Normalisation of 16th and 17th century texts in French and geographical named entity recognition

Eleni Kogkitsidou, Philippe Gambette

NORMALISATION OF 16TH AND 17TH CENTURY TEXTS IN FRENCH AND GEOGRAPHICAL NAMED ENTITY RECOGNITION

Eleni Kogkitsidou
LIGM, Univ Gustave Eiffel, CNRS; LISAA, Univ Gustave Eiffel
eleni.kogkitsidou@univ-eiffel.fr

Philippe Gambette
LIGM, Univ Gustave Eiffel, CNRS; LATTICE, CNRS, ENS / PSL, Univ. Sorbonne Nouvelle
philippe.gambette@univ-eiffel.fr

http://doi.org/10.1145/1122445.1122456

ABSTRACT

Both statistical and rule-based methods for named entity recognition are quite sensitive to the type of language used in the analysed texts. Former studies have shown for example that it was harder to detect named entities in SMS or microblog messages where words are abridged or changed to lowercase. In this article, we focus on old French texts to evaluate the impact of manual and automatic normalisation before applying five geographical named entity recognition tools, as well as an improved version of one of them, in order to help building maps displaying the locations mentioned in ancient texts. Our results show that manual normalisation leads to better results for all methods and that automatic normalisation performs differently depending on the tool used to extract geographical named entities, but with a significant improvement on most methods.

1 Introduction

In the context of the *Cité des Dames* project, which aims at revealing the presence of women creators in cities along centuries, geographical named entity recognition (geoNER) algorithms play an important role. They are used to automatically find urban locations in French texts written by women, from the 16th to the 20th century, in order to build guided walks of cities, in the steps of women who wrote about these locations. Some of these corpora are easily handled by state of the art geoNER algorithms, such as novels written after the 18th century, or biographies or autobiographies of women creators. However, geoNER tools have been shown to obtain worse results for older texts [1, 2] or for texts where characters have been transformed to lowercase or uppercase [3].

Some of the geoNER methods may be trained directly on texts similar to the ones we wish to analyse, but this requires some effort at least to provide gazetteers, to build corpora providing a more appropriate language model or even to label geographical named entities.

Another strategy consists in building or using existing normalisation tools to preprocess the texts before applying geoNER recognition [4, 3], even though this approach was shown to have limited interest for old texts in English [5]. Note that these normalisation tools may be useful on their own, for example in order to obtain easier to read editions of these old texts. Furthermore, they do not need to be specific to the corpus we are analysing, as it is possible to combine in the preprocessing step several tools which handle/tackle specific linguistic or stylistic processes, such as "modernisation" for old texts, case normalisation for verse texts (theater plays or poems), abbreviation resolution, etc.

Named entity recognition is an important process for many natural language processing applications, especially for information extraction. Indeed, information extraction aims to detect relevant information, such as extracting specific events. Most researches on named entity recognition are based on modern text corpora using a formal language such as scientific, encyclopedic or journalistic texts which are relatively long with well-structured sentences, a proper use of

uppercase and without misspelled words. Fewer works deal with toponymic named entities from French literary texts. For example, for the Renom project, 16th century texts were analysed to automatically extract geographical proper names with the Unitex and CasSys open source software [6]. The combination of an hybrid approach based on the PERDIDO platform [7] with queries from the TXM system [8] is proposed in [9] to extract occurrences of odonyms in novels of 19th century taking place in Paris. Another approach based on the SEM system [10] and REDEN[1] [11] for named entity linking was designed in order to recognise geographical named entities in texts from 19th century French literature [12].

These works deal with documents that are relatively recent, mostly from the 19th century. PERDIDO was also applied (as well as the Edinburgh Geoparser) on geographical entries from 18th Century Canonical Encyclopedia, where it showed a lower performance than on 19th century texts [2]. Dealing with even earlier texts, from the 16th and 17th century, in their original version, is not a trivial natural language processing task since the French language is evolving during this period, with a general spelling standardisation [13]. It has been noted that it presents an extreme graphic variability (*scauoir/sauoir/savoir*, alternating *u/v* and *i/j*). Indeed, it retains lot of archaisms (*amy/ami*), the flexional system is not yet established (*amiz/amis*, *lieus/lieux*) and the accentuation is not very regular [14]. Speaking about irregularities we may dare to compare the language of this period with the computer-mediated communication[2] (CMC), a more recent one. In [3], it has been proven that automatic normalisation of SMS would increase the performance of traditional methods. We can conclude that the normalisation step is essential to perform named entity recognition in SMS and by extension in texts with similar characteristics.

