Automated Geoparsing of Paris Street Names in 19th Century Novels

Ludovic Moncla, Mauro Gaio, Thierry Joliveau, Yves-François Le Lay

Automated Geoparsing of Paris Street Names in 19th Century Novels

Ludovic Moncla
Naval Academy Research Institute
Brest - Ecole navale, France
ludovic.moncla@ecole-navale.fr

Mauro Gaio
Laboratoire LIUPPA
Université de Pau et des Pays de l'Adour, France
mauro.gaio@univ-pau.fr

Thierry Joliveau
Laboratoire EVS
Université de Saint-Etienne, France
thierry.joliveau@univ-st-etienne.fr

Yves-François Le Lay
Laboratoire EVS
ENS Lyon, France
yves-francois.le-lay@ens-lyon.fr

ABSTRACT

Our project involves building a platform able to retrieve, map and analyze the occurrences of place names in fictional novels published between 1800 and 1914 and whose action occurs wholly or partly in Paris. We describe a proof of concept using queries made via the TXM textual analysis platform for the extraction of street names. Then, we propose a fully automatic process using the named entity recognition (NER) components of the PERDIDO platform. This paper describes some encouraging initial results obtained by combining NLP approaches (NER methods) with textometric tools for the automated geoparsing of street names.

KEYWORDS

Named Entity Recognition, Geographical Information Retrieval, Geoparsing, Digital Humanities

1 INTRODUCTION

"Spatial turn" is the term currently used to describe a general movement, observed since the end of the 1990s, that emphasizes the reinsertion of place and space in social sciences and humanities [32]. In literature, a spatial turn corresponds to a new interest in places and landscapes described in fictional texts and in maps as a good way to reveal the spatial structure of a narrative. Franco Moretti is a precursor of this movement, with his famous maps of Balzac's Paris, Dickens' London or Austen's England among other authors and works [23, 24]. Beyond a traditional use of literary mapping, cartography is now used to produce a novelistic representation of a place with the aim of supporting new interpretations of novels. New cartographic solutions to new problems, such as imprecision in location, imaginary spaces, links between geospace and textual spaces, have to be found [26]. The widespread use of computing and digital technologies in social sciences and the emergence of Digital Humanities generalize the question to the global processing of information, offering not only new techniques for visualization but also new ways to collect and analyze literary texts spatially. While Moretti drew his maps by hand, it is now possible to use geographical information systems, spatial analysis tools, or multimedia, augmented reality or virtual reality tools [7, 11]. Digital techniques can now be applied to automatically collect or extract information from the text. Before mapping a novel, it is necessary to locate the places the author mentioned, which is a dull and time-consuming task that is only possible for one or two novels, as it is more difficult to consider for all the novels of an author or all the novels published during a certain period of time.

Natural language processing (NLP) offers some solutions that can be customized to extract localized events mentioned in a novel or a poem but are not so easy to perform properly. The aim of this paper is to present some encouraging initial results obtained by combining a named entity recognition (NER) method and textometric tools. These results were obtained by applying the process to a corpus of French novels centered on Paris.

2 MOTIVATION AND BACKGROUND

2.1 Geoparsing literature

Laura Hill [15] defines geoparsing as a two-step process: 1) identifying geographic references in text and 2) assigning geospatial coordinates to these references. In the first step, the challenge is to recognize and extract the places mentioned in a text. This procedure can be tricky for fictional texts because a novelist has multiple ways of evoking a place: directly (by giving an explicit name) or more elusively (by using relative references, e.g., near, behind, or two blocks further, relative to other places mentioned before). Some places can be deliberately disguised and others can be completely imaginary. These different cases, among many others, can be found in the same novel. Automatizing the recognition of all these kinds of places is a real scientific problem, which is made even more difficult when referring to ancient texts [18]. Explicit place names are obviously the easiest type to automatically retrieve from texts, even though some cases can be difficult to resolve because of place name ambiguities and homonymy with other names. Place names are clearly inadequate when the goal is to obtain a deep understanding of the role of places and the meaning of space in a specific book. As Heuser et al. argue [14], place naming can be useful because it plays an important cultural role in fiction, "suturing narrative and geographic space whilst also calling upon and contributing to connotations that have accrued through wider cultural circulations" [14]. It makes sense as well to try to collect all the place names found in a vast corpus of texts about a specific place, such as a big city. Moreover, place names can be useful for finding unnamed places mentioned relative to the first ones.

.

The second step, i.e., assigning geospatial coordinates to the references found in a text, is also a problem, even if the places are not imaginary but have explicit and real names. The geographic coordinates of these places must be found in existing gazetteers or in other text sources that can help to locate these place. For ancient literary texts, it is necessary to have access to geohistorical gazetteers. When a fictional place name found in a text cannot be located, it is very hard to ascertain if the place is imaginary or if the location has been lost and forgotten over time, especially if the numbers of places and texts are very large. There are different ways to cope with these constraints. One can work with a sample of the places that can be found in novels and located. This solution was adopted by [14] when they sought to map the emotion in London using 4,862 texts of fiction from the 18th and 19th centuries. They chose 161 frequently mentioned places that were spatially representative of the wide extent of London. Another option is to cover, as completely as possible, both a corpus of texts and a geographical area to permanently link textual and topographic places. This was the approach of [3] for poetry texts in Edinburgh and of [10], who considered a diverse corpus of texts about the Lake District in England.

