A Theatre of Places

Mapping 17th c. French Theatre

Simon Gabay
Institut de littérature française
Université de Neuchâtel
Neuchâtel Switzerland
simon.gabay@unine.ch

Giovanni Pietro Vitali
Department of Italian
University College Cork
Cork Ireland
giovanni.vitali@ucc.ie

ABSTRACT

As a form of distant reading, mapping texts allows scholars to read classical works anew. Using 17th French theatre as a test case, we describe an easily reproducible and fully open-source workflow used for extracting and mapping place names, then reach conclusions on literary influences and the strength of genre during the Grand Siècle based.

CCS CONCEPTS

• Applied computing → Arts and humanities → Performing arts
• Applied computing → Computers in other domains → Cartography

KEYWORDS

17th c. France, Theatre, Literary genre, Geocriticism, Distant reading, GIR, NER, map

ACM Reference format:

Simon Gabay and Giovanni Pietro Vitali. 2019. Mapping 17th French Theatre. In *Proceedings of the 13th Workshop on Geographic Information Retrieval (GIR'19), 28th-29th November 2019, Lyon, France*, ACM, New York, NY, USA, 4 pages. DOI: 10.1145/3371140.3371146.

1 Introduction

After Schöch's stylometric [10] and thematic [9] studies, we would like to return to French classical theatre from another perspective: spatial analysis. Places mentioned in text are indeed highly informative regarding genre, but also literary influences, and are therefore worth considering. Using a corpus of 36 plays written by prominent authors across three genres (*comédie*, *tragédie* and *tragi-comédie*) and three periods (first, second and third generation of authors), we clean texts and extract data to project toponyms on a map, which has been conceived as a dynamic research interface for literary analysis. The workflow used is composed solely of simple open-source software to maximize its reusability by other scholars who would like to pursue our work.

ACM ISBN 978-1-4503-7260-2/19/11…$15.00
https://doi.org/10.1145/3371140.3371146

2 Corpus

The plays are written by nine of the most important playwrights of 17th France[1]:

1. Four were active in the first half of the century: Pierre Du Ryer (1628-1655), Gorges de Scudéry (1631-1643), Paul Scarron (1648-1658).
2. Four were active in the second half of the century: Claude Boyer (1646-1697), Thomas Corneille (1651-1696), Molière (1655-1673), Jean Racine (1664-1675).
3. One lies between these two groups: Pierre Corneille (1629-1675).
4. Only four plays by each author have been used, for a total of 9 *tragi-comédies*, 17 *tragédies* and 10 *comédies*. Spelling and capitalization have been normalized for all the plays.

3 Creating the gazetteer

As all the plays were already encoded in XML-TEI[2] [3], the texts have been cleaned with XSLT to retain only the content of the <l> element – *i.e.* the content of the verses. It must be noted here that place names are not only present within these dialogic parts, but also in stage directions (<stage>, <set>…). Removing the latter is therefore a literary choice to map not the play itself, but the imaginary world created by the characters' speeches perceived by the audience.

As for place names not tagged with <placeName>, the challenge is to find a way to extract a maximum of them for each text. We decided to use HER [5], an active learning-based Named Entities Recognizer (NER) toolkit, which accelerates the process of manual annotation by pre-selecting sentences that are maximally

[1] Dates are those of the first and the last play of each author.

[2] http://dramacode.github.io

informative to the model. It operates iteratively: a first limited wave of sentences are annotated to produce a first gazetteer, from which a second bigger wave is prepared for annotation, *etc*. In other words, HER does not guarantee the creation of an exhaustive gazetteer, but maximizes its exhaustiveness without reading entirely large corpora.

From an initial list of 3,000 sentences (c. 59,000 tokens) manually annotated, 69 place names were identified. Based on this data, a second list of sentences is created, and ranked according to the probability that they contain a toponym. This time the 100,000 first tokens (out of 694,000, *i.e.* 14.4% of the corpus) were manually annotated, which was enough to identify 924 place names (162 different types) in the corpus.

