Automated Geoparsing of Paris Street Names in 19th Century Novels

L. Moncla, M. Gaio, T. Joliveau, and Y-F. Le Lay

ECOLE NAVALE

UNIVERSITÉ
DE PAU ET DES
PAYS DE L'ADOUR

EVS
Environnement
Ville Société

L. Moncla
ludovic.moncla@ecole-navale.fr

GeoHumanities'17

OUTLINE

1. Introduction
2. Background
3. Methodology
4. Conclusion

Introduction

Objectives

Objectives of the project

- Retrieve, map and analyze the occurrences of place names in fictional novels

Corpus of novels

- 31 French novels
- Published between 1800 and 1914
- Action occurs wholly or partly in Paris

OUTLINE

1 Introduction

2 **Background**

- Digital humanities
- Named Entity Recognition

3 Methodology

4 Conclusion

Background

Digital humanities

Spatial analysis of literary texts

- Generation of data sets
- Spatial representation of social or spatial relationships
- Visualisation and interpretation of texts (historical, novels, . . .)

Use of NLP for text mining

- Geoparsing texts
- Named Entity Recognition (NER) and Toponym resolution

Background

Named Entity Recognition

NER approaches

- Data-driven : machine learning
- Knowledge-based : heuristic and handcrafted rules

Named Entity (NE)

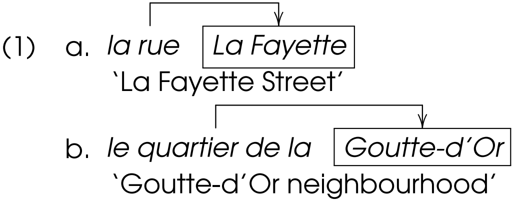
- Pure and descriptive proper names
- Absolute and relative spatial named entities

Background

Named Entity Recognition

Extended Named Entity (ENE)

- Entity built with a proper name and may be composed of one or more concepts
- Several levels of overlaping



(1) a. *la rue* *La Fayette*
'La Fayette Street'

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

Related work

Extended Named Entity

$$\text{ENE}\left[\begin{array}{ll} \text{level} & 1 \\ \text{type} & \textit{location} \\ \text{cat.} & \textit{descriptive} \\ \text{comp.} & \text{NN,IN,} \\ \text{ENE} & \left[\begin{array}{ll} \text{level} & 0 \\ \text{type} & \textit{location} \\ \text{cat.} & \textit{pure} \\ \text{comp.} & \text{NNP} \\ \text{lex.} & \textit{Rivoli} \end{array}\right] \\ \text{lex.} & \textit{rue de Rivoli} \end{array}\right]$$

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

FIGURE – Feature structure representing an ENE of level 1

OUTLINE

1. Introduction
2. Background
3. **Methodology**
   - Overview
   - Results
   - Combine NLP and textometric analysis
4. Conclusion

Methodology

Overview

Main objective

Automatically retrieve street names from novels published in the 19th century and whose action occurs in Paris.

Three steps

- Extract spatial named entities
- Locate these entities using historical sources and gazetteers
- Create maps adapted to the specificities of literary spaces

Methodology

Extracting street names via CQL requests

TXM platform

- `http://sf.net/projects/txm`
- Implements lexicometric methods for content analyses of text corpora
- CQL request find occurrences of specific entities and TXM produces a concordancer

CQL requests used in 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."]
```

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

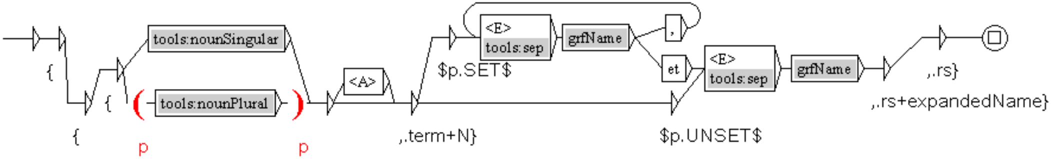
Methodology

Named entity recognition and classification

PERDIDO NER processing chain

- `http://erig.univ-pau.fr/PERDIDO/`
- A geographically oriented NER system
- Retrieve, tag and extract extended named entities
- Produce annotated XML files and a concordancer

Cascaded finite-state transducers



- Transducers are developed and processed with the Unitex platform

Methodology

Named entity recognition and classification

XML/TEI ouput produced by the PPC

