The Foggia Occupator Corpus: Digitisation, Annotation, and Computational Analysis of an Occupation-Era Newspaper (1945–1946)

Michele Ciletti
University of Foggia
Via Arpi 176, 71121 Foggia (FG), Italy
michele.ciletti@unifg.it

Abstract

Historical newspapers are crucial sources yet often remain undigitised or lack machine-readable text. We present the Foggia Occupator corpus, a linguistically enriched, openly licensed resource built from twenty-two issues (Dec 1945–Aug 1946) of a weekly newspaper produced by U.S. personnel in occupied Foggia, Italy. High-resolution scans were processed via OCR with LLM-assisted correction (GPT-4o) and full human verification, then segmented into 874 articles ($\approx$ 216k tokens). We annotate topics, named entities and typed relations via a semi-automatic pipeline with manual reconciliation, and perform argument mining on civics- and conflict-related content, yielding 1,735 arguments. The entity–relation layer supports network analyses that reveal sparse, modular structures linking military units, civic bodies, and social life. We release TEI-XML with entity spans, JSON article files with metadata, CSVs of entities/relations with temporal counts, and an arguments JSON, all under a Creative Commons 4.0 licence. Beyond documenting an in-between moment of reconstruction, the resource enables benchmarking for OCR-robust NER/RE and studies of framing, stance, and community structure in post-war local media.

Keywords: historical newspapers, corpus creation, digitisation, cultural analytics

1. Introduction

Archival newspapers serve as vital assets for linguistic and historical investigations. Yet, a significant number remain untouched by digitization efforts or Optical Character Recognition (OCR), meaning vast amounts of cultural heritage remain restricted to specific physical libraries and museums. The *Foggia Occupator* represents one such example: a weekly periodical authored and circulated by American troops stationed in Foggia, Italy, during the final stages of the Second World War. Prior to this work, the contents of this publication had remained largely unextracted and unexamined. The current study introduces a new linguistic dataset derived from the *Foggia Occupator*, made publicly available via a Creative Commons 4.0 License. This corpus features detailed annotations covering named entities, relationships, topics, and arguments. The article further offers an initial computational exploration of this annotated collection and outlines numerous directions for future application, such as opportunities in benchmarking, diachronic linguistics, frame analysis, and network analysis. Prior to detailing the construction of the dataset, a short overview of the historical background is essential.

1.1. Historical Context

As Allied forces traversed the Strait of Messina on September 3, 1943, and arrived at Salerno six days later, they encountered a nation where core political structures were rapidly collapsing. The dismissal of Benito Mussolini on July 25 preceded a covert armistice (signed in Cassibile on September 3 and announced on September 8). Subsequently, the royal administration led by Marshal Pietro Badoglio declared hostilities against Germany, gaining recognition from Washington and London as the exclusive legitimate power south of the German lines (Harris, 1957). Effectively, Badoglio's jurisdiction was limited to the "Kingdom of the South," a geographical band extending from Sicily up to the Neapolitan outskirts, whereas officials of the Allied Military Government of Occupied Territories (AMGOT) wielded actual administrative authority across the majority of the localities (Fisher, 1950). The Allied administration held several urgent objectives. They aimed to maintain the flow of military logistics toward the Gustav and, eventually, Gothic defensive lines, while simultaneously averting societal breakdown in districts deeply depleted by fascist seizures, German destruction, and intense aerial attacks (Gribaudi, 2005). Everyday life during the occupation involved a mix of collaboration and friction. Gestures of charity, including troops sharing food provisions, occurred frequently and cultivated positive relations (Daniele and Ghezzi, 2017). Servicemen of Italian-American descent fulfilled a vital function as cultural and linguistic bridges. They frequently visited family members and ancestral towns, a practice that helped put a human face on the foreign troops (Pretelli, 2020). Conversely,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6731–6739
11–16 May 2026.