The aim of our project is very similar to the Palimpsest project, which used natural language processing technology in order to text mine literary works set in Edinburgh. The first version of Palimpsest has already been completed, and the results are accessible online[1]. The project combined an adapted geoparser [2] with a "fine-grained" gazetteer focused on Edinburgh at the street and building level but with a slightly different goal. The main objective was to assist curators in identifying texts that were mainly set in Edinburgh from a huge collection of books [1].

2.2 Named entity recognition (NER)

NER approaches cover a large variety of strategies and methods and are generally classified into two main categories: data-driven approaches (based on machine learning methods) and knowledge-based approaches (which use heuristic and handcrafted rules). Knowledge-based approaches were very popular at the end of the 1990s and then neglected in favor of quantitative or data-driven approaches, particularly those based on machine learning methods. Thus, many current proposals are based on machine learning methods in the context of rediscovery of these methods due to the very high computing capacity now available. These proposals use different approaches such as hidden Markov models [33] or conditional random fields [19]. However, many knowledge-based approaches have been proposed in the field of NER, one of the earliest being based on heuristics and handcrafted rules [29]. Many different methods are used, such as cascades of finite-state transducers that produce tree-like representations [8, 28]. Because regular languages and relations can be encoded as finite automata, they can be more easily manipulated than more complex languages; cascades of transducers have therefore turned out to be very useful for many approaches based on domain-specific corpus analysis and rules are described in a readable way and are easy to modify and maintain. The rapid growth of computing capacity combined with the considerable amount of annotated data sets made available enabled quantitative approaches to achieve tangible results. However, so far, for many tasks, knowledge-based approaches have not yet been superseded. This is particularly true when there is not yet an adequate amount of annotated data or when a task require an explanatory phase. At the very most the two approaches are progressively combined into hybrid approaches [4].

A considerable amount of work in NER research takes the language factor as a parameter. In this body of work, a significant part is devoted to the study of English, but French is also considered [8, 27], as well as some other languages. The impact of the literary genre (e.g., narrative, memoir or journalism) and domain (e.g., supply of raw materials, market or economic intelligence, or politics) is a problem that has been more recently addressed in the NER literature.

Starting with the detection of proper names, NER aims at identifying and classifying them into categories such as persons, organizations, or locations. According to Jonasson [16], there are two categories of proper names: pure and descriptive. Pure proper names can be simple (i.e., composed of a single lexeme) or complex (i.e., composed of several lexemes) and include proper names only. Descriptive proper names refer to a composition of proper names and common names (i.e., expansion). In other words, descriptive proper names overlap with pure proper names. Descriptive proper names refer to a a named entity (NE) built with a pure proper name and a descriptive expansion. This expansion can change the implicit type (e.g., location, person or organization) of the initial pure proper name.

(1) a *la rue* [*La Fayette*]
'La Fayette Street'

b *le quartier de la* [*Goutte-d'Or*]
'Goutte-d'Or neighbourhood'

In [9], the authors have introduced the concept of extended named entity (ENE). Based on the Jonasson's definition, an ENE refers to an entity built with a proper name (pure or descriptive) and may be composed of one or more concepts (see example 1). Each ENE may have several levels of overlapping (0, 1, 2, etc.) for the encapsulation of different concepts.

Whereas most of NER works (Standford NER[2], OpeNER[3], Open Calais[4]) usually only consider only pure proper names, the concept of an ENE appears essential for a fine-grained task, such as marking, classifying and disambiguating NEs. An ENE of level 0 refers to the usual NE concept, and an ENE of level 1 (Figure 1) refers to descriptive proper names composed of a pure proper name (i.e., an entity of level 0) and a noun (i.e., expansion). The descriptive expansions specify the nature or the feature type of the entity and can change its implicit or default nature. For instance, example 1a shows a street name built using a person's name.

As stated by [12] the new generation of geoparsers needs to use more information for the understanding of the meaning of the context. Thus, this concept of ENE implemented with cascades of

[1] https://litlong.org/welcome

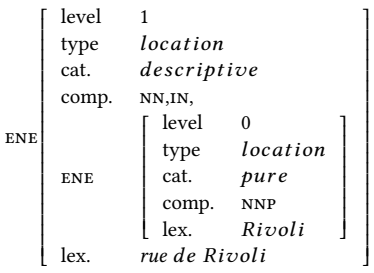
[2] https://nlp.stanford.edu/software/CRF-NER.shtml