In this article, we study the impact of text normalisation on geographical named entity recognition in 16th and 17th century texts in French. We apply several existing geoNER methods which were not trained on such corpora and test them on the original texts as well as on normalised texts, both automatically and manually.

In the following, we focus on geographical named entities corresponding to proper nouns. For example, in "chasteau de Fleurines" ("Fleurine's castle") we only label "Fleurines". Indeed, we consider that proper nouns are easier to find in gazetteers or open linked databases and that the output of automatic methods currently needs to be revised by humans, who may provide more accurate labels if necessary, in case the proper noun is included in a larger expression.

Furthermore, when a geographical proper name is used inside a named entity describing one person (for example in "monsieur de Montigny", "monsieur le comte de Lalain", "messieurs les evesques de Lyon, d'Ambrun et de Vienne" or "le roy d'Espaigne"), we do not consider it as a geographical named entity. This will be useful for the "Cité des Dames" project in order to extract significant relationship between the texts and the locations we automatically find inside. We expect to be able to handle a corpus of more than 500 texts in French written by women, gathered from online digital libraries such as Gallica (only the ePub files, where the texts obtained from OCR have been manually corrected), the Gutenberg project, Wikisource, Theatre-Classique.fr, Labex OBVIL and Bibliothèques Virtuelles Humanistes.

2 Material and methods

2.1 Corpus

In order to evaluate the impact of manual or automatic normalisation, we have built an annotated corpus of three texts, both in the original version and in a "modernised" version, with a total of 264 geographical named entities. This corpus is available under the LGPL-LR license at `https://github.com/PhilippeGambette/GeoNER-corpus`.

Looking for texts already available in their original version (without any of the modernisation practices which may be present in digital or paper editions) as well as texts labeled with geographical named entities, we identified a parallel corpus aligning original and modernised texts, PARALLEL17[3], and a corpus of theater plays from the 17th century where place names have been identified, CARTO17[4] [15, 16]. Both are available on the GitHub repository of the e-ditiones project, supervised by Simon Gabay. We selected the two texts available in both corpora, that is two theater plays, one by Racine, *Bérénice* (1676), and one by Molière, *Dom Garcie* (1694). As the plays were available in different editions in the two repositories, we chose the version available on the PARALLEL17 repository and manually transferred the geographical named entity labels (26 in *Dom Garcie*, 69 in *Bérénice*) from the CARTO17 files in order to get our reference annotated files, containing respectively 12011 and 15385 words (in the modernised version).

[1] `https://github.com/cvbrandoe/REDEN`

[2] This term describes digital communication that involve receiving and transmitting written messages using an electronic devise (e.g., instant messaging, email, chat-bots, social network services, etc.).

[3] available at `https://github.com/e-ditiones/PARALLEL17`

[4] available at `https://github.com/e-ditiones/CARTO17`

We also used the edition (in modernised French) provided by Éliane Viennot of the *Mémoires* by Marguerite de Valois. Geographical named entities were manually annotated by Isabelle Qin and checked by Caroline Trotot in the part of the memoirs about Marguerite de Valois's trip to Flanders. We manually transferred these annotations to a non-modernised XIXth century edition of this text by François Guessard, which we made available on the French Wikisource. We also modernised the text manually, using MEDITE [17] to compare the obtained results with the edition by Éliane Viennot in order to spot and correct a few oversights. Finally, we obtained a text of 12251 words (in the modernised version), containing 169 geographical named entity tags.

Finally, the whole corpus was automatically modernised by Jonathan Poinhos with his ABA software (version of September 4th, 2020) described below[5]. An example of sentence in the three versions of the corpus is given in Table 1.

Table 1: Illustrations of the three versions of the corpus.

Original	Des *Vaiffeaux* dans <placeName>*Oftie*</placeName> *armez* en diligence, *prefts* à quitter le port de *momens* en *momens*, n'attendent pour partir que vos *commandemens*.
Automatically modernised with ABA	Des Vaisseaux dans <placeName>Ostie</placeName> *armez* en diligence, prêts à quitter le port de moments en moments n'attendent pour partir que vos commandements.
Manually modernised	Des Vaisseaux dans <placeName>Ostie</placeName> armés en diligence, prêts à quitter le port de moments en moments, n'attendent pour partir que vos commandements.