clashes emerged due to cultural misinterpretations stemming from divergent societal customs, rivalry over limited goods, and the fundamental power disparities inherent in an occupation (Williams, 2013). A particularly severe aspect of this period involved pervasive sexual assaults committed by Allied troops against local women, leaving behind a deep-seated trauma and enduring societal repercussions (De Paola, 2018). Furthermore, American biases regarding Italians—frequently characterizing them as incompetent and excessively emotional, or alternatively as naive, childlike sufferers of Fascist rule—unavoidably influenced their perspectives (Buchanan, 2008). Military propaganda occasionally bolstered such prejudices, although it generally distinguished the Italian populace from the presumed innate aggression attributed to the Germans or Japanese (Buchanan, 2008). Foggia held significant strategic value due to an extensive network of airbases situated across the neighboring Tavoliere plain. These aviation facilities saw use by Axis and Allied militaries alike at various phases of the conflict, elevating the municipality to a prime tactical target. Before falling to the Allies in 1943, the region suffered intense aerial bombardments that led to severe civilian death tolls (Tempesta, 1995). Following the loss of roughly twenty thousand citizens to these air raids, the town was subsequently garrisoned by a vast array of American flight personnel, ground crews, quartermaster divisions, and Red Cross volunteers.

1.2. The *Foggia Occupator*

Following Germany's surrender in 1945, staff at the Foggia Army Air Base started a weekly English-language newspaper called the Foggia Occupator. A local printing shop produced the paper, and it circulated across the base and the surrounding town (Berardi, 2022). The publication was mostly aimed at military personnel waiting to return home. However, it regularly covered civic notices, employment issues, supply shortages, and regional celebrations. Through crime updates, classified ads, sports announcements, and news about schools reopening or theater performances, the newspaper became closely tied to the city's everyday life. Its generally lighthearted style stood out against the stricter political and military decrees. The very first edition of the Foggia Occupator clearly shows the publication's many roles and its wide variety of topics. The article "A Day at the EM Red Cross" offers a highly detailed description of a visit to the Enlisted Men's Red Cross Club by photographer Giacomo Scirpoli, who worked for the paper. Using a series of vivid images, the piece highlights the club's social and leisure activities, ranging from piano playing to bridge games. This provides proof of how the American troops tried to build a comfortable, familiar setting far away from their own country. The text additionally mentions a fashion show hosted at the club. This points to an effort to bring back everyday routines and provide chances for recreation with a distinct social and stylish focus. The Christmas 1945 edition, which has a mostly cheerful mood, is similarly fascinating. The cover itself shows a soldier wearing a Santa Claus suit next to a young girl, representing how the interactions between the Americans and the people of Foggia could be friendly and helpful. An important article titled "It Was the Week Before Christmas" describes the holiday spirit through various scenes, including children posing while waiting for Santa, locals praying at Christmas Mass, and a musical band. This holiday special gives attention to religious practices as well. Along with the Mass, it includes a brief story about the nativity scene, a tradition that was mostly unknown in the United States. This shows that military leaders valued keeping familiar religious customs alive, and it highlights an exchange between cultures in a setting of shared traditions that was never guaranteed to happen. Figure 1 displays the cover page of this important edition. Negative events are also present in the

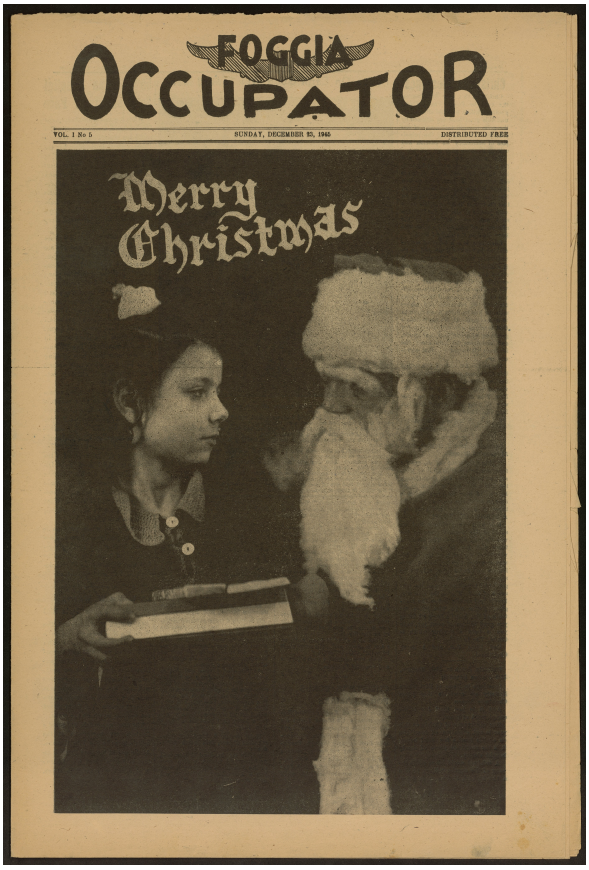

Figure 1: Front cover of a *Foggia Occupator* edition, published on December 23, 1945.

