Mapping the Historical Ecology of the Cyclades: A Diachronic Natural Language Processing Analysis of Travel Narratives (1700–1920)

Aikaterini Christopoulou[1,2,3,4], Vassilis Detsis[1], Basilis Gatos[2,3,4]

[1]Harokopio University of Athens, [2]National Centre for Scientific Research "Demokritos",
[3]AI4DOC P.C., [4]Archimedes Unit - Athena Research Centre
[1]El. Venizelou Ave. 70 17676 Kallithea Greece, [2,3]Patr. Gregoriou E & 27 Neapoleos Str 15341 Agia Paraskevi Greece, [4]Artemidos Street 1 15125 Marousi Greece
achristopoulou@hua.gr, detsis@hua.gr, bgat@iit.demokritos.gr

Abstract

Historical texts can be valuable for the study of a place's ecological history but reading and extracting information from them can be a tedious and time-consuming task. Natural Language Processing can help in order to extract the most important information of the text in a quick, effective and reproducible way. In this study, travel narratives for the Cyclades Islands from 4 different time periods (1700-1920) have been chosen for analysis. The first step, the quantitative part, includes the semi-automatic detection of geographical entities in the texts and their connection to predefined keywords in order to enable temporal and spatial statistical analysis. The output of this procedure is then inserted in a Retrieval-Augmented Generative Synthesis pipeline in which the text segments with the connected place and keyword are processed by a locally orchestrated Large Language Model. The final output is used for the understanding and interpretation of the original text. Even though the study focuses mainly on the coherence and repeatability of the workflow, an effort is made to interpret and connect the results to the past ecological profile of these islands. The dataset/supplementary material is provided via an open access repository.

Keywords: Name Entity Recognition, Large Language Model, Historic Narratives

1. Introduction

Travelers of the 17th to 20th century left an invaluable legacy of historic texts which they wrote in their voyage journals. Some of these travellers who have travelled in Greece during this time are William Martin Leake, Sir William Gell, François Pouqueville, James Theodore Bent and Joseph Pitton de Tournefort. Either archaeologists, historians, geographers, botanists, scholars or even militaries, all of them had a passion for discovery, using their sharp observation and keen curiosity to understand, interpret and transfer landscape and events in such a way, that even 200-400 years later, their texts can be object of academic research. In most cases their journeys included mainland Greece, e.g., Attika and the Peloponnese, and insular Greece, e.g., the Cyclades and Crete Island. In their texts they included various information about the places they visited, such as the description of landscape features, fauna and flora, as well as, accounts of activities, folklore and the way of life of the inhabitants of these places, attributed with each author's personal style. Thus, these texts are well suited for the historic study of past landscapes, both from an ecological and a linguistic point of view. Their value lies in the fact that they are some of the very few depictions of what the Greek territory was like in this time, especially during the Greek occupation by the Ottomans, as well as the 19th and early 20th century.

Despite of their value and importance, historical texts of travellers have been overlooked especially by geographers and ecologists, while in fact, they can be used for the evaluation of historic ecosystem services' provision and landscape reconstructions (Tomscha et al., 2016). They might be scarce and very few in number, but they are still lengthy and dense texts, whose analysis can be a time-consuming task if done manually. The use of modern technologies for textual understanding and analysis, also known as Natural Language Processing (henceforth NLP), can be a very useful and efficient solution to this task. NLP can be used for a quick, systematic and structured screening of large volumes of such texts, in order for the researcher to (1) get the general idea of each narrative, (2) sort out or prioritise the most important, rich and promising sections, (3) understand the views and personal style of the writer and, most importantly, (4) achieve all of the above in a repeatable and reproducible way.

Although NLP has been successfully applied in various linguistic and environmental research works, it has never, to our knowledge, been applied in historic texts of travellers, especially, with the focus on ecological research. Thus, this study aims to bridge the gap between these historical narratives, temporal and spatial ecological research and current state-of-the-art text understanding methods. The goal of this work is the quantitative and qualitative analysis of 4 historic travellers' texts, from 4 different time periods (1700, 1830, 1880 and 1920), using NLP tools in R programming environment. The quantitative analysis was implemented using Name Entity Recognition (NER) techniques and the qualitative analysis was carried out by using local Retrieval-Augmented Generation (RAG)

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 105–114

Large Language Models (henceforth LLMs). The 4 texts that have been chosen for the analysis, supplementary material and results (tables and high-resolution figures) are included in an open access dataset/supplementary material provided[1] and refer to Cyclades as a study area. Each Cycladic Island holds its own unique physical, natural and environmental characteristics, which lead to equally unique activities, production and way of life of its inhabitants. Additionally, they have a strategic location as they are part of the maritime route that connects mainland Greece to Asia Minor, Constantinople and the Holy Land. The Cyclades is one of the few examples of Greek areas that remained more or less the same during the turbulent times of the 17th to 19th century, making them a living laboratory, suitable for the study of landscape and ecological change and human-environment interaction impacts. Due to all of the above, the texts that refer to the Cyclades were considered a more suitable case study for this kind of analysis. Even though the study focuses mainly on the coherence and repeatability of the workflow, an effort is made to interpret and connect the results to the past ecological profile of these islands.

