



An Exploratory Data Analysis of Space in Spanish-Language Literature

Jennifer Isasi (Pennsylvania State University)

Joshua Ortiz Baco (University of Tennessee, Knoxville)

The study of space in literature with computers, in English. One Example:

- Wilkens (2021): “Are American authors homers? Do they devote too much of their attention to American concerns and settings? Is American literature as a whole different from other national literatures in its degree of self-interest?”
 - 100,000 volumes, American, British, and other English-language fiction from 1850 to 2009
 - “Collectively, they show that American literature [...] was consistently and significantly more domestically oriented than was British fiction of the same period” (77) with many “but”
 - It began to show more attention to global locales after 1945

Why? Digital Humanities/Computational methods allows us to extent readings to large(r) groups of texts



**Can we do it on a smaller scale
with literature in Spanish?
That is, within the *humanidades digitales*?**

The study of space in literature with computers, in Spanish. Some approaches:

- One work: Calarco utilizes Recogito and Visone to georeference the text, developed an interactive map, and other visualizations of the *Libro de Alexandre* (a medieval Spanish epic poem about Alexander the Great), in order “to bring the cultural world and the cosmovision of the geographic space [...] to the modern public ” (2022)
- Several works: *Desenrollando el cordel (2020–2024)* “is a research project of the Spanish Unit of the Department of Romance Languages and Literatures of the University of Geneva” that is studying 915 chapbooks from the 19th Century and from all over Spain and is fine tuning several models to detect placenames in their collections



How can we start at bigger scale?

Exploratory Data Analysis

“Exploratory data analyses are strategies that summarize or otherwise **reveal features of interest within a dataset** which are not likely visible through traditional close reading. [...] researchers can make more informed decisions when selecting a method or approach for tackling their research question, and it may help to identify new research questions altogether.” (Wilkinson Saldaña 2018)

We can ask, for example:

- Which spaces do authors pay more attention to in Spanish-language (travel) literature?
- Which spaces are not explored in this corpus?
- How do historical events affect the swift in spaces mentioned?

- Can we actually use a method to respond to our questions at scale?