[3] http://www.opener-project.eu/

[4] http://www.opencalais.com/opencalais-demo/

Table 1: List of the 31 novels of the corpus

Title	Author	Date	Title	Author	Date
La maison du Chat-qui-pelote	Honoré de Balzac	1830	L'Assommoir	Emile Zola	1877
Ferragus	Honoré de Balzac	1833	Une belle journée	Henry Céard	1881
La fille aux yeux d'or	Honoré de Balzac	1835	Pot-Bouille	Emile Zola	1882
Le père Goriot	Honoré de Balzac	1835	Au Bonheur des dames	Emile Zola	1883
Grandeur et décadence de César Birotteau	Honoré de Balzac	1837	Le vingtième siècle	Albert Robida	1883
Les mystères de Paris	Eugène Sue	1842	Sapho	Alphonse Daudet	1884
Sans Cravate ou les Commissionnaires	Paul de Kock	1844	Bel-ami	Guy de Maupassant	1885
L'envers de l'histoire contemporaine	Honoré de Balzac	1848	L'œuvre	Emile Zola	1876
M. Choublanc à la recherche de sa femme	Paul de Kock	1856	La vie électrique	Albert Robida	1892
Les misérables	Victor Hugo	1862	Paris	Emile Zola	1897
Les demoiselles de magasin	Paul de Kock	1863	La charpente	J.H. Rosny jeune	1900
Paris au XXème siècle	Jules Verne	1863	Mr. Bergeret è Paris	Anatole France	1901
L'éducation sentimentale	Gustave Flaubert	1869	La Maternelle	Léon Frapié	1904
La Curée	Emile Zola	1871	La Vague rouge	J.H. Rosny aîné	1910
Le ventre de Paris	Emile Zola	1873	Dans les rues	J.H. Rosny aîné	1913
Jack	Alphonse Daudet	1876			



NN=Noun, IN=Preposition, NNP=Proper noun, singular

Figure 1: Feature structure representing an ENE of level 1

transducers [9] is very appropriate for improving the efficiency of literary geoparsing, especially if the task is to retrieve explicit place names, as we attempt to demonstrate in our project dealing with French novels mainly set in Paris.

3 METHODOLOGY

3.1 Overview

In this paper, we propose a method for automatically retrieving the street names from novels published between 1800 and 1914 and whose action occurs mostly in Paris. This work is part of a project that aims to locate, map, analyze and communicate the places of Paris mentioned in novels, targeting historians, town planners, cultural or literary tourists, inhabitants and curious people interested in visiting places described in novels while reading a book or analyzing the fictional description of lost or still existing places. The three main stages of this project are:

(a) extract spatial named entities,
(b) locate these entities by using historical sources and building an appropriate gazetteer,
(c) offer mapping forms adapted to the specificities of literary spaces.

This paper focuses specifically on stage (a). We start by describing a proof of concept using queries via a textual analysis platform (TXM [13]). The results obtained using this textual analysis platform are compared with the results of a fully automatic NER method (PERDIDO [25]). We also describe how these two methods can be combined to provide interoperability between NLP approaches and textometric analysis. Indeed, PERDIDO and TXM provide complementary tools for the annotation of NE and textual analysis.

Our experimental corpus is composed of 31 French novels covering different periods of the nineteenth century (Table 1) centered on Paris. Some of them were written by famous authors (e.g., Balzac, Zola, Hudo, and Sue), while others, by more confidential ones (e.g., Frapié and de Kock) that had a certain amount of success at the time of publication. A manual preprocessing step was carry out to prepare the files of each novel and correct the errors introduced by the OCR process. For instance, we corrected some spelling mistakes (such as replacements of letters wrongly recognized by the OCR), deleted white spaces after apostrophes and corrected hyphenations.

3.2 Extracting street names via CQL requests

The main objective of our proposal is to exhaustively retrieve, extract and annotate spatial named entities from these novels (Table 1) and to use them to locate and map the spatial imprint of these novels or their authors. For that purpose, we first build a solution combining different IT tools with human interactions. This is used as a proof of concept to show the feasibility of our proposal and its interest.

To quickly retrieve spatial named entities from the novels we made some queries using the TXM platform. TXM[5] is an open-source software platform that provides tools for qualitative and

[5] http://sf.net/projects/txm

quantitative content analyses of text corpora. It implements lexicometric methods based on the corpus query processor (CQP) and R for corpus search and statistical text analysis [13]. The lexical patterns are expressed in the corpus query language (CQL) and based on words and structure-level properties to define linguistic patterns (one or several words) and properties (lemma, part-of-speech). These requests allow us to find the occurrences of several categories of spatial named entities (mainly streets, waterways, and buildings).

```
[lemma="rue"%cd][word!="\.|\,|\;|\!|\?|\...|-|une|\-|où
\ainsi|et|aurait|-l"%c]? [word!="\.|\,|\;|\!|\?|\...|-
|une|\-|où\ainsi|et|aurait|-l"%c]?  [word!="\p{Lu}.*"&
word!=\Ça|Ah|O|Venez|Et|M|L."]
```

Figure 2: First example of a CQL request to extract street names

Figure 2 shows the first CQL query used to extract street names in French novels. Writing a query adapted for every category of spatial named entities is a complex task. First, we built a query specifically for street names. To match as many occurrences as possible, we addresses case sensitivity (%c), diacritical marks (%d) and plurals (lemma). Then, the query specified that the category "*rue*" (i.e., "street") must be followed by one or several words beginning with a capitalized letter ([word!="\pLu.*"]). Additionally, we specified that there may be one or two words between the category and the capitalized word to take into account cases with articles and prepositions, similar to example 1. The query was also tuned to exclude special cases such as ends of phrases and punctuation marks. Finally, to avoid too many false positives, some words found in the results were specifically excluded from the answer. This enable generating a small number of false positives but is too closely associated with the corpus and lacks generality.

```
[frlemma="rue"%cd] [word!="\p{P}+"] ? [word!="\p{P}+"]
? [word!="\p{P}+"] ? [word="\p{Lu}.*"]
```