2.2 Geographical named entity recognition methods

2.2.1 Combining existing geoNER tools

We developed the `GeoNER` tool[6], a comparator of five geographical named entity recognition tools using different approaches. The tool responds to a number of needs that have been defined at the outset : a) an output compatibility that would enable an automatic evaluation, b) a uniform file output format since several formats are possible (XML-TEI, HTML, IOB, text with tags, as illustrated in Table 2), c) a capability to process several dozen of texts at once, d) a toolkit for splitting long texts and gathering the output texts together and partial normalisation of texts (long s, ampersand, etc.), e) a naming file and directories produced to enable an automatic evaluation thereafter and f) alignment of detected geographical named entities in the reference and hypothesis texts

`GeoNER` was developed to extract and evaluate geographical named entities calling CasEN, CoreNLP, Perdido, SEM and spaCy tool. CasEN[18, 19], is a deterministic method integrated on Unitex[7] using an analysis cascade composed by transducers, lexical resources and patterns. CoreNLP[8] a set of natural language analysis tools written in Java such as POS tagging, NER, constituency parsing, dependency parsing etc., provided by the Stanford NLP group. It contains Stanford NER - v4.0.0, a high-performance machine learning named entity recognition system based on linear chain CRF sequence models [20]. Perdido[7] aims at extracting and retrieving displacements from textual documents. This method is based on a hybrid approach combining both POS tagging, cascade transducer application, as well as querying web resources for data and the visualization of geographic information. SEM[9] is a supervised machine learning system for named entity extraction using CRFs, we used its online version [10] through automatic queries with the selenium library in Python. Spacy[10] is a free open-source library for NLP Processing in Python providing convolutional neural network models for POS, text categorization and NER and dependency parsing.

2.2.2 Post-processing of CasEN's output

In light of the obtained results with the methods cited above, we also set up `GeoNER_repair`, a script that combines CasEN's output with supplementary Unitex's graphs that we created by mixing up rules and some CasEN's graphs, as

[5] available at `https://github.com/johnseazer/fren`

[6] available at `https://github.com/kogkitse/geoner`, under the GPL v3.0 license

[7] Unitex is an open source, cross-platform, multilingual, lexicon- and grammar-based corpus processing suite available at `https://unitexgramlab.org/fr`

[8] available at `https://github.com/stanfordnlp/CoreNLP`

[9] available at `https://github.com/YoannDupont/SEM`

[10] available at `https://spacy.io/`

Table 2: An example from the *Memoirs of Marguerite de Valois* in each tool's output before applying the post-processing rules.

tool		format
CasEN	Des domestiques de dom <persName><surname>Juan</surname> </persName>, n'y en avoit de nom ny d'apparence qu'un Ludovic de Gonzague, qui se disoit parent du duc de <placeName>**Mantoue**</placeName>. {S}	TEI
CoreNLP	Des/O domestiques/O de/O dom/O Juan/I-PER ,/O n/O '/O y/O en/O avoit/O de/O nom/O ny/O d'apparence/O qu'un/O Ludovic/I-PER de/I-PER Gonzague/I-PER ,/O qui/O se/O disoit/O parent/O du/O duc/O de/O Mantoue/I-LOC ./O	IO
Perdido	<w lemma="de" type="PREP" xml:id="w28">de</w> <rs type="place" subtype="no" start="127" end="134" startT="28" endT="29" id="en.11"> <name type="place" id="en.3"> <w lemma="null" type="NPr" xml:id="w29">**Mantoue**</w> <location><geo source="bdnyme">2.302553 48.759255</geo></location></name>	XML
SEM	Des domestiques de dom Juan, n'y en avoit de nom ny d'apparence qu'un Ludovic de Gonzague, qui se disoit parent du duc de **Mantoue**.	HTML
spaCy	dom Juan 20 28 PER Ludovic de Gonzague 72 91 PER Mantoue 125 132 LOC	IOB

well as, lexical resources such as DiTex-cities-fr[11]. At this step, choosing CasEN as tool it was anything but a random choice, as a matter of fact, it got higher precision scores, it is easy to access, add and modify an output but also it allows us to control every of its processing steps.