2. Related Work

The interest in the use of Natural Language Processing (NLP) among several disciplines of environmental studies is gradually increasing. The subjects that have been assessed using NLP methods include landscape or biodiversity (Schimanski et al., 2023; Abdelmageed et al., 2022; D'Souza et al., 2025), climate change (Grasso and Locci, 2024), ecosystem services provision (Havinga et al., 2024; Kong et al., 2023; Luo et al., 2025; Zhang et al., 2026), natural hazards (Avcıoğlu et al., 2025; Lai et al., 2022; Sodoge et al., 2023) and sustainability-related regulations (Villacampa-Porta et al., 2025). The majority of these studies use as data sources either social media content (Havinga et al., 2024) or modern newspaper and journal articles (Avcıoğlu et al., 2025; Sodoge et al., 2023). Current works can also be distinguished in three distinct categories, those that use mainly statistical methods, such as Name Entity Recognition (Abdelmageed et al., 2022), those that integrate Large Language Models in their workflow (Zhu et al., 2023; Grasso and Locci, 2024; Luo et al., 2025; Zhang et al., 2026) and those that combine the aforementioned methods (Nundloll et al., 2022). Noteworthily, there are numerous studies that incorporate geographic or spatial features in their pipeline, showcasing the opportunities for connection and interdisciplinarity between geographic, linguistic and environmental subjects (e.g., Avcıoğlu et al., 2025; Gregory et al., 2015; Grossner et al., 2026). What is also worth mentioning is a strong connection between sentiment analysis and environmental studies (Havinga et al., 2024; Barz et al., 2025; Huai and Van De Voorde, 2022). However, the studies that involve the analysis of historical data, such as censuses (e.g., Haider et al., 2025), records (e.g., Li and Shi, 2025) or texts (e.g., Nundloll et al., 2022; Brando and Frontini, 2017) are scarce. Thus, our work focuses on historical travel accounts, as they combine temporal and spatial scale, needed in ecological studies, accompanied by the personal writing and narrative style, which can influence the text understanding and interpretation.

3. Dataset

The corpus is divided into 4 parts as it consists of 4 seminal travelogues covering most of the Cyclades (e.g. Amorgos, Andros, Melos, Naxos, Santorini and Syros) and dated at 1700, 1830, 1880, and 1920.

The first and oldest part, includes text transcriptions from 233 pages of the book «A voyage into the Levant», translated from the original French book «Relation d'un voyage du Levant»[2], by Joseph Pitton de Tournefort, a French doctor, botanist and traveler. The book consists of 3 volumes and it was published in 1741, in London. These texts have some peculiarities that need to be taken into consideration. The first involves the use of older place names some of which have no nomenclatural connection to the modern name of the island. The second is the use of the archaic long s (shown as «ſ» or «f» in the text). The third refers to the language itself where in most cases words are in the form of «parch'd». All the above make the texts difficult for an LLM to read, understand and analyse.

The second part includes text transcriptions from 21 pages of the book «Observations Upon the Peloponnesus and Greek Islands»[3], by Rufus Andrson, an American minister. The book was written in 1829 and it was published in Boston in 1830. Anderson travelled in the Peloponnese, Aegean and Ionian islands, gathering information concerning the geography, history and culture of the places he visited, reflecting in social and political issues of the time. As his main mission during his overseas travel was the spread of Christianity, the main theme of this book is occupied by matters of faith and the Greek church.

The third part includes text transcriptions from 321 pages of the book «The Cyclades, or; Life among the insular Greeks»[4], by James Theodore Bent an

[1] Dataset/Supplementary material accessed from the link: https://zenodo.org/records/19226329.
[2] https://catalog.hathitrust.org/Record/001240678
[3] https://catalog.hathitrust.org/Record/000649704
[4] https://hdl.handle.net/2027/mdp.39015028327800

English explorer, archaeologist, and author. The book was published in London in 1885 and it is a personal account of two tours made by Bent and his wife in the Greek islands between 1882 and 1884. It has been characterised as a classic guide to the Cyclades. His narrative includes a very personal, quirky tone, as he describes the various pleasant and unpleasant events while visiting the Cyclades, describing the landscape, social life, production and activities in a detailed manner.

The fourth part includes text transcriptions from 80 pages of the book «A Handbook of Greece»[5], which was published in 2 volumes in London in 1919 and it was compiled by the Geographical Section of the Naval Intelligence Division, Naval Staff, Admiralty. The selected pages include text that refer to physical, natural and environmental characteristics of the Cyclades in general, as well as social and economic aspects of the Islands. This book was added to the corpus as it is more scientific and data-driven, in comparison, especially, with the journals of Tournefort and Bent which are more narrative, descriptive and even poetic.

