Expérimentation de méthodes d'extraction d'informations géographiques pour les documents historiques

Katherine McDonough[1], **Ludovic Moncla**[2], **Matje van de Camp**[3]

1. *Department of History and Center for Interdisciplinary Digital Research, Stanford University, USA*
   kmcdono2@stanford.edu
2. *INSA Lyon, CNRS, LIRIS UMR 5205, France*
   ludovic.moncla@liris.cnrs.fr
3. *De Taalmonsters, Tilburg, Netherlands*
   matje@taalmonsters.nl

RÉSUMÉ. *Dans cet article, nous nous intéressons à deux aspects peu étudiés dans les travaux de recherche en TAL : traiter des documents historiques en français et traiter des structures textuelles complexes au-delà du texte courant ou des listes de noms de lieux. Notre méthodologie s'appuie sur l'évaluation des résultats de deux outils de reconnaissance d'entités nommées spatiales dans le cadre de l'analyse de documents du début de l'époque moderne et structurés à la manière de dictionnaires.*

ABSTRACT. *In this article, we address two gaps in NLP research: working with historical French and working with complex textual structures moving beyond running text or lists of place names. Our methodology is based on the evaluation of the results of two spatial named entity recognition tools in the context of early modern document analysis structured as dictionaries.*

MOTS-CLÉS : *recherche d'informations géographiques, traitement automatique des langues (TAL), reconnaissance d'entités nommées, humanités numériques*

KEYWORDS: *geographic information retrieval, natural language processing (NLP), named entity recognition, digital humanities*

1. Introduction

Les recherches sur l'analyse de dictionnaires géographiques se sont jusqu'ici concentrées sur des projets utilisant des corpus en anglais moderne, publiés principalement à partir de la fin du XVIIIe siècle. Ces projets dépendent de lexiques et de ressources géographiques spécifiques à l'époque considérée qui permettent d'améliorer l'identification des noms de lieux et leur localisation. Contrairement aux érudits du monde classique et du monde moderne, les chercheurs contemporains travaillant sur le début de la période moderne (1400-1800) manquent de telles ressources. Ce projet est à l'interface entre les sciences humaines (histoire, géographie), le traitement automatique des langues (TAL) et les sciences de l'information géographique. Nous nous appuyons sur des travaux récents sur les espaces et les lieux du siècle des Lumières (Safier, 2014 ; Withers, Mayhew, 2011) ainsi que sur divers projets en humanités numériques portant sur la période des Lumières (Edelstein, 2016 ; Comsa *et al.*, 2016). Contrairement aux études sur cette période qui sont souvent axées sur l'analyse de petits groupes d'élites (dirigeants politiques, scientifiques, philosophes, voyageurs), cette recherche est une étape vers la compréhension de la mobilité de communautés locales et cosmopolites à travers l'information géographique.

Dans cet article, nous nous intéressons à deux aspects peu étudiés dans les travaux de recherche en TAL : 1) traiter des documents historiques en français et 2) traiter des structures textuelles complexes au-delà du texte courant ou des listes de noms de lieux. Le texte numérisé de l'Encyclopédie [1] de Diderot et d'Alembert (Morrissey, Roe, 2017) édité entre 1751 et 1772 est un exemple de corpus historique du genre dictionnaire que l'on se propose d'étudier.

2. Comparaison d'outils d'annotation pour l'adaptation à des corpus historiques français

L'adaptation d'un outil de reconnaissance d'entités nommées spatiales pour le français nous permettra d'explorer la structure et le contenu de l'information géographique au sein de documents du début de l'époque moderne de manière automatique. Cela nous permettra également de développer des techniques répondant aux challenges de la recherche d'informations spatiales communs à tous les états anciens des langues : identification des variantes de noms de lieux, association de variantes de noms à un même lieu et désambiguïsation de différents lieux, et détermination des types de relations entre ressources géographiques (officielles ou participatives) et descriptions textuelles de lieux historiques.

