Classification des entités nommées dans l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres (1751-1772)

Denis Vigier, Ludovic Moncla, Alice Brenon, Katherine Mcdonough, Thierry Joliveau

Classification des entités nommées dans *l'Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres* (1751-1772)

Denis Vigier[1,*], *Ludovic* Moncla[2], *Alice* Brenon[1], *Katherine* McDonough,[3] et *Thierry* Joliveau[4]

[1]Univeristé Lyon 2, CNRS, ICAR UMR 5191, Lyon, France
[2]INSA Lyon, CNRS, LIRIS UMR 5205, Lyon, France
[3]The Alan Turing Institute and Queen Mary, University of London, UK
[4]Université de Saint-Etienne, CNRS, EVS UMR 5600, Saint-Etienne, France

Résumé. Nous présentons la méthode que nous avons suivie pour améliorer notre annotation automatique des entités nommées dans l'*Encyclopédie* de Diderot et d'Alembert. L'outil d'annotation sémantique PERDIDO que nous utilisons a été initialement développé pour l'annotation d'informations géographiques et la reconstruction d'itinéraire. Nous proposons d'y implémenter de nouvelles règles élaborées manuellement à partir d'une étude des cotextes co-occurrentiels des noms propres du corpus accomplie au moyen d'une plateforme automatique d'exploration et de calcul.

Abstract. Named entity classification in the Encyclopédie ou dictionnaire raisonné des sciences et des métiers par une société de gens de lettres (1751-1772). We present a method for improving rule-based named entity recognition for Diderot and d'Alembert's *Encyclopédie*. We use PERDIDO, a semantic annotation tool initially developed for the annotation of geographic information and the reconstruction of itineraries. We improve PERDIDO's recognition of named entities by implementing new rules developed from an a exploratory and statistical analysis of word co-occurrences for proper nouns in the Encyclopédie corpus.

1 Introduction

Nous présentons ici un premier bilan de travaux conduits en interdisciplinarité (linguistique, histoire, géographie et informatique) qui s'inscrivent dans le cadre d'un projet[1] sur l'analyse du discours géographique dans les encyclopédies françaises entre le dix-huitième siècle et nos jours. L'objectif général vise à répondre au mieux à cette question : « Quelles représentations géographiques du monde les encyclopédies françaises véhiculent-elles à travers leurs discours, et que nous disent ces représentations sur chacune des époques où ces encyclopédies ont été écrites et publiées ? ». Une encyclopédie en tant

* Corresponding author : denis.vigier@ens-lyon.fr

qu'inventaire-synthèse du savoir à une époque donnée peut être en effet considérée comme une sorte de microcosme dans lequel se condensent et s'articulent les connaissances de son temps.

Dans cet article, nous décrivons en particulier notre méthodologie pour l'amélioration de la reconnaissance et de la classification des entités nommées dans l'*Encyclopédie* de Diderot et d'Alembert (désormais, EDDA) dont les vingt-huit volumes ont été édités entre 1751 et 1772. Par entité nommée (EN) on entend des unités informationnelles telles que les noms de personnes, de lieux, d'organisations ou encore certaines expressions numériques (Nadeau et Sekine, 2007). Il existe différentes typologies pour la classification des EN provenant de conférences ou de campagnes d'évaluation (Nobata et al., 2002 ; Gravier, Bonastre et Geoffrois, 2004 ; Sundheim, 2005). Dans cette première partie de notre travail nous nous concentrons sur deux catégories : les noms de personnes (individus, groupes, organisations humaines) et de lieux (villes, pays/royaumes, régions et autres). Les systèmes de reconnaissance et de classification des EN sont répartis en trois familles (Sekine et Eriguchi, 2000 ; Poibeau, 2005) : symbolique, statistique et hybride. Les approches symboliques mettent en œuvre des règles mises au point par des experts et mobilisent des ressources linguistiques (dictionnaires, lexiques, données d'ordre typographique, collocationnelles, …). Les approches statistiques utilisent des méthodes d'apprentissage automatique qui nécessitent le plus souvent des corpus annotés. Elles construisent des modèles (arbres de décision, modèles probabilistes, …) de manière supervisée mais dont les résultats obtenus, les règles apprises, ou les décisions prises, peuvent être complexes à comprendre. Enfin, les systèmes hybrides combinent méthodes symboliques et statistiques le plus souvent adaptées pour des méthodes semi-automatiques avec des interactions entre un opérateur et les algorithmes. Dans cet article, le système présenté repose sur une approche symbolique. Notre objectif est d'enrichir une approche à base de règles développées suite à une exploration et à une analyse du corpus. La constitution de ressources linguistiques en vue d'une tâche de reconnaissance d'EN mettant en jeu une approche symbolique pose pour les textes anciens des difficultés spécifiques. On peut citer pour illustration les tâches de segmentation et d'annotation de corpus - les modèles de langue disponibles étant moins performants que pour les états contemporains de la langue-, la disponibilité souvent problématique de dictionnaires ou de *gazetteers*, l'absence de listes préconstruites de termes tels que les mots classifiants (titres, fonctions, statuts, découpages administratifs du territoire au dix-huitième siècle, …) qui sont aujourd'hui inusités ou rares (« dom », « prévôt », « palatinat », …) et qui constituent des indices externes (McDonald, 1996) permettant d'améliorer la reconnaissance et la classification des EN.

Le corpus sur lequel nous travaillons réunit les dix-sept volumes de texte de l'œuvre. Les onze volumes de planches qui en forment l'autre partie sont pour le moment exclus du champ de notre enquête, même s'ils ne sont pas exempts de texte (contenu dans les légendes et les commentaires associés à chaque planche). Dans les sections qui suivent, nous présenterons d'abord les étapes du traitement appliqué à notre corpus; nous exposerons ensuite les indices linguistiques que nous avons élaborés à partir d'une exploration du texte de l'*Encyclopédie* au moyen d'une plateforme automatique, en proposant de distinguer des indices forts et des indices faibles.

2 Le corpus de l'*Encyclopédie* et son traitement actuel

Nous sommes partis de l'édition numérique d'un premier tirage de l'édition originale de Paris de l'EDDA telle qu'elle nous a été communiquée par l'*American and French Research on the Treasury of the French Language* (ARTFL)[2] de l'Université de Chicago. Les dix-sept volumes de texte comptent 24 617 707 mots et 74 198 articles. A l'intérieur de ce corpus, nous avons constitué un sous-corpus dit « de géographie » formé de 14 327 articles (soit 961 844 mots) relevant du domaine « géographie » (Géographie, Géographie moderne, Géographie ancienne, etc.) à partir duquel nous avons extrait des motifs (voir §3.1.2).