This process of annotation has been repeated using NLP tools to automatically detect proper nouns, cross-validate the results and minimize omissions. To select the most efficient technique, we tested three POS taggers:

1. TreeTagger [8] a relatively old POS tagger. Though outperformed by many others, it is currently the most commonly used by literary scholars [1].
2. MElt [4] a recent POS tagger already trained on French.
3. Talismane [11] a recent POS tagger and lemmatizer trained on French.

Tests were conducted on Boyer's *Agamemnon*, which contains c. 16,500 tokens and a high number of toponyms (116 occurrences, 18 types, 2 compound ones):

	Proper nouns	Toponyms	%
TreeTagger	983	16	1.63%
Melt	777	18	2.32%
Talismane	785	18	2.29%

Only MElt and Talismane properly tagged all 17 place names as proper nouns. Because MElt identified slightly fewer proper nouns than Talismane, the former is slightly more efficient than the latter.

Recent research has however shown the superiority of NER taggers for the extraction of anthroponyms and toponyms (Brando *et alii* 2016). We therefore tested GROBID nerd, a new named entity recognizer for English and French with a knowledge base of 37 million entities from Wikidata [6].

	Entities	Positive	%
GROBID nerd	419	18	4.30%
GROBID nerd, location type	83	17	20.48%

Since GROBID nerd seems to offer not only perfect precision but also the best recall, it has been used to control the places tagged with HER in each text and update the gazetteer and offer a first evaluation of HER:

	HER	HER+GROBID	HER accuracy after 100,000 tokens annotated
Toponyms	162	188	86.2%
Occurrences	924	1,032	89.5%

4 Geo-referencing and categorization

All occurrences of places have as much metadata as possible, such as the author's name, the generation he belongs to, the play's name and genre, longitude and latitude, *etc*.

To find the coordinates, we have first used *georeference*, a geolocation R package developed by J. L. Losada [7] that automatically queries *Geonames*[3]. Due to the complexity of the names mentioned in our plays, it was only effective with the most important ones (Madrid, Paris, *etc*.), and we have decided to process the data manually as it relies on two databases: *Genonames* for early-modern places, and *Pleaides*[4] for ancient ones.

Disambiguation was not done automatically for several reasons:

1. Many foreign place names are "francized", usually following French customs of the 17th century (*Gaeta* is translated *Gaïette* and not *Gaète* like today).
2. There is a superposition of chronological period (ancient world & classical era).
3. Not all places actually exist: some are imaginary places such as *Chicuchiquizèque* or Gaula (from the novel *Amadis de Gaula*) in Scarron's *Dom Japhet d'Arménie*. Some are legendary places such as Hell or underworld rivers like the Cocytus in Racine's *Phèdre*.

For the latter, we have tried to rely on actual geography when possible (Mt Olympus as the home of the Greek gods, the Acheron as the entrance to the underworld), thinking that approximate locations were better than none. Regarding rivers, regions or kingdoms, the coordinates are those of the aforementioned databases despite, once again, inevitable approximations – as no polygons were available, or even feasible, for regions like Hyrcania (a historical region composed of the land south-east of the Caspian Sea) or Scythia (a region of Central Eurasia in classical antiquity).

When needed, disambiguation was done with the following criteria:

1. Existence of coordinates.

[3] http://www.geonames.org.

[4] https://pleiades.stoa.org.

2. Common sense: Thessalia is most probably the region in continental Greece (Pleiades n°1332) rather than the city in Peloponnese (Pleiades n°991374).
3. Context: Rome in Geonames (n°3169070) is the modern city, Rome in Pleiades (n°423025) is the ancient one.
4. Historical importance: Troia is the city at its apex (Pleiades n°550595) and not the first settlement in Kumtepe (Pleiades n°210709924).

In order to deal with these issues, but also to strengthen our final visualization, we grouped our results in different macro zones: France, Italy, the Iberian Peninsula (Portugal and Spain), Italy, England, Germany, Greece, Asia Minor/Anatolia, the Levant (Syria, Palestine/Israel, Lebanon), East (Irak, Iran, central Asia), Far East (China, India, Japan), Africa (Maghreb, Libya, Egypt, Ethiopia) and America.

