Mapping urban fingerprints of odonyms automatically extracted from French novels

Ludovic Moncla, Mauro Gaio, Thierry Joliveau, Yves-François Le Lay, Noémie Boeglin, Pierre-Olivier Mazagol

Mapping Urban Fingerprints of Odonyms Automatically Extracted from French Novels

Ludovic Moncla[a], Mauro Gaio[b], Thierry Joliveau[c], Yves-François Le Lay[d], Noémie Boeglin[c] and Pierre-Olivier Mazagol[c]

[a]INSA Lyon, CNRS, LIRIS UMR 5205, France; [b] Université de Pau et des Pays de l'Adour, CNRS, LMAP UMR 5142, France; [c]Université de Saint-Etienne, UMR EVS, France; [d]ENS Lyon, UMR EVS, France

ARTICLE HISTORY
Compiled February 16, 2019

ABSTRACT

In this paper, we propose and discuss a methodology to map the spatial fingerprints of novels and authors based on all of the named urban roads (i.e., odonyms) extracted from novels. We present several ways to explore Parisian space and fictional landscapes by interactively and simultaneously browsing geographical space and literary text. Our project involves building a platform capable of retrieving, mapping and analyzing the occurrences of named urban roads in novels in which the action occurs wholly or partly in Paris. This platform will be used in several areas, such as cultural tourism, urban research, and literary analysis. The paper focuses on extracting named urban roads and mapping the results for a sample of 31 novels published between 1800 and 1914. Two approaches to the annotation of odonyms are compared. First, we describe a proof of concept using queries made via the TXM textual analysis platform. Then, we describe an automatic process using a natural language processing (NLP) method. Additionally, we mention how the geosemantic information annotated from the text (e.g., a structure combining verbs, spatial relations, named entities, adjectives and adverbs) can be used to automatically characterize the semantic content associated with named urban roads.

KEYWORDS
geographical information retrieval; geoparsing; digital humanities; mapping; named entity recognition; odonyms; Paris; novels

1. Introduction

Literary maps and atlases are not recent inventions, but the pioneering work of Moretti gave a new impetus to exploring spatial structures in fictional stories with maps and graphs (Moretti 1999, 2005). New questions have emerged about the relationship between literature and cartography (Engberg-Pedersen 2017). Digital technologies greatly changed the way one can extract spatial information from literary texts and display places and space in narratives (Gregory *et al.* 2015, Cooper *et al.* 2016). Locating the places mentioned in a book is conceivable for one or two novels but difficult to consider for all novels of an author or all novels published during a certain period. Digital methods make a big difference. Improved recognition techniques in parsing

CONTACT Ludovic Moncla. Email: ludovic.moncla@liris.cnrs.fr

and recognition of geographical information resulting from named entity recognition (NER) methods may facilitate the extraction of place names, and GIS offers new ways to spatially interact with data extracted from a large corpus of books.

This paper describes an efficient method to retrieve and map odonyms (i.e., named urban roads) found in literary texts. In the first part of the paper (Section 2), we develop an approach based on the retrieval of lexical patterns via a textual analysis platform and propose a solution to analyze and map odonyms found in a corpus of French novels centered in Paris. In the second part (Section 3), we explore a more automatic approach based on natural language processing (NLP) methods to perform the task. Then, the experimental results of a comparison of the two alternative approaches are discussed. The second part ends by revealing the great interest in combining the two approaches. The highlights of this work and some challenging goals are presented in the concluding Section 4.

2. Mapping the space of Parisian novels

Our first aim in developing this methodology was to support a historical analysis of urban modernity in Paris in the 19th century based on the study of novels from that time (Boeglin 2018). By following the novels' characters across the city and by looking through their eyes (or the author's eyes), it is possible to see the city in a new light. For that purpose, it was necessary to read the selected books of the corpus, collect information about urban facts and descriptions and draw maps to locate the urban roads mentioned in the novels. These are time-consuming and arduous tasks when done manually. The risk of error and omission can be reduced by integrating automatic retrieval methods in a semi-automatic approach that allows the user to interact both with the text in context and the results of text mining.

2.1. *The project*

Our project aims to build a platform capable of retrieving, locating, mapping and analyzing the places in Paris mentioned in novels for a large public audience that includes town planners, historians, literature experts, cultural tourists and inhabitants curious about the lost and extant places described in novels. Our project is similar to the Palimpsest project, which used NLP technology to mine literary works set in Edinburgh (Alex *et al.* 2016), but the current project has a slightly different goal. The main objective of Palimpsest was to assist curators in identifying texts about Edinburgh from a large corpus of books (Alex *et al.* 2015). To create their map of emotion in London, Heuser *et al.* (2015) work with a sample of urban names extracted from a huge number of narratives. Similar to Anderson and Loxley (2016) for Edinburgh and Gregory and Donaldson (2016) for the Lake District in England, we aim to link textual and topographic places in a specific geographical area.

The first challenge is to recognize and extract the places mentioned in a fictional text. There is a large variety of places: towns, villages, buildings, hills and mountains, rivers, lakes, etc. There are many ways for a novelist to evoke a place: through an explicit name or by using relative references, e.g., "near", "behind", "at home". Some places can also be disguised or imaginary. The same novel can combine all of these different ways to evoke a place. Explicit place names are easy to retrieve if a preexisting list of place names exists. However, some cases can remain difficult to resolve because of place name ambiguities and homonymy.

We focus on odonyms for several reasons. Road names are a common way to locate a story in urban areas. At the scale of a city, the street network structures the textual space as well as the urban space. The name of a street refers to a precise urban element and denotes a neighborhood or an area of the city. Monuments or landmarks are a supplementary way to highlight the spatial frame of a novel but are more difficult to recognize automatically in a text when the structure of the place name offers a way to systematize its retrieval. Additionally, relative to explicit street names, odonyms can be useful to find unnamed urban roads or places mentioned in the novel.

To assign geographic coordinates to the references found in a text, it is necessary to refer to gazetteers or to consult historical sources. It is not always easy to determine if an unknown odonym is fictional, infrequent or forgotten. However, in a big city like Paris, abundant information exists about ancient place names and changes in street denomination. Undefined odonyms are quite rare. Visualization and analysis of odonyms may seem like a classical topic. However, because of the original nature of the data and the multiple ways to summarize and display information, it is worthwhile to address the question.

2.2. *Odonym recognition and location*

Several criteria were considered to build our experimental corpus of 31 French novels (Table 1). First, the books had to tell a story located mainly in Paris and describe urban life and space with a sufficient level of detail. We chose famous classic authors (Balzac, Flaubert, Hugo, Zola, etc.) and authors who were successful in their time but are quite forgotten today (e.g., Céard, Frapié, de Kock, Robida). The novels have been published during different historical periods, from the July Monarchy to the Third Republic, and represent different literary movements (romanticism, realism and naturalism). A manual preprocessing step was implemented to prepare the digital files of each novel and correct the errors introduced by the OCR process: spelling mistakes, deleted white spaces after apostrophes and corrected hyphenations.

To quickly retrieve odonyms from the novels, we use the open-source software platform TXM[1] as described in (Moncla *et al.* 2017). TXM provides tools for qualitative and quantitative content analyses of text corpora and implements lexicometric methods for corpus search and statistical text analysis (Heiden 2010). To search for specific lexical patterns (such as urban road names), we use the corpus query language (CQL) based on words and structure-level properties. A query can catch all text segments where the category "rue" (i.e., "street") is followed by one or several words beginning with a capital letter ([word!="\pLu.*"]). The query is described in detail in (Moncla *et al.* 2017), and Figure 1 shows its complexity, addressing case sensitivity (%c), diacritical marks (%d), plurals (lemma), and special cases to avoid too many false positives. This query generated a small number of false positives that were too closely associated with the corpus but lacked generality. 14 categories of odonyms were selected: *allée*, *avenue*, *boulevard*, *cour*, *galerie*, *impasse*, *parvis*, *passage*, *place*, *pont*, *port*, *quai*, *rue*, *square*. Writing a query adapted to each of these categories is a complex task.

