Extracting Volcanological Knowledge from Historical Texts: A Language-Technology Pipeline for Diachronic Geovisualization

Costanza Marini[*], **Gianluca Casagrande**[†], **Alessio Palmero Aprosio**[†], **Claudia Principe**[‡]

[*] University of Pavia, costanza.marini@unipv.it

[†]University of Trento, gianluca.casagrande@studenti.unitn.it, a.palmeroaprosio@unitn.it

[‡]Institute of Geosciences and Earth Resources - CNR, c.principe@igg.cnr.it

Abstract

This paper presents the first results of the CorVo project, a transdisciplinary project combining volcanology and computational linguistics to extract and structure volcanological knowledge from historical documents concerning Mount Vesuvius. We introduce the CorVo corpus, a multilingual diachronic corpus of 180 digitized texts (16th–20th centuries), selected to represent the main eruptive scenarios of the volcano. The digitization workflow integrates image pre-processing, OCR, and LLM-based post-correction to address challenges posed by degraded pages, historical typefaces, and orthographic variation. A domain-aware information extraction pipeline was developed to identify both standard toponyms and fine-grained spatial entities, which are typically overlooked by traditional NER systems. Extracted entities undergo human-in-the-loop validation and georeferencing through a dedicated annotation interface supporting multiple spatial geometries. The resulting dataset enables temporally normalized diachronic geovisualization of the textual-spatial footprint of Vesuvian eruptions across centuries.

Keywords: Historical Corpora, OCR, LLM, Named Entity Recognition, Diachronic Geovisualization

1. Introduction

Due to its long history of civilisation and high density of population, Italy is one of the few countries in the world, next to Japan (Kudo and Hoshizumi, 2006) and Indonesia (Voight et al., 2000), to possess an extensive historical record of the eruptive activity of its volcanoes (Giannini and Paolillo, 2018) – a record extensive enough to be exploited for the automatic extraction of useful information for volcanic eruption forecasting, risk assessment, and resilience planning.

Up until now, the use of historical sources by the geological community has been scarce due to the limited availability of digitized documents, the small number of research teams with both historiographic and volcanological expertise, and the time-consuming nature of manual data extraction from written sources (Principe, 1990).

Moreover, data extraction based solely on extensive reading of document collections cannot be considered fully objective or complete. It reflects the reader's bias, and goals and cannot be replicated without repeating the entire reading process (Guidoboni and Ebel, 2009). Today, the application of automated text analysis techniques to written sources can provide volcanologists with access to organised information on the historical activity of volcanoes in a time-efficient, replicable, and reliable manner, while volcanology can provide computational linguistics with a new testing ground for language technologies.

In this paper, we present the first findings of the CorVo project, which relies on a synergy of these two disciplines and makes three main contributions.

- First, it introduces CorVo corpus, a multilingual diachronic corpus of 180 digitized historical documents (16th–20th centuries) concerning the eruptive activity of Mount Vesuvius, designed to support both volcanological research and computational linguistic analysis.
- Second, it presents a domain-aware NLP pipeline that integrates OCR preprocessing, LLM-based post-correction, and structured geographic entity extraction tailored to fine-grained volcanological categories that are typically overlooked by standard NER systems.
- Third, it describes a human-in-the-loop geospatial annotation framework and a temporally normalized diachronic geovisualization system that enables systematic exploration of the textual-spatial footprint of Vesuvian eruptions across centuries.

Together, these contributions demonstrate how language technologies can operationalize historical sources for scientific risk assessment while advancing NLP research in low-resource, domain-specific contexts.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 38–48.

2. The Project

The CorVo project[1] is a transdisciplinary project bringing together volcanology and computational linguistics to support research on volcanic activity and its cultural impact, as well as to improve volcanic risk forecasting, impact assessment, and resilience planning. The project relies on a digitised corpus of historical documents (the CorVo corpus) and state-of-the-art NLP tools to facilitate its exploration and analysis. The documents contained in the CorVo corpus describe the past activity of one of Italy's most high-risk volcanoes: the Vesuvius. By querying the corpus, Civil Protection units, volcanologists, and other stakeholders will be able to quickly obtain important information concerning past eruptions, their social impact and institutional reactions to the related events. With data on eruption precursors, the actual sequence of eruptive phenomena, deposit distribution, and damages at hand, they are going to be able to update the current eruptive scenarios of the volcano. This information is particularly valuable because it cannot be derived from field data alone and, if properly integrated, may contribute to saving lives.

At present, the CorVo corpus interface enables exploration of 180 of the project's already digitised documents, the spatial data they contain, and their curated metadata.

The CorVo corpus, its metadata and the source code interfaces described in this paper will be released open source upon acceptance. See Section 8 for further details.