Notre méthodologie s'appuie sur l'évaluation des résultats de deux outils de reconnaissance d'entités nommées spatiales dans le cadre de l'analyse de

1. *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres. http://encyclopedie.uchicago.edu*

documents du début de l'époque moderne et structurés à la manière de dictionnaires. La reconnaissance d'entités nommées spatiales combine les tâches de reconnaissance d'entités nommées et d'association de ces entités à une localisation (coordonnées géographiques). Nous avons testé l'outil Edinburgh Geoparser (EG) (Alex *et al.*, 2015) et l'outil Perdido (Gaio, Moncla, 2017). Le premier a été développé pour analyser des documents en langue anglaise et le second pour des documents rédigés en langue française, tout les deux pour les langues modernes. Nous avons sélectionné EG car il s'agit d'un outil reconnu et très utilisé dans les projets d'annotation d'informations géographiques. Nous étudions ici ces deux outils en parallèle non pas pour critiquer leur performances, mais pour montrer aux spécialistes des sciences humaines 1) à quel point les outils d'annotation automatique peuvent être différents selon leur conception et 2) pourquoi adapter un tel outil au contexte issu du texte est une préoccupation méthodologique fondamentale pour l'analyse géographique de textes.

EG comme Perdido sont conçus comme une chaîne de traitement comprenant une étape de pré-traitement (tokenization, lemmatisation et analyse morpho-syntaxique) et un système de règles (patrons linguistiques, lexiques. ...) pour l'annotation des entités nommées. Comme (Won *et al.*, 2018) nous avons obtenu la version d'EG préparé pour le projet *Reassembling the Republic of Letters* mais avec un analyseur grammatical du français. EG prend donc en compte les catégories grammaticales du français, mais implémente la reconnaissance des entités nommées à partir de règles développées pour l'anglais. C'est un problème reconnu par les chercheurs travaillant avec des langues autres que l'anglais (voire même pour l'anglais d'avant le vingtième siècle). (Il est, bien sûr, possible de modifier les règles et les lexiques de EG.) Perdido, au contraire, est conçu spécifiquement pour la langue française. Par ailleurs, les noms de lieux construits autour d'un ou plusieurs mots peuvent être intégrés dans d'autres types d'entités (personnes, fonctions, ...). Dans le cas d'EG, le balisage interne du nom de lieu n'est pas conservé une fois que l'entité est imbriquée et l'information sur la nature spatiale du nom est perdue. À la différence, Perdido a été spécialement développé pour annoter les entités nommées étendues (ENE) telles que définies par (Gaio, Moncla, 2017) et conserver l'information de chaque niveau d'imbrication.

3. Expérimentations

Le corpus étudié comprend les 14445 articles de l'Encyclopédie appartenant à la catégorie Géographie (Morrissey, Roe, 2017). Ceux-ci sont fournis au format TEI, un premier pré-traitement supprime l'en-tête et la structuration afin de faciliter le travail avec les outils d'annotation. Nous n'avons ni modifié ni modernisé le langage étant donné les recherches antérieures sur les textes anglais de l'époque moderne pour lesquels la modernisation est jugé superflus (Won *et al.*, 2018). Nous avons constitué un corpus d'évaluation composé de 100 articles sélectionnés aléatoirement parmi l'ensemble du corpus. Ces articles

ont été annoté manuellement par K. McDonough en utilisant l'outil GeoViz [2] (McDonough, Camp, 2017). Le corpus d'évaluation comprend environ 30000 mots pour 2151 occurrences de noms de lieux.