Figure 3: CQL request used to extract street names

Figure 3 shows the more simple and generic query that we finally adopted. With this method, it is difficult to build a query covering an exhaustive list of the linguistic patterns that can be found in our corpus of 31 novels without too many false positives. Indeed, according to the experiments false positive errors are due to special cases not covered by the query. Thus, before using the results for the geocoding step of the project, we need to manually remove false positives. The different types of errors are described in Section 4.1.

The TXM platform provides a concordancer (that exports to a CSV file), which shows every occurrence of street names for each novel of which the corpus with the context before or after the name, the length can be parametrized. It also provides access to the corresponding position in the text. This tool is very useful for exploring the corpus via the spatial named entities and facilitates the analysis of the literary content and the descriptions associated with the different places. The first results obtained using TXM are encouraging. By writing some general queries, we managed to correctly retrieve most of the place names present in the texts. However, a manual process is necessary for eliminating the false positives of the results. Thus, we propose extending and improving the method using a fully automatic and more generic process to retrieve and semantically tag named entities.

3.3 Named entity recognition and classification

To build a fully automatic processing chain to retrieve, semantically tag and extract named entities from the French novels we propose the use of the NER system designed and implemented in the platform called PERDIDO[6]. It is a geographically oriented NER system for multilingual text documents (currently available in French, Spanish and Italian) combining the notions of retrieving, tagging, and extraction of ENEs and geospatial information (e.g., spatial relations and motion verbs), through the use of local grammar and external resources.

It is implemented as a hybrid approach combining several components. The preprocessing component of the PERDIDO NER processing chain (PPC) transforms and pretags raw texts via different processes: sentence splitting, tokenization, lemmatization, and POS tagging. These shallow linguistic tasks are language dependent and are done using standard POS taggers.

The main component of the PPC addresses with the automatic retrieving and tagging of ENEs and geospatial information via cascaded finite-state transducers [9]. These cascaded finite-state transducers were developed using the CasSys program [8] available in the Unitex platform [31]. Figure 4 shows the transducer retrieving and tagging of ENEs of level 1 using a lexicon of geographical features and morphosyntactic patterns

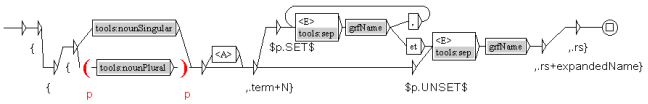


Figure 4: Transducer annotating ENE of level 1

ENEs and their associated spatial relations are semantically tagged using a multilayer markup language [20] following the TEI guidelines [6]. The cascaded finite-state transducers produce a generic annotation of ENEs (i.e., ENE boundaries are identified but not classified). Then, the last component of the PPC implements a gazetteer lookup method to classify them. As described by [30], the PPC uses the local linguistic context (i.e., feature type within ENEs), when available, to identify subtypes associated with toponyms (e.g., city, street, and church) to classify ENEs and, more specifically, to identify spatial ones (i.e., ESNEs). This purpose is closely related to the problem of toponym resolution, and the concept of ENE can be applied to identify spatial entities thanks to information contained within an ENE, such as feature types that can be found in geographical ontologies or lexicons.

The PERDIDO system also implements also some other components for to geocode places, such as toponym disambiguation [22] and itinerary reconstruction [21]. However, in the current work, the geocoding of place names is done in a further step of the process using dedicated geohistorical gazetteers focused on the city of Paris in the nineteenth century. Thus, at this stage, we propose using

[6] http://erig.univ-pau.fr/PERDIDO/

```
<placeName>
  <geogName type="R" subtype="ST">
    <geogFeat>
      <w lemma="rue" type="N">rue</w>
    </geogFeat>
    <w lemma="de" type="PREP">de</w>
    <name>
      <w lemma="Rivoli" type="NPr">Rivoli</w>
    </name>
  </geogName>
</placeName>
```

Figure 5: XML/TEI annotation of street names using the PPC

the PPC to transform the generic XML/TEI markups of ENEs into ESNEs when the feature type of the ENEs matches a concept of our geographical lexicon. Figure 5 shows an excerpt of the XML/TEI output of the PPC for a street name. The values of attributes *type* and *subtype* of the *geogName* element refer to GeoNames feature codes[7].

Additionally, XML/TEI files provided by the PPC are associated with an external style sheet file (CSS). This file contains simple instructions that describe how XML/TEI elements should be displayed on a Web browser. Figure 6 shows the XML/TEI result of one sentence from the novel *L'assommoir* in a Web browser. Street names are highlighted in blue, and the other spatial ENEs are highlighted in green.