(a) threatens. We therefore reſolv'd to wait for a *French* Bark: by good luck there was at *Canea* one of thoſe which your Lordſhip has forbid pickeering from Iſland to Iſland for Plunder. I promis'd the Maſter not to inform againſt him, and ſo he convey'd us to *Argentiere*, the firſt of *Auguſt*. ΚΙΜΩΛΟΣ. This Iſland, by the *Greeks* call'd *Chimoli* [*b*], took the name of *Argentiere* at the time when the Silver Mines were firſt diſcover'd there: there are ſtill to be ſeen the Work-houſes
(b) We therefore refolv'd to wait for a French Bark: by good luck there was at Canea one of **thoſe** which your **Lordſhip** has forbid pickeering from **Iſland** to **Iſland** for Plunder. I promis'd the **Maſter** not to inform **againſt** him, and **ſo** he convey'd us to Argentiere, the first of August. **Thif Iſland**, by the Greeks call'd Chimoli [b], took the name of Argentiere at the time when the Silver Mines were **firſt diſcover'd** there
(c) We therefore resolv'd to wait for a French Bark: by good luck there was at Canea one of **those** which your **Lordship** has forbid pickeering from **Island** to **Island** for Plunder. I promis'd the **Master** not to inform **against** him, and **so** he convey'd us to Argentiere, the first of August. **This Island**, by the Greeks call'd Chimoli [b], took the name of Argentiere at the time when the Silver Mines were **first discover'd** there

Table 1: A sample of the oldest book of the corpus «A voyage into the Levant», (a) original text, (b) the result of the automatic OCR and (c) OCR corrections marked in red.

4. Methodology

This study introduces a multi-phase Natural Language Processing (NLP) framework which was designed (a) to transform unstructured historical journals of travellers of the 17th to 20th century into structured data relevant for ecological studies and (b) to analyse and semantically interpret the ecology of the landscape. The framework integrates statistical text mining with retrieval-augmented neural generation, enabling both quantitative pattern detection and qualitative interpretation. The proposed architecture integrates rule-guided linguistic annotation, lexicon-based semantic extraction, and neural generative modelling (Figure 1). The pipeline consists of two sequential components: (a) A statistical information extraction module which performs linguistic preprocessing name entity identification, lexicon-guided semantic matching and context-based relationship extraction. (b) An LLM based retrieval-augmented neural synthesis module which generates semantically coherent ecological descriptions grounded in retrieved textual evidence. The framework follows a corpus-driven paradigm, combining deterministic extraction with neural interpretation while preserving traceability between generated outputs and source text.

4.1 Data Preparation and Lexicon Construction

The corpus consists of digitized historical travel journals containing descriptions of landscapes, ecological conditions, activities, products and the way of life. Given the linguistic variability and orthographic inconsistencies typical of historical documents, the oldest text (1700) was corrected concerning the automatic Optical Character Recognition (OCR) which was originally provided (Table 1) and all texts were normalized through a preprocessing stage including tokenization, sentence segmentation, and lemmatization. To guide the extraction process, a domain-specific ecological lexicon was constructed comprising 130 keywords organized into thematic categories such as activities (e.g., mining, pottery), harvest (e.g., wheat, olive), products (e.g., wine, jam), infrastructure (e.g., harbour, bridge), natural resource (e.g., marble, emery), water (e.g., stream, river) and landscape (e.g., grove, cave). The lexicon was developed through expert-driven selection based on ecological relevance and expected occurrence in historical descriptive narratives. This lexicon serves as the primary semantic anchor for subsequent statistical and generative analysis.

4.2 Statistical Keyword Pipeline

The first phase of the framework performs statistical extraction of ecological information through entity identification and proximity-based

[5] https://catalog.hathitrust.org/Record/008881832

relationship analysis. This phase combines automated linguistic annotation with expert-guided refinement to ensure accuracy in historical contexts characterized by spelling variation, archaic syntax, and toponymic evolution.

4.2.1 Entity Refinement

Geographical entities were initially detected using part-of-speech tagging and dependency parsing (Wijffels et al., 2018). However, historical corpora present unique challenges, including obsolete place names, spelling variations, and ambiguous entity references. To address these challenges, a semi-automated Human-in-the-Loop refinement stage was introduced. The automatically extracted location candidates were reviewed and curated by domain experts to resolve toponymic ambiguities, normalize historical place names to standardized equivalents (e.g., mapping “Argentiere” to its modern equivalent “Kimolos”) and remove false positives. This process resulted in a validated gazetteer of geographical entities, which was subsequently used as a reliable reference for downstream relationship extraction and retrieval operations. This hybrid human-machine approach significantly improves entity precision while preserving scalability.

4.2.2 Proximity-Based Relationship Extraction

Following entity validation, the framework identifies semantic relationships between geographical locations and ecological concepts using a sliding window co-occurrence algorithm (Silge and Robinson, 2016). This approach captures both explicit and implicit associations within narrative contexts. Two complementary extraction strategies were employed:

4.2.2.1 State-Persistent Contextual Tagging

This strategy models the narrative flow of the text by propagating the most recently mentioned geographical entity across subsequent tokens until a new location is encountered. Ecological keywords appearing within this propagated context are associated with the active spatial anchor. This method captures extended descriptive passages in which ecological features are discussed without repeated explicit mention of the location.

Formally, given a sequence of tokens $t_1,t_2,\ldots,t_n$, and detected location tokens $L_i \subset T$, each ecological keyword occurrence k_j is assigned to the nearest preceding location entity L_i, provided no intervening location entity exists. This approach enables reconstruction of narrative-level geographic-ecological associations.