publication. Another edition features a piece called

"Civilian Assaulted By Soldier's In Off Limits Bar," detailing an event at a local business where American troops badly injured an Italian man named Gino Bagano. It is significant that the newspaper's editors did not avoid covering the frictions and disputes that sometimes emerged between the residents and the foreign forces. Following the attack, a group of people assembled and nearly started a riot, though the military eventually broke up the crowd. The newspaper gives plenty of attention to major political milestones in Italian history. For instance, it covers the June 2, 1946 referendum and the very first elections of the Italian Republic, including pictures of women in Foggia voting for their first time. The main story in that edition, "Italy Votes For Republic," documents the outcome of the referendum, highlighting the significance of this historic event and sharing a desire for future national harmony. An additional significant article is the report "Italy's Fourth King Is Its Last." This piece looks back at the reign of Umberto II, known as the "May King," who had to give up his throne to the temporary head of state after the referendum results. The story acts as a short summary of a period in the nation's history that was officially ending. Readers can gather information about the status of women as well. A unique perspective found in multiple editions is the "Girl About Town" column, penned in a cheerful and easygoing style by young female writers. They discuss social gatherings in the city and share gossip regarding well-known locals. The paper features happy pages dedicated to marriages, honoring the weddings of American servicemen and young Italian women. When the last edition of the Foggia Occupator came out on August 30, 1946, the official occupation systems were already scaling back. The Allied Control Commission, which had changed its name to the Allied Commission, had handed most of its administrative duties over to the Italian authorities. Civilian prefects took back their standard powers, and the initial republican referendum on June 2, 1946, had ended the monarchy. However, on a smaller scale, the newspaper captured the feel of a transitional period. During this time, a heavily damaged provincial city hosted the operational center of an international air campaign along with the early signs of post-fascist public life. The publication documented the interactions between troops and locals as they navigated issues regarding wages, housing, safety, and entertainment. Some of these dealings were peaceful, while others were tense. Ultimately, the paper gave readers at the time a way to picture an Italy characterized by the everyday compromises of rebuilding, rather than simply by loss or rescue.

1.3. Research Questions

Currently, twenty-two physical copies of the Foggia Occupator are kept in local libraries, with the oldest dating back to December 1945 (Berardi, 2022). While these editions have been scanned, the actual text from the articles has never been processed or extracted. The newspaper provides a highly unique viewpoint. It is heavily biased and sometimes unclear, and it serves as a firsthand observer of events that rarely receive attention. Because of this, we established several precise research goals. The purpose was to use the newspaper's material to better understand the social and political environment of Southern Italy right after the war. In particular, the questions included:

- RQ1: How can the Foggia Occupator be turned into an easily accessible linguistic resource, enabling both computational studies on language and historical research?
- RQ2: How is the nascent Italo-American community in Foggia portrayed in the newspaper? What power dynamics, representations of persons, and framings of news took shape? Which individuals, institutions, and topics were represented?

1.4. Background and Related Work

Numerous efforts currently focus on digitizing historical newspapers to improve public access and lay the groundwork for detailed cultural research. For instance, *Chronicling America*—a joint initiative of the Library of Congress and the National Endowment for the Humanities (NEH)—is a massive project aimed at digitizing important American newspapers from 1763 to 1963 (Culpepper, 2007). Furthermore, the *Atlas of Digitised Newspapers and Metadata* (Beals et al., 2020) represents a valuable undertaking that catalogs various digital archives while highlighting the differences in their metadata, layout structures, and licensing rules, thereby facilitating comparative studies. The technical phase of digitizing newspapers generally requires scanning physical pages or microfilms, indexing the content, and using Optical Character Recognition (OCR) to turn the images into searchable text. Ensuring high precision is still difficult, especially when dealing with older publications that feature degraded print or outdated typefaces. Initiatives such as *NewsEye* (Doucet et al., 2020) have specifically targeted the enhancement of OCR outputs for these historical documents. Once established, these digital archives frequently serve as comprehensive textual datasets that researchers analyze using computational approaches in fields like digital history, cultural analytics, and computational social science.