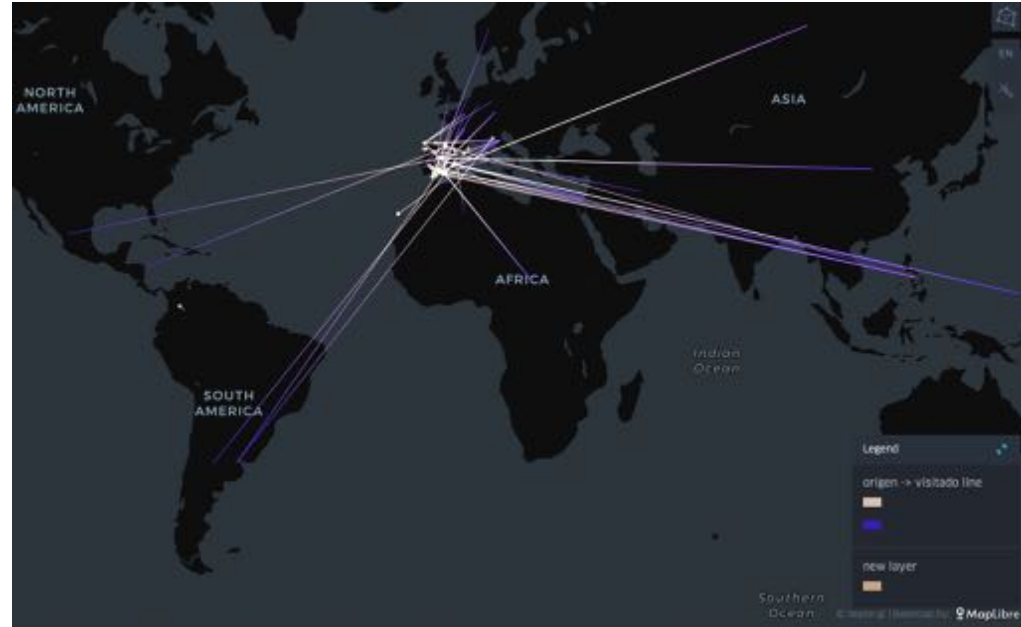
Available travel literature datasets

- “Bibliography of electronically available Spanish–language travel narratives”
 - Assembled by Jörg Lehmann and Konstantin Krechting at the University of Tübingen (Germany)
 - 350 texts; published between 1522 and 2016; by 278 authors
 - Texts are available in HTML (a few) or links to PDF files (most)
- “The Spanish Travelers” (*Viajeros españoles*) collection
 - Created and hosted by the Biblioteca Virtual Cervantes (Spain)
 - “Literary texts about trips made by Spaniards throughout the world with an exploratory, scientific, sociological, cultural, political, evangelizing or recreational objective”
 - 85 texts, published from 1612 to 1950; by 57 authors
 - Texts are available in HTML (a few) or PDF (most)

Exploratory data analysis of corpora metadata: Is Hispanic literature “as a whole” interested on its own spaces?



The Tübingen dataset (1833 to 1999)
Literary mobility happened from Europe to
America, and within the American continent



The CVC dataset (1805 to 1950)
Literary mobility happened within Europe
mostly (mainly in Spain) but also to the
Americas, and other colonized spaces

Corpus for our preliminary study/experiment



The Tübingen dataset (1833 to 1999)

Why? Even though the majority of writers are identified as Europeans, given the diversity of the genre, we can consider this a Hemispheric Spanish-language literature corpus, that represents linguistic and cultural diversity of the Spanish-speaking world - and presents a good challenge for *humanidades digitales*

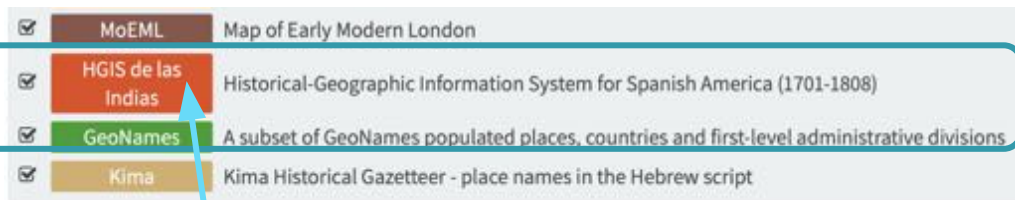


How can we start at bigger scale?
**Three approaches to
exploratory analysis on texts-as-data**

GNER with Recogito

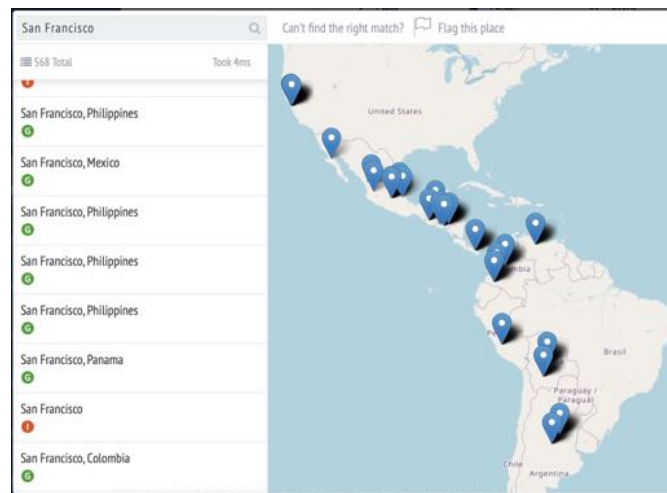
Annotation of 2 clean texts from Project Gutenberg

1. Run the NER included with the integrated Stanford CoreNLP with the Spanish-language model, that “identif[ies] entities against all available 9 authority files:”



Collaboration between Pelagios, WHG (PA), CAICYT-HD (Argentina) and LLILAS Benson (TX), HGIS de las Indias (Austria) and Brumfield Labs (TX)

2. Manually confirm the location of each named entity that is identified (and add, or delete)



GNER with Recogito

Exploratory data analysis of two texts:



Del Plata al Niágara by Paul Groussac (1879)