Figure 1 illustrates a sample of a graph which modify misplaced geographical entity tags preceded by civilities and function contexts, for example:
roi de <placeName>France</placeName> → <persName>roi de France</persName>

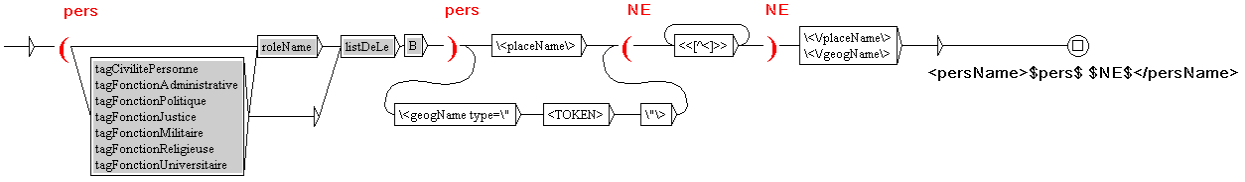


Figure 1: Unitex graph modifying placeName and geogName tags to persName used as a post-processing step after applying the CasEN method.

As this method is actually a post-processing of results obtained by CasEN, we denote it by "CasEN+R" below.

2.3 Normalisation of pre-processing of texts

After a few tests of the methods described above on the two theater plays of our corpus, we noticed that most of them were not able to handle verses properly, not only because of uppercase characters in the beginning of each verse but also the presence of special characters such as the long s, "ſ", an archaic form of letter "s" and the ampersand, "&", representing the coordinating conjunction "and". We therefore developed preprocessing scripts to remove these special characters and to replace uppercase characters in the first word of a verse following a comma or no punctuation sign, unless this word appears in a proper noun dictionary. More precisely, the script first reads the input file and substitutes the special characters in order to create a reference file for further step processing. In a second step, another script deals with words with initial uppercase characters where, with the aid of a CasEN's named entity dictionary

[11] available at https://github.com/martinec/DiTex-GeoTex/tree/master/ditex

(`Prolex-Unitex_1_2.dic`[12] and `Prolex-Unitex-BestOf_2_2_fra.dic`[13]), the transformation to lowercase is validated or not.

In order to automatically normalise the texts of our corpus, we use the ABA software created by Jonathan Poinhos during an internship that we supervised [21]. This Python script, available at `https://github.com/johnseazer/fren` is an alignment-based approach, which uses an aligned learning set of texts in the original and in the normalised version in order to build a transformation dictionary. It also implements a few transformation rules based on character transformations learned from an analysis of the PARALLEL17 corpus. Both transformations are combined with a dictionary of modern French words in order to decide whether the transformed words are kept (if they appear in the dictionary) or not. Tested on the same corpus as the neural network method described in [22], it seems to obtain better results.

3 Results

An evaluation script (`./eval.sh`) has been implemented so as to evaluate the performances of the six geoNER systems called by `GeoNER`, in terms of precision, recall and F-measure, in a strict manner, meaning that the boundaries of tagged named entities must be equal in the reference text and in the one obtained by one of the systems. The script compares a reference tagged file to an hypothesis output file, both of them strictly aligned at sentence level thanks to preprocessing scripts. It then computes the precision, recall and F-measure scores displayed in Table 3 and in the histograms of Figure 2.

Table 3: Precision, recall and F-measure obtained by the 6 geoNER tools on respectively the original corpus ("original"), the automatically modernised corpus ("auto") and the manually modernised corpus ("modern").

	precision			recall			F-measure		
tool	original	auto	modern	original	auto	modern	original	auto	modern
CoreNLP	0,232	0,271	0,267	**0,799**	**0,814**	0,792	0,360	0,407	0,400
spaCy	0,363	0,412	0,443	0,701	0,731	0,754	0,479	0,527	0,558
SEM	0,485	0,652	0,649	0,610	0,602	0,636	0,540	0,626	0,642
PERDIDO	0,590	0,629	0,644	0,523	0,534	0,610	0,554	0,578	0,626
CasEN	0,743	0,723	0,742	0,602	0,602	**0,830**	0,665	0,657	0,784
CasEN+R	**0,843**	**0,814**	**0,879**	0,610	0,614	0,826	**0,708**	**0,700**	**0,852**