Figure 2 shows the simpler and more generic query that was finally adopted. This query generates many false positives that must be manually removed before using the results for the geocoding step of the project. We come back to this issue in Section 3.

All valid odonyms have been located by checking old street atlases and historical

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

```
[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 1. First example of a CQL request to extract street names (Moncla *et al.* 2017)

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

Figure 2. CQL request to extract street names (Moncla *et al.* 2017)

websites. We found and located 3433 references to urban roads in the 31 novels, for a total of 712 unique roads that are either extant (634) or extinct (78). A GIS was built using two sets of data for the first stage of the project. The Plan Vasserot (1810-1836), digitized by the Alpage project[2] (Noizet *et al.* 2013), is the reference for the roads that disappeared after 1850. The present street network available on the ParisOpendata website[3] is the basis for the roads present for the entire period or built after 1850. Automatically retrieving a fair geometry of Parisian streets extracted from historical maps will be easier with the tools developed by (Cura *et al.* 2018). This dataset was used to analyze the distribution of odonyms in the novels and produce a first attempt to draw the spatial fingerprints of novels and authors based on all named urban roads extracted from the texts.

2.3. *Quantitative indicators*

Are there major differences in the way our 31 novels cite Parisian urban roads? Several indicators have been calculated to answer this question (Table 1).

The size of the books in the sample varies from less than 50,000 words (*La fille aux yeux d'or*, *La maison du Chat-qui-pelote*, *Paris*, *Paris au XXème siècle*) to "monsters", like Hugo's *Les Misérables* or Sue's *Les Mystères de Paris* with more than 640,000 words. There is a high correlation between novel size and the number of references ($R^2 = 0.86$), as indicated by the 755 urban road references of *Les Misérables*, which is almost 3 times more than the number of references in *Les Mystères de Paris* (263), which is followed closely by Zola's *Le ventre de Paris* (246). In contrast, some novels mention very few roads: Robida's two anticipation novels mention less than 5 streets, while the average in the 31 novels is 111; *La maison du Chat-qui-pelote* and *La fille aux yeux d'or* by Balzac, *Une belle Journée* by Céard, and *La Maternelle* by Frapié mention fewer than 20 odonyms. Without *Les Misérables*, the correlation remains important ($R^2 = 0.4$), and *Le ventre de Paris* appears to cite a lot of odonyms relative to its size.

To account for the size effect, we calculate an odonym density index, which is the number of roads cited per 1000 words. With this indicator, *Les Mystères de Paris* falls far below the average (0.7) value with a density of 0.4. *Les Misérables* (1.2) is far exceeded by *Le Ventre de Paris*, which has the highest density of the novels examined (1.8 citations per 1000 words); many other novels, such as *M. Choublanc à*

[2] http://alpage.huma-num.fr/
[3] https://opendata.paris.fr/

Table 1. Indicators for evaluating the differences in the way the 31 novels cite Parisian urban roads. The novels are ordered chronologically by publication date. Authors are grouped by colours and values greater than the column average are in red.

ID	Author	Pub. date	Nbr. of words	Nbr. of ref.	Nbr. of unique streets	Nbr. of ref. per unique street	Nbr. of ref. for 1000 words
Maison	Honoré de Balzac	1830	24635	14	5	2.8	0.6
Ferragus	Honoré de Balzac	1833	53723	63	26	2.4	1.2
Fille	Honoré de Balzac	1835	32577	19	12	1.6	0.6
Goriot	Honoré de Balzac	1835	112004	62	23	2.7	0.6
Grandeur	Honoré de Balzac	1837	121384	126	44	2.9	1.0
Mystères	Eugène Sue	1842	686439	263	70	3.8	0.4
Sanscravate	Paul de Kock	1844	177229	136	46	3.0	0.8
Envers	Honoré de Balzac	1848	87198	102	41	2.5	1.2
Choublanc	Paul de Kock	1856	80156	96	24	4.0	1.2
Miserables	Victor Hugo	1862	641244	755	225	3.4	1.2
Demoiselles	Paul de Kock	1863	144823	70	32	2.2	0.5
Paris XX	Jules Verne	1863	49030	51	40	1.3	1.0
Éducation	Gustave Flaubert	1869	182937	156	92	1.7	0.9
Curée	Emile Zola	1871	131364	129	49	2.6	1.0
Ventre	Emile Zola	1873	139688	246	70	3.5	1.8
Jack	Alphonse Daudet	1876	208287	55	17	3.2	0.3
Assommoir	Emile Zola	1877	202078	175	44	4.0	0.9
Journée	Henry Céard	1881	54113	18	14	1.3	0.3
Potbouille	Emile Zola	1882	175098	96	32	3	0.5
Bonheur	Emile Zola	1883	188926	157	43	3.7	0.8