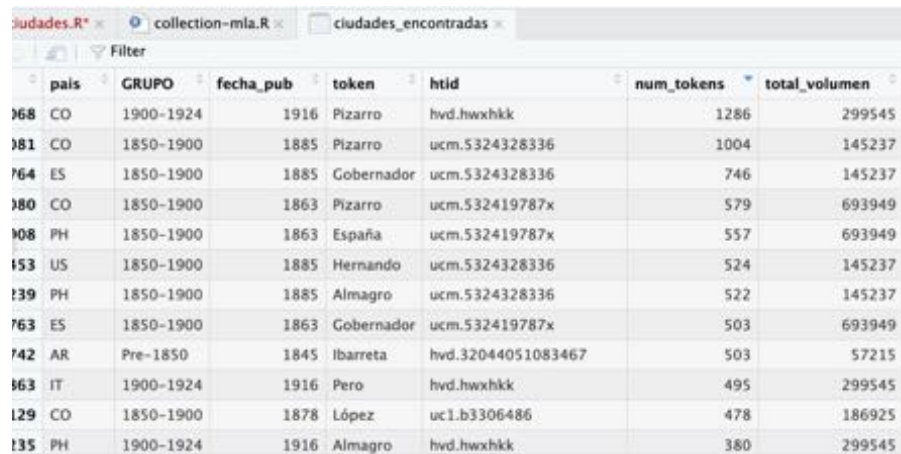


Viajes por Europa y América by Gorgonio Petano y Mazariegos (1858)

Geonames with R for HathiTrust Collection

Extraction of places from 55 OCREd works available in HathiTrust

1. Create a collection of texts
2. With R scripts
 - a. Call to HathiTrust token based data
 - b. Find places based on regex searches that contrast a list of places to the list of tokens
 - c. Map if desired



	pais	GRUPO	fecha_pub	token	htid	num_tokens	total_volumen
068	CO	1900-1924	1916	Pizarro	hvd.hwxhkk	1286	299545
081	CO	1850-1900	1885	Pizarro	ucm.5324328336	1004	145237
064	ES	1850-1900	1885	Gobernador	ucm.5324328336	746	145237
080	CO	1850-1900	1863	Pizarro	ucm.532419787x	579	693949
008	PH	1850-1900	1863	España	ucm.532419787x	557	693949
053	US	1850-1900	1885	Hernando	ucm.5324328336	524	145237
039	PH	1850-1900	1885	Almagro	ucm.5324328336	522	145237
063	ES	1850-1900	1863	Gobernador	ucm.532419787x	503	693949
042	AR	Pre-1850	1845	Ibarreta	hvd.32044051083467	503	57215
063	IT	1900-1924	1916	Pero	hvd.hwxhkk	495	299545
029	CO	1850-1900	1878	López	uc1.b3306486	478	186925
035	PH	1900-1924	1916	Almagro	hvd.hwxhkk	380	299545

- Little to no way to correct (except discarding obvious mistakes)
- Works well with a reduced corpus with locations already identified
 - González, 2024, Analysis of places named in 100 Ecuadorian novels from 1861 to 1949



**Too small scale or
nearly impossible to correct**

Spanish Language NER Models

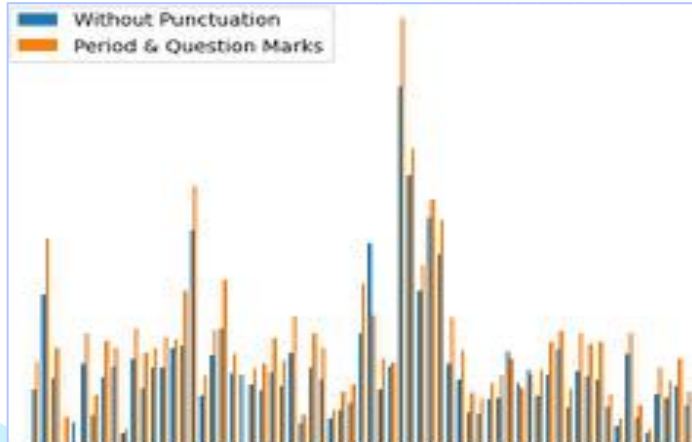
NLP Frameworks:

- a. SpaCy:
 - i. Three NER capable models trained on news data.
- b. Flair NLP Framework:
 - i. Spanish NER using Document-Level Features trained and evaluated on CoNLL-03 shared task datasets
- c. BERTIN:
 - i. RoBERTa-base model trained in Spanish using “the multilingual variant of the C4, the Colossal, Cleaned version of Common Crawl's web crawl corpus.”

