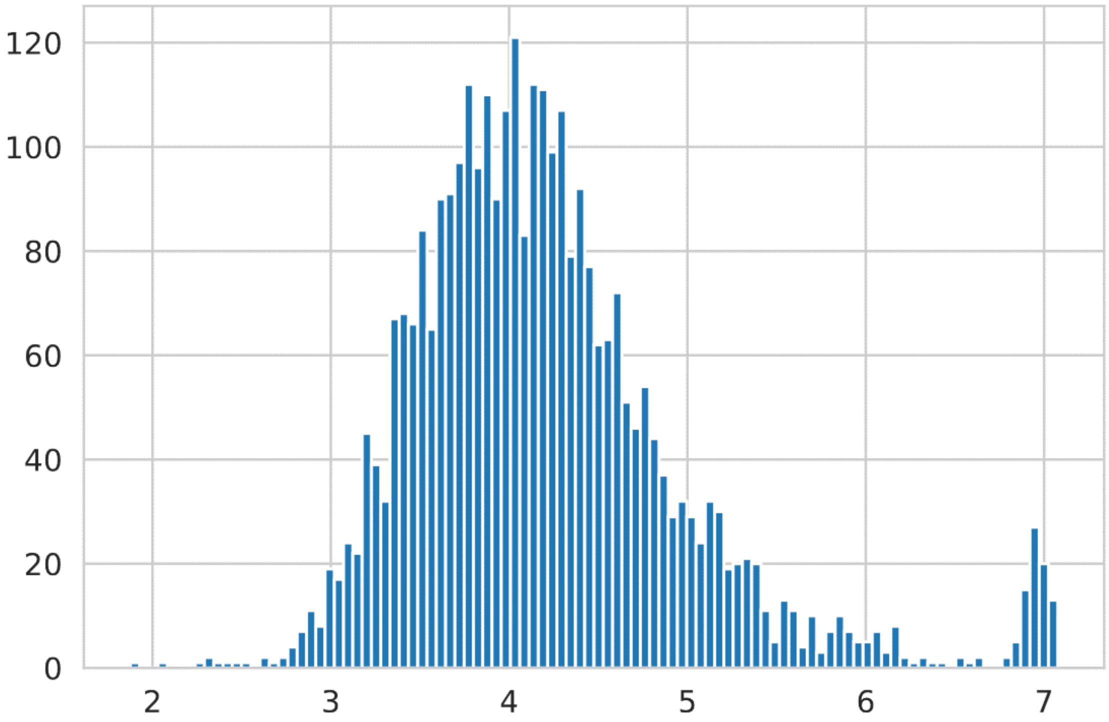

Figure 3. Histogram of BPC scores for the beginnings of the recognised texts.

We used BPC as a quality indicator, with lower values indicating recognition of high quality. As can be seen in Figure 3, the score is normally distributed around 4.2 (st.d. 0.78) with outliers on the right side. These are mainly texts with many recognition errors, which occur across our corpus (Figure 4) and are meaningless to the rest of our study. Hence, we discarded any text with a BPC that was greater than two units of the standard deviation above the mean, resulting in 3,010 texts that we used for information extraction and analysis. To the best of our knowledge, language modelling has not been used before in published studies to filter out low-quality recognition, but our study shows that this could be a promising direction. 18

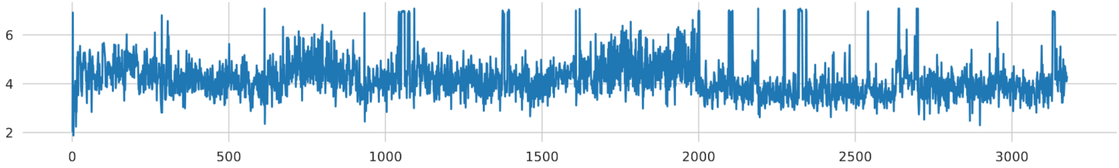

Figure 4. BPC achieved per recognised text. The text's index is shown horizontally and lower scores indicate better quality.