Ce sous-corpus a fait l'objet de traitements informatiques avec la chaîne d'annotation PERDIDO (Moncla et Gaio, 2018 ; Gaio et Moncla, 2019). Il s'agit d'un outil d'annotation sémantique initialement développé pour l'annotation d'informations géographiques et la reconstruction d'itinéraires (Moncla et al., 2016). Le traitement réalisé se décompose en plusieurs tâches. La première effectue la lemmatisation et l'étiquetage morphosyntaxique

avec le logiciel Treetagger. La deuxième tâche consiste à repérer les EN, les entités nommées étendues (ENE), les relations spatiales et les expressions de déplacement. Les EN sont catégorisées en trois principales classes : les noms de lieux, les noms de personnes et les autres. Les EN sont définies comme une expression linguistique faisant référence à une entité concrète et unique construite à partir d'un nom propre (Chinchor et Marsh, 1998). Les ENE ont été définies comme une extension des EN (Gaio et Moncla, 2019). L'objectif étant de capturer au sein d'une même structure l'ensemble des informations permettant de décrire une entité. Cette structure se construit comme une imbrication hiérarchique. Dans le cas de l'EDDA, nous distinguons deux catégories d'ENE. La première concerne celles dont l'extension précise la nature de l'entité noyau (voir l'exemple (1)) et la deuxième celles qui décrivent la nature d'une autre entité (voir l'exemple (2)). Chaque nouveau niveau ajoute une extension (groupe nominal) au nom propre qui permet de préciser la nature de l'entité décrite. Cette extension (ou succession d'extensions) permet par exemple de simplifier le traitement lors de l'étape de classification des EN.

(1) a. ville de Paris

b. rivière de Zacarat

(2) a. ville de France

b. peuple d'Asie

PERDIDO utilise une méthode symbolique implémentée par des cascades de transducteurs dans la plateforme Unitex (Maurel et al., 2011) selon les principes des grammaires de construction (Yannick Mathieu, 2003) et produit des annotations au format XML-TEI contenant les informations sémantiques liées aux EN et ENE. La version standard de PERDIDO et les outils d'annotation automatique de manière générale ne sont pas adaptés pour le traitement de documents anciens ou historiques (McDonough, Moncla et Van de Camp, 2019), à cause principalement des ressources utilisées telles que les modèles de langues ou les jeux d'étiquettes (utilisés par exemple par Treetagger) ou les ressources géographiques utilisées pour la localisation et la désambiguïsation des noms de lieux (Moncla et al, 2019).

La cascade d'analyse se compose de 12 transducteurs principaux qui implémentent les grammaires permettant de repérer et d'annoter sémantiquement les EN et ENE mais également un certain nombre d'informations spatiales (coordonnées géographiques, relations topologiques, orientations, distances, etc.). Un transducteur est un graphe agissant sur le texte par insertion, suppression ou remplacement et produit donc une sortie (annotation). Chaque transducteur peut utiliser les annotations ajoutées par les transducteurs précédents. Les grammaires implémentées formalisent des règles syntaxiques et permettent de reconnaître des motifs linguistiques en accédant à des dictionnaires. Par exemple dans le cadre de ce projet, plusieurs dictionnaires ou lexiques ont été ajoutés à PERDIDO pour la reconnaissance et la classification des EN et ENE tels qu'un dictionnaire de prénoms (auteurs classiques, auteurs médiévaux, auteurs modernes, souverains, etc.), un lexique de termes faisant référence à une personne (« roi », « pape », « chevalier », etc.) et le lexique de termes géographiques a également été enrichi. Une évaluation des résultats est proposée dans la section 3.1.4.

3 Élaboration d'indices linguistiques pour l'amélioration de l'annotation des entités

Un des objectifs de ce travail est l'amélioration de l'annotation des EN dans l'*Encyclopédie*. En effet, l'annotation réalisée par la chaîne de traitement PERDIDO n'est pas adaptée à cette œuvre, les règles d'annotation implémentées au sein de cette plateforme ayant été élaborées dans le cadre d'un projet différent et pour un corpus particulier de description de randonnées (Moncla et al, 2014). Ce qui ressort des différentes méthodes existantes dans la littérature (Nadeau et Sekine, 2007 ; Provani et al, 2009 ; Maurel et al, 2011 ; Gaio et Moncla, 2019), c'est qu'elles sont systématiquement adaptées à un type de corpus particulier selon la structure, la longueur et la nature du document (news, tweets, ...). Notre objectif dans ce projet est donc de proposer une amélioration et une version d'un

système d'annotation d'EN adapté pour l'*Encyclopédie* faisant suite à une précédente étude (McDonough, Moncla et Van de Camp, 2019).

La tâche d'annotation des EN se décompose en deux principales sous-tâches : la détection et la classification. Dans ce travail, nous nous intéressons en particulier à l'amélioration de la classification des EN pour deux catégories : noms de lieux et noms de personnes. Notre proposition consiste à développer un système de classification d'EN reposant sur une pluralité d'indices. Pour cela nous proposons d'une part, d'enrichir les grammaires existantes implémentées au sein de PERDIDO, et d'autre part d'ajouter une étape de post-traitement pour les cas les plus difficiles. Cette méthodologie s'accompagne de l'identification d'indices permettant d'aider à interpréter le contexte associé aux EN et le plus souvent contenu dans les ENE pour l'étape de classification ou de désambiguïsation.

La figure 1 présente l'ensemble des traitements réalisés dans la chaîne de traitement pour l'annotation et la classification des EN de l'EDDA. Notre méthodologie se décompose en deux phases, la première est manuelle et consiste d'une part (1) à analyser le corpus et à repérer des indices linguistiques à l'aide de l'outil TXM[3] (Heiden, 2010), d'autre part (2) à implémenter des règles permettant d'annoter ces indices de manière automatique dans la plateforme PERDIDO. La seconde étape est automatique et consiste à utiliser les règles d'annotation enrichies lors de la première phase pour l'annotation. Le résultat produit se compose des fichiers annotés au format XML-TEI et d'un concordancier au format CSV.

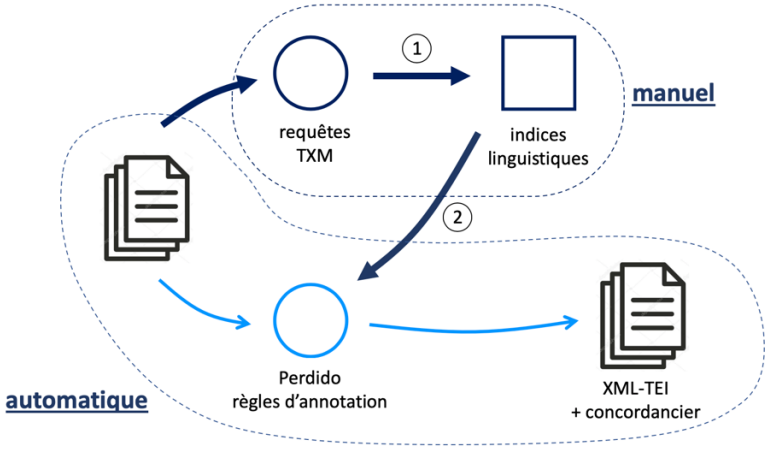

Fig. 1. Schéma du processus d'amélioration de l'annotation des EN dans EDDA.

Dans la suite de ce travail, nous distinguerons deux types d'indices : les indices forts pour lesquels une décision peut être prise avec une probabilité proche de 100% et les indices faibles qui nécessitent la présence de plusieurs indices pour la prise de décision (voir section § 3.2).

Dans la section 3.1, nous allons exposer la méthode que nous avons suivie pour explorer manuellement au moyen de la plateforme TXM un sous-corpus d'EDDA, dans l'objectif d'y détecter des motifs[4] (Longrée, 2013). Ces motifs constituent des indices linguistiques forts à même de nous aider dans l'annotation des EN. On verra que leur structure et leur lieu d'apparition (immédiatement après le mot vedette) sont étroitement liés au genre discursif que constitue l'*Encyclopédie*. Autrement dit, les contraintes qu'exerce un genre de discours sur la linéarisation des informations dans la phrase et dans le texte peuvent apparaître comme des auxiliaires précieux pour la détection et la classification des EN.

