Extraction automatique d'un lexique à connotation géographique à des fins ontologiques dans un corpus de récits de voyage

Marie-Noëlle BESSAGNET (1), Mauro GAIO (2), Eric KERGOSIEN (2), Christian SALLABERRY (1)

(1) Laboratoire LIUPPA, UPPA, IAE, Avenue du doyen Poplawski, 64012 PAU marie-noelle.bessagnet@univ-pau.fr, christian.sallaberry@univ-pau.fr
(2) Laboratoire LIUPPA, UPPA, Faculté des Sciences, Département Informatique, 64000 PAU, mauro.gaio@univ-pau.fr eric.kergosien@univ-pau.fr

Résumé Le but de ces travaux est d'extraire un lexique en analysant les relations entre des syntagmes nominaux et des syntagmes verbaux dans les textes de notre corpus, essentiellement des récits de voyage. L'hypothèse que nous émettons est de pouvoir établir une catégorisation des syntagmes nominaux associés à des Entités Nommées (EN) à l'aide de l'analyse des relations verbales. En effet, nous disposons d'une chaine de traitement qui extrait, interprète et valide des EN dans des documents textuels. Ce travail est complété par l'analyse des relations verbales associées à ces EN, candidates à l'enrichissement d'une ontologie.

Abstract The aim of this research work is to extract a lexicon by analyzing the relationship between nominal syntagms and verb construction within our corpus, namely travel stories. We would like to establish a categorization of nominal syntagms linked to Named Entity (NE) thanks to verbal relationships analysis. In fact, we develop a process flow in order to extract, to interpret and to validate NE in textual documents. This research work is completed by the analyze of verbal relationships linked to these EN which could enrich our ontology.

Mots-clés : Entité nommée, ontologie, relations verbales, patrons linguistiques
Keywords: Named Entity, ontology, verbal relations, language patterns

1. Introduction

Les fonds documentaires patrimoniaux mobilisent de gros efforts de numérisation. Leur valorisation se fait généralement via des outils de gestion documentaire classiques (notices descriptives, catalogues thématiques) intégrant généralement des moteurs d'indexation plein-texte. Le travail que nous présentons s'intègre dans un tel cadre visant la constitution et la gestion de fonds documentaires composés de versions électroniques de documents. Ce fonds documentaire du XIXème siècle est consacré aux Pyrénées (en particulier des récits de voyages). Ces documents contiennent de très nombreuses références au territoire pyrénéen. Ainsi, une chaîne de traitement complète de construction d'index particulièrement adaptée à l'aspect géographique des contenus a été proposée (Sallaberry et al, 2007).

Au sein de cette chaîne, une des premières étapes est constituée par l'annotation automatique d'Entités Nommées (EN) conceptualisées particulières : les toponymes. En accord avec (Karen et al, 2009), « *Au regard du contenu tout d'abord, il importe de se focaliser non pas tant sur comment annoter, mais sur quoi annoter, en fonction de l'application visée* ». Dans nos applications liées à la constitution d'index géographiques, nous sommes confrontés à l'ambigüité référentielle (Leidner Jochen L., 2004). Aussi, pour lever cette ambigüité, nous nous intéressons à la problématique du marquage des toponymes selon un couple Nom propre et expressions antéposées décrivant des relations spatiales ou « indirections ». Par exemple, dans la phrase, «je traversais un affleurement de roches carbonatées au centre de la partie méridionale de la Montagne Pelée», la chaîne de traitement automatique que nous avons mise en œuvre (Sallaberry et al, 2007) permet de construire, après détection des «noms toponymiques» : ici «Montagne Pelée», et après interprétation sémantique de l'indirection : ici l'expression «au centre de la partie méridionale», une représentation géométrique, brique essentielle pour notre processus de construction des index spatiaux.

L'objectif de ce papier est de proposer une étape supplémentaire dans ce processus afin d'exploiter l'information contenue dans l'expression «affleurement de roches carbonatées» qui comme les syntagmes nominaux exprimant une indirection peut être considérée comme faisant partie intégrale du toponyme que nous souhaitons interpréter. Cette nouvelle étape nécessite d'extraire automatiquement ce type d'expression et de vérifier si elle est porteuse d'un sens géographique. Nous proposons, d'une part, une nouvelle chaîne de traitement automatique permettant d'isoler le syntagme nominal pouvant se trouver antéposé à la première forme toponymique marquée, représentée par le couple `[indirection ?, nom propre]`, et d'autre part une première proposition afin de vérifier s'il y a un rapport suffisamment fort entre sa participation à une relation linguistique particulière du type `[verbe de déplacement, préposition ?, syntagme candidat, toponyme`[1]`]` et sa capacité à évoquer un sens géographique. L'objectif est d'utiliser ce syntagme nominal, après validation de son sens géographique, dans deux options : soit comme terme précisant le type lors de la récupération de la géométrie du toponyme (par exemple dans des ressources de type BD géographique ou gazetteer) car une correspondance a été trouvée via l'ontologie géographique ; soit comme proposition à l'enrichissement de cette même ontologie si celui-ci y est absent.