5 Information Extraction

To extract historical information from the corpus resulting from the recognition, we should be able to detect occurrences of dates and names. This is a Named Entity Recognition (NER) problem that is often addressed with supervised learning. Existing solutions, however, are not trained on noisy text and do not perform well. For instance, spaCy[6] found no dates, while names were returned with many false positives. Therefore, we opted for recognising names and dates using regular expressions. 19

An exploration of the proceedings revealed patterns capturing named entities of speakers, applause, and dates. For instance, when a speaker started speaking, their name was printed with spaces between the characters, to indicate the beginning of a new speech and to highlight the name. Sittings of the parliament were identified by an enumeration, followed by the date. Furthermore, an annotation (“χειροκροτήματα” / “applause”, “παρατεταμένα χειροκροτήματα” / 20 “extended applause”) reflected the listeners' approval, enthusiasm, or positive feedback on a subject or opinion. This annotation can be found at the end of every sentence that was applauded. We encoded these observations into the regular expressions presented in Table 2, which we applied to each text of our corpus.

Type	Named Entity	Regular Expression
APPLAUSE	χἰροκροτήματα	.(∗(κροτ)*).
SPEAKER	Θ ε ο τ ό κ η	[\s]*([Α-ΩἈἉἊἘἙἚἸἹἺὈὉὊ᾿Υ῾Υ῝ΥὨὩὪ]\s([α-ωάᾶὰἇέὲύϋῦὺὗὒίΐῖῗόὸώῶὼὧὢ]{1,3}\s){1,8})[\s]*
DATE	Πέμπτη 16 Μαΐου 1946	([Α-ΩΆΈΎΊΌΏἈἘἸὈὨἉἙὙἹὉὩἊἚἺὊὪα-ωάέύίήόἰὐἀἐὀἠῖῦᾶῆ]+\s+[0-9]+\s+[Α-ΩΆΈΎΊΌΏἈἘἸὈὨἙὙἹὉὩἊἚἺὊὪα-ωάέύίήώόἰὐἀἐὠὀῖῦᾶῆϊϋΐΰ]+\s+[0-9]+*)

Table 2. Regular expressions capturing named entities of applause, speakers, and dates. Example entities captured by the respective expression presented in the middle.

To verify that the found named entities are valid, we performed a manual investigation of the results. During this investigation, we removed 52 date entities and 36 applause entities that were mistakenly detected. These mistakes were mainly due to OCR mistakes, leading to words that were falsely recognized. For example, “τοῦ ἀπὸ 13 Μψρόυ 132ῆ'” matches our regular expression for the date but it is actually a noisy string (see Subsection: Quality Assessment above). Another frequent mistake type, regarding dates, stems from law codes. Laws are often written in the form of a year followed by a left parenthesis and then a number, yielding terms that can be matched by our date regular expression, especially if OCR mistakes alter the text further. Mistaken detection of applause concerned mainly words comprising “κροτ”. On the other hand, we expect to have missed named entities, for example in cases of spelling errors, such as in the following text (Page ID: main22_2): “Δν πρέπει νὰ λγηντειαύτά είς παραμονὰς ἒκλογον καὶ ἀπο εκενους ες τοῦς ὁποίους εἶχεν ἀνατεθῇ ἡ δακυβίρνησῃ τῆς ὑρας (χεροκροήματα)”. The translation is: “Those words shouldn't be said in the eve of elections and from the those tasked with governing the country (applause)”. The text in Greek has many spelling errors, including “χειροκροτήματα”, which has been recognised as “χεροκροήματα”. With “τ” missing from the word, regular expressions would not capture it. 21