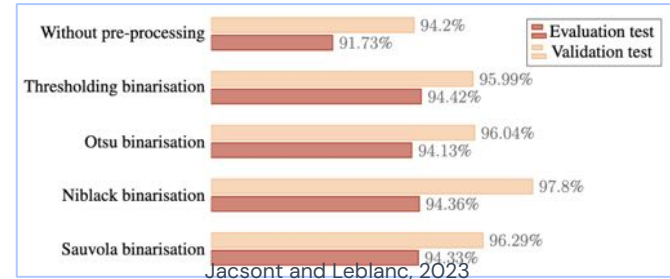
Beyond “eval_accuracy”

Preprocessing Images/Texts/OCR

- Recent research suggests that preprocessing Spanish-language with Niblack binarization can (slightly) improve character and word recognition.
- Sensitivity to noisy OCR, diacritics, and punctuation.
- End of sentence punctuation and capital letter “cleaning” reduces unique words and subsequent location entities.



Unique Entities from the Travel literature corpus



“What is in a name?” Quite a lot!

Classes or Entity Labels

- The majority of NER models in Spanish are trained from data labeled with the PER, ORG, MISC, and LOC classes.
- In contrast, the smallest SpaCy English language model contains data labeled with 18 classes.

LABEL	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.

Disambiguating with GeoNames

Impact of additional NER labels:

a. FAC and NORP

```
{ "text": "Megicano", "start_pos": 1429, "end_pos": 1437, "type": "LOC" }  
{ "text": "colegio de San Gregorio", "start_pos": 1454, "end_pos": 1477, "type": "LOC" }  
{ "text": "colegio de Gregorio", "start_pos": 1666, "end_pos": 1685, "type": "LOC" }  
{ "text": "Nueva España", "start_pos": 2351, "end_pos": 2363, "type": "LOC" }
```

↓

```
[ "colegio de gregorio", { "geonameid": 8981838, "name": "De Gregorio", "asciiname": "De Gregorio", "alternatenames": ["De Gregorio"], "latitude": 41.08511, "longitude": 14.92212, "feature_class": "P", "feature_code": "PPL", "country_code": "IT", "cc2": [""], "admin1_code": "04", "admin2_code": "BN", "admin3_code": "062012", "admin4_code": "", "population": 12, "elevation": null, "dem": 236, "timezone": "Europe/Rome", "modification_date": null}], ]
```

b. FAC and GPE

```
{ "text": "Sierra del Afta", "start_pos": 820, "end_pos": 835, "type": "LOC" }  
{ "text": "Cofre de Perote", "start_pos": 846, "end_pos": 861, "type": "LOC" }
```

```
[ "sierra del afta", { "geonameid": 344703, "name": "Afta", "asciiname": "Afta", "alternatenames": ["Afta", "afta", "أفت", "latitudo": 15.28075, "longitudo": 39.64508, "feature_class": "P", "feature_code": "PPL", "country_code": "ER", "cc2": [""], "admin1_code": "06", "admin2_code": "302", "admin3_code": "", "admin4_code": "", "population": 0, "elevation": null, "dem": 22, "timezone": "Africa/Asmara", "modification_date": null}], ]
```

Other Filtering Methods

GeoNames “alternativenames”

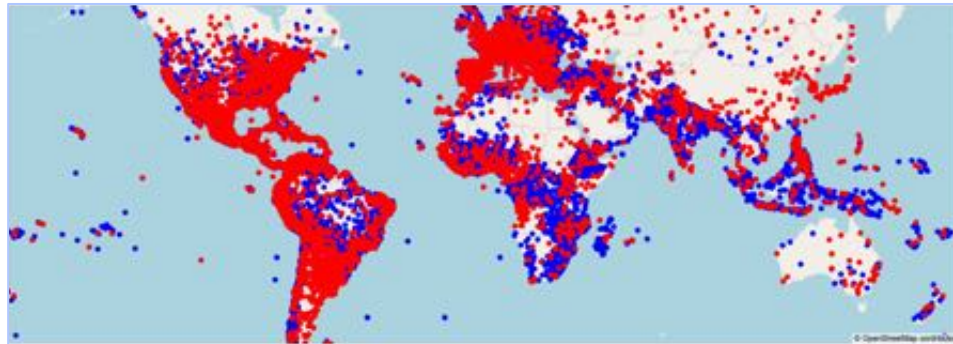
- a. Using full GeoNames data dumps to yield more results.