Vingtième	Albert Robida	1883	101449	5	4	1.3	0.0
Sapho	Alphonse Daudet	1884	70755	34	16	2.1	0.5
Belami	Guy de Maupassant	1885	133873	57	26	2.2	0.4
Oeuvre	Emile Zola	1886	170080	185	81	2.3	1.1
Vie	Albert Robida	1890	65152	2	1	2	0.0
Paris	Emile Zola	1897	227001	140	61	2.3	0.6
Charpente	J.H. Rosny jeune	1900	98424	3	2	1.5	0.0
Bergeret	Anatole France	1901	68440	29	20	1.5	0.4
Maternelle	Léon Frapié	1904	86443	19	9	2.1	0.2
Vague	J.H. Rosny aîné	1910	182605	93	52	1.8	0.5
Rues	J.H. Rosny aîné	1913	89773	79	46	1.7	0.9

la recherche de sa femme by Kock and several novels by Balzac (*Ferragus*, *L'Envers de l'histoire contemporaine*) and Zola, have a density close to 1.2. The odonym density index barely shows an authorial style. Among the authors with several books in the sample, de Kock and Zola, and (to a lesser extent) Balzac tend to use a lot of odonyms in some but not all of their novels. However, we can notice that 10 of the 14 novels with an odonym density index above average were published between 1837 and 1877. Among the novels of our sample published between 1844 and 1873, only one has an index below average. This indicator, built on a larger sample, could help to explore the idea developed by Benjamin (1993) about the "panoramic literature" in Paris in the middle of the 19th century. Although he mentioned mainly documentary texts, such as "tableaux de mœurs", "physiologies", and various kinds of codes and guides, it could be possible to extend this category to the field of literary fiction.

Do the novels tend to cite many different roads at low frequencies or few roads at high frequencies? Does the novel deeply explore a local topography, or does it cover many roads in a superficial way? On average, the novels of our sample mention 41 different odonyms. *Les Misérables* is the champion with 225 unique roads named, followed by *Les Mystères de Paris* with 70 unique urban roads, Flaubert's *L'Education sentimentale* (92), *L'œuvre* (81) and *Le Ventre de Paris* (70) by Zola. This absolute indicator is of course also related to the size of the novel. The calculation of the ratio between the total number of roads and the number of unique roads cited provides an estimation of the relative odonym concentration of a novel. On average, there are 2.5

citations per unique road in the sample. The values of some novels are much higher. *L'Assommoir* and *"M. Choublanc..."* have a concentration of 4, *"Au Bonheur..."* and *"Le ventre ..."* have a concentration of 3.7, and *Les Misérables* has a concentration of 3.4. This concentration index decreases clearly after 1883 (when there are no novels above average). If there is no correlation between the two indexes, some novels appear to be more odonym-oriented than others, with both high odonym density and concentration, including two novels from Balzac, two from de Kock, *Les Misérables*, and five from Zola.

2.4. *Mapping*

Mapping odonyms, which is the best way to display and analyze the geographical distribution of street references in Paris, is a fundamental task of the project. In the usual Google-style map, a pin is associated with the centroid of the road and displayed on a Google or OpenStreetMap background map. It is a clear oversimplification of the information to conceal the fact that the number of references goes from 1 (for 248 roads) to 86 for the most cited road *La rue du Temple*. There are several ways to represent the quantitative distribution of odonym occurrences. We opt for a punctual map with proportional symbols (Figure 3).

Figure 3. Number of urban road references in all of the novels

If the global distribution is rendered well, the accumulation of symbols in the center of Paris and the disparity in the values make the map difficult to read, even when using classes of symbols instead of a strict proportionality. Moreover, rendering streets by using punctual symbols is debatable, given that precise addresses are very rare in the novels.

Since the entities of reference are roads, a linear visualization with line symbols

proportional to the number of occurrences appears to be a good solution. To respect the Bertin rule (Bertin 1967) of representing quantity by size, we use both the color and the thickness of the lines to denote quantity (Figure 4). Thus, it is possible to respect proportionality while keeping thinner symbols visible.

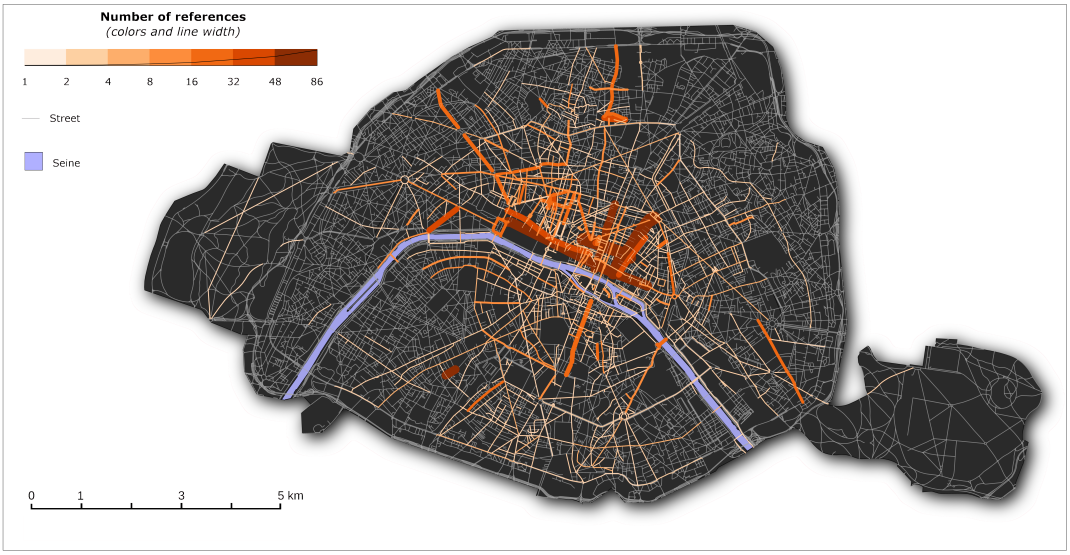

Figure 4. Number of novels that mention a specific street