As an example, a broad analysis of lynching reports in American newspapers identified the "Ida B. Wells effect" by combining regular expressions with Latent Dirichlet Allocation (Wells et al., 2025). Comparable methods were used on Dutch periodicals to measure the growing connection between past and present using the phrase "n years ago," supporting concepts of "everyday memory" (Huijnen, 2025). Additionally, visual analyses are common in this domain: research by Fyfe and Ge (2018) examined 140,000 illustrations from nineteenth-century British papers using pixel clustering to uncover previously unnoticed visual trends, including syndicated portrait styles and nighttime crime scenes. The methodology applied to the Foggia Occupator shares a similar twofold objective: applying modern digitization methods to build a freely available linguistic dataset, while concurrently using the resulting texts to investigate historical and social research questions via computational techniques.

2. Methodology

2.1. Data Collection

Carrying out computational research on the text of the Foggia Occupator (RQ1) and making these contents publicly readable (RQ2) required acquiring high-resolution scans of the original papers. We achieved this by partnering with the "La Magna Capitana" library, which preserves all twenty-two remaining editions (Berardi, 2022). The length of these issues ranges from 4 to 15 pages, featuring a mix of text and images, including headlines, articles, captions, and cartoons. We had to establish a working definition for our textual units. Consequently, we segmented all printed material into "articles," defined here as continuous text blocks grouped by a shared subject and physical closeness on the page. This approach produced extremely brief items, like image captions, alongside extensive multi-page features. Occasionally, the dividing lines between articles are ambiguous and subjective. We must emphasize that this division was performed strictly to organize the data and may not align with the original editors' layout decisions. Ultimately, this segmentation yielded a total of 874 articles. Our digitization workflow utilized Optical Character Recognition (OCR) (Stockman and Shapiro, 2001), supported by a two-step error correction phase. The initial text refinement was handled by OpenAI's (2024) GPT-4o using a few-shot prompting approach (Brown et al., 2020). Subsequently, human reviewers manually checked the text to guarantee precision. We chose this hybrid method largely to save time. Because human reviewers evaluated every article anyway, the language model's contribution might seem redundant. Still, having the human annotators review a partially corrected dataset greatly sped up the workflow while maintaining a high standard of quality. We selected GPT-4o based on an internal test comparing multiple language models against a manually corrected baseline of two newspaper issues. GPT-4o outperformed the alternatives, successfully fixing over 90% of the text errors. We note that this evaluation represents a particular point in time, meaning more recent models could yield even greater accuracy. Additionally, we found that feeding raw OCR output to the model for correction worked better than relying on the language model to perform the OCR task directly. Because the final dataset is intended for open distribution under a flexible license, using a proprietary AI model did not present any ethical issues. A recognized drawback of our OCR pipeline was its difficulty with sports score tables, which were sometimes lost or distorted. The final text corpus contains 216,934 tokens and is stored as JSON files. Each entry includes fundamental metadata, such as the publication date, issue number, page, and the specific position of the article on that page. We combined these metadata attributes to generate a distinct identifier for every article.

2.2. Topic Classification

In order to guide as judiciously as possible the construction of the corpus and its computational analysis, an initial investigation was carried out into the principal topics present in the Foggia Occupator's articles. Such an approach ensured greater precision in identifying categories of entities and relations, as well as in interpreting the broader contexts in which they are embedded. After preliminary coding sessions, several categories of articles were identified, namely "politics," "war operations," "sports," "international news," "Italian news," "social events," "military life," "letters," "fiction," "celebrations and festivities," "cinema and theatre,"humanitarian operations," "conflicts," "miscellaneous". Furthermore, additional categories were created for articles belonging to specific recurring columns: these were "Local Items," "Static from your A.E.S.," "Nights 'round Foggia," "Girl About Town," and "A Lire for Your Thoughts". The first two mainly covered the daily life of U.S. soldiers, while the third one focused on entertainment events happening in the city. The fourth one commented on social events and the lives of renowned members of the community, almost like a "gossip" section. The fifth one collected short thoughts of passersby on a variety of topics. Each article was assigned one or more of these categories by two independent annotators. An agreement of $\kappa = 0.89$ was reached, and any difference was reconciled after a joint discussion.