3. Related Work

In NLP research, volcanology can be currently considered a low-resource domain. The only Italian volcanology-focused corpus we found in the literature contains 2,839 daily and weekly Stromboli monitoring bulletins (2015-2021) and was compiled by Berardi et al. (2022) to extract numerical parameters for time-series reconstruction. More geology-focused corpora have been built in other languages, mostly for NLP tasks such as Named Entity Recognition (NER), Information Extraction (IE) and Information Retrieval (IR). For instance, for Chinese, Fan et al. (2020) created a geological-hazard corpus of 4,548 sentences annotated for NER and knowledge-graph construction, while Chen et al. (2022) released a fine-grained geological NER corpus annotated with 21 entity labels (e.g., sedimentary rock). Ma et al. (2023) developed an ontology-based BERT model for IE on geological hazard reports, while Qiu et al. (2020) worked on spatiotemporal IE from unstructured geoscience reports. For Portuguese, Lima de Oliveira et al. (2021) created REGIS (Retrieval Evaluation for Geoscientific Information Systems) a test collection containing over 20,000 scientific documents, 34 query topics and the corresponding relevance judgments for geological information retrieval evaluation, an important field often forgotten. On the other hand, Nunes et al. (2024) introduced GeoCorpus3, a geological NER resource containing 30 fine-grained entity classes including rock types (e.g., magmatic, metamorphic and sedimentary), as well as stratigraphical elements (e.g., sedimentary basin and geotectonic unit). For English, Lawley et al. (2022) have retrained language models on geoscientific documents and obtained better performing domain-specific geoscience language models, while Padarian and Fuentes (2019) introduced the GeoVec corpus, a corpus of 280,764 English geoscience articles and Wikipedia pages for word embeddings. Together, these studies define the current landscape of volcanology-specific and broader geological textual corpora for NLP.

As one can see, NLP research has not been focusing on extracting volcanological knowledge and reconstructing past eruptions from historical sources. On the other hand, this is exactly what several Italian volcanologists and geologists have been doing since the 1980s while compiling catalogues of geological disasters to extract this information from, as in the case of both the catalogue of Italian earthquakes by Boschi et al. (1995) and the BIBV database of Italian active volcanoes (Giannini and Paolillo, 2018), albeit without the help of computational methodologies. This being said, due to the enormous amount of time required to manually collect and analyse historical documents, the extraction of volcanological information from the available sources has always been restricted to the most documented eruptions and oriented towards specific phenomena, such as an eruption precursor (Bertagnini et al., 2006) or a given territory occupied by a lava flow (Branca et al., 2013).

Thanks to an extensive document collection, which could benefit from decades of expertise in the manual consultation of the relevant sources, and the help of LLMs, the CorVo project overcomes these limitations.

4. The Corpus

The CorVo corpus contains a balanced selection of historical documents, chronicles and monographs describing the three possible eruptive scenarios Vesuvius may experience in case of reactivation: (1) Plinian-type eruption, (2) fissural eruption, and (3) mixed (explosive-effusive) eruption.

[1] https://www.corvo-project.eu/

4.1. Document Selection

The document selection process was facilitated by the BIBV database (Giannini and Paolillo, 2018), which contains the bibliographical metadata (e.g., manuscript title, author, number of pages, exact location in the Italian National Libraries, etc.) of over 3,400 documents describing the historical eruptive activity of all active Italian volcanoes. Since the CorVo project is a case study restricted to Vesuvius, 180 Vesuvius-related documents – 176 of which from BIBV – were selected, traced back to a physical Library, scanned and transformed into machine-readable text format using OCR technology. The difficulties encountered at this stage will be dealt with in section 5.

In terms of additional criteria used for the selection of these 180 documents, we decided to favour descriptions of eruptive events as contemporary as possible to the eruption year, authored by eyewitnesses or educated individuals quoting eyewitnesses. The cultural backgrounds of said authors can vary greatly and range from well-known scholars of the time to administrators and clergymen. Furthermore, since not all authors have described the same phenomena from the same geographical point of view, the author's awareness of their geolocation and perspective was an important component in the selection process. In addition to documents and chronicles concerning the three main eruptions/scenarios of the Vesuvius, a small amount of treatises on the history of the volcano have been selected to obtain more detailed geographical information on the location.

This being said, the bulk of the CorVo corpus consists of documents that have already been individually used in previous volcanological studies to compare manually extracted historical data from written sources with volcanological data from fieldwork and deposits – see Sigurdsson et al. (1985) on the Vesuvian eruption of 79 AD and Cole and Scarpati (2010) on the eruption of 1944. We see the existence of these studies as an added value, since they will allow us to validate the findings emerging from our work with something we already know well.

However, unlike these works, the effort made in the CorVo project differs in both extent and methodology. For one, previous works usually exploited only a few historical sources at a time due to the limitations of manual consultation: an obstacle erased by the use of NLP tools. Moreover, most of these documents had not been digitised: one of our achievements. Another hurdle was identifying the right physical locations where the digitization could take place, minimising the number of Libraries and movements of the digitisation company, while acknowledging the restrictions of the equipment in terms of time, space and accessibility. Only printed (i.e., non-handwritten) texts were selected for the project to minimize the difficulties in producing OCR formats. Despite this, the quality of the documents varied greatly depending on their age, state of preservation and binding. Therefore, some compromises had to be made, and only texts from which sufficiently high-quality images could be obtained were selected.