This kind of map may tend to overstate the visual impact of the longest roads. For instance, the long *rue de Rivoli*, which is mentioned 65 times, is much more visible than the short *rue Plumet* in the southwest of Paris with its 57 references. However, there is a relationship between the length and the importance of an urban street in the real world as well as in literary representation. Thus, it is expected that the map shows this prominence.

The complex geometry of urban roads must be simplified before adopting a linear representation. For example, double roads and side paths must be removed because they contribute to the overstatement of the linear symbols of a single road. Additionally, there is a distortion related to public squares, which are represented as lines in this case. Another solution is to regroup the data at the road level by areal unit, which are, in this case, the 80 official neighborhood areas (quartiers) of Paris, by dividing the number of occurrences of road names by the part of the road length in every area. A classical GIS intersection is applicable if roads and area limits match perfectly. In most cases, roads act as partial or complete boundaries between two areas. Thus, this method can be tricky if the two sets of data do not match geometrically, which is precisely the case for our data. To shorten this explanation, we regroup the punctual data instead of the linear data (Figure 5).

Despite this approximation, the result more clearly shows the global spatial structure of the place names in our novels, which is barely visible in the previous maps. The number of occurrences decreases gradually from the center to the periphery, with a secondary pole in the north of the city.

A map of the density of occurrences on a regular grid is a good compromise between the linear nature of the information and the aim of visualizing the spatial structure of the phenomenon. The road density index is calculated for every cell per 1-Ha square pro rata of the part of the length of each road in the cell. Then, the indices of every road

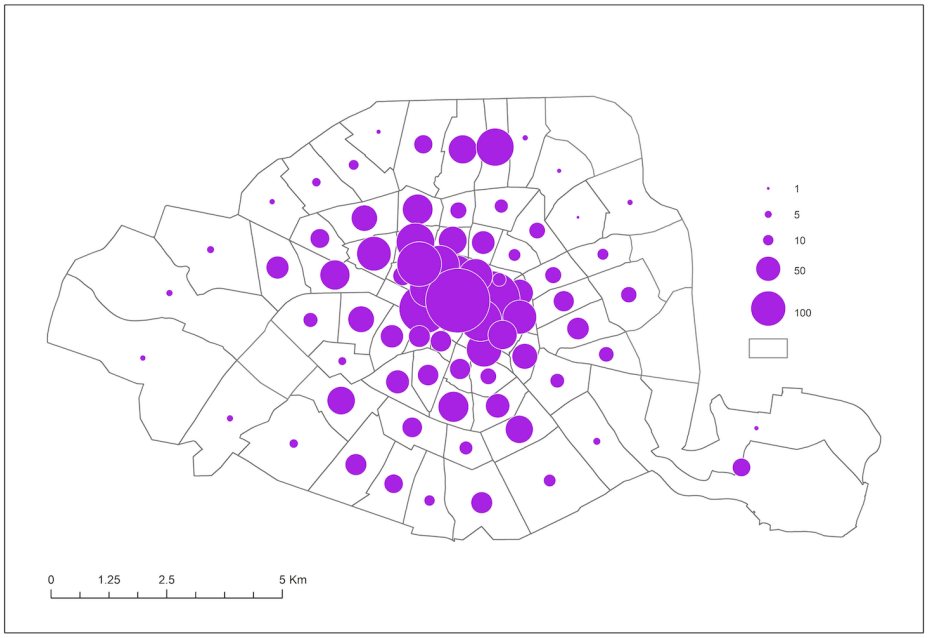

Figure 5. Number of urban road references by neighborhood area

are summed in the cell (Figure 6). The absolute quantitative values are lost, but we get rid of the bias related to the road length and the accumulation of punctual symbols in dense areas. The map still has a linear dimension while providing a smooth and continuous representation of the distribution of road references across space. However, the map produces a certain attenuation of the impact of long roads, the values of which are reduced by the largest number of cells that the road crosses. One problem of this technique that is often cited is a slight exaggeration of the crossroads between two streets. One advantage of this kind of map is the ease for adding other information.

For instance, we chose to show the Paris limits at the following three moments:

(1) the last surrounding medieval wall,
(2) the twelve Paris arrondissements before the annexation of peripheral localities in 1859, which led to
(3) the twenty arrondissements that are still in place today in the city.

This map shows very clearly the asymmetry between the two banks of the Seine. A high density of references tends to be located near the "*Rive Droite*" in the three extension areas. Four main poles can be observed. The first pole corresponds to the medieval city on the right bank of the Seine, with the addition of the *Ile de la Cité*. The second pole is the *Chaussée d'Antin* area, which is not far from the Opera and is a new bourgeois area built in the first part of the 19th century. The third pole, which has been already mentioned, is located in the north of the city beyond the limits of 1860 and includes the working-class neighborhoods of *La Goutte d'Or* and *Montmartre* described in *L'Assommoir* and *L'Oeuvre* by Zola. A final isolated pole outside the 1860 south limit is the *Rue Plumet*, which is mentioned frequently in *Les Misérables* and in *Les Mystères de Paris*.

The line map of the number of novels that cite a specific road (Fig. 7) completes the image of the main streets shared by the writers of the time and brings to light the traditional orthogonal axial structure of the road network inherited from the Roman

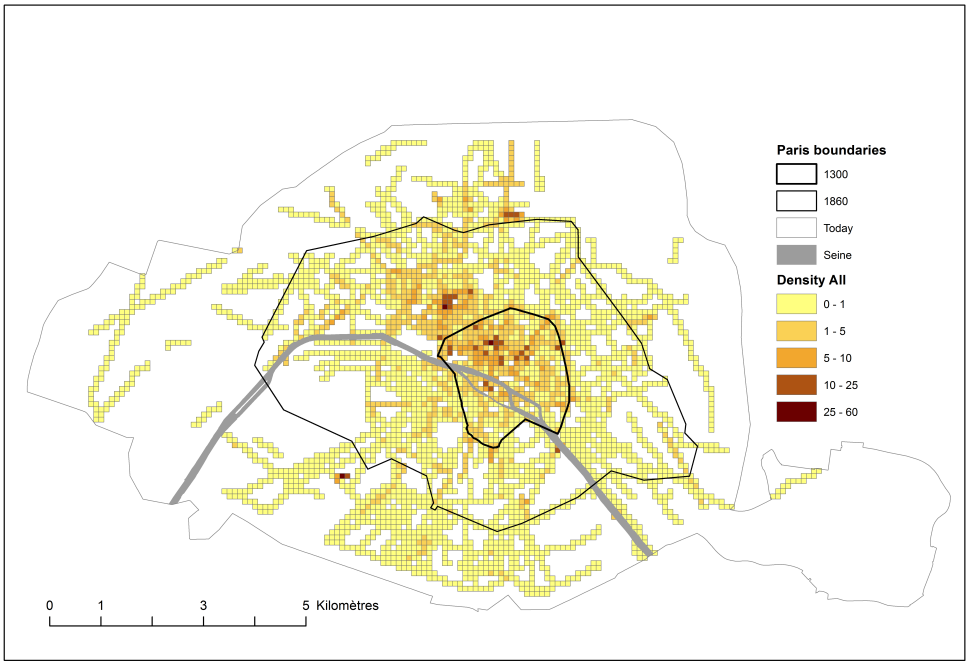

Figure 6. Density of urban road references in all of the novels

foundation and strengthened by the "*Grande croisée*" (the great crossing) project of Haussmann. Among the odonyms cited by one third of the novels, the majority are old streets with north-south orientation, such as *Rue Saint-Jacques*, *Rue de Richelieu*, *Rue Saint-Denis*, or more modern ones with west-east orientation, such as *Rue de Rivoli*, *Rue Saint-Honoré*, *Avenue des Champs-Elysée*.