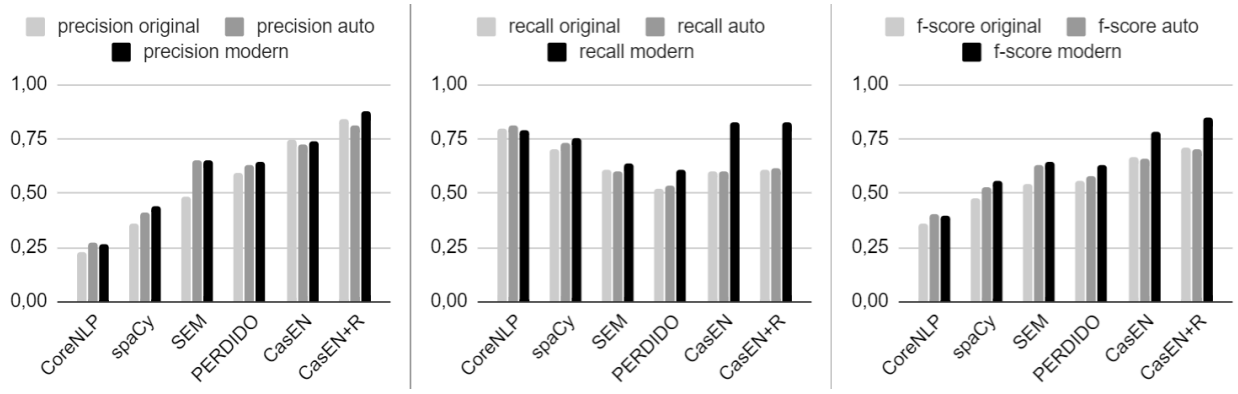


Figure 2: Precision, recall and F-measure obtained by the 6 geoNER tools on respectively the original corpus ("original"), the automatically modernised corpus ("auto") and the manually modernised corpus ("modern").

[12] available at `https://tln.lifat.univ-tours.fr/medias/fichier/prolex-unitex-1-2_1562935068094-zip?ID_FICHE=321994&INLINE=FALSE`

[13] available at `https://tln.lifat.univ-tours.fr/medias/fichier/casen-fr-1-4_1596032302677-zip?ID_FICHE=332027&INLINE=FALSE`

4 Discussion

4.1 About the obtained results

The results of the evaluation show, as we could have expected, that rule-based methods generally obtained better precision than statistical methods, but the latter obtain better results than the former, at least on the original corpus.

Overall, CasEN+R, our new method combining CasEN with a post-processing step, obtains the best F-measure results on the original corpus as well as on the automatically or manually modernised versions. It performs significantly better than other methods in terms of precision. Using specific gazetteers from the targeted period, including historical variations such as "Espaigne" (used instead of "Espagne" for "Spain") could potentially improve even more the performance of this method.

We note that most methods generally obtain better results on the automatically modernised version of the corpus, and even better on the manually modernised version of the corpus. This is however not the case of CasEN with the automatic modernisation. It will be interesting to investigate further the significant difference in terms of recall for this method after automatic or manual modernisation, which could also lead to improve the modernisation algorithm, or at least to identify its mistakes which have a big impact on the results of CasEN or other geoNER methods.

The diversity of the results obtained by statistical and rule-based methods, especially the opposite trends that we can observe on precision and recall, is a good motivation to try to combine them, using machine learning algorithms on the output results of the `GeoNER` tool. Another option to take advantage of the good recall scores of the statistical methods would be to allow `GeoNER` to display the results obtained by any of the 6 methods, highlighting the ones found by several methods, with a confidence score computed to take into account the number of the methods which found them.

4.2 Mapping the extracted locations

Another perspective of this work, in view of the applications of this tool for the *Cité des Dames* project, would be to add a mapping tool allowing the users to quickly remove some place name tags, or refine them, for example to identify the locations visited by the narrator of the analyzed texts. This would be useful in order to help building itinerary maps from memoirs, like in Figure 3 from the *Memoirs* by Marguerite de Valois[14] and Catherine de La Guette[15].

Although this mapping step can be helped by automatic techniques, especially in the most simple cases, it seems relevant to humans check the results at this step, for example with Recogito [23] or GeoViz [24]. To build the itineraries of Figure 3, we simply used the tool *Batch geocoder for journalists*, available at `https://geocode.localfocus.nl`, to quickly find the correct location of most of the cities cited in the two *Memoirs*, but we carefully checked and corrected several errors. However, as the density of geographical named entities is rather small, this task should be faster, with the appropriate tools, than checking the found named entities in the whole text. Furthermore, if the users are provided with the context surrounding each named entity during this step, they can take into account some precisions around the tagged proper noun (for example "la porte du château de Namur", that is "the gate of Namur's castle" instead of just "Namur"), in order to locate the place name more accurately.