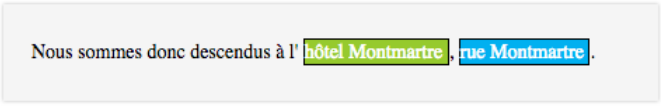


Figure 6: XML/TEI display of ESNE on a Web browser

As with TXM, the PPC also creates a concordancer that lists every occurrence of an ENE. This file also contains a column with the feature type of each ENE. Then, we can easily sort or filter results to obtain all the street names for each novel.

4 RESULTS

In this section, we describe the preliminary results obtained using TXM for the extraction of street names in the 31 novels of our corpus. We use these results as a basis and compare them to the results obtained using the PPC. Then, based on the results of this first task, we show how NER and textometric methods can be utilized together in order to process a text analysis.

4.1 Extraction of street names

As described in Section 3.2 and shown in Figure 2, building a CQL query for TXM may be very complex and time consuming, as every instance of an elaborate linguistic pattern must be matched without introducing too many false positives. For the current work, we built a query specifically to extract street names and not all types of place names. This request returns 2607 results from the 31 novels of our corpus, among which there are 56 plurals (i.e., enumeration of several street names), as shown in example (2). We manually reviewed these results and found that 44 do not refer to street names (i.e., false positives), among which 32 are plurals and refer to all the streets of a city or a neighborhood (see example 3).

(2) rues Godefroy et Mouffetard
'Godefroy and Mouffetard streets'

(3) rues de Paris
'streets of Paris'

Additionally, most of the false positive errors are due to special cases not covered by the query, such as punctuation marks (example 4a) or words that do not refer to a street name listed after the word *rue* (see example 4b). The boxes in examples (4) and (5) mark the pivots and the words outside the boxes are part of the right or left context.

(4) a. à l'entrée de cette [rue. . . Jean]
'at the entrance of this street. . . Jean'

b. en passant dans la [rue de Mme] Arnoux
'passing in the street of Mrs. Arnoux

However, the majority of errors are due to incorrect boundary detection; indeed, 127 results are malformed (i.e., only a part of the street name is correctly found according to the pivot column of the concordancer). This represents 5% of the total number of results returned by the query. Those results also refer to special cases not covered by the query and are due to hyphens, apostrophes or the number of connecting words between two capitalized words (see examples 5a-5c).

(5) a. [rue de la Tour-] d'Auvergne
'Tour-d'Auvergne Street

b. [rue Notre-] Dame-des-Champs
'Notre-Dame-des-Champs Street'

c. [rue de la Barrière] des Gobelins
'Barrière des Gobelins Street'

All these errors must be manually removed from the results before the geocoding step of the project. However, in this paper, we describe how we can build an automatic process. Thus, the results provided by TXM are used as a comparison basis to evaluate the results obtained using the PPC.

As described in Section 3.3, the main component of the PPC executes some cascaded finite-state transducers to semantically tag place names. Based on the concept of ESNE, the PPC implements a construction grammar foreseen for the retrieval of several geographical feature types, including street names. The feature types are stored in a local lexicon, but the PPC can easily be linked with a thesaurus, an ontology or any other linked open data. According to the results provided by the PPC, 112 feature types are found in the corpus and Table 2 shows the list of the most frequent geographical feature types found in the ESNEs (i.e., terms associated with a place name) of the 31 novels. This table shows that street names are the most used geographical feature in this corpus, which confirms the great interest in these novels for the cartographic analysis of Paris.

For the evaluation of the results produced by the PPC, we filtered the lines of the concordancer that only refer to street names. Then, we manually reviewed and compared each result with those previously obtained using TXM.

[7] http://www.geonames.org/export/codes.html

Table 2: Most frequent geographical feature types

Feature type	Occurrences	Feature type	Occurrences
rue	2583	quartier	106
boulevard	257	porte	105
maison	200	place	81
faubourg	149	bois	75
hôtel	134	avenue	68
pont	123	barrière	62
quai	122	route	58

Table 3: Number of results and errors obtained using TXM and PPC

	TXM	PERDIDO
# of results	2607	2583
false positive	44	7
false negative	39	26
malformed	127	4

Table 3 shows the number of results and errors returned by TXM and the PPC. The concordancer provided by the PPC lists 2583 results referring to street names (containing 21 plurals). Among these 2583 occurrences, 7 do not refer to a street name (i.e., false positive results), and 4 are malformed. Then, by comparing the concordancers of TXM and the PPC, we noticed that 39 street names are missing for TXM and 26 for the PPC (i.e., false negatives).

We measure the performances of the two approaches using three metrics commonly used in information retrieval: *precision* (P), *recall* (R) and F_1-*score* (F_1). The precision P is the ratio of the number of relevant street names annotated divided be the total number of street names annotated (i.e., both true positives and false positives). The recall R is the ratio of the number of relevant street names annotated to the number of relevant street name instances existing (i.e., both true positives and false negatives). The F_1-*score* F_1 is the harmonic mean of the precision and recall and is defined as:

$$F_1 = \frac{2 \times P \times R}{P + R}$$

According to the evaluation results shown in Table 4, the two methods obtain very high scores for the extraction of street names from our corpus. These results validate the choice of the PPC for the automatic extraction of place names. We obtain comparable results in terms of precision and recall (F_1-score of 99.3 versus 98.4), but more importantly, we obtain many more precise results by reducing the number of malformed occurrences by 97% (4 malformed occurrences obtained using the PPC versus 127 obtained using TXM). Although the number of malformed occurrences obtained using TXM can be improved by modifying the request, TXM was not designed for developing a fully automatic process for the annotation of place names.