4.2. Corpus languages

In terms of languages, the CorVo corpus is a multilingual diachronic corpus skewed towards Italian, as you can see from Table 1. According to Renzi and Salvi (2010), a distinction can be made between Old and Modern Italian around 1525, the year in which Bembo published his *Prose della volgar lingua*. Therefore, we can state that the vast majority of the documents of the CorVo corpus is in Modern Italian, which optimizes the trade-off between the amount of linguistic variation in our data and potential processing problems cause by such variation.

As for the other languages in the corpus, it is well-known that the activity of Italian volcanoes has been described over time by numerous observers of different native languages. Therefore, the CorVo project includes: eleven Latin texts published between the 16th and 17th century; two documents in 17th-century Spanish written by members of the Spanish bureaucratic-governmental apparatus during the period of Spanish domination of Southern Italy (17th-18th centuries); and finally six accounts in French, four in German, and five in English by various travelers on their Italian Grand Tour (18th-19th centuries).

To facilitate access to the corpus to Protection units, volcanologists and other stakeholders interested in obtaining information concerning past Vesuvian eruptions, the documents in foreign languages were translated using TranslateGemma (Finkelstein et al., 2026), with the exception of text in Latin, translated using ChatGPT 5.2.[2] Toponym extraction was performed on original-language texts; translations are provided only for accessibility (see Section 6).

5. Optical Character Recognition

The first phase of the digitization workflow consisted of Optical Character Recognition (OCR), aimed at converting scanned page images into machine-readable text. Given the heterogeneous nature of the document collection, the OCR process required careful configuration and post-processing

[2] https://openai.com/index/introducing-gpt-5-2/

	Century (publication year)					
Lang	16th	17th	18th	19th	20th	21st
it	-	26	61	38	26	1
en	-	-	4	1	-	-
de	-	-	-	4	-	-
fr	-	-	-	5	1	-
es	-	2	-	-	-	-
la	2	9	-	-	-	-

Table 1: Number of documents, collected by year of publication and language.

to address several structural and linguistic challenges.

5.1. Image Pre-processing

Prior to text recognition, the scanned images underwent a pre-processing stage to improve OCR accuracy. Historical documents often exhibit physical degradation such as ink stains, paper discoloration, bleed-through, and scanning artifacts. To mitigate these issues, image enhancement techniques were applied, including:

- Binarization and adaptive thresholding to increase contrast between text and background;
- Noise reduction filters to remove ink spots and speckle artifacts;
- Deskewing and geometric correction to address page misalignment;

The Python library OpenCV[3] has been used to perform the above-described pre-processing tasks.

5.2. Layout Analysis and Segmentation

A layout analysis phase was conducted to identify and segment distinct structural components of each page, including main text blocks, footnotes, marginal notes, tables, and other paratextual elements. Accurate segmentation was particularly important for:

- Separating footnotes from the main body text to preserve logical structure;
- Identifying tables and maintaining row–column relationships;
- Distinguishing text regions from non-textual elements such as decorative initials or graphical separators.

These steps have been mainly carried forward using Mistral OCR API[4] and were essential to reduce recognition errors introduced by degraded page conditions.

5.3. OCR issues

Although OCR is a well-established technology, the characteristics of the document collection significantly affected transcription accuracy. Apart from physical degradation, already described in Section 5.1, handwritten marginal notes and interlinear additions further complicated the process due to their variability and overlap with printed text.

Historical typefaces, obsolete characters, and non-standard orthographic conventions reduced recognition performance, especially when combined with print deterioration. The corpus' multilingual and diachronic nature—spanning modern and historical languages—introduced additional variability in vocabulary, syntax, and character sets, increasing the likelihood of errors (Dereza et al., 2024). Complex layouts also posed challenges: footnotes and tables were prone to segmentation errors and structural misinterpretation.

To improve textual quality, Large Language Models (LLMs) were employed in post-processing (Thomas et al., 2024). Rather than applying blind correction, the model performed contextual disambiguation to distinguish OCR artifacts from historically valid spellings, enabling selective corrections that reduced noise while preserving philological authenticity.

6. Geographic Entity Extraction

A key task in the information extraction pipeline developed within the CorVo corpus project consists in the identification and structuring of geographic entities (toponyms) mentioned in historical documents related to volcanic phenomena, and in particular to the eruption of the Vesuvius. These sources, digitized through scanning and processed with Optical Character Recognition (OCR), exhibit the typical challenges of historical textual data, including recognition noise, inconsistent spelling, and fragmented layout, all of which complicate automatic geoparsing.