3.1 Mise au jour d'indices linguistiques forts

3.1.1 Détection d'une liste de mots classifieurs des EN dans l'EDDA

De manière générale, les articles de l'EDDA suivent un modèle de composition linéaire relativement stable et proche à certains égards d'un article de dictionnaire. On trouve d'abord un ou plusieurs mots vedettes graphiés en lettres capitales, suivis le plus souvent (mais non systématiquement) d'une mention généralement en italiques, placée entre parenthèses, et qui désigne le domaine de savoir dont relève ce mot. Les formulations pour un même domaine peuvent être extrêmement variées. Suit ensuite l'article proprement dit, signé dans un peu plus d'un cas sur deux. Voici un exemple :

(3) ZYGIANA, (*Géog. anc.*) contrée de l'Asie mineure, dans la Bithynie, selon Ptolomée, l. V. c. I. (D. J.) (XVII, 750)

Afin d'identifier des indices linguistiques suffisamment forts pour nous permettre d'enrichir les annotations déjà existantes produites par PERDIDO, nous avons restreint notre corpus de travail au seul sous-corpus « Géographie ». Sur les 74 198 articles que comptent l'EDDA, ce domaine apparaît le plus représenté avec plus de 14 000 articles dont 8300 articles signés du seul Chevalier de Jaucourt[6]. L'ensemble des mots vedettes qui y figure désigne presque toujours des lieux (réels ou mythologiques), parfois[8] des collectivités de personnes (peuples, tribus, …). On peut faire l'hypothèse que la première phrase (il s'agit parfois de la seule phrase de l'article) qui suit immédiatement le ou les mot vedettes figurant en entrée des articles contient des informations clefs pour la classification de l'EN en nom de lieux ou nom de personnes. Pour vérifier cette hypothèse, nous avons exploré au moyen du calcul cooccurrentiel dans TXM le voisinage aval des mots vedettes des articles de notre corpus de travail. La requête CQL[7] (Corpus Query Language) destinée à extraire les vedettes des articles du sous-corpus « Géographie » nous a permis d'extraire 12 110 articles sur les 14 327. Les faux-négatifs s'expliquent le plus souvent soit par la présence d'un très grand nombre de mots entre le premier nom de lieu figurant en vedette de l'article et la parenthèse ouvrante (4), soit par une formulation du domaine géographique de l'article différente de celle couverte par notre requête ((5) et (6)).

(4) ARCHIPEL ou ARCHIPELAGE, quoique cette derniere dénomination ne soit que peu en usage, subst. m. (*Géogr.*) *terme de Géographie*, qui (…) (I, 615)
(5) AMPHISCIENS, s. m. pl. *terme de Géographie & d'Astronomie*, se dit des peuples qui habitent la Zone torride (…) (I, 377)
(6) ANTECIENS, *Antoeci*, adj. pl. m. du Grec ἀντὶ *contre*, & d'οἰκέω, *j'habite*. On appelle en Géographie *Antéciens*, les peuples placés sous le même méridien & à la même distance de l'équateur (…) (I, 477)

Quant aux faux-positifs que nous avons observés, ils s'expliquent par un triple problème : i) la présence de chiffres romains non reconnus (car formés avec des lettres), ii) le fait que certains empans de textes puissent ne pas être traités par notre chaîne du fait d'un balisage erroné, iii) le fait enfin qu'avec notre étiquetage actuel nous ne puissions pas manipuler la propriété de structure « article » dans TXM, de sorte que la requête opérée peut franchir les bornes de l'article. Par exemple, dans le pivot faussement positif présenté en (7), i) le mot « XII » a été extrait comme une séquence de lettres, ii) le segment de texte qui suit les deux points dans l'article a été ignoré lors de la segmentation (voir sur le site ARTFL la fin de l'article PARNES, volume 12, p.73), iii) la requête a extrait comme dernier mot du pivot la vedette de l'article suivant (PARNI).

(7) XII. Vers. 620. Dit : PARNI, (*Géog. anc.*)

Afin d'évaluer la fiabilité des occurrences de nos extractions, nous avons tiré un échantillon aléatoire de 100 occurrences sur les 12 110 extraites : 99 étaient bonnes, la centième est celle présentée dans (7).

Pour les seuls noms de lieux, nous avons observé que pour les 12 110 extractions obtenues par notre requête CQL (soit 85% de tous les mots vedettes du sous-corpus Géographie), dans 79% des cas (soit 9 520 articles) apparaît parmi les trois mots qui suivent

la vedette un mot classifieur qui permet de classer l'EN. Par exemple (nous avons graissé dans (8)-(10) la séquence de mots extraite par la requête CQL stipulée dans la note 7 et souligné le mot classifieur) :

(8) **HAINGEN, (*Géogr.*)** petite ville d'Allemagne (…) (VIII, 26)
(9) **ZERGUE, (*Géog. mod.*)** petite riviere de France (…) (XVII, 706)
(10) **VINDERIUS, (Géogr. anc.)** fleuve de l'Hibernie. (…) (XVII, 307)

Voici, pour les vedettes désignant un lieu, la liste des lemmes à rôle classifieur identifiés et dont le nombre d'occurrences est supérieur à 10 (et précisé entre parenthèse) :

> « *ville (5696), rivière (753), île (521), province (346), fleuve (206), pays (202), bourg (194), royaume (194), montagne (190), lieu (122), comté (90), port (85), village (66), lac (65), contrée (60), bourgade (55), promontoire (55), golfe (51), forteresse (47), duché (38), château (34), canton (29), cap (28), place (24), capitale (22), palatinat (20), vallée (18), forêt (14), principauté (14), fontaine (11), chaîne (10), maison (10)* »

Certains de ces mots seraient parfaitement attendus dans une encyclopédie contemporaine, *modulo* leur graphie (par ex. pour le lemme « île », on trouve au singulier les formes « ile », « île », « isle ») qui pose le problème de la couverture des occurrences avec une chaîne d'annotation entraînée uniquement sur des textes contemporains. D'autres appartiennent à un lexique qui n'a plus cours aujourd'hui : « municipe »[8] (8 occ.), « bailliage »[9] (8 occ.), « capitainerie »[10] (5 occ.), …

(11) **SCAMBONIDOE, (*Géog. anc.*) municipe** de l'Attique (…) (XIV, 739)
(12) **SCHENKCBERG, (*Géog. mod.*) bailliage** de Suisse (XIV, 762)
(13) **TAMARACA, ou Tamarica (*Géog. mod.*) capitainerie** du Brésil (…) (XV, 871)

Enfin, certains de ces mots classifieurs semblent inattendus dans une définition géographique, soit parce qu'ils dénotent une portion d'espace réduite rarement utilisée comme situeur dans les définitions d'ouvrages géographiques aujourd'hui, soit parce qu'ils constituent au contraire des noms de lieu extrêmement génériques et vagues (« lieu », « endroit », …).