In addition to the good results stored in the concordancer file, we are also interested in the XML/TEI annotation files produced by the PPC. Indeed, the XML/TEI files automatically generated contain not only the annotations of place names but also some geospatial and semantic annotations such as expressions of motion or perception in which a place name is involved. Our objective is to use the XML/TEI annotations with TXM for a textometric analysis.

Table 4: Evaluation scores obtained using TXM and PPC

	Precision	Recall	F_1-score
TXM	98.3	98.5	98.4
PERDIDO	99.7	99.0	99.3

This kind of analysis can be extended to compare the spatial footprints of different authors or different literary movements, e.g., Flaubert's realism and Zola's naturalism.

4.2 Textometric analysis

The use of a PPC upstream of a textometric tool such as TXM is therefore very appreciable because of the automatism and the rapidity that it brings in retrieving place names in French novels. The XML/TEI files produced by the PPC are compatible with TXM [17], which will make it possible to complete this analysis with additional ad hoc queries and to produce ultimately statistical analyses. Indeed, TXM implements an import module responsible for interpreting corpora encoded in the XML/TEI standard, and it provides innovative analytical corpus tools. We describe hereafter some examples of the results obtained during this first stage of the project.

Figure 7 shows the distribution of the number of occurrences of street names (blue) and the number of distinct streets mentioned (red) over the 31 novels of our corpus. One can notice the great disparity in the level of use of street names between the novels. Les Misérables by Victor Hugo obtains the highest score with more than 600 occurrences, followed by Le Ventre de Paris by Emile Zola (200 occurrences). Additionally, the number of distinct occurrences gives an idea of the richness of the references. Les Misérables is still the novel with the vastest realm of streets (200 distinct names), but l'Education sentimentale is now second ahead of le Ventre de Paris.

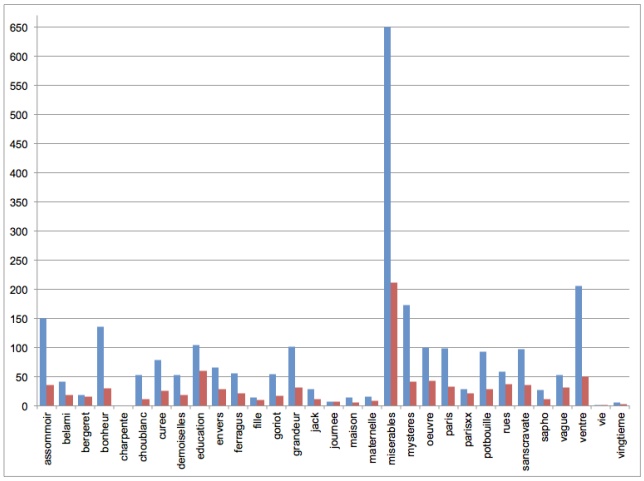


Figure 7: Number of street name occurrences (all occurrences in blue and distinct in red) per novel

Figure 8 illustrates how the presence of the four most frequently quoted streets differs. For instance, Rue Plumet and rue du Temple are mentioned mainly in one unique novel, i.e., in Les Misérables and in Les Mystères de Paris by Eugène Sue, respectively, while Rue de Rivoli and Rue Saint-Denis are distributed among many novels.

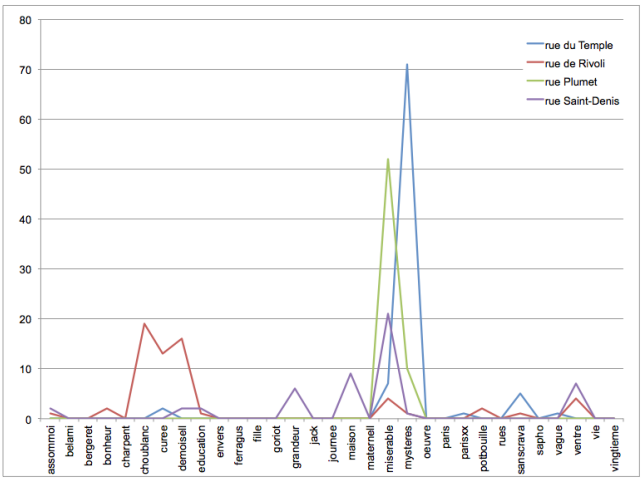


Figure 8: Number of occurrences of the four most frequent street names per novel

Figure 9 shows the distribution of the number of street names over the 31 novels normalized by the number of words in each novel. It is a way to correct the great variability of novel size. We calculate an indicator of the density of street names for each novel as follows: index = number of streets occurrences / 10,000 ∗ number of words (NbStr/10KWds). With its index of 10, the much longer Les Misérables is clearly overtaken by the shorter Ventre de Paris (with an index of 15), while Les Mystères de Paris, a book almost as big as Les Misérables, is left far behind. Some other shorter novels, such as Ferragus by Balzac and M. Choublanc by de Kock, also have an index over 10 and the same density of street names as that of Hugo's book. This very rapid and superficial analysis shows how different authors are in the use of toponymy in their writing.