6 Historical Analysis

The extracted information can unlock historical analysis of the political status quo of Greece in the studied period. The topics that were more often applauded, as these were identified by manual inspection by a historian that was part of the team, revolved around anticommunist speech and comments (Figure 5). These mainly refer to the left-wing organisations of the National Liberation Front (E.A.M.), a diverse alliance that fought to liberate Greece from the Axis Occupation, and its military branch, the Greek People's Liberation Army (E.L.A.S.). It had been created by the Communist Party of Greece (K.K.E.) and at the time they were fighting against the official Greek state in the Civil War. During this war, as well as for the following decades, representatives of the communist party were not present in the parliament.[7] This situation not only facilitated anticommunist discourse but also embraced the idea that communism was something hostile to Greece that should be fought. For instance, the MP T. Tourkovasilis was applauded after stating that “The enemies of Greece should know, that our Motherland was the grave of Hitlerism, it was the grave of Mussolinism, our motherland will also be the grave of Communism without doubt”.[8] 22

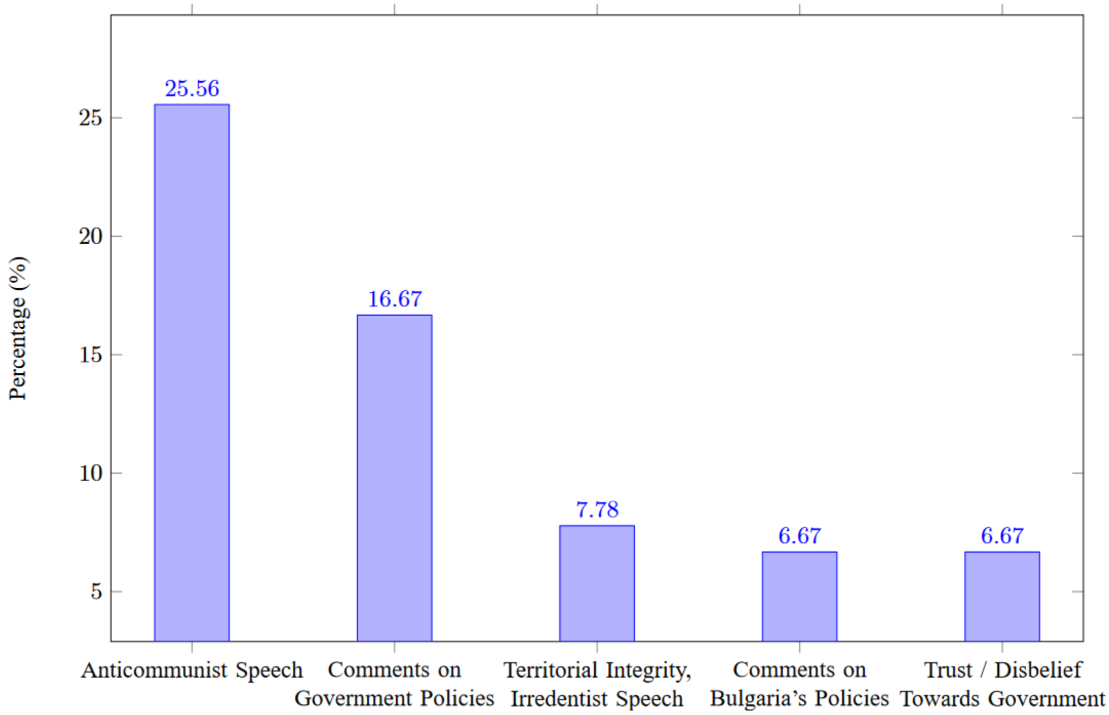

Figure 5. Five most applauded topics in a sample of 50 pages.