4.2.2.2 Keyword-in-Context (KWIC) Extraction

To ensure high-precision extraction suitable for downstream neural processing, a localized context window approach was also applied. For each detected geographical entity (e.g., Andros), a symmetric window of ±20 tokens, was extracted. These contextual segments were scanned for ecological keyword occurrences (e.g., olive, silk) using boundary-aware lexical matching. This method produces high-confidence, localized evidence of ecological-geographical relationships (e.g., consistent presence of silk in 1700 and 1880 in Andros Island) and serves as the primary evidence source for retrieval-augmented generation.

4.3 Retrieval-Augmented Generative Synthesis Pipeline

While statistical keyword extraction effectively identifies candidate relationships, it cannot fully capture the semantic richness and implicit ecological descriptions present in historical narratives. To address this limitation, the second stage employs a Retrieval-Augmented Generation (RAG) framework that combines evidence retrieval with neural language modeling. The output of the statistical pipeline is transformed into a structured knowledge index containing temporal metadata, ecological keyword and contextual text segment. Rather than processing the entire corpus, the system selectively retrieves high-density textual segments where ecological and geographical entities co-occur. This targeted retrieval improves computational efficiency while maximizing semantic relevance. Each retrieved segment serves as grounded input for generative interpretation.

The retrieved evidence segments are processed by a locally orchestrated Large Language Model (Llama 3) to generate structured ecological descriptions. The model was prompted to synthesize information across multiple textual fragments, producing coherent interpretations of historical landscapes focusing on human-environment interactions. Unlike purely generative approaches, the model operates within a constrained retrieval-augmented setting, ensuring that all generated content is grounded in verifiable textual evidence rather than unconstrained inference. An example of a prompt provided is this: “*Analyze this historical snippet mentioning Andros. Extract four ecological data points in exactly this format, separated by pipes: LANDSCAPE | EXPLOITATION | WATER | PRODUCTION. Consider as mentioned if 1-4 words are included for each category. If not mentioned, write 'NA'*”.

4.4 Implementation Details

The proposed pipeline was implemented using a modular architecture designed to support corpus preprocessing, linguistic annotation, lexicon-based information extraction, and retrieval-augmented neural generation. The implementation was developed in the R programming environment, enabling reproducible corpus processing and seamless integration of statistical and neural components.

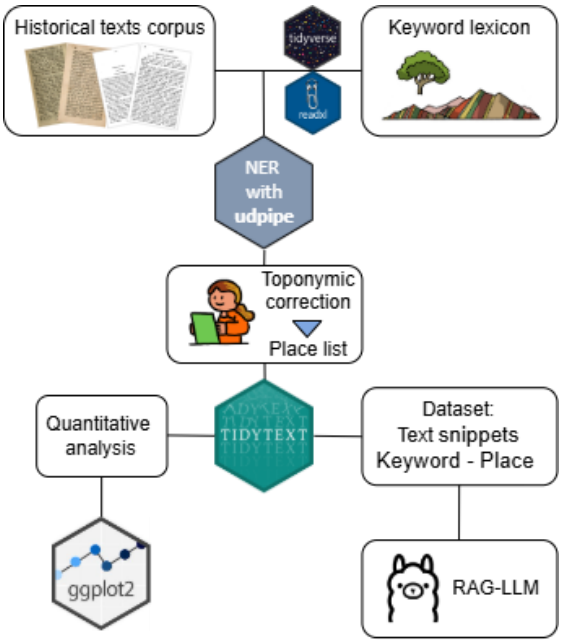

Figure 1: Hybrid NLP Framework for Historical Ecology Analysis.

4.4.1 Linguistic Annotation and Corpus Processing

Linguistic preprocessing, including tokenization, sentence segmentation, lemmatization, and morphosyntactic annotation, was performed using the **udpipe** package (Wijffels et al., 2018). This framework implements models based on the Universal Dependencies formalism, providing standardized annotations including lemma normalization, part-of-speech tagging, morphological feature annotation and dependency parsing. These annotations enabled robust lexical matching and improved resilience to orthographic and morphological variation present in historical texts. The annotated output was stored in structured tabular format, allowing efficient downstream querying and contextual extraction.

4.4.2 Lexicon-Based Extraction and Contextual Analysis

Lexicon-driven ecological concept detection and contextual co-occurrence analysis were implemented using the **tidytext** package (Silge and Robinson, 2016) which provides corpus-oriented text mining functionality based on token-level data representations. The tidytext framework enabled token-level corpus representation using tidy data principles, efficient keyword matching at the lemma level, implementation of sliding window context extraction, Keyword-in-Context (KWIC) analysis and context propagation for entity-keyword association. Regular expression matching with boundary constraints was used to ensure precise detection of ecological terms while avoiding spurious partial matches.

Corpus transformation, filtering, and aggregation operations were implemented using the **tidyverse** package (Wickham et al., 2019b), which provides efficient and reproducible data manipulation capabilities. These operations included corpus normalization, token filtering and aggregation, context window construction, knowledge base assembly.