Finally, Figure 10 shows the first example of a map built with an extract of 240 streets already located to understand the spatial logic of literary place naming in the 19th century Paris. Once the geohistorical gazetteer covering the area of Paris from 1800 to 1914 is complete, different types of maps will be produced using several conventional indicators of spatial statistics such as convex envelopes, weighted barycenters, standard dispersion ellipses and spatial interpolations [5]. Eventually, it will be possible to propose a platform for facilitating the cartographic analysis of other texts evoking Paris in the 19th century. This platform will be based on the automatic process of street name retrieval, as the gazetteer will be completed dynamically using the new streets mentioned in the texts.

5 CONCLUSIONS

In this paper, we proposed a method for retrieving place names from French novels. The preliminary results described in this paper highlight the great interest in combining NLP approaches (such as

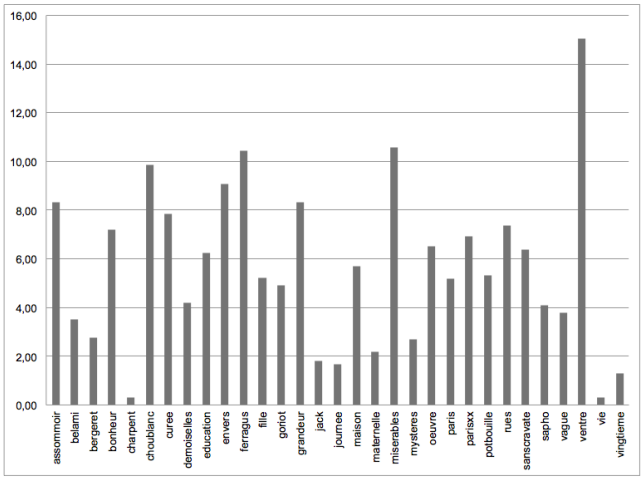


Figure 9: Number of street names normalized by the number of words per novel

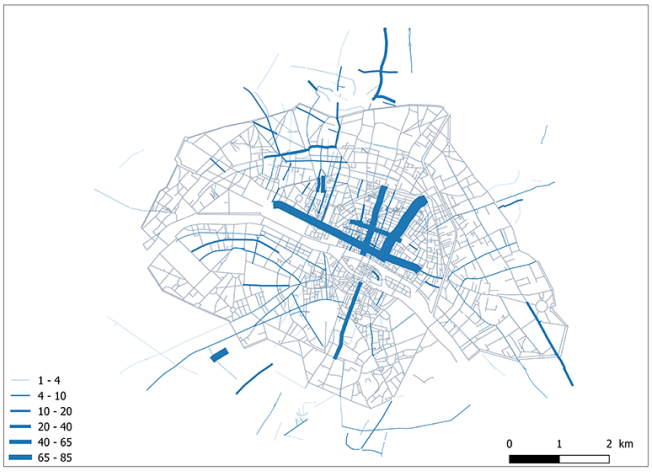


Figure 10: Number of occurrences of 240 street names represented using proportional line symbols

NER methods) with textometric analysis tools. We first described a proof of concept using CQL queries with the TXM platform. Then, the results were used to evaluate those obtained using the NER component of the PERDIDO platform. The high results in terms of precision and recall are comparable and the NER method reduces the number of malformed occurrences by 97%. These scores validate the choice of the NER method for the automatic process. Additionally, we showed the great interest in combining NLP approaches with a textometric tool to provide automated analysis of novels based on spatial named entities.

Since the geographical indexing of places in Paris in the 19th century has not been completed, it is not yet possible to validate the interest of geovisualizations for the analysis of the spatial fingerprints of various novels or authors. However, even without maps or comprehensive indicators, it remains undeniable that direct access to a corpus of text through the use of place names greatly transforms the ways in which Parisian space and fictional landscapes

can be explored. It becomes possible to interactively and simultaneously browse geographical and literary space. Having tools for interrogation, visualization and spatial analysis appears to be useful in making hypotheses and synthesizing results related to the spatialization of novels.

The automatization of the process of information retrieval from novels opens new perspectives regarding future work. For instance, we can process the geo-semantic information annotated using the NER component of PERDIDO to automatically characterize the semantic content associated with the spatial named entities by searching for the qualifiers used to describe them. For example, it will be possible to specify whether a street is large, dark, or animated or which characters are living in a particular place. Finally, a long-term goal is the visualization of displacements of one or several characters during a story and the representation of the temporal dynamics of places in the chronology of a narrative.

REFERENCES

[1] Beatrice Alex, Kate Byrne, Claire Grover, and Richard Tobin. 2015. Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing* 9, 1 (2015), 15–35.

[2] Beatrice Alex, Claire Grover, Jon Oberlander, Tara Thomson, Miranda Anderson, James Loxley, Uta Hinrichs, and Ke Zhou. 2016. Palimpsest: Improving assisted curation of loco-specific literature. *Digital Scholarship in the Humanities* 32, 1 (2016), i4–i16.

[3] Miranda Anderson and James Loxley. 2016. The Digital Poetics of Place-Names in Literary Edinburgh. *Literary Mapping in the Digital Age* (2016), 47.