Traditional Named Entity Recognition (NER) systems typically focus on broad-grained geographic categories such as cities, countries, rivers, mountains, and other major geopolitical or natural entities. While this level of abstraction is sufficient for many general-purpose NLP applications, it proves inadequate in the context of historical volcanological studies, where references often concern highly specific and localized spatial entities.

[3] https://opencv.org/

[4] https://mistral.ai/news/mistral-ocr

Figure 1: Two examples of degraded pages.

In particular, historical accounts of the eruption of the Vesuvius frequently mention micro-toponyms and fine-grained spatial markers that are crucial for reconstructing the spatial dynamics of the event, patterns of destruction, and settlement structures. Standard NER taxonomies do not usually distinguish such detailed categories, nor do they explicitly model the diversity of built, infrastructural, and geomorphological entities described in early modern sources.

A previous study in the volcanological domain (Paolillo et al., 2016) identified a set of entity categories that are especially relevant for this type of historical-spatial analysis. These include architectural structures (e.g., palace, villa, convent, church, basilica, chapel, fortress/tower, hospital, clinic, university, café, restaurant, thermal baths), infrastructural elements (e.g., aqueduct, bridge, road, mills), settlement and administrative units (e.g., district, neighborhood, rione), archaeological or historical sites (e.g., ruins, epitaph), as well as fine-grained geomorphological features (e.g., canal, ravine, ridge (cognoli), ditch, plain, sinkhole, rock formations). Additional categories include entities such as masseria, atrium, rocks (scogli), and other locally specific denominations frequently attested in the corpus.

As a consequence, the extraction step cannot be framed as a generic "extract all named entities" task. Instead, the Large Language Model is prompted with a structured and domain-aware instruction that explicitly enumerates the entity types of interest. The full prompt specification, including the detailed list of categories and formatting constraints, will be reported in the Appendix of the camera-ready. This design ensures that the model's output aligns with the analytical requirements of historical volcanological research, capturing both conventional toponyms and domain-specific spatial entities that would otherwise be overlooked by standard NER pipelines.

6.1. Toponym Extraction with Large Language Models

The first stage of the process focuses on the automatic extraction of geographic named entities, aiming at identifying place names occurring in the OCR-transcribed text. In this work, preliminary toponym recognition is performed using a large

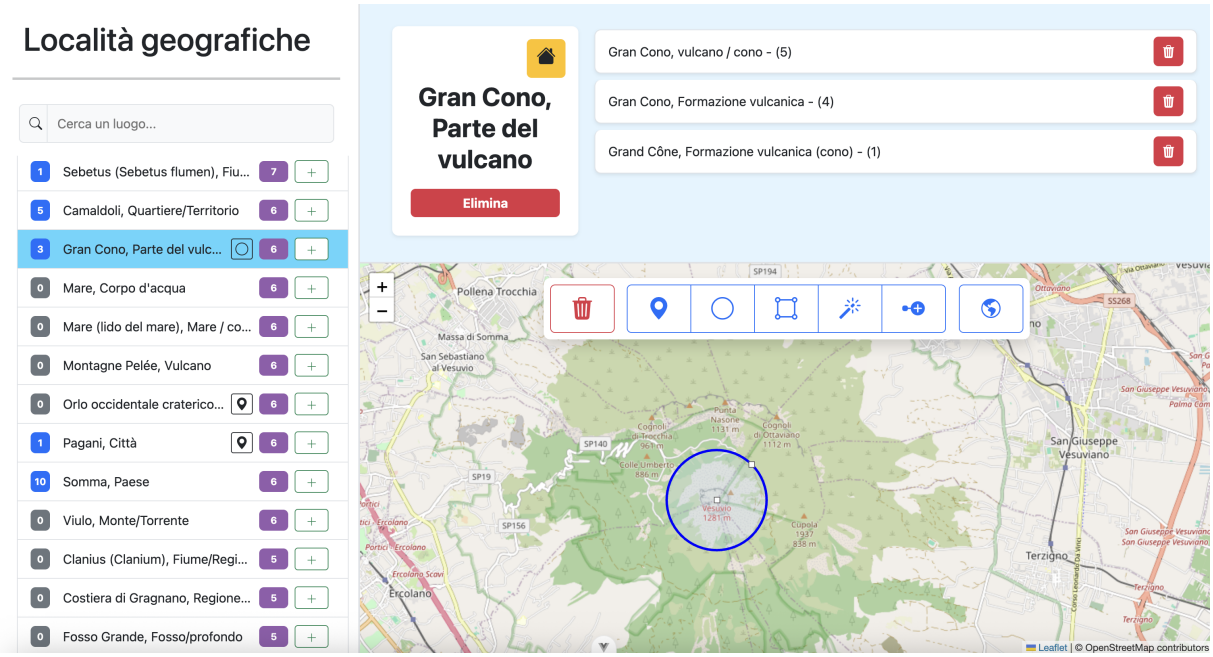

Figure 2: The annotation interface.

language model (LLM), specifically Mistral Large,[5] leveraging its ability to handle noisy textual input and to generalize over variant historical forms.