The Greek state at that period had a democratic government, but a flimsy one. Mainly due to the Civil War, the state excluded part of its population from political life (i.e., the ones who were believed to be communists were persecuted by the state) and the country was characterised by a climate of White and Red Terror,[9] which inflamed the political and social status quo [Hatzis 2019]. When a referendum was held in September 1946 for the return of king George II, political and social divisions peaked. K.K.E. abstained from the referendum and the royal right won. The return of the king symbolised the ideological victory of the latter [Gallant 2017], which can be supported by parliamentary speeches concerning this topic. For example, N. Zervas[10] received extended applause after stating that the return of George II with the referendum would show that the Greek people were against the communists.[11]

Date	Speaker	Topic
Σάββατο 18 Μαΐου 1946 (Saturday 18 May 1946)	_	Applause after a letter from the Association of People of Northern Epirus was read. In the letter, the Association asked the parliament to fight for the union of Northern Epirus with the rest of Greece.
Τρίτη 5 Νοεμβρίου 1946 (Tuesday 5 November 1946)	Δ. Σαρρής (D. Sarris)	Sarris suggested that a government department should be installed in the Macedonia region inorder to enforce the law.
Πέμπτη 15 Οκτωβρίου 1947 (Thursday 15 October 1947)	Π. Ροζάκης (P. Rozakis)	Rozakis stated that Greeks would always claim the revival of Eastern Rumelia, a region belonging to Bulgaria.
Τρίτη 18 Νοεμβρίου 1946 (Tuesday 18 November 1946)	Κ. Τσαλδάρης (K. Tsaldaris)	Tsaldaris tried to clear up a misunderstanding between himself and a member of the opposition, P. Kanellopoulos.
Σάββατο 18 Μαΐου 1946 (Saturday 18 May 1946)	Γ. Παπανδρέου (G. Papandreou)	Papandreou stated that the Greek nation is furious with Bulgaria's claims on Western Thrace.
Πέμπτη 23 Μαΐου 1946 (Thursday 23 May 1946)	Η. Μπάτζιος (I. Mpatzios)	Mpatzios proposed that the Greeks should ask for the Great Powers' compassion for their Greek brothers in Northern Epirus.
Δευτέρα 19 Μαίου 1947 (Monday 19 May 1947)	Α. Δημητράτος (A. Dimitratos)	Dimitratos suggested that there is a need for an organized attempt of all the parties of the parliament in order to improve the life and the purchasing power of the Greek people, mainly the working class.
Παρασκευή 17 Μαΐου 1946 (Friday 18 May 1946)	Κ. Τσαλδάρης (K. Tsaldaris)	Tsaldaris stated that the communists would be fought by the state if they acted like terrorists, meaning their plan of a revolution in Greece.
Παρασκευή 28 Ιουνίου 1946 (Friday 28 June 1946)	Π. Κανελλόπουλος (P. Kanellopoulos)	Kanellopoulos stated that the Greeks could claim the revival of Eastern Rumelia in the peace treaty that they were about to sign.
Πέμπτη 27 Ιουνίου 1946 (Thursday 27 June 1946)	Δ. Μαντούβαλος (D. Mantouvalos)	Mantouvalos falsely predicted that the region of Northern Epirus would unite with Greece.
Παρασκευή 13 Ιουνίου 1947 (Friday 13 June 1947)	Η. Λαγάκος (I. Lagakos)	Lagakos stated that the parliament should show interest and “affection” to the working class and the workers' unions since the Communist Party of Greece (K.K.E.) had tried to influence them yet they remained on the state's side.
Πέμπτη 20 Ιουνίου 1946 (Thursday 20 June 1946)	Ν. Ζέρβας (N. Zervas)	Zervas stated that the referendum for the return of King George II would show to the communists that the people of Greece are against them (if the royalists won).

Table 3. A sample of the topics that were applauded, extracted by a historian that was part of our team, along with information about the speaker and the date.