5 Mapping process

In order to display these macro zones, shapefiles of contemporary countries are downloaded, and we work on the vectors to merge, divide and re-draw the boundaries with QGIS. To avoid slowdown or timeout and to lower the size of the final map, shapefiles were resized with MapShaper[5].

All our data is contained in 14 datasets, 12 of which are csv-based and the 2 others are shp-based. We processed it with R, which allows us to easily publish and share our code. The result is composed of 19 superimposed layers of elements, which allows the reader to select precisely the information he wants to see. Each type of information is displayed with a specific marker to ease the map's readability:

- Genres are displayed with Google Maps pin
- Authors are displayed with circles
- Periods are displayed as a heat map
- Macro zones are displayed with a coloured background

A selector on the right allows the user to choose what he wants to display, and a map legend on the left provides basic information for understanding the results.

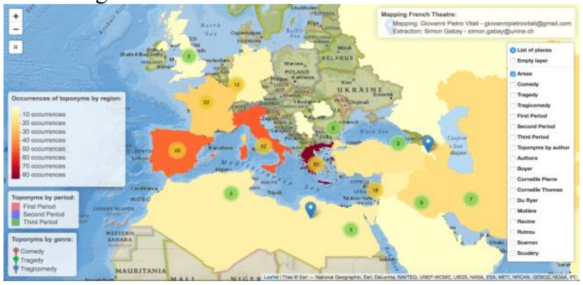

Figure 1: the final map.

6 Preliminary literary conclusions

It is now possible to start exploring our data. As a first observation, French 17th c. theater is already global: it mentions Europe, but also America, Africa and Asia. This distribution is unbalanced, however, since most of the toponyms are concentrated on the Mediterranean basin, which remains the core of the period's imaginary (Figure 2).

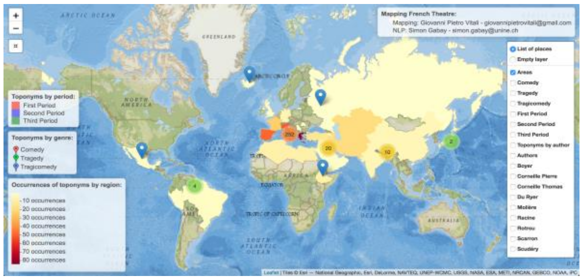

Figure 2: Global distribution of toponyms.

It seems that there is a very clear generic distribution of places, with Italy as a shared border. *Tragédies* are located east, *comédies* west, and *tragi-comédies* both east and west (Figure 2).

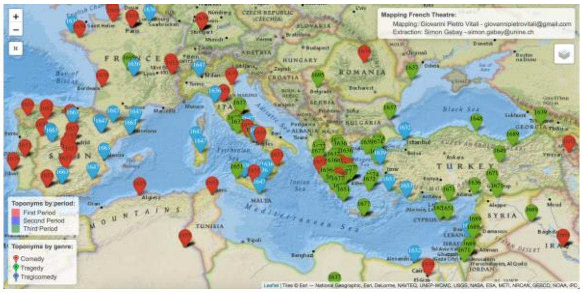

Figure 3: *Comédies* (red), *tragédies* (green), *tragi-comédies* (blue).

This reveals the strong impact of Spanish Golden Age theatre on *comédies*, and of the classical and the biblical world on *tragédies*. *Tragi-comédies*, as its name suggests, appears as a truly intermediary genre, not leaning in either direction.

[5] https://mapshaper.org.

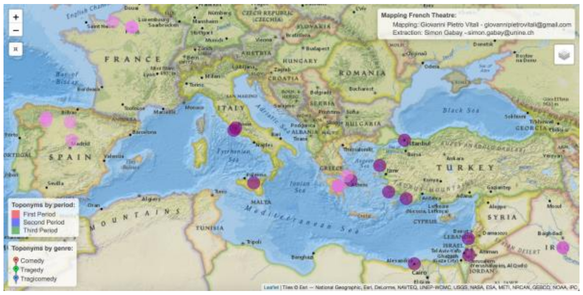

Figure 4: Molière (pink) and Du Ryer (purple).