4.4.3 Gazetteer Curation and External Resource Integration

Supplementary structured resources, including curated gazetteers and lexicon files, were managed using the **readxl** package (Wickham et al., 2019a). This enabled integration of manually curated entity normalization mappings into the automated pipeline. The curated gazetteer was used to normalize historical toponyms, resolve spelling variation and support consistent entity linking. This ensured high precision in geographic entity resolution.

4.4.4 Retrieval-Augmented Neural Generation

The Retrieval-Augmented Generation (RAG) component was implemented using the **ollamar** framework (Lin and Safi, 2024), which provides an interface for local deployment and orchestration of transformer-based large language models. The system performs the following operations: retrieval of contextually relevant evidence segments from the knowledge base, construction of structured prompts combining geographic and ecological information, neural generation of evidence-grounded ecological interpretations. This local deployment approach ensures reproducibility, data privacy, reduced computational latency and full control over model inference.

4.4.5 Visualization and Statistical Analysis

Statistical summaries and visualizations of extracted ecological patterns were generated using the **ggplot2** package (Wickham, 2016), which implements a grammar-of-graphics approach. Visualization outputs included temporal frequency distributions of ecological keywords, geographic distributions of ecological references and co-occurrence frequency plots. These visualizations support exploratory corpus analysis and facilitate interpretation of extracted ecological patterns.

5. Results and Discussions

5.1 Quantitative Analysis: The Keyword Pipeline

The comparison of keyword occurrences belonging in the categories “Landscape”, “Activities”, “Inland Water” and “Infrastructure”, between the 4 time periods, is presented in Figure

2. The description of “Landscape” features is dominant in every text included in the analysis, while keywords describing “Activities” are the least mentioned, presenting the highest frequency

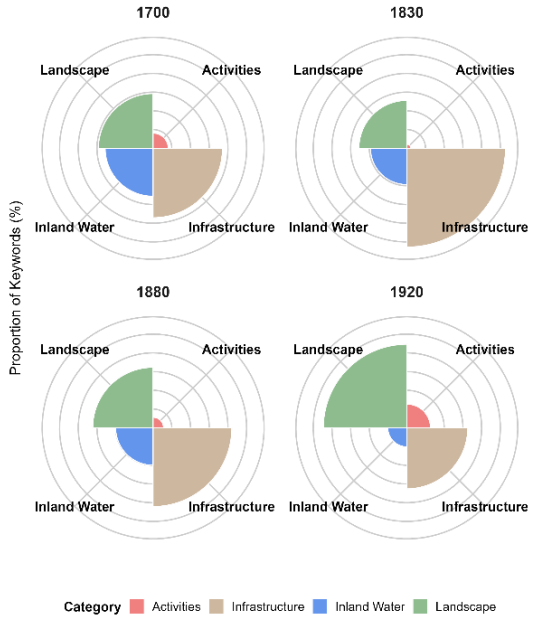

Figure 2: A macro-level thematic context of the keyword-island connection analysis.

in the text of 1920. “Inland Water” shows a gradual decrease, while “Infrastructure” peaks in the 1830 text. The decrease in “Inland Water” occurrences could be probably attributed to the absence of the specific keywords included in the lexicon rather than an actual unavailability of water. These could be counted as false negatives, as the absence of keywords belonging to a category e.g., “Inland Water”, does not necessarily mean the actual lack of availability of water. This could be validated by the results of the LLM analysis (see Section 5.2 and Figure 5).

In Figure 3, a comparison between the density of specific keywords, for 1700 and 1880, is presented. The narrative of 1700 focuses heavily on the presence of “Product” or “Harvest” (high frequency for wine, oil, cotton, etc.), while the narrative of 1880 shows a surge in “Activities” (pottery or mines). This suggests a transition from a consumer/observer perspective, to a more industrial or ethnographic perspective of documenting people’s way of life. The “Product” category undergoes a clear evolution in variety. While in 1700 the narrative was dominated by “wine”, “oil” and “silk”, in 1880 “wine” maintains a high frequency, but the emergence of processed goods like “honey”, “cheese” and “jam” is observed. This could indicate a more sophisticated economic reporting style or an actual diversification of the Aegean export economy during the 19th century.

In Figure 3, the keyword “mines” remains consistent, but its location shifts or intensifies in specific areas like Melos and Naxos across the two periods. A high frequency of occurrences of specific words, e.g., “rock” and “flower” for the text of 1700 or “mountain”, “hill” and “ground” for the text of 1880 can be observed. In 1700, keywords are clustered around specific islands like Delos,

Figure 3: Keyword frequency comparison between 1700 and 1880 for 6 different categories in all the islands included in the texts.