Unlike traditional Named Entity Recognition approaches, which often require substantial annotated corpora and struggle with OCR-induced errors, the LLM-based method provides a flexible alternative for bootstrapping entity candidates in low-resource and domain-specific contexts. The extracted toponyms are then collected as structured mentions, serving as input for the subsequent human-in-the-loop validation and georeferencing stage.

To illustrate the potential of extracted toponyms for volcanological interpretation, we consider a case study based on historical accounts of the 1631 Vesuvius eruption, focusing on the site of the Madonna del Soccorso in Portici. The automatic extraction of place names enables the alignment of textual evidence with known eruptive dynamics. In this case, multiple sources consistently associate the location with the passage of pyroclastic density currents (PDCs), described as fast-moving, high-temperature flows causing widespread destruction but locally diverted by morphology. Furthermore, the spatial clustering of place mentions supports the reconstruction of flow paths and depositional patterns, which can be compared with geological evidence. This example demonstrates how toponym extraction from historical narratives provides complementary data for volcanological analysis, contributing to the reconstruction of eruptive sequences, hazard mapping, and interpretation of past events beyond what can be inferred from deposits alone.

[5] https://mistral.ai/it/news/mistral-large

7. Human-in-the-loop Geospatial Annotation

Given the ambiguity and variability of historical toponyms, automatic extraction alone is insufficient to produce reliable geospatial references. Place names may refer to the same location under different spellings, may correspond to obsolete administrative divisions, or may be used metonymically. For this reason, the CorVo corpus integrates a dedicated annotation interface (see Figure 2) designed to support expert validation, entity consolidation, and geolocation.

The annotation tool provides two core functionalities:

- Toponym merging, allowing annotators to unify multiple extracted mentions that refer to the same geographic entity (e.g., alternate spellings or synonymous references).
- Geospatial grounding, enabling the association of each entity with explicit geographic coordinates.

7.1. Coordinate Selection and Georeferencing

The interface supports several mechanisms for selecting the geographic position of an entity. Annotators may:

- directly select a point on the interactive map,
- query the Google Maps geocoding service for automated suggestions,
- manually enter latitude and longitude coordinates,
- or navigate freely across the map until the correct location is found.

This flexibility allows efficient handling of both well-known locations and obscure or historically shifted place references.

7.2. Multi-geometry Spatial Annotation Modes

In addition to selecting coordinates, the platform allows annotators to represent geographic entities according to different spatial configurations. In particular, the interface supports multiple annotation geometries:

- Single-point annotation, suitable for discrete landmarks such as buildings or archaeological sites.
- Circular area annotation, defined by a center point and radius, typically used for cities or settlements with an approximate spatial extent.
- Polygonal area annotation, where users manually define the vertices of a polygon, appropriate for larger regions or administrative territories.
- Polyline annotation, representing linear geographic structures such as valleys, ravines, or volcanic flows.

By supporting different spatial representations, the system accommodates the diverse nature of geographic references in historical volcanological narratives.

The extraction process yielded 812 distinct geographical toponyms, corresponding to 437 unique locations (such as cities, regions, mountains, rivers).

The second extraction phase, targeting fine-grained spatial entities, resulted in 2,964 annotated locations, categorized as listed in Table 2.

Type	Count	Type	Count
palace	85	thermae	6
villa	88	ruins	62
epitaph	9	canyon	96
convent	95	trench	202
church	291	plain	380
basilica	2	floor	222
aqueduct	7	masseria	41
fortress/tower	87	mill	3
district	390	bridge	39
hospital	9	reef/rock	29
tavern	8	road/railway	311
university	53	chasm	449

Table 2: Number of specific entities found in the dataset.

7.3. From Annotated Toponyms to Diachronic Geovisualization.

Building upon the annotated dataset described above, we construct a diachronic geovisualization that represents how different eruptions of the Vesuvius are reflected in historical sources through the places they mention (see Figure 3). For each annotated entity, all textual occurrences across the corpus are aggregated and grouped according to predefined temporal intervals (e.g., centuries). However, a direct comparison of raw mention counts across periods would be misleading, as the number of available texts varies substantially from one century to another. A period with a larger number of documents would naturally produce higher absolute frequencies, independently of the actual prominence of specific locations.