(14) LACROME, (*Géog.*) écueil au voisinage du port de Raguse (…) (IX, 70)
(15) APHACE, (*Géog. anc.*) lieu dans la Palestine (…) (I, 523)
(16) **MAXIMIACUM, (*Géog.*) endroit** de la Franche-Comté (…) (X, 215)

Dans le cas des noms de personnes, un nombre moindre de mots classifieurs permet de classer l'EN désignée par le ou les mots vedettes de l'entrée de l'article comme une entité humaine collective. Ces mots sont : « peuple », « communauté », « habitant », « race », « peuplade » - soit 495 articles en tout (contre 9 520 articles pour les lieux). Par exemple :

(17) TZCHALATZKI les, & les TZUKTZCHI, (*Géog. mod.*) nom de deux peuples barbares & alliés (…) (XVI, 788)
(18) BARBETS, s. m. pl. (*Géog.*) habitans des vallées du piémont (…) (II, 73)
(19) OPHIOGENES les, (*Géog. anc.*) race particuliere d'hommes (…) (XI, 502)

3.1.2 Extension du calcul cooccurrentiel à des motifs plus larges

Le principe que nous avons appliqué pour mener une exploration cotextuelle des articles du sous-corpus Géographie avec TXM a consisté à utiliser, comme pivot d'une nouvelle recherche cooccurrentielle, le motif détecté à partir de la recherche précédente. Ainsi, après que nous avons identifié la liste exhaustive des mots classifieurs susceptibles d'apparaître

dans une fenêtre de trois mots à droite de la mention de domaine, nous avons relancé un calcul de cooccurrence à partir du pivot constitué comme suit : requête CQL utilisée pour extraire les vedettes des articles (cf. notre note 7) étendue à droite afin d'intégrer la liste des mot classifieurs identifiés[11]. Ce calcul a permis de détecter des motifs plus étendus. Concernant les noms de lieux, nous avons pu identifier dans 7 695 articles un nouveau motif (extension du motif précédent) où le nom classifieur est modifié par un syntagme prépositionnel (= SP) ayant pour tête « de ». Cette préposition a pour complément un nom propre tête d'un syntagme nominal (= SN) et référant à une entité géographique. Cette entité joue sémantiquement le rôle de site englobant (= Sit1) par rapport au nom classifieur (cible). Les exemples (8) à (13) et (16) cités plus haut vérifient ce nouveau motif (étendu). Pour (11) à (13) et (16), nous avons graissé pour information la séquence des mots extraite par TXM *via* la requête CQL donnée en note 7. Autrement dit, 65% des articles de Géographie pour lesquels notre expression régulière a permis d'extraire la vedette dans TXM mettent en œuvre ce motif étendu. Sur le plan sémantique, on peut reformuler les relations existant entre le nom de lieu en mot vedette, le nom classifiant et le nom propre situeur comme suit:

[Nom de lieu vedette] est un-e [*N classifieur*] située en /au /dans le/la/les [N situeur – ou Sit1 - du lieu classifiée].
HAINGEN, (Géogr.) petite ville d'Allemagne (...) => [Haigen] est une [ville] située en [Allemagne].
ZERGUE, (Géog. mod.) petite riviere de France (...) => [Zergue] est une [rivière] située en [France]

Le motif exposé ci-dessus devient à son tour le pivot d'une requête CQL. On identifie ainsi un nouveau motif plus étendu. Le SP modifieur prépositionnel (« de + SN ») y est suivi le plus souvent après virgule d'un nouveau SP jouant le rôle syntaxique d'un ajout, ayant le plus souvent pour tête « dans », « en » ou « à » ; cette préposition est suivie d'un SN complément incluant un nom propre et jouant le rôle de nouveau site (= Sit2). Par exemple :

(20) SCHENAW, (*Géog. mod.*) petite ville d'Allemagne, en Silésie (…) (XIV, 762)
(21) CAMPOLI, (*Géog.*) petite ville d'Italie, au royaume de Naples (…) (II, 581)
(22) VIELITSKA, montagne de, (*Géog. mod.*) montagne de Pologne, dans le palatinat de Cracovie. (…) (XVII, 260)

La relation dimensionnelle (Dim) entre les deux sites (Sit1 et Sit2) apparaissant successivement dans la phrase est de type : DimSit1 > DimSit2. Très souvent en effet, Sit1 réfère à un pays voire un continent (« Allemagne » (20), « Italie » (21), « Pologne » (22), …), Sit2 à un royaume, une province, une contrée etc. inclus dans Sit1 (« Silésie » (20), « royaume de Naples » (21), « palatinat de Cracovie » (22), …).
Cette première analyse nous permet de voir se détacher la structure générale d'un motif propre à la première phrase des articles du sous-corpus Géographie. Cette structure contient : (1) un pattern central formé du syntagme nominal « N1 de N2 » (N1 = nom/syntagme classifieur, N2 = nom propre situeur (Sit1), l'ensemble formant une ENE), (2) un pattern optionnel juxtaposable au pattern central et qui prend la forme suivante : syntagme prépositionnel « Prép SN » dont la tête est une préposition de sens spatial et qui possède dans son complément (= situeur Sit2, possiblement une ENE) un nom propre. Par récursivité, cette structure juxtaposée engendre des motifs de plus en plus complexes.

N1 de N2	[motif 1]
=> N1 de N2 (,) Prép SN	[motif 2]
=> N1 de N2 (,) Prép SN, Prép SN	[motif 3]
=> N1 de N2 (,) Prép SN, Prép SN, Prép SN	[motif 4]
=> N1 de N2 (,) Prép SN, Prép SN, Prép SN, Prép SN	[motif 5]

L'exemple suivant illustre le motif 3 dont nous avons extrait 1342 occurrences (soit 14% des occurrences des vedettes d'articles extraites sur TXM).

(23) AZAMOR, (*Géog.*) petite ville maritime d'Afrique, au royaume de Maroc, dans la province de Duquela. (I, 910)

Voici un exemple illustrant le motif 4 (329 occurrences dans TXM) :

(24) AUBENTON, (*Géog.*) ville de France en Picardie, dans la Thiérache, sur l'Aube. (I, 865)

Enfin, ce dernier exemple illustre les rares occurrences (65 occurrences) du motif 5 :

(25) DENAT, (*Géog. mod.*) petite ville de France au diocèse d'Alby dans le Languedoc, sur l'Assore, à trois lieues d'Alby. (IV, 824)

Concernant les noms de personnes, l'examen des indices forts relatifs à ce type d'entité sera plus rapide pour deux raisons : elles sont beaucoup moins nombreuses à figurer parmi les mots vedettes du corpus Géographie et le principe de détection puis d'extension des motifs est identique à celui exposé précédemment. Ainsi l'exemple ci-dessous illustre le motif 1 [N1 de N2][12] (446 occurrences) :

(26) GARAMANTES, s. m. pl. Garamantae, (*Géogr. ancienne.*) anciens peuples de la Lybie (…) (VII, 477)

Les exemples (27) et (28) illustrent successivement les motifs 2 [N1 de N2 (,) Prép SN] (89 occurrences) et 3 [N1 de N2 (,) Prép SN, Prép SN] (21 occurrences).

