


Using Machine Learning and LLM to Identify Rhetorical Devices in Historical Spanish/Catalan Texts



Exploring 15th-Century Castilian Spanish & Catalan
Women's Writing



Why Study Historical Texts?

01

Preserve cultural and linguistic heritage

02

Explore rhetorical style and authorial intent

03

Underrepresented women's voices in Iberian history

04

Manual annotation is limited; we use AI for scale and insight





Three Part Project



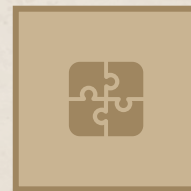
LLM

Used Gemini to generate rhetorical device labels from texts given



NLP

Trained a RoBERTa model to classify rhetorical devices using structured XML data



Clustering

Clustered annotated sonnets to uncover rhetorical patterns across time periods



LLM (Data Generation)

- **Input:** XML/PDF historical texts
- **Preprocessing:** Cleaning, normalization, chunking
- **Prompting:** Custom taxonomy, annotation rules
- **Output:** XML, CSV, JSON with rhetorical labels
- **Runtime:** 397.6 seconds (relatively efficient)
- **Difficulties:** Input-text dependency, fails to generate correctly when English Appears.
- **Accuracy:** correct for 18 out of the 20 samples we've manually looked at

Final Summary:

Files processed: 5
Chunks processed: 50
Total annotations: 628
Total processing time: 397.6 seconds
Average time per chunk: 8.0 seconds

Annotation Distribution:

Rhetoric: 328 (52.2%)
Lexis: 105 (16.7%)
Verb_Functions: 87 (13.9%)
Notes: 54 (8.6%)
Genre: 54 (8.6%)

Output files saved to: /content/drive/MyDrive/project_data

XML: annotations.xml
CSV: annotations.csv
JSON: ml_training_data.json

Pipeline completed successfully!

Pipeline execution ended

A Taxonomy Library

```
TAXONOMY = {  
  "Genre": ["sermon", "vision", "consolatory_treatise", "vita", "confession",  
    "prayer_devotional", "disputation", "performance", "music", "parable",  
    "sacraments", "didactics", "scripture", "autobiography", "dialogue",  
    "exemplum", "epistle", "theater_performance", "poetry", "music_laic",  
    "testament_will", "speculum", "relacion", "genealogy", "testimony",  
    "disputation_laic", "pastoral"],  
  
  "Rhetoric": ["captatio", "colloquialism", "pathos", "logos", "ethos", "allegory",  
    "ekphrasis", "metaphor", "sign_symbol", "exegesis", "parallelism",  
    "didactics_rhet", "invective", "amplification", "anaphora", "antithesis",  
    "apostrophe", "exclamation", "polypoton", "hypophora", "orality_literacy"],  
  
  "Lexis": ["place", "person", "name", "gender", "building", "family", "authority",  
    "body", "soul", "material", "nature", "animal", "time", "age", "food",  
    "festivity", "latin", "sins", "virtues", "vice"],  
  
  "Verb_Functions": ["affirm", "negate", "question", "exhort", "narrate", "describe",  
    "command", "promise", "lament", "praise", "blame", "warn",  
    "supplicate", "advise", "instruct", "prophecy", "invoke", "confess"],  
  
  "Notes": ["hkbtext", "hkbobs", "intertext", "classical_sources", "biblical_sources",  
    "contemporary_sources", "metatextual"]  
}
```

Prompt

```
{taxonomy_desc}
```

ANNOTATION RULES:

1. Only use the exact subtypes listed above
2. Identify spans of 2-50 words that clearly represent each category
3. Focus on significant literary, rhetorical, and semantic elements
4. Spans should not overlap
5. Return ONLY valid JSON array format

EXAMPLES:

- "Teresa de Cartagena" → `{{"span": {{ "text": "Teresa de Cartagena", "start": 0, "end": 18}}, "type": "Lexis", "subtype": "person"}}`
- "en el año de 1425" → `{{"span": {{ "text": "año de 1425", "start": 6, "end": 17}}, "type": "Lexis", "subtype": "time"}}`
- "Dios todopoderoso" → `{{"span": {{ "text": "Dios todopoderoso", "start": 0, "end": 17}}, "type": "Lexis", "subtype": "authority"}}`
- "convento de San Francisco" → `{{"span": {{ "text": "convento de San Francisco", "start": 0, "end": 25}}, "type": "Lexis", "subtype": "building"}}`

TEXT TO ANALYZE (`{{len(text_chunk)}}` characters):

```
{text_chunk}
```

Return JSON array with this exact format:

```
[{"span": {{ "text": "exact_text_from_above", "start": start_index, "end": end_index}}, "type": "Category", "subtype": "exact_subtype"}]
```




	A	B	C	D	E	F	G	H	I	J
1	chunk_index	annotation_text	start_position	end_position	type	subtype	text_length	chunk_length	processing_time	source_files
2	0	A Transcription with	0	96	Notes	metatextual	95	1447	7.851140976	Arboleda1119.xml; M
3	0	el qual con Arboleda	219	280	Lexis	person	61	1447	7.851140976	Arboleda1119.xml; M
4	0	Seyendo apasyoñad	281	317	Rhetoric	pathos	36	1447	7.851140976	Arboleda1119.xml; M
5	0	e espiritual consolaç	476	505	Genre	consolatory_treatise	29	1447	7.851140976	Arboleda1119.xml; M
6	0	porque despedidos d	663	726	Genre	prayer_devotional	63	1447	7.851140976	Arboleda1119.xml; M
7	0	q' e q' u e es verdader	727	756	Rhetoric	logos	29	1447	7.851140976	Arboleda1119.xml; M
8	0	the truth and reueali	758	830	Notes	metatextual	72	1447	7.851140976	Arboleda1119.xml; M
9	0	Teresa refers to the t	846	906	Rhetoric	logos	60	1447	7.851140976	Arboleda1119.xml; M
10	0	virtuosa señora	940	955	Lexis	gender	15	1447	7.851140976	Arboleda1119.xml; M
11	0	Arboleda is a treatise	956	1012	Genre	consolatory_treatise	56	1447	7.851140976	Arboleda1119.xml; M
12	0	Doña Juana de Men	1030	1051	Lexis	person	21	1447	7.851140976	Arboleda1119.xml; M
13	0	Admiración operum	1092	1113	Notes	intertext	21	1447	7.851140976	Arboleda1119.xml; M
14	0	que la niebla de trist	1115	1197	Rhetoric	metaphor	82	1447	7.851140976	Arboleda1119.xml; M
15	0	e con vn espeso toru	1198	1249	Rhetoric	metaphor	51	1447	7.851140976	Arboleda1119.xml; M
16	0	from the very beginn	1250	1410	Notes	metatextual	160	1447	7.851140976	Arboleda1119.xml; M
17	0	me lleuo a vna ynsul	1412	1446	Genre	vision	34	1447	7.851140976	Arboleda1119.xml; M
18	1	Here is the first exam	0	97	Rhetoric	allegory	97	832	5.027142048	Arboleda1119.xml; M
19	1	employing metaphor	99	170	Rhetoric	metaphor	71	832	5.027142048	Arboleda1119.xml; M
20	1	the vse of latin lends	214	273	Rhetoric	ethos	59	832	5.027142048	Arboleda1119.xml; M
21	1	and demonstrates the	274	321	Rhetoric	ethos	47	832	5.027142048	Arboleda1119.xml; M
22	1	This is an example of	323	372	Rhetoric	allegory	49	832	5.027142048	Arboleda1119.xml; M
23	1	Teresa	29	35	Lexis	person	6	832	5.027142048	Arboleda1119.xml; M
24	1	Seidenspinner, Deye	389	424	Notes	contemporary_source	35	832	5.027142048	Arboleda1119.xml; M
25	1	donde tan tos años l	456	523	Rhetoric	pathos	67	832	5.027142048	Arboleda1119.xml; M
26	1	throughout the treati	524	631	Genre	consolatory_treatise	107	832	5.027142048	Arboleda1119.xml; M
27	1	and her resistance to	782	831	Verb_Functions	negate	49	832	5.027142048	Arboleda1119.xml; M
28	2	jamaspude yo ver pe	53	194	Rhetoric	pathos	141	1376	8.719111204	Arboleda1119.xml; M
29	2	me sintiendo alunbra	312	366	Rhetoric	pathos	54	1376	8.719111204	Arboleda1119.xml; M
30	2	Teresa's explicit con	548	624	Genre	vita	76	1376	8.719111204	Arboleda1119.xml; M
31	2	A reference to rebirth	816	919	Notes	intertext	103	1376	8.719111204	Arboleda1119.xml; M
32	2	Y tu, niño, seras llam	923	1138	Notes	biblical_sources	215	1376	8.719111204	Arboleda1119.xml; M
33	2	Zaçarlah	888	896	Lexis	person	8	1376	8.719111204	Arboleda1119.xml; M
34	2	Luke 179	911	919	Lexis	place	8	1376	8.719111204	Arboleda1119.xml; M
35	2	niño	929	933	Lexis	gender	4	1376	8.719111204	Arboleda1119.xml; M
36	2	Señor	994	999	Lexis	authority	5	1376	8.719111204	Arboleda1119.xml; M
37	2	tiñieblas y en sombra	1061	1092	Lexis	body	31	1376	8.719111204	Arboleda1119.xml; M
38	2	jamaspude yo ver	53	70	Verb_Functions	negate	17	1376	8.719111204	Arboleda1119.xml; M
39	2	pudiese llegar	160	174	Verb_Functions	affirm	14	1376	8.719111204	Arboleda1119.xml; M
40	2	me sintiendo alunbra	312	333	Verb_Functions	narrate	21	1376	8.719111204	Arboleda1119.xml; M
41	2	pudiese poder	495	508	Verb_Functions	affirm	13	1376	8.719111204	Arboleda1119.xml; M