To address this imbalance, we apply a *temporal normalization* procedure. For each time interval, the total number of geographic mentions in the corpus is computed. The frequency of each individual location within that period is then divided by the total number of mentions for the same period, producing a relative measure. Formally, if $f_{i,t}$ denotes the number of mentions of location i in period t, and T_t the total number of geographic mentions in that period, the normalized value is computed as:

$$n_{i,t} = \frac{f_{i,t}}{T_t} \tag{1}$$

This normalization yields comparable proportions rather than absolute counts, enabling a meaningful diachronic comparison across periods with heterogeneous textual coverage. The normalized values are subsequently used to generate an interactive cartographic representation in which each annotated location is positioned according to its validated coordinates and visually encoded with its temporal distribution. In the implemented visualization, each location is represented by a georeferenced marker enriched with temporal information, allowing users to perceive at a glance how

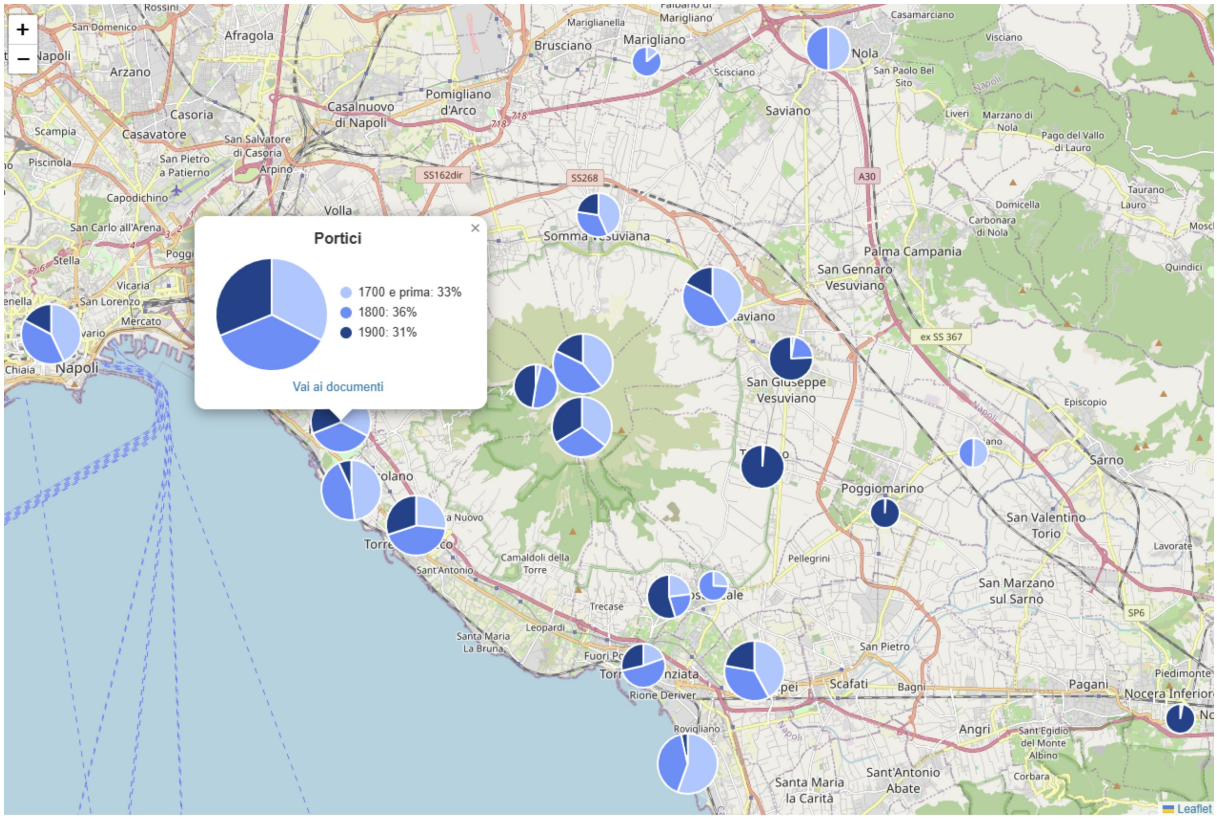

Figure 3: The diachronic visualization.

the relative prominence of that place varies across historical periods.

Through this pipeline, the manual geospatial annotations serve not only as a corrective layer for entity extraction, but as the enabling infrastructure for higher-level interpretative representations. The resulting system does not attempt to reconstruct the physical evolution of eruptions directly; rather, it reconstructs their *textual-spatial footprint* across time, offering an intuitive way to explore how different areas around the Vesuvius emerge, recede, or persist in historical narratives.

The manual annotation process was conducted by a single expert annotator with extensive domain knowledge in volcanology and long-standing fieldwork experience. One of the authors is a specialist in Vesuvian volcanology and has contributed to multiple scientific studies on the area, including the production of geological maps (Principe et al., 2013), providing further domain validation of the annotation process.

8. Release

The CorVo corpus and the software components developed within the project are released on Github[6] under open and permissive licenses to ensure transparency, reproducibility, and reuse by the research community.

8.1. Data

The released dataset includes:

- the OCR-cleaned textual transcriptions of the digitized historical documents contained in the corpus, with the exception of a limited number of recent documents whose copyright has not yet expired;
- the extracted geographic entities, including both toponyms and fine-grained domain-specific spatial entities;
- the associated categorical labels (e.g., architectural, infrastructural, geomorphological, administrative categories);
- the validated geographic information resulting from the human-in-the-loop annotation process, including coordinates and geometry types (point, circular area, polygon, polyline).