(27) CALMOUCKES ou CALMUQUES, (*Géog*.) peuples d'Asie, dans la grande Tartarie (…) (II, 563)
(28) BATUECAS ou LOS BATUECAS, (*Géog.*) peuples d'Espagne, dans le royaume de Léon, au diocese de Coria (…) (II, 161)

Pour ce motif, 14 occurrences sur 21 utilisent la préposition « selon » dans le dernier SP :

(29) SAMBRES, les (*Géog. anc.*) Sambri ; ancien peuple de l'Ethiopie sous l'Egypte, selon Pline (…) (XIV, 595)

Tous ces motifs, qu'ils concernent les noms de lieux ou de personnes, sont considérés comme incluant des indices forts, au sens de décisifs, en cela que leur détection dans le corpus par PERDIDO déclenche des décisions. Ces décisions sont relatives i) à la détection et à la classification des entités nommées, ii) mais également à leur mise en relation. Les mots classifieurs que nous avons extraits permettront également d'enrichir les dictionnaires et lexiques utilisés. Par ailleurs, les motifs que nous avons identifiés permettent de lier les EN évoquées suivant trois types de relation :

- une relation de sous-catégorisation entre les mots vedettes des articles de géographie et les mots classifieurs apparaissant à leur droite ;
- une relation de localisation site/cible entre les noms propres apparaissant dans les patterns juxtaposés et le mot classifieur (pattern noyau) ainsi que la vedette ;
- une relation d'inclusion spatiale entre le site 1 (Sit1) et les autres sites apparaissant à sa suite.

Illustrons notre propos par un exemple :

(26) CANGOXUMA, les (*Géog. anc.*) ville d'Asie de l'empire du Japon, dans l'ile de Ximo, au royaume de Bungo (…) (II, 597)

Dans (26), la reconnaissance du motif 4 permet de décider qu'il existe :

- une relation de sous-catégorisation (orientée du classifieur vers le classifié) entre les mots classifieurs (« ville », « empire », « ile », « royaume ») et les noms propres de lieux auxquels ils s'appliquent – mot vedette (« Cangoxuma ») et autres noms propres (« Japon », « Ximo », « Bungo ») : cette relation est déjà traitée par PERDIDO par le concept des ENE.

ville -> Cangoxuma ; empire -> Japon ; ile -> Ximo ; royaume -> Bungo

- une relation de localisation, non traitée par PERDIDO (orientée du site vers la cible), entre le mot vedette et les noms de lieux suivants :

Asie -> Cagoxuma ; Japon -> Cagoxuma ; Ximo -> Cagoxuma ; Bungo -> Cagoxuma

- une relation d'inclusion non traitée par PERDIDO (site incluant vers le site inclus : DimSit1 > DimSit2, …) entre les noms/SN de lieux qui suivent le mot vedette :

Asie (Sit1) -> empire du Japon (Sit2), ile de Ximo (sit3), royaume de Bungo (Sit4)

Le traitement des noms de personnes est quelque peu différent pour la relation à tisser entre le dernier nom propre et le reste de la phrase pour le motif 3. En effet, le dernier SP y apparaît très souvent du type « selon X » et installe une relation d'évidentialité (Dendale et Tasmowski, 2013) avec non pas l'EN désignée par le mot vedette mais avec tout le contenu de la phrase qui précède. L'identification de cette relation passe par celle de la préposition « selon » (cf. *infra*).

3.1.3 Implémentation des règles dans PERDIDO

Un grand nombre d'informations est déjà présent dans les annotations produites par PERDIDO, comme le montre la figure 2. Par exemple, le concept d'ENE (balises XML <rs>) permet d'annoter les relations entre les mots classifieurs et les noms propres. La figure ci-dessous montre un exemple simplifié du résultat XML-TEI produit par PERDIDO. On note que les relations de localisation et d'inclusion ne sont pas annotées dans la version existante du logiciel. L'étude du corpus décrite dans les sections précédentes nous a permis de développer des nouvelles règles que nous avons intégrées dans le système d'annotation de PERDIDO. L'objectif étant d'automatiser le traitement réalisé et de produire une annotation XML de ces informations qui sera interrogeable et interprétable pour l'analyse des informations géographiques de l'EDDA.

```xml
<s>
  <rs type="place">
    <name type="place" subtype="edda">CANGOXUMA</name>
  </rs> ,
  <term type="articleClass">(Géog.)</term>
  <rs type="place">
    <rs type="place">
      <term type="place">ville d'</term>
      <rs type="unknown">
        <name type="unknown">Asie</name>
      </rs>
    </rs>
    <rs type="unknown">
      <term type="unknown">de l'empire du</term>
      <rs type="place">
        <name type="place" subtype="edda">Japon</name>
      </rs>
    </rs>
  </rs> , dans
  <rs type="place">
    <term type="place">l'ile de</term>
    <name type="unknown">Ximo</name>
  </rs> , au
  <rs type="place">
    <term type="place"> royaume de </term>
    <rs type="place">
      <name type="place" subtype="edda">Bungo</name>
    </rs>
  </rs> .
</s>
```

Fig. 2. Annotation au format XML-TEI simplifié produit par PERDIDO.

La figure 3 montre le nouveau transducteur implémenté au sein de la cascade d'annotation de PERDIDO grâce au logiciel Unitex. Ce transducteur a pour rôle l'annotation du Motif 1. Il insère dans le texte la balise XML <phr type="relationHead"> qui symbolise la relation entre le mot vedette (première balise XML <rs>) et une ENE (élément <rs>) qui suit immédiatement la classification de l'article (<articleClass>). Ces trois éléments (<rs> et <articleClass>) sont des annotations réalisées par les transducteurs déjà existants dans la cascade d'annotation de PERDIDO.

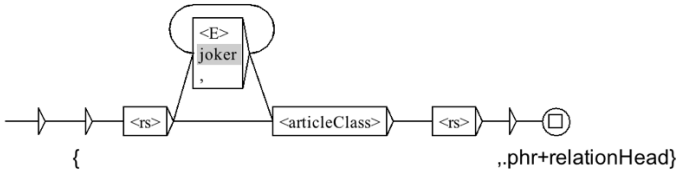

Fig. 3. Transducteur implémenté avec Unitex pour la reconnaissance du motif 1

Nous avons également ajouté un deuxième transducteur qui permet de repérer et d'annoter les motifs 2 à 5 (voir figure 4). Ce transducteur utilise l'annotation réalisée par le transducteur précédent (<phr+relationHead>) et annote les relations successives existantes entre cette entité et les informations de localisation (grâce aux prépositions « dans », « sur », « près » et à l'amalgame « au »).

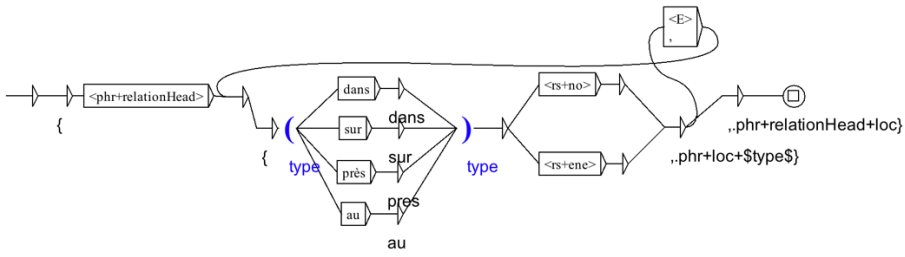

Fig. 4. Transducteur implémenté avec Unitex pour la reconnaissance des motifs 2 à 5

3.1.4 Expérimentations et évaluation

Nos premières expérimentations menées sur le sous-corpus Géographique montrent que l'ajout de ces règles d'annotation a permis de repérer 6 963 motifs (motifs 1 à 5), parmi lesquels 5 627 motifs de type 1, 970 motifs de type 2, 339 motifs de type 3 et 27 motifs de type 4. Les annotations réalisées permettent de déterminer que parmi ces 6 963 articles, 4 432 décrivent des villes, 692 des rivières, 388 des îles, 162 des fleuves, 147 des pays, 137 des montagnes, 121 des contrées, 107 des royaumes, 104 des bourgs, 53 des lacs, etc.