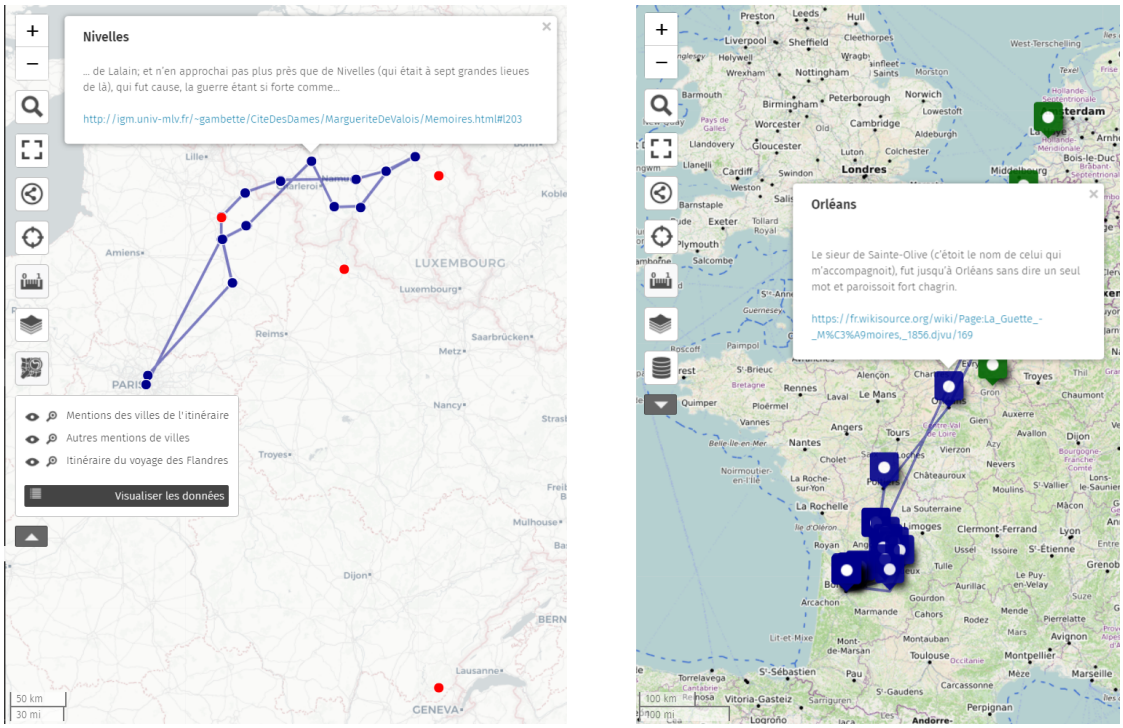
5 Conclusion

Our experiments suggest that it is useful to apply normalisation techniques on 16^{th} and 17^{th} century texts in French before applying geographical named entity recognition. Comparing this technique with alternatives, consisting in training statistical methods with specific corpora similar to the texts we wish to analyse, would be interesting in order to evaluate the benefits of the effort needed for specific model adaptation. Furthermore, it would be interesting to test whether text normalisation has a positive impact on active learning methods such as Prodigy[16]. Indeed, we note that text normalisation often decreases some inherent variability in the texts (for example in the spelling of the words, which may vary even inside a single text), which suggests it could improve the results even with such techniques. Our corpus of 39647 words and 264 proper noun geographical named entities is freely available on GitHub to help other teams develop and test such alternative techniques, and we will update it in the future with other texts written by women analysed in the context of the *Cité des Dames* project.

[14] interactive version available at `https://umap.openstreetmap.fr/fr/map/memoires-de-marguerite-de-valois-le-voyage-des-fla_421450`

[15] interactive version available at `https://umap.openstreetmap.fr/fr/map/memoires-de-mme-de-la-guette_350743`

[16] available at `https://prodi.gy/`



(a) (b)

Figure 3: Mapping (a) the *Memoirs* by Marguerite de Valois (the trip to Flanders of 1577, red dots show cities which are only mentioned in the text, not visited by Marguerite de Valois) and (b) by Catherine de La Guette (the trip to Bordeaux of 1653, during the Fronde, is shown with blue markers) with the uMap online tool.