[6] https://github.com/ziorufus/corvo-project

In total the raw dataset contains around 16,000 pages and 4,160,000 words, in which 2,964 toponyms are annotated with spatial information (437 unique locations).

All data will be distributed under the Creative Commons Attribution 4.0 International license (CC BY 4.0),[7] allowing reuse, redistribution, and adaptation for both research and educational purposes, provided appropriate credit is given to the project.

8.2. Software

In addition to the dataset, the source code of the software tools developed for the project is publicly released in the same Github project. These include, both implemented with a Vue/JavaScript frontend and a Python/SQLite backend:

- the geographic annotation interface, designed to support entity consolidation and geospatial grounding;
- the diachronic geovisualization interface, enabling temporally normalized spatial exploration of the corpus.

The source code is distributed under the Apache License 2.0,[8] to allow modification, redistribution, and integration into other systems, including commercial applications, while preserving attribution and license notices.

9. Conclusion and Future Work

This project aimed to evaluate the financial, technical, and methodological feasibility of using documents describing past volcanic activity to address pressing issues in modern volcanology, such as the assessment and mitigation of volcanic risk in Italy. For this purpose, Vesuvius was selected, due to the abundance and diachronic nature of the descriptions of its activity; the three most representative eruptions of the three possible reactivation scenarios were selected, and 180 related digitized documents were collected, from which geographical entities were extracted, and a graphical interface was developed to represent them.

The use of historical linguistics with the most up-to-date technologies has been shown to be an effective tool for preserving precious linguistic data that survived from old periods. The greatest difficulties have thus far been encountered in building the corpus of documents. This difficulty could be largely overcome in the future through the ongoing digitization of their documentary collections, a process now underway at many libraries. This will enable the specialist to work on select texts to form targeted corpora, such as the one we have built in this project, without the burden of researching and digitizing the original texts.

Beyond the specific case study on Vesuvius, an important outcome of this work lies in the generality of the proposed framework. The combination of OCR preprocessing, LLM-based post-correction, domain-aware entity extraction, and human-in-the-loop geospatial annotation defines a modular pipeline that can be readily adapted to other domains and datasets. In particular, the tools and methodologies described here enable researchers to initiate similar studies from scratch on different volcanic systems (or, more broadly, on other types of geographically grounded historical phenomena) without requiring pre-existing structured resources. By reconfiguring the entity schema and adapting the annotation guidelines to a new domain, the same infrastructure can support the systematic extraction and diachronic spatial analysis of knowledge from heterogeneous historical corpora, thus opening the way to comparative and cross-regional studies.

10. Limitations

The corpus is restricted to printed documents of sufficient physical quality, excluding handwritten materials and severely degraded sources, which limits coverage of the historical record. OCR errors, although mitigated through preprocessing and LLM-based correction, may still propagate to downstream extraction tasks. The toponym extraction pipeline relies on large language models without gold-standard annotated training data, and therefore may introduce inconsistencies or omissions, particularly for rare or highly ambiguous entities. Furthermore, georeferencing depends on expert validation, which, while improving reliability, constrains scalability. Finally, the diachronic visualization reflects the textual footprint of eruptions rather than their physical evolution, and is therefore influenced by historiographical bias and uneven document distribution across centuries.

An additional limitation concerns the absence of a systematic quantitative evaluation of the NLP components in the current phase of the project. While the pipeline integrates OCR preprocessing, LLM-based post-correction, and domain-aware toponym extraction followed by expert validation, no gold-standard annotated subset has yet been constructed to compute standard metrics such as precision, recall, F1-score, or character error rate reduction. As a result, the performance of the automatic extraction stages has not been formally benchmarked against existing NER systems or al-

[7] https://creativecommons.org/licenses/by/4.0/

[8] https://www.apache.org/licenses/LICENSE-2.0

ternative approaches. Future work will address this gap by developing a manually annotated evaluation dataset, enabling rigorous assessment of extraction accuracy, error propagation across pipeline stages, and the overall reliability of the system.

11. Bibliographical References

Pietro Bembo. 1525. *Prose di M. Pietro Bembo nelle quali si ragiona della volgar lingua*. Giovanni Tacuino, Venezia. Editio princeps.

Margherita Berardi, Luigi Santamaria Amato, Francesca Cigna, Deodato Tapete, and Mario Siciliani de Cumis. 2022. Text Mining from Free Unstructured Text: An Experiment of Time Series Retrieval for Volcano Monitoring. *Applied Sciences*, 12(7).

A. Bertagnini, R. Cioni, E. Guidoboni, M. Rosi, A. Neri, and E. Boschi. 2006. Eruption early warning at Vesuvius: The A.D. 1631 lesson. *Geophysical Research Letters*.