The above-mentioned referendum was a powerful ideological tool for the Greek government. Its outcome would not only be useful for resolving the local political situation, but also for the image of the country abroad. During the Civil War, Greece participated in the Paris Peace Conference (1946). The king's return showed that Greece had resolved internal issues and was, therefore, ready to lay claims for reparations and for land that Greeks thought rightfully theirs. Greece's claims, in terms of its borders, mainly referred to four regions: Northern Epirus, Macedonia, Western Thrace, and Cyprus. These four regions, which contained Greek populations, were either part of another country (e.g., Northern Epirus, Cyprus, Macedonia) or were occupied by Bulgaria (i.e., Western Thrace; Bulgaria was allied to Germany) during the war, after the defeat of Greece.

At the Paris Peace Conference, the Greek delegation claimed that the borders of Greece should expand to include Northern Epirus, which had a Greek population and had been liberated by the Greek army from the Italian invading army during the Greek-Italian war that brought Greece in WWII [Χριστίδης [Christidis] 2007].[12] Such claims, then, were deeply integrated into the irredentist speech of Greek politicians. I. Mpatzios's emotional speech was extensively applauded,[13] after stating that even the stones of Northern Epirus are willing to rise and say “we are and we will be Greek”. A few pages later,[14] the same MP was once more applauded after suggesting that they should ask for the Great Powers' compassion for the brothers in Northern Epirus. The speakers were even more passionate after the islands of the Dodecanese were united with Greece.[15]D. Mantouvalos was applauded after (falsely) predicting that since these islands got what they deserved and what they were fighting for (i.e., their unification with Greece), the time would soon come for Northern Epirus that was full of bones and blood of Greek heroes to do the same.[16] 25

Greece also had territorial disputes with Bulgaria. The two states had a history of war based on border issues (Second Balkan War of 1913, Greek-Bulgarian War of 1925). This time, it was the region of Thrace that was the bone of contention. During the Paris Peace Conference, Bulgaria, supported by the USSR, claimed Western Thrace mainly based on ethnological and historical arguments, but also for economic reasons [Χριστίδης [Christidis] 2007]. This infuriated Greece. G. Papandreou was extensively applauded after stating, “Since the shameless leaders of Bulgaria [...] claim Western Thrace, the whole [Greek] Nation revolts and demands from the Allies to bring justice.”[17] In the irredentist (and nationalistic) crescendo of his speech, P. Kanellopoulos was applauded after stating that the current claims of Greece (Western Thrace) are just a small part of what it could claim: the revival of Eastern Rumelia.[18][19] This idea continued to be relevant even when the discussion in the parliament was about the validation of the Peace Treaty with Bulgaria. P. Rozakis gave a speech in order to propose the validation of the Treaty, but he was applauded only after he said that they should take an oath never to resign from their claims in Eastern Rumelia.[20] 26

7 Conclusions

This study proposed a method for recognising and exploiting text from a challenging corpus that is based on text recognition, language modeling-based filtering, and information extraction. We then applied this method to the proceedings of the Greek Parliament after World War II. We used Transkribus to train a text recognition model, which we used to recognise the text from 3,178 scanned double-column pages. We used language modeling to discard records with low-quality recognition and information extraction annotate the speakers' names, dates, and indicators of applause. Using the latter, we used a sample of 50 pages to analyse the topics that were positively received by the house. 27

Since the period of interest was marked by the Greek Civil War and a referendum regarding the nature of the political system (i.e., return of the exiled king versus a republic), this information can be of great importance for any historian who wants to put into test the dominant historical narrative for these politically complex years. Our analysis led to a quantitative estimation of the topics applauded, which has not been presented before, to the best of the authors' knowledge. The most applauded topic was anti-communist comments, followed by comments on the government's policies, territorial integrity and comments on Bulgaria's policies. Without the application of the proposed this would have required an exhaustive manual investigation of thousands of images. 28

Notes

[1] See https://library.parliament.gr/.

[2] All our code and data can be found at https://zenodo.org/record/8302990.

[3] We used the evaluation results returned by the platforms and only the accuracy is reported by eScriptorium, which in this case was 73.1%.

[4] See https://readcoop.eu/wp-content/uploads/2018/11/LEIFERT-CITLAB.pdf.

