A Theatre of Places: Mapping 17th French Theatre
Simon Gabay, Giovanni Pietro Vitali

To cite this version:

HAL Id: hal-02388411
https://hal.science/hal-02388411
Submitted on 1 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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:

1 Introduction

After Ch. 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 softwares to maximize its reusability by other scholars who would like to pursue our work.

2 Corpus

The plays are written by nine of the most important playwrights of 17th France1:

1. Four are active in the first half of the century: Pierre Du Ryer (1628-1655), Georges de Scudéry (1631-1643), Paul Scarron (1648-1658).
2. Four are 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 is in between these two groups: Pierre Corneille (1629-1675).
4. Only four plays by each author have been used, for a total of 9 tragï-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-TEI2 [3], the texts have been cleaned with XSLT to keep 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 have 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 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.

1Dates are those of the first and the last play of each author.
2http://dramacode.github.io
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, all 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 have 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].

Tests have been conducted on Boyer’s Agamemnon, which contains c. 16,500 tokens and a high number of toponyms (116 occurrences, 18 types, 2 compounds ones):

<table>
<thead>
<tr>
<th>Proper nouns</th>
<th>Toponyms</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>TreeTagger</td>
<td>983</td>
<td>16</td>
</tr>
<tr>
<td>MElt</td>
<td>777</td>
<td>18</td>
</tr>
<tr>
<td>Talismane</td>
<td>785</td>
<td>18</td>
</tr>
</tbody>
</table>

Only MElt and Talismane have properly tagged all 17 place names as proper nouns. Because MElt has identified slightly less 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 have therefore tested GROBID nerd, a new named entity recognizer for English and French with a knowledge base of 37 million entities from Wikidata [6].

<table>
<thead>
<tr>
<th>Entities</th>
<th>Positive</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROBID nerd</td>
<td>419</td>
<td>18</td>
</tr>
<tr>
<td>GROBID nerd, location type</td>
<td>83</td>
<td>17</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Entities</th>
<th>Positive</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toponyms</td>
<td>162</td>
<td>188</td>
</tr>
<tr>
<td>Occurrences</td>
<td>924</td>
<td>1,032</td>
</tr>
</tbody>
</table>

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 Genonames\(^1\). Due to the complexity of the names mentioned in our plays, it has only been efficient 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 Pleides\(^4\) for ancient ones.

Disambiguation has not been done automatically for several reasons:

1. Many foreign place names are “francized”, usually following French customs of the 17\(^{th}\) 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 Hycania (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 has been done with the following criteria:

1. Existence of coordinates.
2. Common sense: Thessalia is most probably the region in continental Greece (Pleades n°1332) rather than the city in

\(^1\) http://www.geonames.org.
\(^4\) https://pleiades.stoa.org.
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 loose coordinates, but also to strengthen our final visualization, we have decided to group 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 have been resized with MapShaper.

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.

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 (fig. 2).

![Figure 2: Global distribution of toponyms.](image2)

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 (fig. 2).

![Figure 3: *Comédies* (red), *tragédies* (green), *tragi-comédies* (blue).](image3)

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.

![Figure 4: Mollière (pink) and Du Ryer (purple).](image4)

7  **Preliminary literary conclusions**

---

5 [https://mapshaper.org](https://mapshaper.org)
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 (fig. 4), while others make an intensive use of it (fig. 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 on 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).

<table>
<thead>
<tr>
<th>Toponyms</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>HER crf</td>
<td>17</td>
</tr>
<tr>
<td>HER bi-lstm</td>
<td>12</td>
</tr>
<tr>
<td>NER Stanford</td>
<td>11</td>
</tr>
<tr>
<td>Spacy</td>
<td>13 (+10 false)</td>
</tr>
</tbody>
</table>

We have also developed alternative models with additional training data taken from the Europeana newspapers project6 to ameliorate the results (out of the 12 different types, 8 are in the training data):

<table>
<thead>
<tr>
<th>Toponyms</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER Stanford</td>
<td>13</td>
</tr>
<tr>
<td>Spacy</td>
<td>13 (+3 false)</td>
</tr>
</tbody>
</table>

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


REFERENCES


---