L'étude de notre corpus relatif aux récits de voyages a montré qu'un grand nombre de toponymes sont mentionnés dans des phrases comportant un verbe de déplacement. Il a

[1] Le point d'interrogation indique une présence optionnelle

également révélé que les termes présents entre le toponyme et le verbe ont fréquemment une connotation géographique (environ 50 % des termes sont issus de l'ontologie initiale). Ainsi, nous émettons l'hypothèse que la présence de verbes de déplacement est un bon indicateur quant à la connotation géographique des syntagmes antéposés aux toponymes.

Dans une première partie (&2), nous exposerons les travaux connexes. Dans la partie (&3) nous aborderons notre méthodologie générale et nous l'exemplifierons. Nous présenterons ensuite la chaine de traitement du langage naturel mise en œuvre dans ces travaux en montrant comment elle permet d'analyser les entités nommées et leurs termes associés ainsi que les verbes de déplacement (&4). La quatrième partie (&5) abordera la construction du lexique par le biais de la catégorisation des termes candidats à l'enrichissement de l'ontologie via le patron (V,P?,_,E). Enfin, nous présenterons les limites et perspectives de notre approche (&6).

2. Travaux connexes

De nombreux travaux traitent de l'identification et de l'annotation des entités nommées (Karen et al, 2009). Le plus grand nombre est basé sur des méthodes par apprentissage exploitant des propriétés de type patrons morphosyntaxiques de surface. L'approche ici présentée se focalise sur deux points : l'annotation fine et la désambiguïsation des EN de type géographique d'une part, et l'assistance à l'enrichissement d'une ontologie d'autre part. En accord avec (Erman et Jacquet, 2006), nous distinguerons trois types de travaux : « ceux dont le but est de désambiguïser les EN, ceux cherchant à construire une ressource spécifique pour le traitement des EN, et enfin ceux combinant les deux précédents, c'est à dire cherchant à faire de la désambiguïsation tout en exploitant une ressource spécifique ». Nous nous plaçons dans le troisième groupe.

Dans notre approche, nous tentons de traiter finement les EN en nous appuyant sur des ressources spécifiques. Ceci est proche des travaux de (Bunescu et al., 2006) qui présentent une approche de désambiguïsation d'EN qui exploite la ressource encyclopédique Wikipédia. Nous pourrions également nous rapprocher du travail de (Paşca M., 2004) concernant la construction d'une ressource à partir de corpus pour annoter finement les EN. Cependant, ici, la ressource n'est pas construite mais enrichie.

De plus, le fait d'avoir des toponymes aussi précis nous permet de lever des ambigüités dans le processus de résolution géométrique de ces derniers (Leidner Jochen L., 2004).

3. Méthodologie générale

3.1. Démarche : présentation générale

La méthodologie proposée (figure 1) se décompose selon quatre étapes:

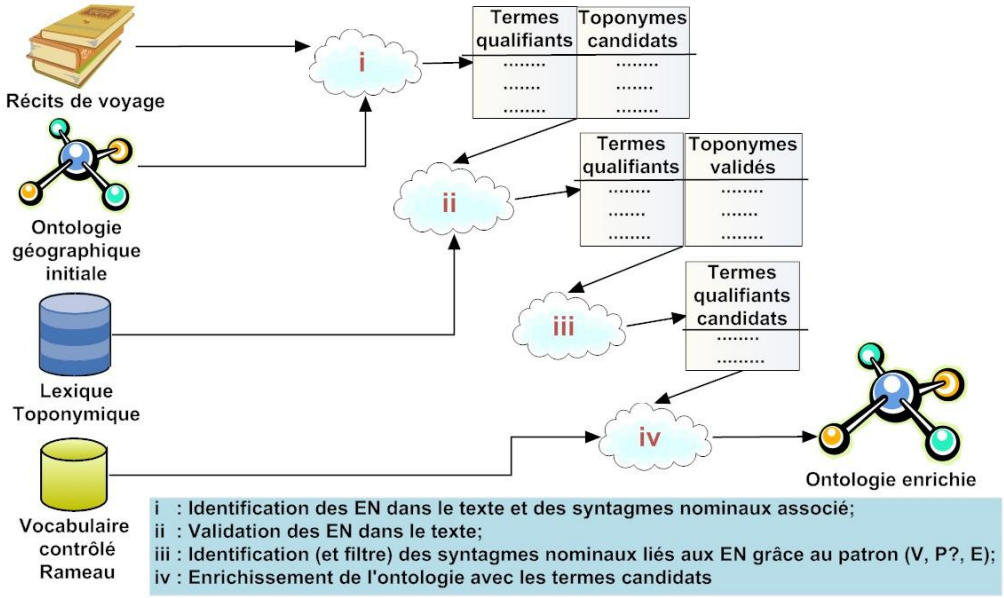

Figure 1. ***La démarche générale***

Le but est d'utiliser une succession de patrons lexico-syntaxiques et un ensemble de transducteurs modélisant les verbes de déplacement afin de nous aider à filtrer des termes candidats à l'enrichissement de l'ontologie géographique selon les propositions décrites dans (Kergosien et al, 2009). A cette fin, nous possédons plusieurs ressources : un lexique des termes exprimant les indirections, une ontologie géographique initiale, un lexique toponymique, le thésaurus Rameau[2]. Nous utilisons ici Rameau comme une source de connaissance généraliste ; elle comporte plus de 400000 termes.