[5] For the layout analysis the CITlab Advanced Method was used with the “Add estimated word coordinates” and “Do polygon simplification” features added. No baseline model was created, as this specific feature was released during our study. No other manual processing methods, such as deskewing or contrast adjusting, were used.

[6] See https://spacy.io/.

[7] There were no representatives of the official Communist Party (being illegal) until 1974. Nevertheless, K.K.E. had created the United Democratic Left (E.D.A.), which was its legal representative in the parliament. In the elections of 1958 E.D.A. became the leading opposition party.

[8] Page ID: main1367_2.

[9] The “White Terror” and “Red Terror” were political purges that took place in Greece after WWII. The “White Terror” targeted left-wing groups and the “Red Terror” targeted right-wing groups, leading to a brutal period of violence and repression.

[10] N. Zervas was an important figure for Greece at that period. Before starting his political career, he was the general of the National Republican Greek League (E.D.E.S.), which was the largest right-wing resistance organisation during the Nazi occupation of Greece.

[11] Page ID: main139_2.

[12] This claim was later withdrawn.

[13] Page ID: main5_2.

[14] Page ID: main41_1.

[15] The Dodecanese islands were under Ottoman rule until 1912, when they were occupied by Italy. Italy retained control until the end of WWII, when the islands became a British protectorate and then officially united with Greece in 1947.

[16] Page ID: main180_1.

[17] Page ID: main14_2.

[18] Eastern Rumelia was a semi-autonomous province of the Ottoman Empire, with many Greek inhabitants. It was located in the region of today's Southern Bulgaria, and it covered the area among the Balkan Mountains, the Rhodope Mountain and Northern Thrace.

[19] Page ID: main20_1.

[20] Page ID: main1245_2.

Works Cited

Gallant 2017 Gallant, T.W. (2017) *Neotere Ellada*. Athens, Greece: Pedio Books.

Gartner 2018 Gartner, R. (2018) “Using structured text corpora in Parliamentary Metadata Language for the analysis of legislative proceedings”. *DHQ: Digital Humanities Quarterly*, 12(2).

Gkoumas et al. 2018 Gkoumas, D. et al. (2018) “Exploring the political agenda of the Greek Parliament plenary sessions”, *Proceedings of the 11th annual language resources and evaluation conference, LREC, 2018* Miyazaki, Japan, 7-12 May. Available at: http://lrec-conf.org/workshops/lrec2018/W2/pdf/8_W2.pdf.

Graves 2013 Graves, A. (2013) “Generating sequences with recurrent neural networks”, *arXiv*, 1308.0850. https://arxiv.org/abs/1308.0850.

Hatzis 2019 Hatzis, A.N. (2019) “A political history of modern Greece, 1821–2018”, in Marciano, A. and Battista Ramello, G. (eds), *Encyclopedia of law and economics.* New York: Springer, pp. 838-845. Available at: https://link.springer.com/referenceworkentry/10.1007/978-1-4614-7883-6_53-1.

Kahle et al. 2017 Kahle, P. et al. (2017) “Transkribus: A service platform for transcription, recognition and retrieval of historical documents”, *Proceedings of the 14th annual IAPR international conference on document analysis and recognition, IEEE, 2017.* Kyoto, Japan, 9-12 November. pp. 19-24. Available at: https://ieeexplore.ieee.org/document/8270253.

Kiessling et al. 2019 Kiessling, B. (2019) “eScriptorium: An open source platform for historical document analysis”, *Proceedings of the 16th annual IAPR international conference on document analysis and recognition workshops, IEEE, 2019*. Sydney, Australia, 22-25 September. pp. 19. Available at: https://ieeexplore.ieee.org/document/8893029.

Konstantina et al. 2022 Dritsa, K. et al. (2022) “A Greek parliament proceedings dataset for computational linguistics and political analysis”, *Proceedings of the 36th conference on neural information processing systems, NeurIPS, 2022*. New Orleans, LA, USA, 28 November–9 December. pp. 28874-28888. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/b96ce67b2f2d45e4ab315e13a6b5b9c5-Paper-Datasets_and_Benchmarks.pdf.