Parmi les 1 336 motifs (2 à 5) repérés, 1 073 impliquent la préposition « dans », 390 l'amalgame *au* et 252 la préposition *sur*. Ces prépositions sont réparties de la manière suivante sur les 1 336 motifs :

— 970 motif 2:
  — 749 « dans », 170 « au », 51 « sur » ;
— 339 motif 3:
  — 131 « dans+sur », 78 « au+dans », 74 « dans+au », 37 « au+sur », 8 « dans+dans », 6 « sur+dans », 3 « sur+au », 2 « au+au » ;
— 27motif 4:
  — 8 « au+dans+sur », 2 « au+dans+au », 1 « au+au+dans », 1 « au+sur+dans », 14 « dans+au+sur », 1 « dans+sur+sur » ;

Par ailleurs, l'ajout de règles et de dictionnaires spécifiques a permis une première étape d'amélioration des performances de la chaîne de traitement PERDIDO. En effet, une première évaluation des étapes de reconnaissance et de classification des EN réalisée sur un échantillon de 20 articles montre une amélioration significative des résultats (voir le tableau 1). Nous avons évalué différents types d'erreurs : EN non détectées (faux négatifs), EN détectées à tort (faux positifs), ainsi que les erreurs partielles : repérage partiel de l'EN (certains mots de l'EN peuvent être oubliés ou repérés en trop), erreurs de classification et les deux à la fois. Ces différentes erreurs sont combinées par la mesure SER (*Slot Error Rate*) définie par Makhoul et al, (1999), qui mesure le taux d'erreur du système d'annotation. A l'inverse de la précision (nombre de bonnes réponses par rapport au nombre total de réponses) et du rappel (nombre de bonnes réponses par rapport aux nombre de réponses attendues), plus ce score est faible meilleur est le résultat.

A titre de comparaison nous avons également évalué les résultats obtenus par les deux versions de PERDIDO avec ceux obtenus par l'outil d'annotation spaCy. Pour un nombre total d'EN de 230 (sur les 20 articles de notre échantillon), la version adaptée pour l'EDDA de PERDIDO repère et classe correctement 174 entités (75,65%) au lieu de 57 (24,78%) avec la version standard et 135 (58,7%) avec spaCy. On note d'après ces premiers résultats que l'ajout de règles spécifiques pour le traitement des articles de l'EDDA améliore significativement les résultats des systèmes d'annotation d'EN. Des erreurs subsistent et en particulier pour la classification. En effet, pour la version améliorée de PERDIDO, les erreurs de classification représente 50,74% du total des erreurs (34 / 67). Cela nous encourage à poursuivre notre analyse pour ajouter de nouvelles règles ou améliorer celles existantes.

Tableau 1. Evaluation des tâches de reconnaissance et de classification.

	Précision	Rappel	SER
PERDIDO standard	29,23%	24,78%	49,78%
PERDIDO EDDA	75,98%	75,65%	21,30%
spaCy	58,19%	58,70%	33,26%

3.2 Mise au jour d'indices linguistiques faibles

3.2.1 Élaboration des règles de détection

Nous avons également commencé l'étude d'indices faibles permettant de distinguer catégories et sous-catégories d'EN. Ces indices ne permettent pas de prendre une décision avec certitude lorsqu'un seul d'entre eux est repéré mais la combinaison de plusieurs indices (et leur pondération) peut nous permettre d'interpréter le contexte de l'EN et de déterminer sa nature selon la probabilité préalablement calculée sur l'ensemble du corpus.

Nous avons cherché à identifier les cotextes préférés d'apparition, dans tout le texte de l'EDDA, des noms propres (désormais NPr) référant à des sous-catégories d'entités géographiques et humaines dont nous avons observé qu'elles étaient les plus nombreuses (nombre d'occurrences > 100) dans le sous-corpus Géographie. En particulier, nous avons observé qu'un très grand nombre de NPr désignant des entités géographiques - quelles que soient leurs sous-catégories (*villes*, *rivières*, *îles*, etc. - cf. *supra*), sont préférentiellement précédées de la préposition « en » (« en France », « en Allemagne », …), des prépositions « dans » et « à » suivies de l'article défini (« dans le Boulonois », « à la Chine », …), ou encore de l'amalgame « au » (« au Maroc », « au Brésil », …). En revanche, les NPr référant à des individus humains sont préférentiellement précédés (rang -1) des prépositions « depuis » (« depuis Homère », « depuis Philippe-Le-Bel », …), « par » (« par César », « par Plutarque », …), « selon » (« selon Virgile », « selon Pline », …), « sous » (« sous Philippe-Auguste », « sous Henri III », …) et « suivant » (« suivant Cassini », « suivant Ptolémée »).

Nous avons aussi identifié que parmi les prépositions sélectionnées préférentiellement par les NPr référant aux étendues d'eau que sont les rivières (« Thamise », « Moselle », « Garonne », « Tage », …), les fleuves (« Rhône », « Danube », « Rhein », « Nil », …), les mers (« Baltique », « Caspienne », « de Marmara »…) et les océans (« Atlantique », « Indien », …), la préposition « sur » suivie de l'article défini leur est plus spécifique. Quant aux prépositions « chez », « par » et « parmi » suivies de l'article défini pluriel « les », elles précèdent préférentiellement les NPr désignant des groupes humains (peuples, tribus, …) tels que « chez les Hollandois », « par les Romains » ou « parmi les Grecs ». Enfin la préposition « dans » suivie directement d'un NPr apparaît préférentiellement devant les noms de villes (« dans Paris », « dans Rome »…) et les noms de personnes (« il est fait mention dans Pline »). Afin de disposer d'une règle susceptible de discriminer ces deux catégories d'EN dans les constructions de type « dans NPr », nous avons cherché à identifier une liste de cooccurrents préférés en amont et en aval de cette structure prise comme pivot. Seuls les cotextes où le nom propre désigne un individu permettent de faire émerger, au moyen du calcul des cooccurrences, des environnements statistiquement significatifs : en amont, les cooccurrents verbaux « lire », « trouver », « parler », « faire mention », « voir », « nommer », « rapporter », « signifier » ; en aval, la présence après virgule de l'abréviation « liv./lib. » et/ou de chiffres romains indiquant les références des citations. Par exemple :

(27) On lit dans Pausanias, I. III. c. x. que lorsqu'on étoit descendu du lieu nommé les Hermes (…) (SCOTIAS, XIV, 811)

Les règles que nous tirons de ces observations sont :

Règle 1 : si le NPr est précédé
— en position -1,
(a) de la préposition « en », (b) de l'amalgame « au »
— en position -2,
(c) de la préposition « dans » suivie de l'article défini, (d) de la préposition « sur » suivie de l'article défini,
— en positions -1 et -2,
(e) de la préposition « sur » suivie de l'article défini,
alors ce NPr se verra affecter un score favorable pour un classement dans la catégorie des noms de lieux, et pour le cas (e), dans la sous-catégorie « nom d'étendue d'eau » .