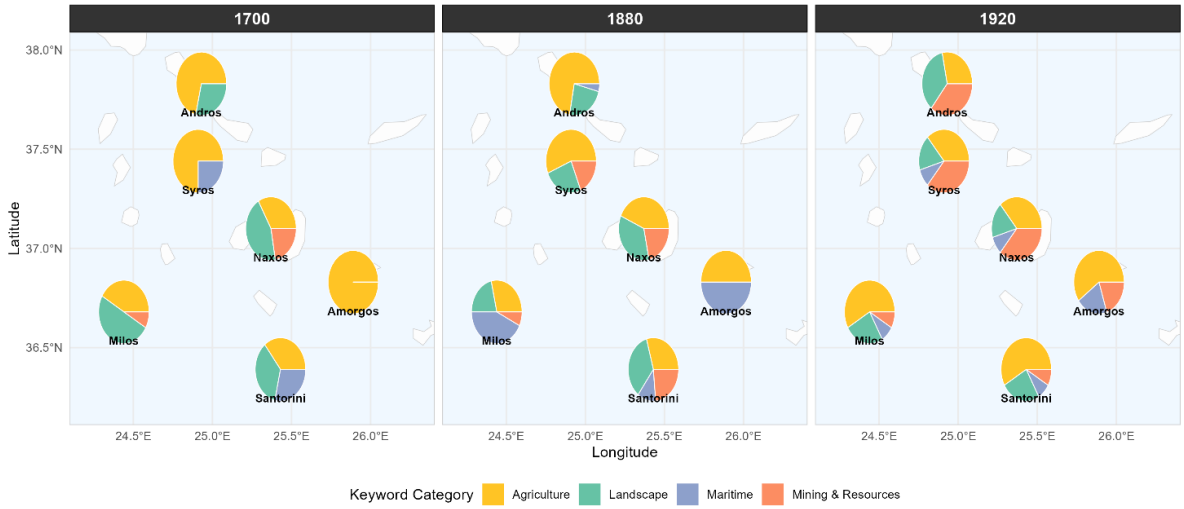

Figure 4: Spatial distribution of 4 ecological themes. Pie charts represent the relative frequency (mentions per 1,000 words) of keywords across 3 historical journals.

Melos, and Naxos. In 1880, the distribution is much more spread across smaller islands like Amorgos, Anaphi, and Keos. Also, there is a noticeable emergence of “cave” mentions in 1880 (specifically in Andros and Antiparos) that barely existed in 1700.

The difference in the text contents can be mainly attributed to the author’s interest and point of view and not at the presence or absence of specific features in the landscape, e.g., it is impossible that hills or ground to be absent from the landscape in 1700. Another reason for the absence of specific features, e.g., “cave”, could be attributed to the difficulty of access that would characterize the journey of a traveller in comparison to a more recent one.

There is a persistent occurrence of specific words across all places and time periods, e.g. the word “wine” or “rock” in Figure 3. These are attributed in most cases to the objective truth, as the Cyclades are, indeed, areas with rocky landscapes with scarce vegetation, especially during the summer. In cases like these, one can export safe conclusions, e.g., for the continuous production or trade of wine across the selected time periods.

There is high density of words for a specific time period or place, e.g. the word “plant” for the 1700’s text or “mountain” for the 1880’s text (Figure 3). These can be attributed to the personal interests of the author of the specific text, as more specifically, Tournefort was a botanist, so his interest in plants justifies the high density of the word in his texts, more than drawing the assumptions that plants had a higher occurrence in the Cyclades in 1700 than in 1880. This can also be observed in Figure 2 where the highest occurrences of keywords belonging in the category “Infrastructure” was attributed in the text of 1830, presumably because of the profession of the author as he was a minister with an interest in religion and education. Thus, occurrences like these cannot be used as evidence for changes, in the absence of further indications.

In Figure 3, there are scarce, low in frequency word occurrences, such as the “corn”, “barley” and “mulberry”. These are the cases that need to be cross-validated with additional historical sources. For example, according to Kolovos (2007), Andros was one of the highest silk producers in Greece the 17th and 18th century, thus the connection of the word with this island is justified. The presence of other agricultural products in the Cycladic Islands can, additionally, be validated with agricultural censuses, the oldest of which was in 1911. Another interesting finding is the striking absence of “figs” from the 1880’s texts which could, also, be validated from Kolovos (2017), who quoted Horden and Purcell (2000) who suggested that «there is no other edible fruit in the Mediterranean that was so unfairly abandoned like figs».

Figure 4 which includes the granular island profiles (Amorgos, Andros, Melos, Naxos, Santorini and Syros) reveals distinct ecological and economic trajectories. A major “mining shift” is observed in all 6 islands. While mining was present only in Naxos and Melos Islands during the 17th century, the occurrences of words related to mining and natural resources increase drastically during the 20th century texts. A more stable presence of “Agriculture”-related keywords, during every period reflects the agrarian focus especially of early travellers. More specifically, Amorgos Island shows a clear transition from 1700's agricultural focus to an interest in mining and maritime activity in 1880. For Andros Island, the data captures the consistent presence of silk in 1700 and 1880 (Figure 3) and a continuation of agricultural activities and production even in 1920. Notably, it

also records a sudden shift toward industrial mineral extraction in 1920. All these changes occurring in Andros Island justify the focus on this island for the LLM analysis pipeline. As for Naxos Island, it demonstrates long-term agricultural stability and it is the only island with a steady presence of mining activity throughout the studied years.

5.2 Qualitative Synthesis: The LLM-RAG Pipeline

The use of the Retrieval-Augmented Generation (RAG) component in the ollamar framework, assisted in the understanding and interpretation especially of those aspects that were either overlooked or misinterpreted by the keyword-spotting pipeline. A total of 130 snippets, that were produced from the previous phase and refer to Andros Island, were used as an input to the LLM framework. Figure 5 shows an example from the enriched text output containing the LLM result.