Figure 7. Number of novels that mention a specific street

The high number of references to the old Paris is of course due to its permanence during the period. If the old center was drastically transformed after 1852 by Hauss-

mann, it remains a place where writers set their stories. Accordingly, a lower density of occurrences in peripheral areas is not unexpected. Some roads mentioned by Zola did not exist when Balzac was alive. We grouped the novels in three periods based on the year of publication between 1830 and 1914. The first period ends in 1860 when Paris expanded to its current limits after the annexation of the surrounding towns. The second period ends in 1880 with the definitive consolidation of the Republic. The maps drawn for every period evidence the temporal dynamic of the spaces named in the novels (Figures 8 - 10). Novels published before 1860 follow the extension of the city outside the medieval city and stay mainly within the 1860 limits. Between 1860 and 1880, the novels of our sample spread out in the new areas of urbanization. After 1880, the novels seem to relocate again in inner territories.

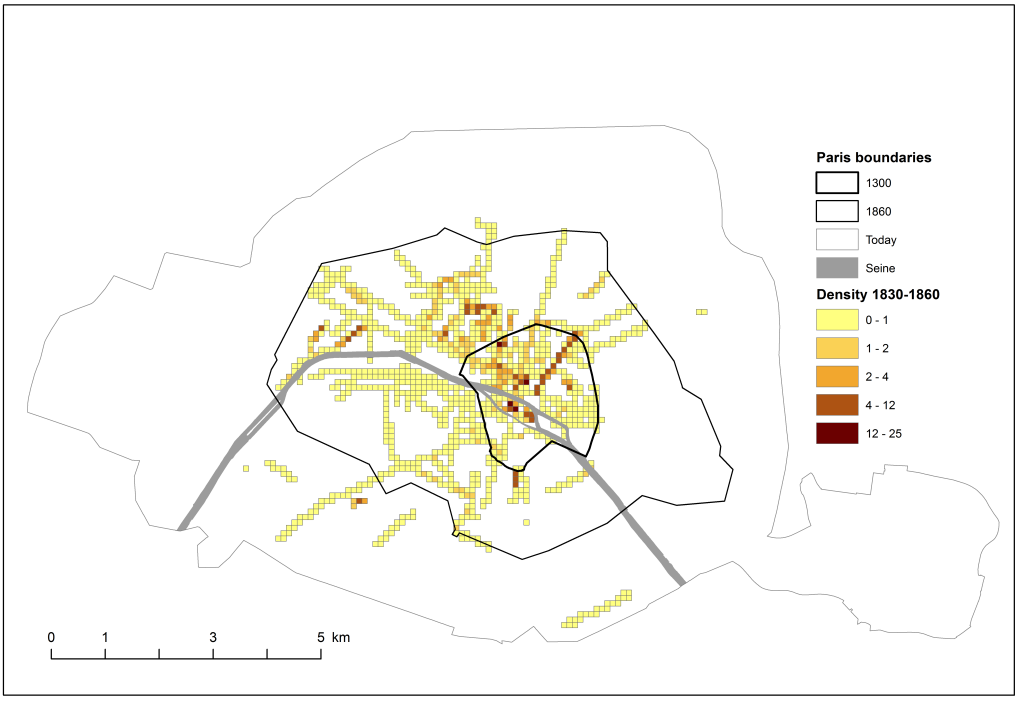

Figure 8. Density of urban road references in novels published between 1830 and 1860

In addition to giving general insight into the places visited by novels, maps can also help compare the spatial extents of different novels or authors. All of the maps described above can be used for this purpose. We propose Balzac and Zola as two examples using the grid system (Figures 11 - 12).

It is also possible to summarize a novel (or an author) by a synthetic view of its urban road name space. This kind of spatial signature combines different measures of the geographic distribution of the named roads, namely, the mean center of the number of occurrence, the standard deviation ellipse that marks the area of the highest density of occurrence, the smaller envelope of all the occurrences delimited by the convex hull polygon and the occurrences themselves (Figure 13). These measures can also be used to produce spatial indicators in addition to those in the list mentioned above.

Those maps must be understood as examples of tools designed to render a fictional odonym urban footprint. Our aim here is primarily methodological and based on a

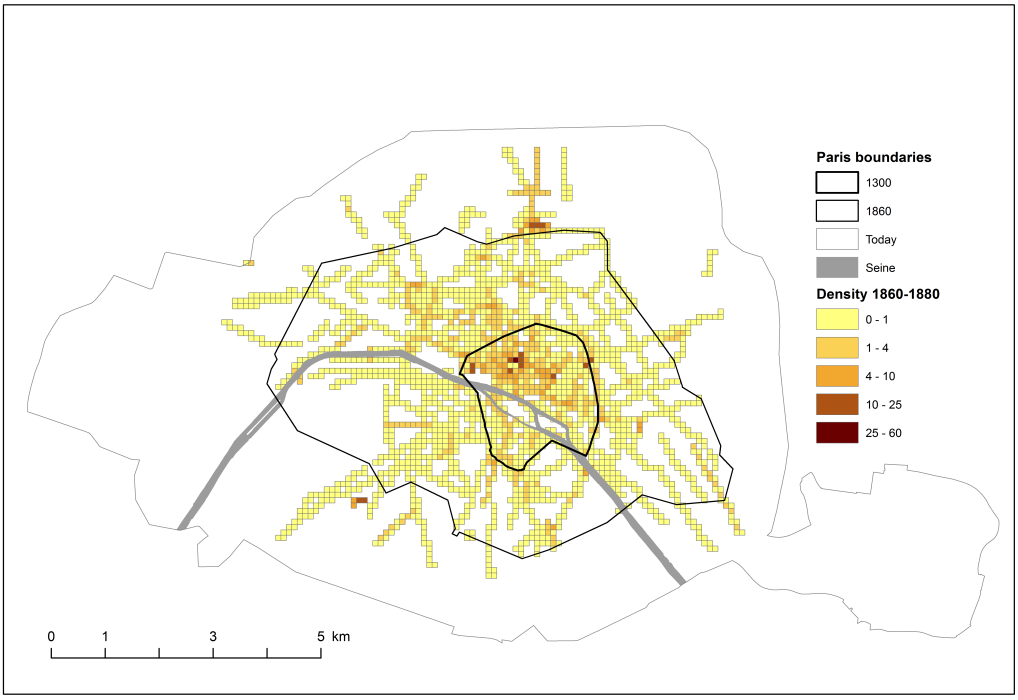

Figure 9. Density of urban road references in novels published between 1860 and 1880

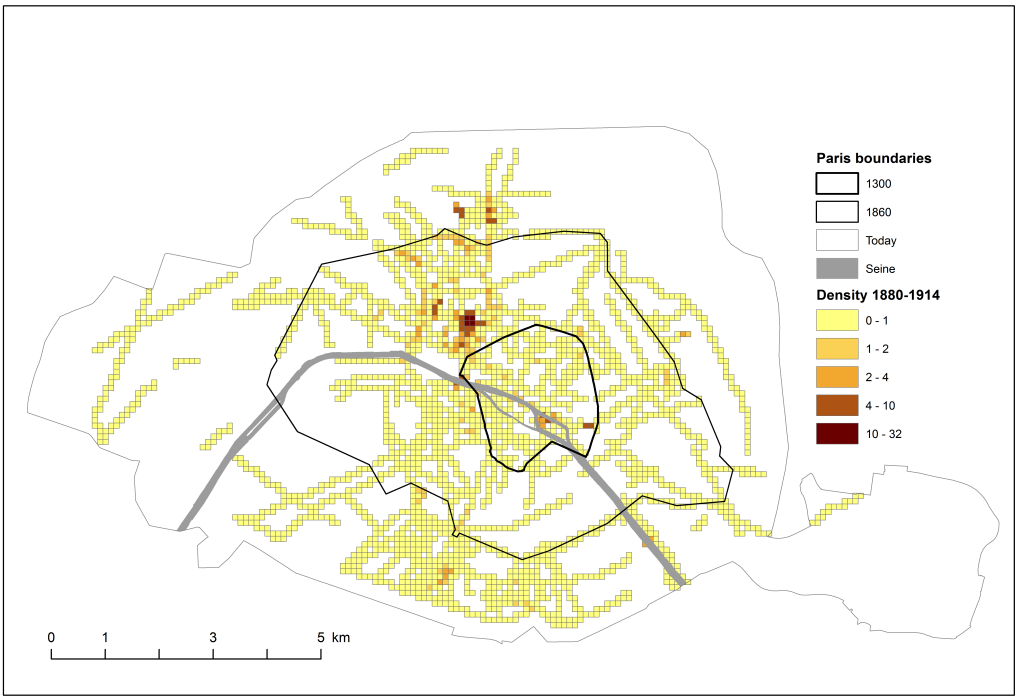

Figure 10. Density of urban road references in novels published between 1880 and 1914

Figure 11. Density of urban road references in Balzac's novels from the sample

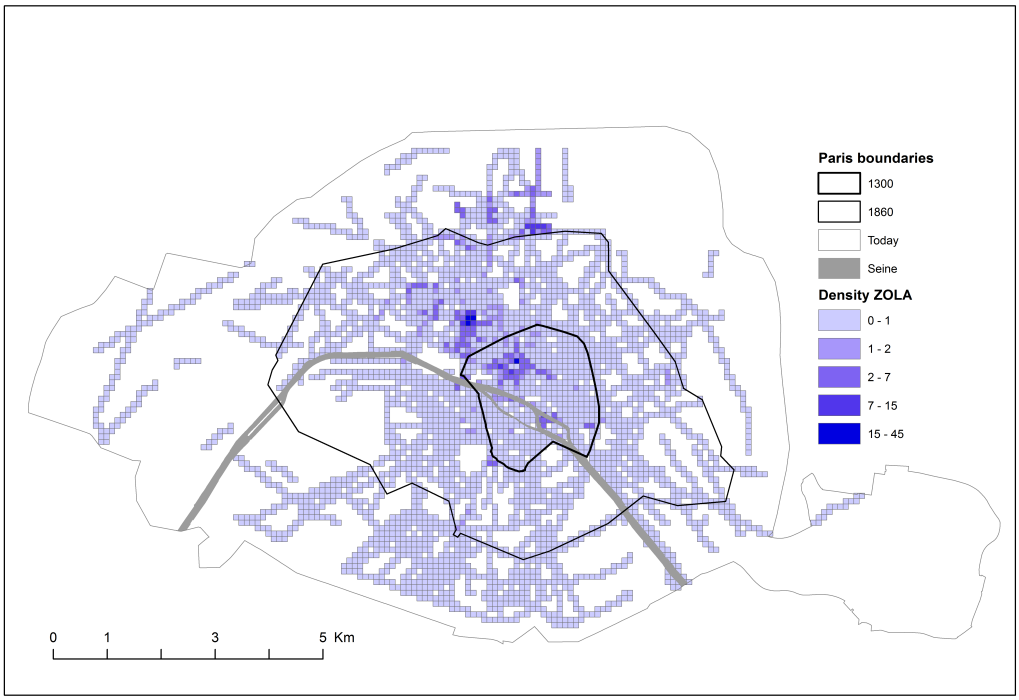

Figure 12. Density of urban road references in Zola's novels from the sample

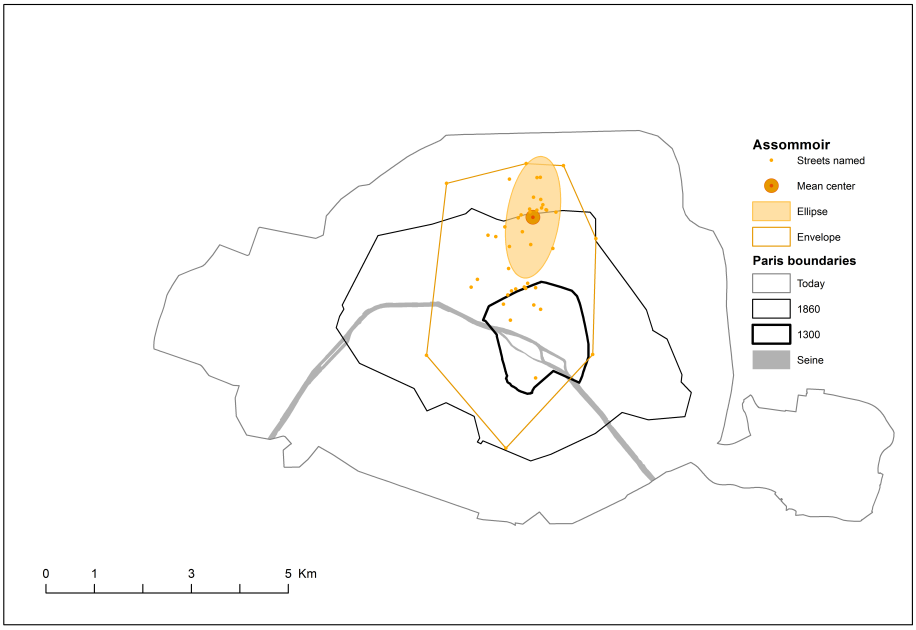

Figure 13. Urban road name space of the novel l'Assommoir (Zola)

small sample of novels. We do not intend to claim that this literary spatial representation reflects or is determined by the complex changes in Paris society and landscape during the 19th century. We think that these tools can be useful for historians and literature experts to explore the intertextual spatial links between novels, and the relationship existing between fictional texts and real space. The simple fact that some well-known structures of the Parisian road network show up in the novels of our sample is promising.

3. Toward an autonomous process

In the previous section, we described a proof of concept for analyzing and mapping named urban roads found in novels. We showed the feasibility of our proposal based on the retrieve of lexical patterns via a textual analysis platform. In the current section, we propose a more automatic approach based on an NLP method to perform the task of information retrieval and, more specifically, for extracting odonyms. The main objectives are to reduce human interactions during the process and to combine the two methods (i.e., NLP and textometric analysis) to improve performance and obtain more complete annotations.

3.1. *Automatically parsing and recognizing place names*

Computational methods for parsing and recognizing place references in textual documents have been specifically addressed in previous and current proposals. Melo and Martins (2017) and Gritta *et al.* (2017) propose two complementary surveys and compile an inventory of the actual knowledge and systems in this field. Despite very strenuous research efforts, several nontrivial challenges are still relevant. Some issues could