Règle 2 : si le NPr est précédé
— en position -1
(a) de la préposition « depuis », (b) de la préposition « par », (c) de la préposition « selon », (d) de la préposition « sous », (e) de la préposition « suivant »,
— en positions -1 et -2
(f) de la préposition « chez » suivie de « les », (g) de la préposition « par » suivie de « les », (h) de la préposition « parmi » suivie de « les »,
alors ce NPr se verra affecter un score favorable pour un classement dans la catégorie des EN et dans la sous-catégorie « personne » dans les cas (a) à (e), dans la sous-catégorie « collectivité » dans les cas (f) à (h).

Règle 3 : si le NPr est précédé
— en position -1 par « dans »
(a) et si cette préposition est elle-même précédée (fenêtre de deux mots en amont) par l'un des lemmes verbaux suivants : « lire », « trouver », « parler », « faire mention », « voir », « nommer », « rapporter », « signifier » ;
(b) ou si le NPr est suivi (fenêtre de 3 mots en aval) d'une virgule puis de la séquence « liv./lib. »,
alors ce NPr se verra affecter un score favorable pour un classement dans la catégorie des EN et dans la sous-catégorie « personne ».
(c) Par défaut - c'est-à-dire si les règles (a) et (b) ne sont pas vérifiées - le NPr se verra affecter un score favorable pour un classement dans la catégorie des lieux et dans la sous-catégorie « ville ».

3.2.2 Combinaison des indices faibles

Plusieurs occurrences d'un même NPr se voient souvent affecter des indices faibles qui peuvent tous être convergents (voir exemple (28), pour le NPr « César » où nous avons graissé les indices). Les règles de combinaison des indices faibles sont pour le moment les suivantes : les indices convergents sont additionnés. Les indices divergents s'annulent deux à deux. A l'issue des additions/soustractions d'indices, s'il demeure deux indices faibles convergents pour le classement d'un NPr dans une catégorie/sous-catégorie d'EN, leur somme aboutit à un indice fort et ce classement est opéré. S'il ne reste qu'un seul indice faible, aucun classement automatique n'est réalisé. L'implémentation des règles pour les indices faibles est en cours dans PERDIDO.

(28) a. Les deux autres lois agraires, dont il est fait mention dans le Digeste, & dont l'une fut publiée **par** César & l'autre par Nerva (AGRAIRE, I, 182)
b. METIOSEDUM, (Géog. anc.) lieu de la Gaule celtique, voisin de Paris, dont il est **parlé dans** César**, lib. VII.** de bello Gallico. (METIOSEDUM, X, 463)
c. On voit par-là combien il s'introduisit de nouvelles noblesses, tant **sous** César & sous Auguste (…) (PATRICE, Patriciat, Patricien, XII, 177)

d. Cette riviere, **selon** César, de Bel. Gal. l. I. faisoit avec la Marne, la séparation entre les Gaulois & les Belges. (SEQUANA, XV, 76)

4 Conclusion

Selon Nadeau et Sekine (2007), "The impact of textual genre (journalistic, scientific, informal, etc.) and domain (gardening, sports, business, etc.) has been rather neglected in the NERC literature". L'article présenté ici propose de faire fond sur des spécificités discursives propres au genre encyclopédique pour mettre au jour des indices forts à même d'aboutir à la catégorisation et à la sous-catégorisation d'entités nommées non encore annotées par la chaîne de traitement PERDIDO. Nous sommes conscients que cette relation avec le genre discursif concerné tend à restreindre la réutilisation des règles tirées de ces indices forts dans d'autres textes que des encyclopédies ou des dictionnaires. "Perhaps unsurprisingly, (…) porting a system to a new domain or textual genre remains a major challenge" (ibid.). On soulignera cependant qu'elles peuvent aussi être utilisées dans des tableaux géographiques, des descriptions de régions ou de contrées, des rapports administratifs… et dans d'autres dictionnaires et encyclopédies. On comparera ainsi les premières phrases de l'article ADEN dans EDDA (33), dans *Le Dictionnaire de Trévoux* (DT)(édition lorraine, 1740) (34) et dans *La Grande Encyclopédie* (1885-1902) (LGE)(35)

(33) ADEN, (Géogr.) ville de l'Arabie heureuse, capitale du Royaume de ce nom. (EDDA)

(34) ADEN, Aden, Adena, Adenium. Nom d'une ville de l'Arabie heureuse, & et non pas de la haute Éthiopie, Cornelius l'a dit dans le Marmol François. (DT)

(35) ADEN. Presqu'ile et ville situées sur la côte méridionale de l'Arabie, par 12° 46' 40'' de lat. et 42° 38' 44'' de long. E . de Paris (à la station télégraphique) à 92 milles (170 kil.) à l'E. du Bab-el-Mandeb. (LGE).

Les indices faibles quant à eux ne sont pas, pour un bon nombre, sensibles au genre discursif et possèdent donc une applicabilité plus vaste. Ajoutons enfin que le repérage et la classification de ces toponymes facilitera leur annotation dans d'autres textes contemporains non encyclopédiques.

Ces indices forts et faibles ainsi que les motifs qui les combinent nous donnent des informations précieuses sur l'histoire du genre encyclopédique français et la construction du discours géographique des Lumières. Les EN du dix-huitième siècle ne sont pas les mêmes qu'à l'époque contemporaine et une meilleure connaissance du contexte discursif où elles opèrent améliore, comme nous l'avons montré, leur reconnaissance. En travaillant avec un système qui récupère plus d'EN et (éventuellement) les classe selon des relations géographiques ou culturelles de l'époque, il devient possible de faire des analyses approfondies de textes de très grande taille qui ne soient plus fondées sur des méthodes anachroniques. L'interprétation du discours géographique dans l'ensemble des dictionnaires et des encyclopédies du dix-huitième siècle (textes situés au cœur de la production imprimée de l'époque) s'approche ainsi progressivement de notre portée.

Dans cet article, nous avons présenté une méthodologie allant de l'analyse manuelle du corpus grâce au logiciel TXM à l'enrichissement des règles d'annotation implémentées dans la plateforme PERDIDO pour la reconnaissance et la classification des EN dans l'*Encyclopédie*. Nos travaux ultérieurs vont consister dans un premier temps à évaluer les résultats obtenus par notre approche et à améliorer la combinaison des indices faibles et forts pour la création d'un indice de confiance sur les annotations réalisées. Dans un second temps, nous nous emploierons à interpréter les informations annotées et à les associer à des ressources géographiques externes afin de déterminer la localisation des lieux décrits.

Projet soutenu par la MSH-LSE, Université de Lyon, CNRS. Ce travail a été réalisé grâce au soutien financier du projet IDEXLYON de l'Université de Lyon, dans le cadre du programme "Investissements d'Avenir" (ANR-16-IDEX-0005)

Références bibliographiques

Chinchor, N., & Marsh, E. (1998). MUC-7 information extraction task definition (version 5.1). In *Proceedings of the 7th Message Understanding Conference (MUC)*

Dendale, P., & Tasmowski, L., (1994). Présentation : l'évidentialité ou le marquage des sources du savoir. *Langue française (102),* pp. 3–7

Gaio, M., & Moncla, L. (2019). Geoparsing and geocoding places in a dynamic space context. *The Semantics of Dynamic Space in French: Descriptive, experimental and formal studies on motion expression*, 66, 353.