```
<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>
```

- PERDIDO annotates also geo-semantic information (spatial relations, motion verbs, . . .)

Results

Extraction of street names

Evaluation scores

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

Classified errors

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

Results

Extraction of street names

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

- Use of a lexicon or a geographic ontology
- 112 feature types are found in the corpus

Combining NLP and textometric analysis

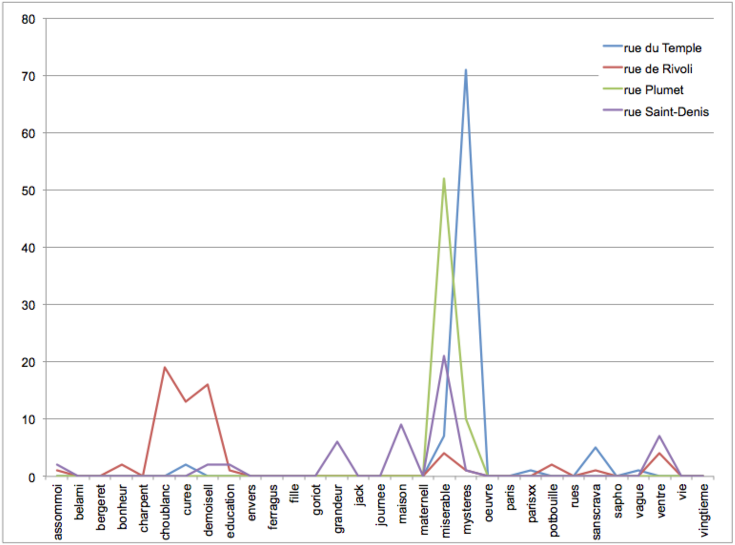
Interoperability between NER and textometric tools

- Building a fully automatic and more generic process
- The XML/TEI files produced by PERDIDO are compatible with TXM
- TXM provides innovative analytical corpus tools

Textometric analysis

Preliminary results

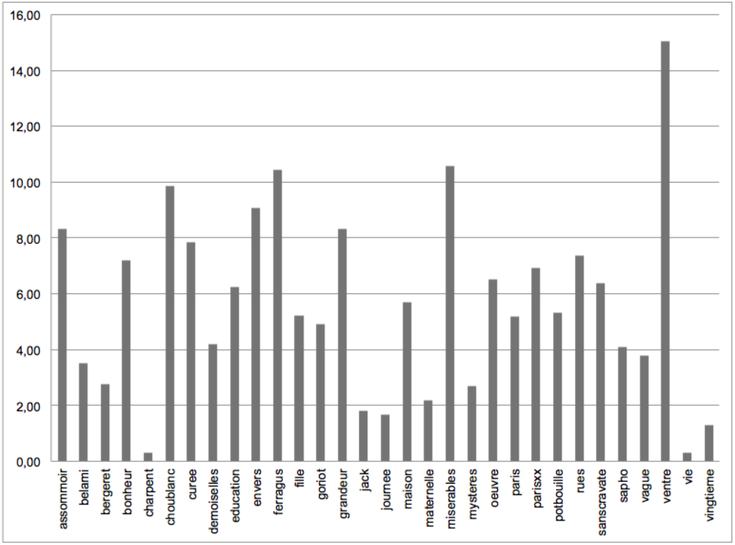
Occurrences of the four most frequent street names



Textometric analysis

Preliminary results

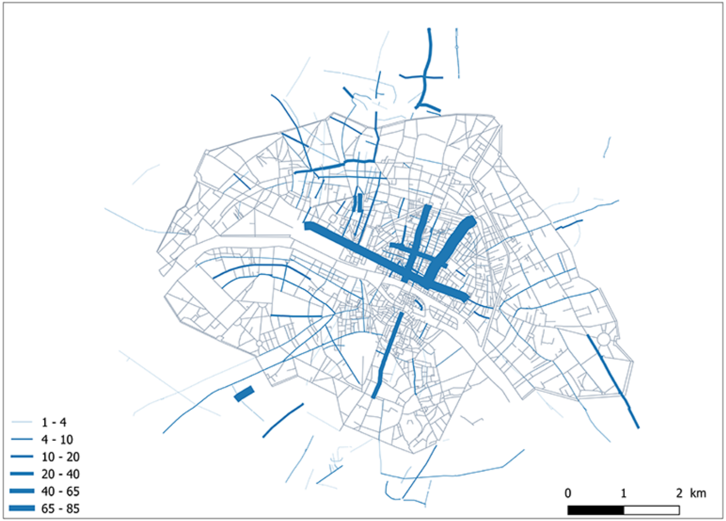
Distribution of street names normalized by number of words



Textometric analysis

Preliminary results

Map showing the number of occurrences of street names



OUTLINE

1. Introduction
2. Background
3. Methodology
4. **Conclusion**

Conclusion

Retrieving street names from novels

- The evaluation validate the choice of the NER method for the automatic process
- Results show the interest of combining NLP approaches and textometric analysis tools

Further work

- Finish the other steps (creating maps once the gazetteer will be finished)
- Use more geo-semantic annotations provided by PERDIDO
  - semantic content associated with the spatial named entities
- Visualization of displacements of a character and the representation of the temporal dynamics of places

Thank you for your attention

CONTACT

Ludovic Moncla (IRENav)
ludovic.moncla@ecole-navale.fr

Mauro Gaio (LIUPPA)
mauro.gaio@univ-pau.fr

Thierry Joliveau (EVS)
thierry.joliveau@univ-st-etienne.fr

Yves-François Le Lay (EVS)
yves-francois.le-lay@ens-lyon.fr

ECOLE NAVALE

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

EVS
Environnement
Ville Société