Open Sources

Project repository:

https://github.com/MehtA-AI-AIMLResearchBootcamp25/Mid-term_Group2

Final Product

Rhetorical Device Annotation Pipeline (Python Package)

This repository hosts a **Python-based code package** designed to automate the identification and annotation of rhetorical devices and other literary elements in Medieval Castilian Spanish and Catalan texts.

Package Contents: The package includes:

- The core pipeline code used to **generate text annotations with an LLM**.
- An importable module (`medieval_annotator.py`) allowing for flexible integration into other Python scripts.

Getting Started: Please refer to the **README.md** file in this repository for detailed instructions on how to set up, use, and run the code.

Future Developments: The current package is limited to plain text input, but future enhancements will expand its capabilities to support diverse document formats commonly found in medieval text archives. Planned developments include PDF text extraction functionality to process digitized manuscripts, support for additional file formats (DOCX, HTML, EPUB, TEI-XML), and OCR integration for scanned documents.



NLP

- Tried RoBERTa Large → RAM issues
- Switched to RoBERTa Base
- Processed and filtered XML data (≥ 10 samples/class)
- Trained for 50 epochs on 6 rhetorical labels

```
training_args = TrainingArguments(  
    output_dir='./results',  
    num_train_epochs=50,  
    per_device_train_batch_size=8,  
    per_device_eval_batch_size=8,  
    warmup_steps=10,  
    weight_decay=0.01,  
    logging_dir='./logs',  
    logging_steps=10,  
    eval_strategy="epoch",  
    save_strategy="epoch",  
    save_total_limit=1,  
    load_best_model_at_end=True,  
    metric_for_best_model="eval_loss",  
    greater_is_better=False  
)
```





Limitations of NLP and LLM

NLP

- Struggles with subtle rhetorical nuance and figurative language.
- Relies heavily on surface-level patterns, not deep context.
- Performs poorly on low-resource, imbalanced datasets.