2.3. Named Entity Recognition

To investigate the community structures and relationships associated with the events covered in the Foggia Occupator, we needed to extract the named entities mentioned throughout the texts. Through expert consultation and a close reading of the material, we established a targeted list of entity types: animal, date, event, journal column, location, person, organization, sports team, and miscellaneous. We similarly defined a set of relationship categories to link these entities. The chosen relation labels were: "works for", "part of", "located in", "travelled to", "reported on", "attended the event", "participated in the event", "occurred at", "occurred on", "interacted with", "related to", "played against", "writes column", and "was in conflict with". We extracted these entities and their connections using another two-step workflow, starting with automated detection and finishing with manual review. Several language models received detailed instructions alongside examples and were evaluated against a baseline of two fully annotated issues. This test measured their ability to identify entities, categorize them, and map the relationships within an article to our established labels. GPT-4o emerged as the most capable tool for this task. Afterward, comprehensive human review allowed us to fix any automated mistakes and unify the spelling of entities appearing under various names. This standardization was particularly important for people (frequently listed with military titles), dates (written in multiple formats), and organizations (occasionally referred to by acronyms). We consistently adopted the most frequent spelling found in the dataset to replace alternative versions. However, our dataset might still contain slight inaccuracies, such as distinct text strings failing to resolve to a single real-world entity. These lingering issues stem from original typographical errors, inconsistent spelling choices by authors with varying levels of formal education, and a minor number of uncorrected OCR mistakes. To facilitate deeper research into the overall community framework and the smaller dynamics shaping these interactions, we modeled the extracted Foggia Occupator entities as a network. In this graph, individual entities function as nodes, and the relationships identified within a single article act as edges. This process generated a comprehensive network comprising 6,297 nodes and 9,094 edges.

2.4. Argument Mining

In order to systematically investigate the arguments and framing of such an inevitably biased news source (RQ2), a further annotation step was conducted. Following the procedure detailed by Oberbichler et al. (2025), an LLM assistant was asked to identify and annotate each argument in a subset of articles—namely, pieces whose topic was previously tagged as either "politics,"war operations," "international news," "Italian news," "military life," "letters," "celebrations and festivities," "humanitarian operations," "conflicts". This screening mainly excluded articles focused on sports, entertainment, and social events, which could have constituted a source of noise. Once again, several models were tested on a manually-annotated sub-corpus, and GPT-4o obtained the best performance. For each argument that was identified, the model was asked to classify it as either implicit or explicit, provide an explanation of its reasoning, and synthesize the main claim being advanced. The model, equipped with a complex system prompt, was instructed to also annotate dubious cases, trying to be as thorough as possible: this was done because, as in previous steps, human annotators reviewed the entirety of the outputs and corrected them, sometimes even deleting altogether an identified argument. This process ensured that no misleading or inaccurate argument was left. However, there is the possibility that a few arguments may have been missed entirely in the initial automated step.

3. Results

3.1. Corpus Creation

Addressing RQ1, the project generated a primary, tangible result by building an open-access dataset that includes the complete text of all surviving articles from the Foggia Occupator—an archive that was previously unavailable. Structured in JSON, the collection features various metadata attributes (such as publication date, article ID, volume, page number, position on the page, and assigned topics) and was shared on Zenodo using a Creative Commons 4.0 license. (Ciletti, 2026) Information concerning named entities and their connections was also made public. Specifically, the entities and their linking relationships were detailed in CSV files, providing data on their overall frequency in terms of quantity and time (including total appearances and the specific issues they show up in). This setup is intended to enable an easy reconstruction of the community networks found within the publication. Additionally, a distinct JSON file was compiled to hold all the annotated arguments. Each argument includes metadata detailing its source article, exact text, explanation, primary claim, and a unique identifier. These documents were subsequently stored under the exact same access conditions as the main collection.

3.2. Statistics and Insights

Turning to RQ2, an in-depth review of the gathered articles and arguments, combined with a numeri-

cal summary of the categorized topics and entities, yielded various statistics and observations regarding the Occupator's material. Table 1 provides a broad summary of the data within the collection. Table 2 details the figures for each categorized topic, including the quantity and kinds of arguments they contain. Table 3 presents overall data concerning the annotated arguments. Table 4 offers details on the categories of entities present in the text, Table 5 provides similar data regarding relationships, and Table 6 highlights the general architecture of the network.

Statistic	Value
Number of Articles	874
Earliest Publication Date	1945-12-16
Latest Publication Date	1946-08-30
Total Tokens	216,934
Total Types	17,166
Type-Token Ratio	0.079
Mean Tokens per Article	248.21
Median Tokens per Article	177.5
Total Sentences	16,205
Mean Sentence Length (Tokens)	15.62
Median Sentence Length (Tokens)	15.33
Number of Unique Topics	19
Mean Label Cardinality	1.34
Median Label Cardinality	1.0
Mean Label Density	0.071

Table 1: Descriptive statistics of the article corpus.