When comparing the use of place names mentioned by author, it also appears that some authors, like Molière or Du Ryer, do not use geography as a stylistic marker (Figure 4), while others make an intensive use of it (Figure 5).

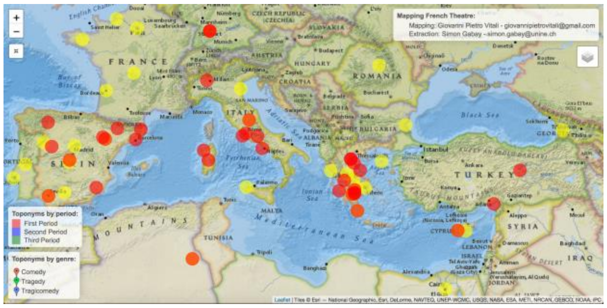

Figure 5: Pierre Corneille (yellow) and Rotrou (red).

7 NER models

HER offers the possibility to train NER models (CRF or Bi-LSTM) based the previously annotated data. To keep our promise of an easily reproducible workflow, we have decided to compare this model with two other popular solutions using two other architectures: Stanford NER (CRF) and Spacy (CNN.) It has been trained on all the texts of our corpus but one, kept for testing the model: Scudéry's *Le Prince déguisé* (c. 15,935 tokens, 17 place names, 12 different types, 7 in the training data).

	Toponyms	Types
HER crf	17	12
HER bi-lstm	12	7
NER Stanford	11	6
Spacy	13 (+10 false)	8 (+3 false)

We have also developed alternative models with additional training data taken from the Europeana newspapers project[6] to ameliorate the results (out of the 12 different types, 8 are in the training data):

	Toponyms	Types
NER Stanford	13	8
Spacy	13 (+3 false)	9 (+2 false)

ACKNOWLEDGMENTS

Names are in alphabetical order, as the paper was equally written by the two authors. Merci à Jean-Baptiste Camps pour le corpus, Alex Erdmann pour le TAL et Patience Haggin pour la relecture.

DATA

Gabay Simon, & Giovanni Pietro Vitali. (2019). CARTO17: Data and Scripts for Mapping 17th c. French Theatre (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3530791

REFERENCES

[1] Boeglin Noémie, Depeyre Michel, Joliveau Thierry, Le Lay Yves-François (2016). Pour une cartographie romanesque de Paris au XIXe siècle. Proposition méthodologique. *Spatial Analysis and GEOmatics*, Nice, France. https://hal.archives-ouvertes.fr/hal-01619600.

[2] Brando Carmen, Frontini Francesca, Ganascia Jean-Gabriel. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly*, 2016, pp.60 – 80. DOI: 10.7250/csimq.2016-7.04.

[3] Burnard Lou (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. OpenEdition Press, Marseille. DOI: 10.4000/books.oep.426.

[4] Denis Pascal, Sagot Benoît (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, DOI: 10.1007/s10579-012-9193-0.

[5] Erdmann Alexander, Wrisley David Joseph, Allen Benjamin, Brown Christopher, Cohen Bodénès Sophie, Elsner Micha, Feng Yukun, Joseph Brian, Joyeaux-Prunel Béatrice and Marneffe Marie-Catherine de - (2019). Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of North American Association of Computational Linguistics (NAACL 2019)*. Minneapolis, MN, DOI: 10.18653/v1/N19-1231.

[6] Lopez, Patrice (2019). *Entity-fishing documentation 0.0.3*. https://nerd.readthedocs.io.

[7] Losada Palenzuela, José Luis. Georeference: Geolocation R package, http://editio.github.io/2018/03/27/georeference-geolocation-r-package.html.

[8] Schmid, Helmut (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

[9] Schöch, Christof (2017) Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly* 11 (2), §1-53. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

[10] Schöch Christof (2013). Fine-tuning our stylometric tools: Investigating authorship, genre, and form in French classical theater. *Digital Humanities 2013: Conference AbstractsUniversity of Nebraska–Lincoln*, 383–86.

[11] Urieli, Assaf (2013) *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Université Toulouse le Mirail, 2013. https://tel.archives-ouvertes.fr/tel-01058143.

[6] https://github.com/EuropeanaNewspapers/ner-corpora.