Nos expérimentations ont été conçues principalement pour identifier et évaluer les faiblesses des deux méthodes testées. Nous avons intentionnellement utilisé des méthodes qui n'ont pas été conçues spécifiquement pour ce contexte (une pratique de plus en plus courante à mesure que les chercheurs en sciences humaines cherchent à utiliser des méthodes d'extraction d'informations géographiques). Nous avons ainsi acquis des informations précieuses pour guider les futures adaptations de Perdido en réfléchissant à ce qui a fonctionné, à ce qui n'a pas fonctionné et pourquoi. Les résultats de notre évaluation mesurent le rappel, la précision et le F1-score pour EG et Perdido par rapport à l'annotation manuelle (tab. 1). Le rappel mesure le nombre d'entités identifiées par l'outil comme étant un lieu par rapport à tous les noms de lieux existants. La précision est le nombre d'entités correctement identifiées sur le nombre total d'entités de lieu trouvées par l'outil. Le score F1 est la moyenne pondérée de ces deux mesures.

TABLE *1. Evaluation de la reconnaissance des entités nommées spatiales*

	Recall	Precision	F1-score
EG	9.20%	94.64%	16.78%
PERDIDO	55.58%	75.71%	64.10%

Comme attendu le rappel de EG est faible compte tenu de l'absence d'adaptations spécifiques à la langue et au format (9,2 %). Mais Perdido n'obtient pas un rappel aussi haut qu'espéré (seulement 55,58%) par rapport à d'autres tests réalisés avec des romans français du XIX^{e} siècle (rappel à 99,7% pour les noms de rues parisiens) (Moncla *et al.*, 2017). La précision des résultats obtenue par EG est élevée, 94,64% contre seulement 75,71% avec Perdido et s'explique par le grand nombre d'entités non identifiées. En guise de comparaison, l'utilisation de EG sur un corpus de correspondances anglaises (lettres Hartlib) a permis d'obtenir une précision de 53,8%, un taux de rappel de 52,4% et un F1-score de 53,1% (Won *et al.*, 2018). Comme pour les lettres Hartlib, les noms de lieux peuvent apparaître dans l'Encyclopédie en plusieurs langues (grec, espagnol, anglais, allemand et autres), ce qui perturbe le processus de reconnaissance. Perdido produit un nombre important d'erreurs de catégorisation des entités nommées, et a tendance à privilégier la catégorie spatiale. Ce problème se pose en particulier lorsqu'il y a très peu de contexte pouvant servir à la désambiguïsation et que le nom identifié existe dans les ressources géographiques (ici, il est questions de l'Alexandria Digital Library). En revanche comme l'illustre l'exemple (1), lorsque des éléments du contexte, et en particulier ceux compo-

2. http://geoviz.taalmonsters.nl

sants l'ENE, sont présents alors la catégorisation des noms de personnes par exemple pose moins de problème.

(1) a. *le roi de* *Chypre*

b. *le chevalier de* *Jaucourt*

Les lexiques, les priorités et les règles d'interrogation des ressources géographiques nécessitent d'être améliorés afin d'adapter Perdido à l'analyse de documents historiques ne comprenant pas exclusivement des informations géographiques. Un problème anticipé et confirmé par nos expérimentations est celui de la couverture spatiale mais également temporelle des ressources géographiques interrogées. En revanche, les expérimentations ont également permis de mettre en évidence l'importance dans le classement des résultats retournés par les ressources géographiques qui peut être basé sur des heuristiques qui ne sont pas adaptées au corpus étudié.

4. Identification des entités complexes et imbriquées

Comme ont pu le montrer nos expérimentations, le contexte associé aux entités nommées et en particulier les informations impliquées dans la construction des ENE est primordiale pour la catégorisation et la désambiguïsation des entités nommées. Par ailleurs, les ENE peuvent nous permettre de désambiguïser et d'associer un nom de lieu à un lieu spécifique ou de tenir compte de l'ambiguïté grâce à certaines expressions locatives. Elles capturent la portée de l'information géographique telle qu'elle a été façonnée dans une période où les localisations précises étaient difficiles à mesurer et pas nécessairement utiles.