be related to the underspecified nature of textual spatial expressions (e.g., east side of the city, in the city's suburban west-end), while other issues are caused by the context-dependent nature of the place names, i.e., ambiguity. Smith and Mann (2003) have described three main categories of ambiguity: (1) referent ambiguity (i.e., the same place name is used for several locations); (2) reference ambiguity (i.e., the same location can be referred to by various names); (3) referent class ambiguity (i.e., the place name can be used in a nongeographical context (i.e., the name of a person or an organization). Another ambiguity, called structural ambiguity (Wacholder *et al.* 1997), may be due to the polysemy of some words constituting the place name (e.g., in "*rue couverte des Halles*" the lexeme "couverte" - covered - is part of the place name and adds precision to the nature of the referenced object).

The disambiguation of place names is considered by Leidner (2007) as a subtask of the toponym resolution in text and consists of associating a well-categorized named spatial entity with the most suitable fingerprint. Given these requirements, the quality of the categorization appears as the first essential step in such a process. To reduce the risk of the misclassification of named spatial entities, it is essential to focus on their conversational contexts. Accordingly, the first effort must focus on tagging the lexemes constituting the place name.

As stated by Gritta *et al.* (2017), the new generation of geoparsers needs to use more information to understand the meaning of the context. Under the principles of Jonasson's proposition (Jonasson 1994), which distinguishes pure and descriptive proper names, Gaio and Moncla (2017) jointly suggested the concepts of an extended named entity (ENE) and an extended spatial named entity (ESNE). ENE and ESNE may additionally encapsulate the various components of the named entity and its expansion in several overlapping levels (see examples 1a, 1b and 1c). Additionally, they also introduced the VT concept, and a VT grammar was previously proposed for the annotation of motion expressions related with ESNE.

(1) a. *le marché des* [*Halles*]
"the Halles Market"

b. *rue des* [*Halles*]
"Halles Street"

c. *à la fontaine de la* [*rue des* [*Poissoniers*]]
"at the fountain of Poissonniers Street"

(2) [*il la*] <u>reconduisit</u> [*à la station de fiacres de la* [*rue de Rome*]]
"he took her back to the cab station of Rome Street"

In this paper, in addition to the use of the VT grammar for the annotation of phrases like the one in example 2, we propose an extension of the VT concept to annotate human activities related to ESNE. In the examples 3a, 3b and 3c, it can be observed that, to describe a human activity, the structure of the sentence can have a shape that is close to that expressing a movement. To achieve this extension, it has been necessary to make an inventory of verbs used in the corpus to express human activity. The behavior of each verb within the corpus was then manually analyzed. Finally, a grammar adapted to this category of verbs was designed to allow automatic annotation.

(3) a. *Fumichon* possédait *un hôtel* *rue Saint-Dominique*
"Fumichon owned a hotel on St-Dominique Street"

b. *il ne pouvait* peindre son grand tableau dans *le petit atelier de la* *rue Douai*
"he could not paint his large painting in the small workshop of the rue de Douai"

c. il était installé maintenant dans *le petit rez-de-chaussée de la* *rue de Constantinople*
"he was now based in the small ground floor of Constantinople Street"

3.2. *Comparison of NLP and textometric methods*

As stated in Section 2.2, the use of CQL queries generates many false positives that need to be manually corrected. To reduce the number of human interactions in the process of retrieving place names, we propose the use of an automatic NLP approach in combination with the TXM platform. Based on the principles of grammar construction and improving the NER (named entity recognition) tool of the Perdido platform[4] (Moncla *et al.* 2017), our proposal implements the automatic annotation of odonyms. This solution semantically annotates ENE and ESNE and their associated spatial relationships using a multilayer markup language (Moncla and Gaio 2015) following the TEI guidelines (TEI 2016). Figure 14 shows an excerpt of the XML/TEI output of the odonym annotation for an urban road name. The values of attributes *type* and *subtype* of the *geogName* element refer to GeoNames feature codes[5], and the *rs* elements refer to unclassified ENE.