Limiting regions

- a. Specifying limited areas for matching, not generalizable.

An embarrassment of riches

- a. Span and scale are also limits.



Travel Literature GNER is a Complex Task

Historical Corpora (Blouin, et. al. 2021):

- a. “What annotation effort [is needed] in the target historical domain?”
- b. “Is it worth adapting initial pre-trained word representations?”
- c. “What is the impact of OCR errors on transfer performance?”

Spanish-Language Models (de la Rosa, et.al. 2021):

- a. “How much data is enough to train a well-performing monolingual language model?”
- b. “When more than enough data exist, how to select the documents that enable a more efficient training?”
- c. “How does data quality affect training times?”

Future work

With this, we have:

- An original, unprecedented inventory of named locations in a collection of Spanish-language literary works
- The beginning of a study into the geographic creativity and/or visits within literary and historical narratives (the line is blurry)

We now have to:

- Manually annotate enough works relevant to literary texts in Spanish from the 19th Century on to fine tune existing models to this particular cultural artifact
- With the hopes of:
 - Developing context and expanded set of labels for nationalities and groups, historical events, geopolitical entities, and GeoNames feature codes.
 - Identifying historical periods or events that present a shift on the geographical imaginaries of these (or other) corpora

Cited references

- Aldana-Bobadilla, Edwin, et al. "Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text." *Remote Sensing*, vol. 12, no. 18, Sept. 2020, p. 3041. DOI.org (Crossref), <https://doi.org/10.3390/rs12183041>
- Calarco, Gabriel. "La visualización del espacio geográfico en las éfrasis del Libro de Alexandre con Recogito y Visone." *Publicaciones de la Asociación Argentina de Humanidades Digitales*, vol. 3, Nov. 2022, p. e035. DOI.org (Crossref), <https://doi.org/10.24215/27187470e035>.
- DesenrollandoElCordel/Pliegos-Ner: Experiments with Named Entity Recognition for Spanish Chapbooks*. <https://github.com/DesenrollandoElCordel/pliegos-ner/tree/main>. Accessed 17 Dec. 2023.
- Jacsont, Pauline, and Elina Leblanc. "Impact of Image Enhancement Methods on Automatic Transcription Trainings with eScriptorium." *Journal of Data Mining and Digital Humanities*, vol. Historical Documents and..., Sept. 2023. doaj.org, <https://doi.org/10.46298/jdmdh.10262>.
- Lehmann, Jörg, and Konstantin Krechting. *Bibliography of electronically available Spanish-language travel literature*. <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/85802>. Accessed 19 Nov. 2020.
- Pelagios. "Final Report on LatAm: A Historical Gazetteer of Colonial Latin America and the Caribbean." *Pelagios*, 14 June 2019.
- "Presentación del portal Viajeros españoles - Viajeros españoles." *Biblioteca Virtual Miguel de Cervantes*, https://www.cervantesvirtual.com/portales/viajeros_espanoles/presentacion/.
- Recogito, an initiative of Pelagios Commons, <http://recogito.pelagios.org/> (accessed 17 November 2023)
- Wilkens, Matthew. "Too Isolated, Too Insular: American Literature and the World." *Journal of Cultural Analytics*, vol. 6, no. 3, June 2021. DOI.org (Crossref), <https://doi.org/10.22148/001c.25273>.

The slide features a white background with decorative hexagonal shapes in the corners. The top-left and bottom-right corners have overlapping cyan and light blue hexagons. The bottom-left corner has overlapping light blue and cyan hexagons. The text '¡Gracias!' is centered in a bold, dark blue font.

¡Gracias!