[4] Frédéric Béchet, Benoît Sagot, and Rosa Stern. 2011. Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In *TALN'2011 - Traitement Automatique des Langues Naturelles*. https://hal.inria.fr/inria-00617068/document

[5] Noémie Boeglin, Michel Depeyre, Thierry Joliveau, and Yves-Francois Le Lay. 2016. Pour une cartographie romanesque de Paris au XIXe siècle. Proposition méthodologique. In *Actes de la conférence SAGEO'2016 - Spatial Analysis and GEOmatics*. Nice, France, 76–90.

[6] TEI Consortium (Ed.). 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/ (accessed July 2017). P5, version 3.1.0. Last updated on 15th December 2016.

[7] David Cooper, Christopher Donaldson, and Patricia Murrieta-Flores. 2016. *Literary mapping in the digital age*. Routledge.

[8] Nathalie Friburger and Denis Maurel. 2004. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science* 313, 1 (2004), 93–104. https://doi.org/10.1016/j.tcs.2003.10.007

[9] Mauro Gaio and Ludovic Moncla. 2017. Extended Named Entity Recognition Using Finite-State Transducers: An Application to Place Names. In *9th International Conference on Advanced Geographic Information Systems, Applications, and Services*. Nice, France.

[10] Ian Gregory and Christopher Donaldson. 2016. Geographical text analysis: Digital cartographies of Lake District literature. *Literary Mapping in the Digital Age* (2016), 67–87.

[11] Ian Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. 2015. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing* 9, 1 (March 2015), 1–14. https://doi.org/10.3366/ijhac.2015.0135

[12] Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. What's missing in geographical parsing? *Language Resources and Evaluation* (07 Mar 2017). https://doi.org/10.1007/s10579-017-9385-8

[13] Serge Heiden. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, Otoguro Ryo, Ishikawa Kiyoshi, Umemoto Hiroshi, Yoshimoto Kei, and Harada Yasunari (Eds.). Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan, 389–398. https://halshs.archives-ouvertes.fr/halshs-00549764

[14] Ryan Heuser, Mark Algee-Hewitt, Van Tran, Annalise Lockhart, and Erik Steiner. 2015. Mapping the emotions of London in fiction, 1700-1900: A crowdsourcing experiment. *Proceedings of the Digital Humanities* (2015).

[15] Linda L Hill. 2006. *Georeferencing: The geographic associations of information*. Mit Press.

[16] Kerstin Jonasson. 1994. *Le nom propre*. Duculot, Belgique, Louvain-la-Neuve.

[17] Alexei Lavrentiev, Serge Heiden, and Matthieu Decorde. 2013. Analyzing TEI encoded texts with the TXM platform. In *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*.

[18] Monica Matei-Chesnoiu. 2015. *Geoparsing early modern English drama*. Springer.

[19] Andrew McCallum and Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 188–191.

[20] Ludovic Moncla and Mauro Gaio. 2015. A Multi-layer Markup Language for Geospatial Semantic Annotations. In *Proceedings of the 9th Workshop on Geographic Information Retrieval (GIR '15)*. ACM, New York, NY, USA, Article 5, 10 pages. https://doi.org/10.1145/2837689.2837700

[21] Ludovic Moncla, Mauro Gaio, Javier Nogueras-Iso, and Sébastien Mustière. 2016. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science* 30, 2 (2016).

[22] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for Texts with Fine-grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus. In *22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*. ACM, Dallas,TX, USA, 183–192. https://doi.org/10.1145/2666310.2666386

[23] Franco Moretti. 1999. *Atlas of the European novel, 1800-1900*. Verso.

[24] Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.

[25] PERDIDO. 2017. Extended Named Entity Annotation Service. http://erig.univ-pau.fr/PERDIDO/api.jsp. (2017). [accessed 2017-07-9].

[26] Barbara Piatti, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright. 2009. Mapping literature: Towards a geography of fiction. *Cartography and art* (2009), 1–16.

[27] Thierry Poibeau. 2003. In *Extraction automatique d'information: du texte brut au web sémantique*. Hermès Lavoisier.

[28] Thierry Poibeau. 2011. *Traitement automatique du contenu textuel*. Lavoisier.

[29] Lisa F. Rau. 1991. Extracting Company Names from Text. In *Artificial Intelligence Applications*. IEEE, Miami Beach, 29–32. https://doi.org/10.1109/CAIA.1991.120841

[30] Erik Rauch, Michael Bukatin, and Kenneth Baker. 2003. A Confidence-based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1 (HLT-NAACL-GEOREF '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 50–54. https://doi.org/10.3115/1119394.1119402

[31] Unitex. 2017. Unitex/GramLab: an open source, cross-platform, multilingual, lexicon- and grammar-based corpus processing suite. http://www-igm.univ-mlv.fr/~unitex/. (2017). [accessed 2017-01-12].

[32] Barney Warf and Santa Arias. 2008. *The spatial turn: Interdisciplinary perspectives*. Routledge.

[33] GuoDong Zhou and Jian Su. 2002. Named Entity Recognition Using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 473–480. https://doi.org/10.3115/1073083.1073163