```
<placeName n="2">
<geogName type="R" subtype="ST">
<geogFeat>
<w lemma="avenue" type="N">avenue</w>
<w lemma="de" type="PREP">du</w>
</geogFeat>
<rs n="1">
<w lemma="bois" type="N">Bois</w>
<w lemma="de" type="PREP">de</w>
<rs n="0">
<name>
<w lemma="Boulogne" type="NPr">Boulogne</w>
</name>
</rs>
</rs>
</geogName>
</placeName>
```

Figure 14. XML/TEI annotation of named urban roads using the odonym annotation method

In (Moncla *et al.* 2017), we described a comparison of the results for street name extraction obtained using the CQL queries and those obtained with the Perdido NER tool. The automatic NER approach described in (Moncla *et al.* 2017) obtains comparable results in terms of precision and recall as those obtained with CQL queries after manual correction. The F1-score obtained after corrections with TXM is 98.4 and that for the fully automatic NER method is 99.3. Additionally, the NER method reduced

[4] http://erig.univ-pau.fr/PERDIDO/
[5] http://www.geonames.org/export/codes.html

by 97% the number of malformed occurrences. Results were very encouraging; thus, we have extended the comparison to all odonyms.

The number of false negatives (i.e., odonym references not found) is based on the comparison between the two methods and not with the actual number of odonym references that were not retrieved from the corpus. The aim of our proposal is not to build a fully automatic process (i.e., from text to map) but to offer new tools for experts (geographers, historians, . . .) to explore a corpus of novels. Thus, results show (Table 2) that some steps of the process (such as retrieving odonym references) may be semiautomatic, using NLP methods completed by human interactions.

Table 2. Results

	CQL-TXM	Perdido
# not odonyms found (false positive)	286	88
# odonyms not found (false negative)	11	117
# odonyms (true positive)	3573	3467
Total # results	3859	3555
Precision	0.926	0.975
Recall	0.997	0.967
F-score	0.960	0.971

The CQL query gives 3859 results against 3555 for Perdido NER (Table 2); 3871 results are identified either by Perdido NER or by TXM. Of the 3871 results, 31 are located out of Paris: 8 identified by the TXM request and 23 by the two tools. It is possible that other names that appear are not related to Paris. We do not address this question because this sort of task is still done via human interaction, and the objective here is to compare two ways of extracting odonyms. Of these 3871 results, 3512 (90%) are identified by the two tools in identical form, except for very rare exceptions related to the presence of dashes that can be interpreted differently in the original text (i.e., encoding issues, such as in example 4). A total of 117 results are not found by Perdido NER and 11 are not found by TXM. In the case of Perdido NER, some omissions are borderline cases, such as the example 5, due to the concept of ESNE. In this specific case, the odonym is annotated in the XML file as an ESNE of level 2 (see Figure 14), whereas most odonyms are annotated as ESNE of level 1 (see example 1). The problem is thus simply the way we analyze the XML files produced by Perdido NER for building the concordance method. Additionally, most errors are due to part-of-speech analysis[6], where proper names are categorized as common nouns or adjectives (e.g., *allée des Veuves*, *Quai aux fleurs*, *rue de l'Ouest*). Another problem is the integration of adjectives between the type of odonym and its name (e.g., *cour dite du Bâtiment-Neuf*, *rue couverte des Halles*). Other false-negative errors obtained with Perdido are due to plural cases. Indeed, Perdido does not annotate enumeration of entities (e.g., example 6). This limit of the rules implemented in the Perdido NER tool also shows the interest of the combination of the two approaches. Thus, the objective is to complete the results obtained using Perdido with new CQL queries. These queries may use the semantic information already annotated by Perdido and can be used to annotate more information or add new relationships between already annotated entities.

(4) rue Basse-Saint-Pierre
"Basse-Saint-Pierre street"

(5) *avenue du* Bois de *Boulogne*

[6]In this experiment, we use Treetagger (Schmid 1994) for part-of-speech analysis

"Bois de Boulogne Avenue"

(6) rues Saint-Honoré, Croix-des-Petits-Champs et du Bouloi
"Saint-Honoré, Croix-des-Petits-Champs and Bouloi streets"

Additionally, 286 results found by TXM and 88 found by both TXM and Perdido NER are false positives (i.e., not referring to odonyms). A significant number are due to ambiguities of construction around the terms "*cour*" (e.g., *cour royal*, *cour imperiale*, *cour criminelle*), "*passage*" (e.g., passing of persons), and "*place*", or constructions of sentences where the word street (i.e., "*rue*") is followed by a capitalized word that is not a street name but a name of character (e.g., *dans la rue où Frédéric...*). In this case, words such as "*cour*" and "*rue*" refer to something different than an odonym. The other important error is the case in which the result is not referring to a specific odonym but to several odonyms located inside a larger geographical entity (e.g., neighborhood, city) such as *les rues de Paris* or *les boulevards de Paris*.

Even if improving the Perdido tool or building more complex CQL queries is possible to optimize the results, a combined use of the two tools is necessary to automatically identify the odonyms and annotate them in a standard format such as TEI/XML. Indeed, TEI/XML can be loaded and analyzed in TXM with complementary queries to interactively explore the corpus and build ad hoc queries. By its annotation system of query results, TXM can complete the first annotation produced by Perdido. This combined use of the two tools involves an indexed management at the level of the word, which allows duplicates to be managed when there is a combination of requests and also allows the distribution of the odonyms to be analyzed in the space of the text to locate the nuclei at the different places of the narration. Lexicometric and statistical functions of TXM thus make it possible to complete the analysis.

4. Realizations and prospects

In this paper, we proposed a method to geographically retrieve and display odonyms from novels.

4.1. *Results*

A proof of concept based on a lexicometric method shows its advantages in terms of simplifying and accelerating the task of gathering information for the cartographic process. These methods implemented in the open-source software platform TXM allow lexical patterns in the corpus to be queried. The lexical patterns can be based on simple words or by combining words or group of words with structured properties.

The process developed for the proof of concept clearly shows that it would be interesting to have some of the treatments carried out using NLP methods. An experiment comparing the two processes was performed to evaluate the feasibility of using NLP methods for odonym recognition. The experiment confirms the preliminary results published in (Moncla *et al.* 2017). These results highlight the great interest in combining NLP approaches based on the principles of grammar construction (an improved version of the NER component of the Perdido platform) with textometric analysis tools. The high precision and recall of the results are comparable, and the NLP approach reduces the number of malformed occurrences by 97%.

Computer-assisted or automated geoparsing truly transforms the interactions between texts, maps and locations. The combination of automatic tools and manual and

visual interactions with the text is essential for exploring the place named footprint of a corpus of novels. We proposed an original mapping approach to efficiently visualize and analyze the results. New texts can be quickly added to our corpus for mapping. Part of the work remains manual, for instance, verifying that a road is truly located in Paris and looking for the exact location of ancient roads that are unknown in our gazetteer. As we progressively complete this gazetteer with new novels, the difficulty will progressively decrease. Once the corpus is digitized and stored in the right format, the new textometric tools largely facilitate the handling of the corpus. It becomes possible to automatically browse the text of the novels to search for new items (e.g., the famous Paris passages).

Geoparsing and recognizing automatically named entities and their discursive context in novels also opens new perspectives of future work.

4.2. *Using NLP to characterize places and movements*