E. Boschi, G. Ferrari, P. Gasperini, E. Guidoboni, G. Smriglio, and G. Valensise. 1995. *Catalogo dei forti terremoti in Italia dal 461 a.C. al 1980*. ING-SGA.

S. Branca, E. De Beni, and C. Proietti. 2013. The large and destructive 1669 AD eruption at Etna volcano: reconstruction of the lava flow field evolution and effusion rate trend. *Bulletin of Volcanology*, page 1–16.

S. Chen, W. Hua, X. Liu, X. Deng, X. Zeng, and J. Duan. 2022. Chinese Fine-Grained Geological Named Entity Recognition With Rules and FLAT. *Earth and Space Science*, 9(12).

P.D. Cole and C. Scarpati. 2010. The 1944 eruption of Vesuvius, Italy: combining contemporary accounts and field studies for a new volcanological reconstruction. *Geological Magazine*, 147(3):391–415.

Runyu Fan, Lizhe Wang, Jining Yan, Weijing Song, Yingqian Zhu, and Xiaodao Chen. 2020. Deep Learning-Based Named Entity Recognition and Knowledge Graph Construction for Geological Hazards. *ISPRS International Journal of Geo-Information*, 9(1).

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report.

Emanuela Guidoboni and John E Ebel. 2009. *Earthquakes and tsunamis in the past: A guide to techniques in historical seismology*. Cambridge University Press.

Takashi Kudo and Hideo Hoshizumi. 2006. Construction of a new catalog of eruptive events during the last 10,000 years in Japan (V101-P029) (poster session) (abstract). In *Abstracts, Japan Geoscience Union Meeting (CD-ROM)*, volume 2006, pages V101–P029, Japan. Japan Geoscience Union. Poster abstract.

Christopher J.M. Lawley, Stefania Raimondo, Tianyi Chen, Lindsay Brin, Anton Zakharov, Daniel Kur, Jenny Hui, Glen Newton, Sari L. Burgoyne, and Geneviève Marquis. 2022. Geoscience language models and their intrinsic evaluation. *Applied Computing and Geosciences*, 14:100084.

Kai Ma, Miao Tian, Yongjian Tan, Qinjun Qiu, Zhong Xie, and Rong Huang. 2023. Ontology-Based BERT Model for Automated Information Extraction from Geological Hazard Reports. *Journal of Earth Science*, 34:1390–1405.

Rafael O. Nunes, Andre S. Spritzer, Dennis G. Balreira, Carla M. D. S. Freitas, and Joel L. Carbonera. 2024. An Evaluation of Large Language Models for Geological Named Entity Recognition. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 494–501.

J. Padarian and I. Fuentes. 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. *SOIL*, 5(2):177–187.

Annarita Paolillo, Claudia Principe, Marina Bisson, Roberto Gianardi, Daniele Giordano, and Sonia La Felice. 2016. Volcanology of the Southwestern sector of Vesuvius volcano, Italy. *Journal of Maps*, 12(sup1):425–440.

Claudia Principe. 1990. ADSVI – Archivio Documenti Storici sui Vulcani Italiani (IGG internal technical report).

Claudia Principe, Daniele Giordano, Marina Bisson, Annarita Paolillo, and Roberto Gianardi. 2013. Volcanological map of the South-Western sector of Vesuvius between Torre del Greco and Ercolaneum.

Qinjun Qiu, Zhong Xie, Liang Wu, and Liufeng Tao. 2020. Automatic spatiotemporal and semantic information extraction from unstructured

geoscience reports using text mining techniques. *Earth Science Informatics*, 13.

Lorenzo Renzi and Giampaolo Salvi. 2010. *Grammatica dell'italiano antico*. Il Mulino, Bologna. 2 vols.

H. Sigurdsson, S. Carey, and W. Cornell. 1985. The Eruption of Vesuvius in A.D. 79. *National Geographic Research*, 1(3):332–387.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

B. Voight, E.K. Constantine, S. Siswowidjoyo, and R. Torley. 2000. Historical eruptions of Merapi Volcano, Central Java, Indonesia, 1768–1998. *Journal of Volcanology and Geothermal Research*, 100(1):69–138.

12. Language Resource References

Oksana Dereza, Deirdre Ní Chonghaile, and Nicholas Wolf. 2024. "To Have the 'Million' Readers Yet": Building a Digitally Enhanced Edition of the Bilingual Irish-English Newspaper an Gaodhal (1881-1898). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 65–78, Torino, Italia. ELRA and ICCL.

S. Giannini and A. Paolillo. 2018. BIBV V: "Bibliografia storica dei vulcani attivi italiani". Database (IRIS CNR). Accessed via IRIS CNR Institutional Research Information System.

Lucas Lima de Oliveira, Regis Kruel Romeu, and Viviane Pereira Moreira. 2021. REGIS: A Test Collection for Geoscientific Documents in Portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2363–2368, New York, NY, USA. Association for Computing Machinery.