Nikolaos and George 2008 Nikolaos, T. and George, T. (2008) “Document classification system based on HMM word map”, *Proceedings of the 5th annual international conference on soft computing as transdisciplinary science and technology, ACM, 2008*. Cergy-Pontoise, France, 28-31 October. pp. 7-12. Available at: https://dl.acm.org/doi/abs/10.1145/1456223.1456229.

Nockels et al. 2022 Nockels, J. et al. (2022) “Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research”, *Archival Science*, 22(3), pp. 367–392.

Pančur and Šorn 2016 Pančur, A. and Šorn, M. (2016) “Smart big data: Use of Slovenian parliamentary papers in digital history”, *Contributions to Contemporary History*, 56(3), pp. 130–146.

Pavlopoulos and Konstantinidou 2023 Pavlopoulos, J. and Konstantinidou, M. (2023) “Computational authorship analysis of the homeric poems”, *International Journal of Digital Humanities*, 5(1), pp. 45–64.

Puren et al. 2022 Puren, M. et al. (2022) “Between history and natural language processing: Study, enrichment and online publication of French parliamentary debates of the early third republic (1881–1899)”, *Proceedings of the 13th annual language resources and evaluation conference, ACL, 2022*. Marseille, France, 20-25 June. Available at: https://aclanthology.org/2022.parlaclarin-1.3/.

Stefanidis 2016 Stefanidis, I. (2016) *Stirring the Greek nation: Political culture, irredentism and anti-Americanism in post-war Greece, 1945–1967*. New York: Routledge.

Ströbel et al. 2022 Ströbel, P.B. et al. (2022) “Evaluation of HTR models without ground truth material”, *arXiv*, 2201.06170. Available at: https://arxiv.org/abs/2201.06170

Walter 2021 Walter, T. et al. (2021) “Diachronic analysis of German parliamentary proceedings: Ideological shifts through the lens of political biases”, *Proceedings of the ACM/IEEE joint conference on digital libraries, IEEE, 2021*. Online, 27-30 September. pp. 51-60. Available at: https://ieeexplore.ieee.org/document/9651887.

Μπουρνάζος [Bournazos] 2009 Μπουρνάζος [Bournazos] Σ. (2009) “ Το κράτος των εθνικοφρόνων: αντικομμουνιστικός λόγος και πρακτικές” [“The nationalists' state: Anticommunist speech and practices”], in “Ιστορία της Ελλάδας του 20ού αιώνα: 1945–1952, Ανασυγκρότηση-Εμφύλιος-Παλινόρθωση” [*History of Greece in the 20th century: 1945–1952, reconstruction-civil war-restoration*]. Athens, Greece: Bibliorama, pp. 9-49.

Νικολακόπουλος [Nikolakopoulos] 2009 Νικολακόπουλος [Nikolakopoulos], Η. (2001) “Η Κακεκτική Δημοκρατία: Κόμματα και Εκλογές, 1946–1967” [*The sickly democracy: Parties and elections, 1946–1967*]. Athens, Greece: Patakis.

Χατζηβασιλείου [Chatzivasiliou] 2011 Χατζηβασιλείου [Chatzivasiliou], E. (2011) “Η πρό” [*The reception of the left by its opponents, 1949–1967: Observations on the workings of Greek anticommunism*]. Corfu, Greece: Ionian University, Department of History.

Χριστίδης [Christidis] 2007 Χριστίδης [Christidis], Π. (2007) “Οι ελληνικές εθνικές διεκδικήσεις στη συνδιάσκεψη για την ειρήνη στο Παρίσι, 1946” [*The Greek national claims in the Paris peace conference, 1946*]. PhD thesis. Aristotle University of Thessaloniki.