Gravier, G., Bonastre, J., & Geoffrois, E. (2004). Ester, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. Actes de JEP-TALN

Heiden, S. (2010). « The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme », in *24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japon : ENS-Lyon, p. 10 p. http://halshs.archives-ouvertes.fr/docs/00/54/97/64/PDF/paclic24_sheiden.pdf.

Longrée, D., & Mellet, S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages* (189), pp. 65–79

Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. (1999). Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pp. 249-252.

Maurel, D., Friburger, N., Antoine, J.Y., Eshkol, I., & Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées, *Traitement automatique des langues* 52, 69

McDonald, D.D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. In *Corpus processing for lexical acquisition*, pp 21-39

McDonough, K., Moncla, L., & van de Camp, M. (2019). Named entity recognition goes to old regime france: geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science, 33 (12), pp. 2498-2522*

Moncla, L., & Gaio, M. (2018). Services web pour l'annotation sémantique d'information spatiale à partir de corpus textuels. *Revue Internationale de Géomatique* (28), 439

Moncla, L., Gaio, M., Nogueras-Iso, J., & Mustière, S. (2016). Reconstruction of itineraries from annotated text with an informed spanning tree algorithm, *International Journal of Geographical Information Science* 30, 1137

Moncla, L., McDonough, K., Vigier, D., Joliveau, T., & Brenon, A. (2019). Toponym disambiguation in historical documents using network analysis of qualitative relationships. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities*, pp. 1-4.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., & Gaio, M. (2014). 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*, pp. 183-192

Nadeau, N., & Sekine, S. (2007). A survey of named entity recognition and classification, *Lingvisticae Investigationes*, 30, 3

Nobata, C., Sekine, S., Isahara, H., & Grishman, R. (2002). Summarization System Integrated with Named Entity Tagging and IE pattern Discovery. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), pp. 1742-1745

Pirovani, J., Alves, J., Spalenza, M., Silva, W., da Silveira Colombo, C., & Oliveira, E. (2009) Adapting NER (CRF+LG) for Many Textual Genres. In *Proceedings of the 35th Conference of the spanish society for natural language processing*, pp. 421-433.

Poibeau, T. (2005). Sur le statut référentiel des entités nommées, in *Conférence Traitement Automatique des Langues*, pp. 173-183

Sekine, S., & Eriguchi, Y. (2000). Japanese named entity extraction evaluation : analysis of results. In *Proceedings of the 18th conference on Computational linguistics*. Volume 2 (ACL), pp. 1106–1110

Sundheim, B.M. (2005). Overview of results of the MUC-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding* (ACL), pp. 13-31

Yannick Mathieu, Y. (2003). La grammaire de construction, *Linx. Revue des linguistes de l'université Paris X Nanterre, (48),* pp. 43-56

---

[1] Projet soutenu financièrement par la MSH Lyon-St-Etienne pour la période 2019-2020 (https://www.msh-lse.fr/projet19/geodisco)

[2] https://encyclopedie.uchicago.edu

[3] http://textometrie.ens-lyon.fr

[4] Sur le plan fonctionnel, le motif est un « cadre collocationnel » accueillant un ensemble d'éléments fixes et de variables, susceptible d'accompagner la structuration textuelle et, simultanément, de caractériser des textes de genres divers.

[5] Les deux principaux contributeurs aux articles de géographie dans EDDA sont L. de Jaucourt (signature (D.J) et D. Diderot (signature par la marque *). Les 1271 articles signés de ce seul dernier pour ce domaine apparaissent essentiellement dans les volumes 1 et 2, alors que c'est à partir du volume 7 que les contributions du Chevalier de Jaucourt deviennent significatives. Elles montent en charge dans les volumes suivants pour atteindre plus de mille articles dans chacun des quatre derniers volumes de texte (14 à 17). Enfin les articles géographiques non-signés se concentrent surtout dans le volume 2. Nous avons cherché à déterminer si ces critères d'auteurs influaient sur les résultats présentés ici. Nous avons observé que les articles non-signés sont ceux qui recourent le plus souvent aux motifs mis au jour, tous les auteurs y recourant cependant.

[6] « Parfois » seulement car il convient de rappeler que dans l'EDDA, il n'existe presque aucune vedette d'article où figure un nom propre référant à une personne, sinon inclus éventuellement dans un syntagme dont il n'est pas la tête (voir par ex. l'article: *SPINOSA, philosophie de*). Aucune trace parmi les mots vedettes de grandes figures historiques comme Néron, Charlemagne, Christophe Colomb etc. C'est là une position idéologique. Il n'est plus question de valoriser des hommes ou des femmes (comme dans les dictionnaires de Bayle ou de Moréri), mais d'étendre la lumière sur les pratiques des arts et sciences et sur les nouvelles découvertes historiques.

[7] [word="[A-Z].*[A-Z]"] [] ?[] ?[] ?[] ?[] ?[] ?[] ?[] ?[] ?[word="\("] [word="Gé.*"] [word="."] [] ?[] ?[] ?[] ?[word="\)"]

[8] « C'est le titre que portoient les Villes du Latium et de l'Italie, dont les habitans participoient au droit de bourgeoisie Romaine, sans qu'elles cessassent de former des Cités à part. » (*Dictionnaire de l'Académie*, 1798, art. « Municipe »)

[9] « (...) se dit aussi De certaine étendue de pays qui est sous la Juridiction du Bailli. Ce bourg est d'un tel Bailliage. » (*Dictionnaire de l'Académie*, 1798, art. « Bailliage »)

[10] Forme d'administration territoriale des colonies dans l'empire colonial portugais.

[11] Deuxième requête CQL : en gras figure la première requête (note *supra*) employée pour l'extraction des mots vedettes sur TXM: **[word="[A-Z].*[A-Z]"] [] ?[] ?[] ?[] ?[] ?[] ?[] ?[] ?[] ?[word="\("] [word="Gé.*"] [word="."] [] ?[] ?[] ?[] ?[word="\)"]** [] ?[] ?[] ?[word="ville|Ville|riviere|rivieres|ile|Ile|île|isle|iles|îles|province|fleuve|bourg|Bourg|montagne|montagnes|lieu|royaume|Royaume|pays|village|port|bourgade|promontoire|Promontoire|comté|lac|lacs|forteresse|golfe|golphe|cap|capitale|canton|vallée|place|principauté|château|fauxbourg|fauxbourgs|fontaine|forêt|forêts|gouvernement|municipe|maison|nation|palatinat|Palatinat|campagne|duché|bailliage|bois|capitainerie|contrée|état|marais|cercle|district|eaux|écueil|écueils|paroisse|plaine|quartier|champ|endroit|forum|Forum|havre|passage|pont|ruisseau|terre|torrent|volcan|abbaye|baronie|capitainie|champ|champs|chef|chemin|cité|colline|désert|empire|détroit|entrepôt|fauxbourg|grotte|habitation|isthme|marquisat|mont|mur|palais|péninsule|préfecture|Province|rade|région|rocher|route|ruines|salines|seigneurie|station|territoire|hameau|mer|rue"]. Nous ne donnerons pas le détail des requêtes CQL à venir, faute de place

[12] [word="[A-Z].*[A-Z]"] [] ?[] ?[] ?[] ?[] ?[] ?[] ?[] ?[] ?[word="\("] [word="Gé.*"] [word="."] [] ?[] ?[] ?[] ?[word="\)"] [] ?[] ?[] ?[word="peuple|peuples|communauté|habitans|race|peuplade"] [] ?[word="de|d'|du|des"]