4. Discussion

The corpus offers a compact view of an in-between period, in which everyday routines and the legacies of war coexisted. The topic distribution, argument profiles, and interaction networks all point to a newspaper that balanced entertainment and information for servicemen with a sustained, if selective, engagement with local civic life. The topical mix is skewed toward "Military Life" (341 articles) and "Sports" (217), with a second tier comprising "Social Events," "Politics," "Italian News," and "Cinema and Theatre." This is consistent with a publication primarily aimed at a military audience awaiting repatriation. Yet the presence of political coverage and Italian civic affairs, as well as columns embedded in the city's social rhythms, indicates a porous boundary between base culture and city life. The descriptive statistics reinforce this picture: short, notice-like pieces coexist with longer features (median 177.5 tokens per article; mean 248.2), while the low type–token ratio (0.079) reflects both corpus size and the formulaic patterns of periodical writing; at the same time, the range of topics and named entities suggests a heterogeneous content base. The argument inventory (1,735 total) leans toward implicitness (59.1%), a finding in line with the paper's generally breezy tone and reliance on narrative framing. Argument density varies by topic. Only a subset of articles had arguments annotated 2.4, but it is interesting that many articles cover widely different topics, which means that even articles on sports and social events sometimes contain political arguments. "Letters" also exhibits elevated argumentative density (3.72), often with a personal, situational stance rather than editorial generalisation - these articles represent one of the few cases where civilians' voices were reported.

The balance of explicit and implicit reasoning resonates with the newspaper's position within an occupied city: much persuasion is achieved through selection, tone, and juxtaposition instead of formal editorialising. The Christmas issue described in the historical overview is a case in point. Photographs, captions, and vignettes establish an argument about normalcy and conviviality without prolonged discursive claims. Where conflict is addressed—such as the assault reported in an "off-limits" bar—assertions move closer to the surface.

4.1. Entities, Network structure and community boundaries

People make up the largest part of the entity collection (49.2%), followed by organizations (16.0%) and locations (11.4%). The list of relationships is primarily driven by institutional activities and events: "Participated in Event" (21.2% of connections), "Part Of" (17.2%), "Located In" (13.2%), and "Works For" (12.7%) constitute the majority of the ties. These connections align with the publication's role in delivering news about military postings, assignments, formal ceremonies, sporting events, and official notices. Edges indicating disputes—such as "Conflict With" (0.2%) and "Interacted With" (2.9%)—are quite rare in terms of numbers. Tensions may still exist within the text; however, these low figures simply reflect how events are described and which connections are formally captured during annotation. The small percentage of "Conflict With" is additionally due to the limited scope of the relationship categories and the fact that extraction was done at the article level, a process that inherently favors consistent and standardized ties. Sports reporting makes a clear impact. The "Played Against" relationship represents a minor portion of the connections (1.6%) but remains highly focused and repetitive, aligning with the 217 articles on the subject. Specifically, teams and match schedules create dense, fairly regular subgraphs with distinct two-part structures. A noticeable percentage of entities

Topic	Articles	Total Arguments	Explicit Rate (%)	Avg. Arguments per Article
Military Life	341	1,141	38.7	3.35
Sports	217	70	28.6	0.32
Social Events	95	100	35.0	1.05
Politics	66	313	48.9	4.74
Cinema and Theatre	66	82	42.7	1.24
Italian News	54	252	43.7	4.67
Humanitarian Operations	49	165	37.0	3.37
Celebrations and Festivities	41	102	25.5	2.49
International News	38	172	47.1	4.53
Conflicts	36	135	36.3	3.75
Miscellaneous	34	17	17.6	0.50
Fiction	30	6	50.0	0.20
Girl About Town	23	25	32.0	1.09
Letters	18	67	40.3	3.72
A Lire For Your Thoughts	18	56	33.9	3.11
Static From Your AES	16	104	51.0	6.50
Local Items	13	32	53.1	2.46
Nights Round Foggia	13	5	40.0	0.38
War Operations	4	5	60.0	1.25

Table 2: Argument distribution and characteristics across different article topics, sorted by the number of articles per topic.

Statistic	Value
Number of Arguments	1,735
Mean Arguments per Article	4.05
Median Arguments per Article	3.0
Explicit Argument Count	709
Implicit Argument Count	1,026
Explicit Argument Percentage	40.86%
Mean Argument Length (Tokens)	39.98
Median Argument Length (Tokens)	33.0

Table 3: Descriptive statistics of arguments identified within the corpus.