Acknowledgements

This work has received support under the program "Investissement d'Avenir" launched by the French Government and implemented by ANR, with the reference ANR-16-IDEX-0003. We thank Isabelle Qin for her annotations of geographical named entities in the *Memoirs of Marguerite de Valois* and Caroline Trotot for her annotations and discussions about geographical named entities. We thank Éliane Viennot for allowing us to use and publish as an annotated corpus her edition of the *Memoirs of Marguerite de Valois* and Simon Gabay for telling us about the PARALLEL17 corpus. We thank Ludovic Moncla for extending the daily limitations for our use of the Perdido API and Jonathan Poinhos for running a preliminary version of his ABA normalisation software on our corpus.

References

[1] Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1343–1346, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

[2] Katherine McDonough, Ludovic Moncla, and Matje van de Camp. Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science*, 33(12):2498–2522, 2019.

[3] Eleni Kogkitsidou. *Communiquer par SMS: Analyse automatique du langage et extraction de l'information véhiculée*. PhD thesis, Grenoble Alpes, 2018.

[4] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

[5] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. Ensemble named entity recognition (NER): Evaluating NER tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5:2, 2018.

[6] Denis Maurel, Nathalie Friburger, and Iris Eshkol-Taravella. Enrichment of Renaissance texts with proper names. *INFOtheca : Journal of Information and Library Science*, 15(1):15–27, September 2014.

[7] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSpatial International Conference on Advances in Geographic Information Systems*, pages 183–192, 2014.

[8] Serge Heiden. The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398, 2010.

[9] Ludovic Moncla, Mauro Gaio, and Thierry Joliveau. Cartographier les odonymes de Paris cités dans les romans du XIXe siècle. In *Atelier Humanités Numériques Spatialisées (HumaNS'2018) - SAGEO 2018*, 2018.

[10] Yoann Dupont. Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. In *TALN 2017*, 2017.

[11] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. REDEN: named entity linking in digital literary editions using linked data sets. *Complex Systems Informatics and Modeling Quarterly*, (7):60–80, 2016.

[12] Aicha Soudani, Yosra Meherzi, Asma Bouhafs, Francesca Frontini, Carmen Brando, Yoann Dupont, and Frédérique Mélanie-Becquet. Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle. In *Atelier Humanités Numériques Spatialisées (HumaNS'2018)*, Montpellier, France, November 2018.

[13] Yvonne Cazal and Gabriella Parussa. *Introduction à l'histoire de l'orthographe*. 2015.

[14] Gilles Souvay and Jean-Marie Pierrel. LGeRM Lemmatisation des mots en Moyen Français. *Traitement Automatique des Langues*, 50(2):21, 2009.

[15] Simon Gabay. PARALLEL17: a parallel corpus (diplomatic vs normalised) of 17th c. French texts, 2019. `https://github.com/e-ditiones/PARALLEL17`.

[16] Simon Gabay and Giovanni Pietro Vitali. CARTO17: Data and scripts for mapping 17th c. French theatre, 2019. `https://github.com/e-ditiones/CARTO17`.

[17] Zied Sellami, Jean-Gabriel Ganascia, and Mohamed Amine Boukhaled. MEDITE : logiciel d'alignement de textes pour l'étude de la génétique textuelle. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Démonstrations*, pages 1–2, Caen, France, June 2015. ATALA.

[18] Nathalie Friburger and Denis Maurel. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104, 2004.

[19] Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol, and Damien Nouvel. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues*, 52(1):69–96, 2011.

[20] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[21] Jonathan Poinhos. Approches automatiques pour la modernisation des textes du XVe au XVIIIe siècle, 2020. Msc Thesis, Université Gustave Eiffel, Institut Gaspard-Monge (IGM), Marne-la-Vallée, France.

[22] Simon Gabay and Loïc Barrault. Traduction automatique pour la normalisation du français du XVIIe siècle. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 213–222, 2020.

[23] Elton Barker, Rainer Simon, Valeria Vitale, Rebecca Kahn, and Leif Isaksen. Revisiting linking early geospatial documents with Recogito. *e-Perimetron*, 14(3):150–163, 2019.

[24] Katherine McDonough and Matje van de Camp. Mapping the Encyclopédie: Working towards an early modern digital gazetteer. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'17, page 16–22, New York, NY, USA, 2017. Association for Computing Machinery.