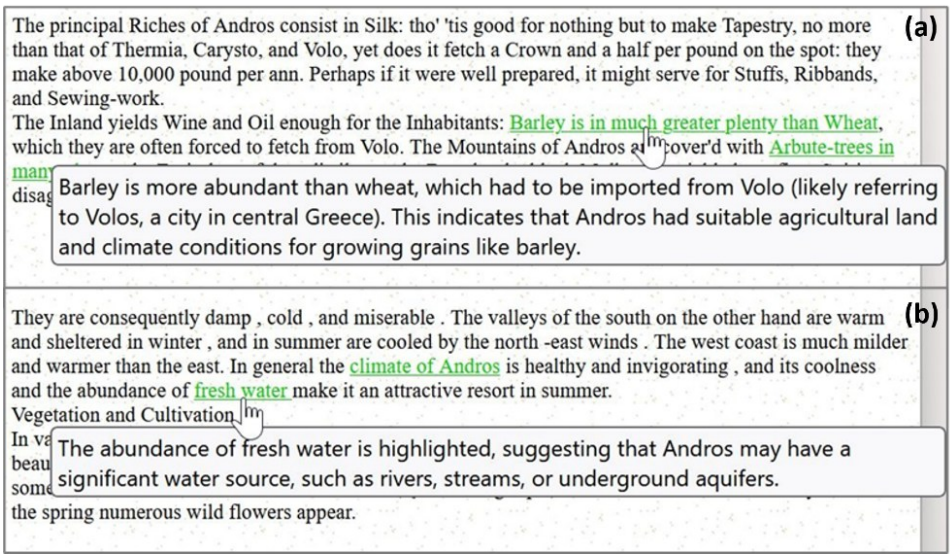

Figure 5: The LLM output incorporated in the original text of (a) 1700 and (b) 1920.

The first important issue that the LLM handles well is the competing entities in the same sentence like in the case of the occurrences of the words “barley” and “wheat” in the sentence «Barley is in much greater plenty than Wheat», which is included in the input dataset to the RAG model. The occurrence of these words is counted as equal when it comes to keyword frequencies but according to the aforementioned sentence, the context of comparison is missing. A second issue that was handled with the LLM, refers to false positives of the first phase, e.g. “silk” in Andros Island, that was detected in the sentence «the silk industry, once flourishing, has now almost disappeared from Andros» of the 1920 text. A third example of the disagreement between the two phases has to do with “water”. According to Figure 2, the category “Inland Water” shows a gradual decrease in team, however, the output of the LLM, seen in Figure 5, does not imply such a decrease.

6. Conclusion

The combination of the keyword pipeline and the LLM-RAG pipeline is a holistic approach towards understanding and analysis of old text contents, whose study can be very useful in the field of historical ecology but very challenging and time-consuming when done manually. The keyword pipeline provides the statistical validity and longitudinal scope, while the RAG pipeline provides the qualitative depth necessary to understand human-environment interactions. This dual approach mitigates the risk of “keyword spotting” without context, while also preventing the “hallucination” risks associated with running LLMs on unfiltered, massive text volumes.

6.1 Limitations

Despite of the usefulness and repeatability of the indicated methodology there are some limitations that should be addressed. The first refers to the construction of the keyword lexicon, which in the present study was compiled independently of the corpus, leading to possible omissions of keywords, especially when it comes to older texts. A wider, more appropriate and more focused keyword lexicon can lead to a deeper insight and understanding of the texts. A second limitation refers to the content of the text, as in many cases it is the result of personal perspective or “agenda” of the author. This is addressed by the evaluation with additional historical texts and published works. A third limitation involves the limited size of the dataset used for this analysis, with the results being influenced by the authors’ personal views. Using a richer and more diverse dataset can minimize the aforementioned effect.

6.2 Future Work

The first aim of further work is the creation of a solid evaluation process which will not only assess the framework created but also lead to the development of ecological conclusions. The corpus used in this study can be enriched with even older texts, or texts of travellers written in other languages (e.g., German or Greek), or description of other areas for deeper temporal or spatial comparison, even with more recent texts. As the perspective of the authors of these texts play a very important role in understanding, a thorough sentiment analysis could be insightful. The NLP analysis can also be connected to an in-depth spatial analysis with the use of satellite images or historical aerial photographs for land cover change study.

7. Acknowledgments

This work was supported by the Project "D-AI-LECT – Digital Analysis and Recognition of Handwritten Documents of Greek Dialects" (Archimedes, Athena Research Centre, Greece) and the GREEN TALENT Horizon Europe project (Grant Agreement No. 101217375). GREEN TALENT is funded by the European Union's Horizon Europe research and innovation programme; however, the views expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the granting authority can be held responsible for them. The authors also thank the anonymous reviewers for their valuable comments.

8. Bibliographical References

Abdelmageed, N., Löffler, F., Feddoul, L., Algergawy, A., Samuel, S., Gaikwad, J., and König-Ries, B. (2022). BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, 10:e89481.