Entity Type	Count	Percentage (%)
Person	3,099	49.21
Organization	1,007	15.99
Location	717	11.39
Event	344	5.46
Miscellaneous	326	5.18
Date	317	5.03
Unknown	288	4.57
Sports Team	158	2.51
Journal Column	26	0.41
Animal	15	0.24

Table 4: Distribution of identified entity types.

Relation Type	Edge Count	Mean Weight	Percentage of Edges
Participated In Event	1,985	1.10	21.19
Part Of	1,610	1.20	17.19
Located In	1,238	1.15	13.22
Works For	1,185	1.12	12.65
Related To	743	1.03	7.93
Occurred On	600	1.06	6.40
Travelled To	528	1.13	5.64
Occurred At	428	1.11	4.57
Reported On	337	1.04	3.60
Interacted With	273	1.04	2.91
Attended Event	239	1.00	2.55
Played Against	148	1.24	1.58
Writes Column	35	1.89	0.37
Conflict With	19	1.00	0.20

Table 5: Distribution and weight statistics of relation types in the network.

fall under "Unknown" (4.6%) or "Miscellaneous" (5.2%), pointing to remaining OCR errors, unclear references, and boundary cases within the labeling system. Efforts to standardize the data successfully reduced the alternative spellings for names and acronyms, though some minor variations persist. The overall network is relatively sparse (average degree 1.44; clustering coefficient 0.034) and

Network Metric	Value
Number of Nodes	6,297
Number of Edges	9,094
Average Degree	1.444
Average Weighted Degree	1.666
Network Diameter	22
Modularity	0.752
Average Clustering Coefficient	0.034
Average Path Length	7.749

Table 6: Key topological properties of the analyzed network.

highly modular (modularity 0.752), featuring a large diameter (22) and an average path length of 7.75. This type of layout is highly predictable: entities appear together inside individual articles and create temporary, tightly knit local connections (like events, team lists, and committees), frequently failing to form closed triangles between different texts. The strong modularity points to distinct, separate communities that likely represent regular newspaper columns, localized sports networks, military units, and city departments operating as independent social spheres. The extended path lengths suggest the presence of a few weak connecting links—such as frequently mentioned officials, organizations, or recurring sections—that allow distant groups to interact across different issues.

4.2. Methodological reflections and limitations

Segmentation at the article level is an operational choice that fits the material but remains interpretive. OCR plus assisted correction yields high-quality text but is imperfect with tables and certain proper names. Topic and argument annotations show strong agreement, yet topic overlap and column idiosyncrasies mean that some assignments remain debatable. Finally, the corpus covers twenty-two issues from December 1945 to August 1946; it offers a specific snapshot rather than a complete run. Inferences about temporal dynamics should therefore be cautious unless paired with explicit time-series analyses.

4.3. Potential Uses and Future Perspectives

The resource enables several strands of research. As an open, TEI- and JSON-structured corpus with entity and relation layers, it can serve as a benchmark for information extraction on historical English with Italian names and institutions, where OCR residue, acronyms, and mixed orthographies create realistic difficulty. The argument layer supports studies of stance and framing in post-war military periodicals, including comparisons between explicit and implicit persuasion across topics and columns. The network export allows social and institutional mapping at city scale, from sports leagues to Red Cross and municipal circuits, with the possibility of overlaying events and administrative reorganisations. Historians of language could examine lexical routines of occupation-era journalism—register, euphemism, and loanwords—while social historians could track the representation of women, Italian–American intermediaries, and civic rituals. Because metadata include dates and page positions, the corpus can be used for diachronic microanalyses of the referendum period, demobilisation, and the fading of formal occupation structures. Methodologically, the dataset is well-suited for evaluating robust NER and relation extraction under OCR noise and for testing domain adaptation of sequence labellers on historical prose. Looking at the future, two directions appear particularly promising. First, enrichment of the current layers: cross-article and cross-issue coreference resolution; geocoding of locations; temporal normalisation of dates; and a broader relation schema including affiliation change, endorsement, and conflictual interactions. Argument annotations could be extended with evidence spans and rhetorical devices, enabling more fine-grained studies. Second, contextual expansion. Integrating page-layout features and images would open the way to multimodal analyses of captioned photographs and visual rhetoric. Comparative corpora from neighbouring localities or from other Allied bases would situate Foggia within a wider communicative ecology, inviting controlled comparisons of tone, topics, and community structure. Where feasible, linking entities to authority files would facilitate prosopographic research and cross-repository discovery.