Nos efforts pour identifier les ENE au sein des articles de l'Encyclopédie reflètent l'importance de la manière dont les lieux peuvent être intégrés dans les relations sociales et spatiales. Cela nous permettra de compléter, voire de remplacer, la recherche de coordonnées géographiques avec d'autres types d'informations qui exploitent les relations contextuelles. Parmi les 1721 entités annotées par Perdido, 396 (23%) sont imbriquées au sein d'une ENE. De plus, 51 ENE font référence à un nom de personne parmi lesquelles 33 étendent un nom de lieu (65%). Les noms de lieux imbriqués dans les noms de personne sont généralement perdus, même lorsque ces lieux peuvent constituer des références significatives au contexte spatial du texte.

5. Conclusion et perspectives

La méthodologie proposée est généralisable : les articles de l'Encyclopédie ne sont pas les seules à contenir des associations culturelles, sociales ou géogra-

phiques entre des lieux et d'autres entités. Notre principale contribution à la conception des outils de reconnaissance des entités nommées spatiales consiste à améliorer l'identification de ces entités complexes. Alors que la plupart des outils n'identifieront pas un nom de lieu en tant qu'entité de lieu s'il est imbriqué dans un autre type d'entité, nous souhaitons de notre côté capturer et conserver ces informations. Cela comprend: a) les noms de lieux associés à des relations spatiales, b) des noms de lieux imbriqués au sein d'une autre entité (par exemple une personne ou une institution), c) les entités ambigues, et d) des entités qui ne peuvent pas être géolocalisées (ex : mythique ou extraterrestre). La conception initiale de Perdido adaptée pour l'annotation des ENE nous semble un point important et les différentes faiblesses identifiées de la méthode nous permettrons de faire évoluer l'outil afin d'obtenir de meilleurs résultats pour l'annotation de corpus historiques. Nos expérimentations ont montré les limites de l'utilisation d'une méthode unique basée sur des règles pour l'extraction d'informations. Les principaux problèmes sont dus aux spécificités de la langue utilisée dans les textes historiques en français classique. En effet, Perdido et EG utilisent tous deux une analyse morphosyntaxique basée sur des modèles non adaptés à ce langage. Ainsi, une amélioration importante consiste à construire de nouveaux modèles à l'aide de l'Encyclopédie et d'autres textes de référence de style classique, que ce soit pour l'analyse grammaticale ou pour la reconnaissance des entités nommées. La combinaison d'approches symboliques et statistiques (apprentissage automatique) semble une perspective intéressante pour une meilleure adaptabilité de l'outil au corpus ainsi que pour l'amélioration des résultats.

Bibliographie

Alex B., Byrne K., Grover C., Tobin R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing*, vol. 9, n° 1, p. 15–35.

Comsa M. T., Conroy M., Edelstein D., Edmondson C. S., Willan C. (2016). The french enlightenment network. , vol. 88, n° 3, p. 495–534.

Edelstein D. (2016). Intellectual history and digital humanities. *Modern Intellectual History*, vol. 13, n° 1, p. 237–246.

Gaio M., 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.

McDonough K., Camp M. van de. (2017). Mapping the Encyclopedie: working towards an early modern digital gazetteer. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, p. 16–22. ACM.

Moncla L., Gaio M., Joliveau T., Le Lay Y.-F. (2017). Automated Geoparsing of Paris Street Names in 19th Century Novels. In *1st ACM SIGSPATIAL Workshop on Geospatial Humanities.* Redondo Beach, CA, United States.

Morrissey R., Roe G. (2017). Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc., eds. Denis Diderot and Jean le Rond d'Alembert. University of Chicago: ARTFL Encyclopédie Project (Autumn 2017 Edition).

Safier N. (2014, février). The Tenacious Travels of the Torrid Zone and the Global Dimensions of Geographical Knowledge in the Eighteenth Century. *Journal of Early Modern History*, vol. 18, n° 1-2, p. 141–172.

Withers C. W. J., Mayhew R. J. (2011, décembre). Geography: Space, Place and Intellectual History in the Eighteenth Century. *Journal for Eighteenth-Century Studies*, vol. 34, n° 4, p. 445–452.

Won M., Murrieta-Flores P., Martins B. (2018). Ensemble named entity recognition (ner): Evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*.