The semantic information embedded in ESNE or VT annotations (in particular, geographic information) may help to automatically characterize places or visualize displacements. For instance, it would be possible to detect whether a street or a neighborhood is popular, commercial, or animated and which characters are living, working or travelling in a particular place. Another more challenging future task is related to the processing and representation of the meaning of motion-based narrations leading to relate their semantic content to reality, such as reconstructing the displacement on a map. The distinction between static descriptions of the landscape and real motions is then a crucial issue for the analysis of passages in novels featuring motion-based narration, in particular when fictive motion are involved in the texts. By using the other components annotated with the Perdido NER, a first analysis was made possible. The objective was to quantify in all of the novels of the corpus how often the odonyms are mentioned with a verbal expression. The second objective was to identify what proportion of these evocations provides a dynamic context or a static context; that is, when the ESNE appears in a VT structure as defined in section 3.1 and when the verb is a motion verb. The results show that, of the 3467 named urban roads mentioned in the 31 novels, 61% are in a VT structure. Among these VT structures, 26% have a movement verb, and the values vary between different novels from 15% to 40%. These first results encourage us to prioritize further work on this issue, acknowledging, of course, that this is only a first approximation, because the VT structure represents only one of many ways to evoke a static or dynamic context.

4.3. *Why map novel place names*

Retrieving odonyms in novels is significant in many fields. Place naming plays an important cultural role in fiction (Heuser *et al.* 2015). Place names are markers in the text as well as landmarks in the referential geographic space. Naming is the most direct way for a novelist to relate his story to a geographical space. By choosing place names, an author may want to anchor the space of the novel in a real existing place. The author may also mean to borrow the atmosphere of a city or a neighborhood for a story or to play with well-known cultural images. In literary analysis, considering only place names without context (i.e. such as a simple list of names) for understanding the role and meaning of space in a specific book is clearly inappropriate. Mapping can be relevant for an intertextual approach, by adopting the principle of distant

reading proposed by Moretti to look for patterns in space and time and compare chapters of books and novels, writers and genres (Moretti 2013). In this field, maps are useful for displaying the urban roads cited in the novels, but spatial interrogation, visualization and analysis tools can be used to make hypotheses and check new ideas about the urban fingerprints of novels or authors and their evolution over time. We reach a better understanding of how novelists drew their inspiration from the city and contributed to the nourishment of the urban imagination.

From a reception point of view, readers can turn themselves into tourists eager to visit the places mentioned in the novels they love (Joliveau 2009). Many catalogs of maps exist for all the places mentioned by an author. Taking an example among French novelists, the locations in the works of Balzac have been documented well for some time (Hoffmann 1965, Guichardet 1999). However, these kinds of data are useful only for literature specialists. With the system we propose, it is easy to not only automatically locate the street mentioned in a novel but also extract a located excerpt of the narrative. A smartphone application can help browsing the text when walking along the streets. In a nutshell, the attempt to exhaustively locate all the places in a big city named in a large number of novels is original and may be useful in many fields, from tourism to urban history, literary analysis to deep cartography (Bodenhamer and al. 2015). Interactively and simultaneously browsing geographical and literary space critically changes the ways in which both urban space and fictional landscapes can be visited. These tools offer an easy way to explore the urban territory through urban road naming in novels. It may help build a geographic framework for literary texts and attach a tiny imaginary aura to mundane topographic objects.

References

Alex, B., *et al.*, 2015. Adapting the edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, 9 (1), 15–35. 2

Alex, B., *et al.*, 2016. Palimpsest: Improving assisted curation of loco-specific literature. *Digital Scholarship in the Humanities*, 32 (1), i4–i16. 2

Anderson, M. and Loxley, J., 2016. The digital poetics of place-names in literary edinburgh. *Literary Mapping in the Digital Age*, 47. 2

Benjamin, W., 1993. *capitale du xixe siècle: le livre des passages*. Paris, France: Les éd. du Cerf. 5

Bertin, J., 1967. *Sémiologie graphique: Les diagrammes-les réseaux-les cartes*. Gauthier-VillarsMouton & Cie. 6

Boeglin, N., 2018. *Représentations romanesques de la modernité parisienne dans le "Grand XIXe siècle", 1830-1913*. Thesis (PhD). Université Jean Monnet Saint-Etienne, Université de Lyon. 2

Cooper, D., Donaldson, C., and Murrieta-Flores, P., 2016. *Literary mapping in the digital age*. Routledge. 1

Cura, R., *et al.*, 2018. Historical collaborative geocoding. *ISPRS International Journal of Geo-Information*, 7 (7), 262. Available from: http://dx.doi.org/10.3390/ijgi7070262. 4

Engberg-Pedersen, A., 2017. *Literature and cartography: Theories, histories, genres*. MIT Press. 1

Gaio, M. and Moncla, L., 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. 14

Gregory, I. and Donaldson, C., 2016. Geographical text analysis: Digital cartographies of lake district literature. *Literary Mapping in the Digital Age*, 67–87. 2

Gregory, I., *et al.*, 2015. Geoparsing, GIS, and Textual Analysis: Current Developments in

Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9 (1), 1–14. Available from: http://www.euppublishing.com/doi/abs/10.3366/ijhac.2015.0135. 1

Gritta, M., *et al.*, 2017. What's missing in geographical parsing? *Language Resources and Evaluation*. Available from: https://doi.org/10.1007/s10579-017-9385-8. 13, 14

Guichardet, J., 1999. *Balzac, archéologue de paris*. Genève, Switzerland: Slatkine. 19

Heiden, S., 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *In*: O. Ryo, I. Kiyoshi, U. Hiroshi, Y. Kei and H. Yasunari, eds. *24th Pacific Asia Conference on Language, Information and Computation*, November, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University, 389–398. Available from: https://halshs.archives-ouvertes.fr/halshs-00549764. 3

Heuser, R., *et al.*, 2015. Mapping the emotions of london in fiction, 1700-1900: A crowdsourcing experiment. *Proceedings of the Digital Humanities*. 2, 18

Hoffmann, L.F., 1965. *Répertoire géographique de la comédie humaine*. Paris, France: J. Corti. 19

Joliveau, T., 2009. Connecting real and imaginary places through geospatial technologies: Examples from set-jetting and art-oriented tourism. *Cartographic Journal*, 46 (1). Cinematic Cartography Special Issue. 18

Jonasson, K., 1994. *Le nom propre*. Louvain-la-Neuve, Belgium: Duculot. 14

Leidner, J.L., 2007. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Universal-Publishers. 14

Melo, F. and Martins, B., 2017. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21 (1), 3–38. 13

Moncla, L. and Gaio, M., 2015. A multi-layer markup language for geospatial semantic annotations. *In*: *Proceedings of the 9th Workshop on Geographic Information Retrieval*, GIR '15, New York, NY, USA. ACM, 5:1–5:10. Available from: http://doi.acm.org/10.1145/2837689.2837700. 15

Moncla, L., *et al.*, 2017. Automated geoparsing of paris street names in 19th century novels. *In*: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, Geo-Humanities'17, New York, NY, USA. ACM, 1–8. Available from: http://doi.acm.org/10.1145/3149858.3149859. 3, 4, 15, 17

Moretti, F., 1999. *Atlas of the european novel, 1800-1900*. London, UK: Verso. 1

Moretti, F., 2005. *Graphs, maps, trees: abstract models for a literary history*. London, UK: Verso. 1

Moretti, F., 2013. *Distant reading*. London, UK: Verso. 18

Noizet, H., Bove, B., and Costa, L., 2013. *Paris, de parcelles en pixels : Analyse géomatique de l'espace parisien médiéval et moderne*. Saint-Denis (Seine-Saint-Denis), Paris, France: Coédition PU Vincennes. 4

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. *In*: *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom. 16

Smith, D.A. and Mann, G.S., 2003. Bootstrapping toponym classifiers. *In*: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, Stroudsburg, PA, USA. ACL, 45–49. 13

TEI, C., 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. 15

Wacholder, N., Ravin, Y., and Choi, M., 1997. Disambiguation of Proper Names in Text. *In*: *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, Stroudsburg, PA, USA. Association for Computational Linguistics, 202–208. Available from: http://dx.doi.org/10.3115/974557.974587. 13