5. Conclusion

The Foggia Occupator corpus provides, for the first time, a machine-readable, openly licensed record of a brief but dense episode in post-war Southern Italy. Its topic distribution, argument profiles, and network structure reveal a newspaper that framed daily life on the base and in the city through events, institutions, and compact, often implicit claims. The data confirm the coexistence of convivial routines and moments of tension, while the sparse, modular interaction graph points to distinct communities bound by columns, units, and civic bodies. Beyond documenting this local world, the resource offers a practical testbed for information extraction, argument mining, and historical network analysis on challenging historical material. It is a foundation that can be extended—across layers, modalities, and neighbouring archives—to support both com-

putational studies and close historical reading of an in-between moment when reconstruction began to take shape.

6. Acknowledgments

The author thanks Dr Nadia Di Leo for helping annotate part of the data. He also thanks the staff of the library *La Magna Capitana* for their support in collecting the textual data of the *Foggia Occupator*.

7. Bibliographical References

Melodee Beals, Emily Bell, Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Mila Oiva, Sebastian Pado, et al. 2020. The atlas of digitised newspapers: Reports from oceanic exchanges.

Gabriella Berardi. 2022. Biblioteca digitale e studi storici locali: il progetto della magna capitana di foggia. *DigItalia*, 17(1):203–212.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Andrew Buchanan. 2008. 'good morning, pupil!'american representations of italianness and the occupation of italy, 1943—1945. *Journal of contemporary history*, 43(2):217–240.

Michele Ciletti. 2026. Mikcil/foggia_occupator: v2.0.0 the foggia occupator dataset (1945-1946).

Jetta Culpepper. 2007. Chronicling america: Historic american newspapers. *Reference Reviews*, 21(7):52–53.

Vittorio Daniele and Renato Ghezzi. 2017. The impact of world war ii on nutrition and children's health in italy. *Investigaciones de Historia Económica*.

Stephanie Lauren De Paola. 2018. *Sexual Violence, Interracial Relations, and Racism During the Allied Occupation of Italy: History and the Politics of Memory*. Ph.D. thesis, Fordham University.

Antoine Doucet, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, Guenter Muehlberger, Eva Pfanzelter, Marie-Eve Therenty, et al. 2020. Newseye: A digital investigator for historical newspapers. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020*.

Thomas R Fisher. 1950. Allied military government in italy. *The Annals of the American Academy of Political and Social Science*, 267(1):114–122.

Paul Fyfe and Qian Ge. 2018. Image analytics and the nineteenth-century illustrated newspaper. *Journal of Cultural Analytics*, 3(1).

Gabriella Gribaudi. 2005. *Guerra totale: Tra bombe alleate e violenze naziste. Napoli e il fronte meridionale 1940-1944*. Nuova Cultura. Bollati Boringhieri, Torino. Brossura.

Charles Reginald Schiller Harris. 1957. *Allied military administration of Italy, 1943-1945*, volume 1. HM Stationery Office.

Pim Huijnen. 2025. Everyday memory: a computational analysis of changing relations between past and present in dutch newspapers in the 20th century. *Digital Scholarship in the Humanities*, 40(Supplement_1):i27–i38.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Sarah Oberbichler, Johanna Mauermann, The Trung Tran, and Carlos-Emiliano González-Gallardo. 2025. Studying model design biases in llms for multilingual historical newspaper extraction; the messina earthquake case study. In *International Conference on Theory and Practice of Digital Libraries*, pages 263–286. Springer.

Matteo Pretelli. 2020. Hollywood's depiction of italian american servicemen during the italian campaign of world war ii. *European journal of American studies*, 15(15-2).

George Stockman and Linda G Shapiro. 2001. *Computer vision*. Prentice Hall PTR.

Odorico Tempesta. 1995. *Foggia: nelle ore della sua tragedia*, volume 11. Edizioni del Rosone.

Rob Wells, Kathy Roberts Forde, Sean Mussenden, Mohamed Salama, and Sasha Allen. 2025. The ida b. wells effect: A novel computational analysis of us newspaper lynching coverage, 1805–1963. *Journalism Studies*, 26(7):854–879.

Isobel Williams. 2013. *Allies and Italians Under Occupation: Sicily and Southern Italy 1943-45*. Springer.