Avcıoğlu, A., Demir, O., and Görüm, T. (2025). An automated approach for developing geohazard inventories using news: integrating natural language processing (NLP), machine learning, and mapping. *Natural Hazards and Earth System Sciences*, 25:2421–2435.

Barz, C., Siegel, M., and Hanss, D. (2025). Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025)*, pages 68–76, Tallin, Estonia, March.

Brando, C. and Frontini, F. (2017). Semantic historical gazetteers and related NLP and corpus linguistics applications. *Journal of Map & Geography Libraries*, 13(1):1–6.

D'Souza, J., Laubach, Z., Al Mustafa, T., Zarrieß, S., Frühstückl, R., and Illari, P. (2025). Mining for species, locations, habitats, and ecosystems from scientific papers in invasion biology: A large-scale exploratory study with large language models. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025)*, pages 16–23, Tallin, Estonia, March.

Grasso, F. and Locci, S. (2024). Assessing generative language models in classification tasks: Performance and self-evaluation capabilities in the environmental and climate change domain. In *International Conference on Applications of Natural Language to Information Systems*, pages 302–313, Cham, Switzerland, June. Springer Nature Switzerland.

Gregory, I., Donaldson, C., Murrieta-Flores, P., and Rayson, P. (2015). Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9(1):1–14.

Grossner, K., Janowicz, K., and Keßler, C. (2016). Place, period, and setting for linked data gazetteers. In *Placing Names: Enriching and Integrating Gazetteers*, pages 80–96, Indiana University Press, United States.

Haider, T., Perschl, T., and Rehbein, M. (2025). Quantification of biodiversity from historical survey text with LLM-based best-worst-scaling. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025)*, pages 61–67, Tallin, Estonia, March.

Havinga, I., Marcos, D., Bogaart, P., Tuia, D., and Hein, L. (2024). Understanding the sentiment associated with cultural ecosystem services using images and text from social media. *Ecosystem Services*, 65:101581.

Horden, P. and Purcell, N. (2000). *The Corrupting Sea: A Study of Mediterranean History*. Blackwell Publishing, Oxford, UK.

Huai, S. and Van de Voorde, T. (2022). Which environmental features contribute to positive and negative perceptions of urban parks? A cross-cultural comparison using online reviews and Natural Language Processing methods. *Landscape and Urban Planning*, 218:104307.

Kolovos, E. (2017). *"There was a garden..." The Economy of the Mediterranean Island of Andros according to the Ottoman Land and Property Survey of 1670*. Crete University Press, Herakleio, Greece.

Kong, I., Sarmiento, F. O., and Mu, L. (2023). Crowdsourced text analysis to characterize the US National Parks based on cultural ecosystem

services. *Landscape and Urban Planning*, 233:104692.

Lai, K., Porter, J. R., Amodeo, M., Miller, D., Marston, M., and Armal, S. (2022). A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures. *Information Processing & Management*, 59(1):102735.

Li, T. and Shi, T. (2025). High-resolution climate reconstruction from historical Chinese weather records using optimized natural language processing. *Scientific Reports*, 15:44447.

Lin, H. and Safi, T. (2024). ollamar: An R package for running large language models. *Journal of Open Source Software,* 10(105): 7211.

Luo, H., Zhang, Z., Zhu, Q., Ameur, N. E. H. B., Liu, X., Ding, F., and Cai, Y. (2025). Using large language models to investigate cultural ecosystem services perceptions: A few-shot and prompt method. *Landscape and Urban Planning*, 258:105323.

Nundloll, V., Smail, R., Stevens, C., and Blair, G. (2022). Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. *Heliyon*, 8(10).

Schimanski, T., Senni, C. C., Gostlow, G., Ni, J., Yu, T., and Leippold, M. (2023). Exploring nature: Datasets and models for analyzing nature-related disclosures. *arXiv preprint arXiv:2312.17337*.

Silge, J. and Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3): 37.

Sodoge, J., Kuhlicke, C., and de Brito, M. M. (2023). Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, 41:100574.

Tomscha, S. A., Sutherland, I. J., Renard, D., Gergel, S. E., Rhemtulla, J. M., Bennett, E. M., and Clark, E. E. (2016). A guide to historical data sets for reconstructing ecosystem service change over time. *BioScience*, 66(9):747–762.

Villacampa-Porta, J., Coronado-Vaca, M., and Garrido-Merchán, E. C. (2025). Impact of EU non-financial reporting regulation on Spanish companies' environmental disclosure: a cutting-edge natural language processing approach. *Environmental Sciences Europe*, 37(1):29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY.

Wickham, H., Bryan, J., Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., and Bryan, M. J. (2019a). Package 'readxl'. Version 1.3.1.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019b). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wijffels, J., Straka, M., and Straková, J. (2018). Package 'udpipe'. CRAN.

Zhang, Z., He, Z., Su, K., Wu, S., and Liu, L. (2026). Global evolution of ecosystem services research in watersheds: Insights from large language models. *Land Use Policy*, 160:107849.

Zhu, J. J., Jiang, J., Yang, M., and Ren, Z. J. (2023). ChatGPT and environmental research. *Environmental Science & Technology*, 57(46):17667–17670.