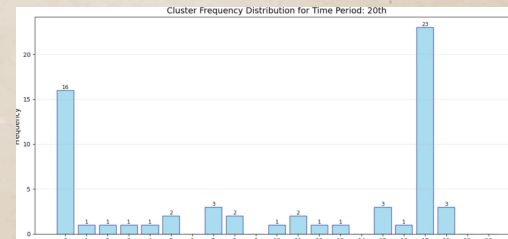
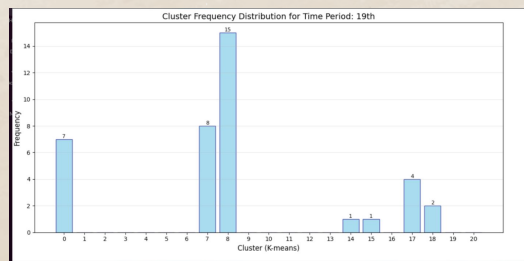
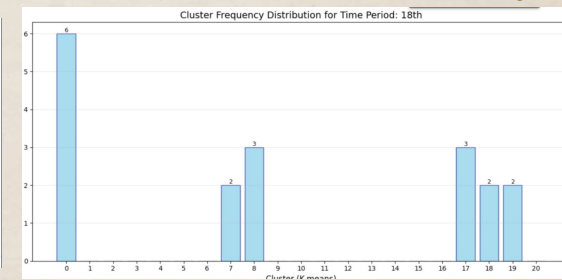
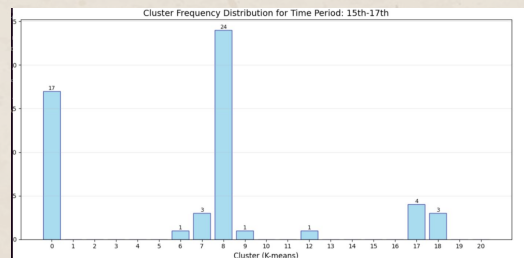
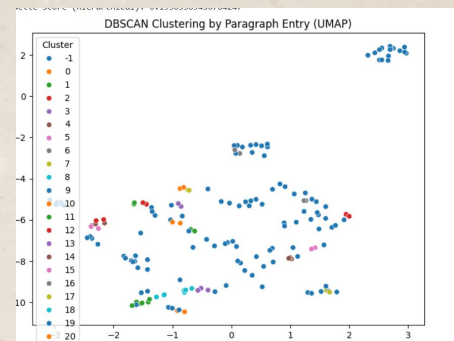
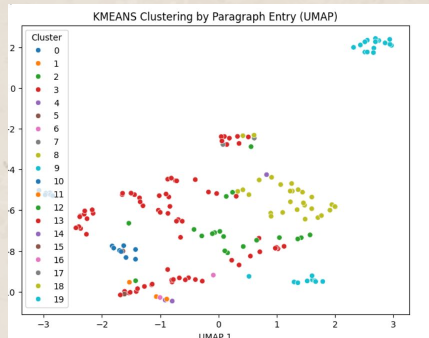
LLM

- May hallucinate or generate inaccurate annotations.
- Limited by context window, missing long-range rhetorical structures.
- General-purpose models lack domain-specific fine-tuning for medieval texts.



Corpus-Wide Analysis

- Used Diachronic Spanish Sonnet Corpus (15th–20th c.)
- Annotated with Gemini for rhetorical sequences
- Applied K-means, Hierarchical, and DBSCAN clustering
- Observed dominant rhetorical patterns by period





What are our **results**?

LLMs **outperform** traditional NLP models when it comes to identifying **complex rhetorical devices**, making them better suited for analyzing historical texts rich in nuance.

Their ability to process large volumes of data enables **scalable annotation**, especially valuable for amplifying underrepresented voices in literary history. This project lays the groundwork for **future enhancements** such as fine-tuning, human-in-the-loop review systems, and integration with TEI standards, marking a significant step forward in the field of digital humanities research.





Thanks!

Do you have **any** questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**