En recherche d'information, les entités nommées (EN) sont généralement séparées en quatre classes : personnes, lieux, organisations et expressions temporelles. Les syntagmes nominaux *Henri IV, Laruns*, *lac de Crabioules, pic d'Ossau, l'ouest des Pyrénées* sont des exemples d'entités nommées de notre corpus. Le premier est une EN de type personne. Les quatre suivants sont des EN de type lieu que nous appelons également entités nommées spatiales (ENs) ou toponymes. Nous distinguerons le toponyme *lac de Crabioules* du nom toponymique *Crabioules*. Ainsi, *lac de Crabioules* et *pic de Crabioules* sont deux ENs qu'il s'agit de distinguer, car leurs représentations géométriques dans une ressource de type base de données géographiques peuvent fortement différer. Comme déjà mentionné, notre chaine de traitement a notamment pour but de repérer et d'analyser les syntagmes associés au nom toponymique, différenciant ainsi au sein de la base de données géographique, le toponyme *lac de Crabioules* de type *hydronyme* du toponyme *pic de Crabioules* de type *oronyme.* Dans le processus proposé, l'ontologie géographique descriptive de la BD géographique est utilisée comme ressource pour cette différenciation des ENs, mais dans certains cas, cette ontologie ne permet pas une telle distinction, il est alors fait appel à des ressources plus génériques (comme le thésaurus Rameau).

---

[2] Répertoire d'autorité-matière encyclopédique et alphabétique unifié. Rameau est utilisé, en France, par la Bibliothèque Nationale de France, les médiathèques régionales ou locales, les bibliothèques universitaires, ainsi que plusieurs organismes privés. Voir http://rameau.bnf.fr/informations/rameauenbref.htm

Prenons les quatre exemples suivants :

1-j'ai remonté à pied la vallée d'Ossau *2- j'ai marché jusqu'au pas des Echelles*
3- J'ai mangé mes provisions à Marsous *4- j'ai traversé le gave de Pau*

Seul le premier est étiqueté grâce à la ressource que constitue notre ontologie géographique. Les termes *pas, provisions, gave* ne figurent pas dans l'ontologie. Or, *pas des Echelles* et *gave de Pau* sont bien des ENs dans lesquelles les termes associés *pas* et *gave* ont un sens géographique et nous devons essayer de les typer à l'aide de ressources externes complémentaires.

Notre méthodologie, de type patrons lexico-syntaxiques, permet, d'une part, de repérer l'entité nommée spatiale c'est à dire l'ensemble du syntagme relatif au toponyme [Indirection?, Nom_Toponymique] et, d'autre part de repérer le syntagme nominal candidat à être lexique. L'analyse de son emploi dans une construction de type [verbe_déplacement/préposition?/expression candidate ?/ entité nommée spatiale], soit (V,P ?,_,E), permet ensuite d'en supposer l'emploi géographique. Alors est mobilisée la ressource de type thésaurus, afin d'associer au syntagme un ensemble d'informations permettant de mieux caractériser son champ lexical.

Détaillons sur des exemples tirés de notre corpus le processus dans sa globalité.

3.2.Exemplification sur un extrait de notre corpus

Rappelons que dans la phrase « ***Le lac d'Artouste*** *est très beau* », l'entité nommée « ***Artouste»*** est considérée comme un **nom_toponymique** et «***Le lac d'Artouste»*** comme un **toponyme ou une ENs**. Nous employons de manière indifférenciée ces dénominations dans notre article. Dans un objectif de simplification, les exemples ne comportent pas d'indirections, mais comme dit précédemment, si celles-ci avaient été présentes, elles auraient été préalablement isolées.

Nous allons détailler les quatre étapes sur un corpus-exemple composé de six phrases :

a. « j'ai remonté à pied la vallée d'Ossau » ;
b. « A droite se trouve la route des Eaux-Chaudes » ;
c. « j'ai marché jusqu'au pas des Echelles » ;
d. « J'ai mangé mes provisions à Marsous » ;
e. « j'ai traversé le gave de Pau » ;
f. « J'ai marché jusqu'à la fontaine de Visos avant le déjeuner »

(i) Identification des EN dans le texte et des syntagmes nominaux associés

Dans une première étape, notre chaine de traitement repère les entités nommées et leurs syntagmes nominaux associés, soient : «vallée d'Ossau», «route des Eaux-Chaudes», «pas des Echelles», «provisions à Marsous», «gave de Pau», «fontaine de Visos». Pour les syntagmes nominaux associés, leur classe d'appartenance est également marquée : ils peuvent être connus dans l'ontologie géographique initiale, ils peuvent être connus dans le thésaurus Rameau, ils peuvent être de source inconnue. Ainsi, nous obtenons une liste de toponymes candidats.

(ii) Validation des EN dans le texte

Dans une deuxième étape, les noms_toponymiques_candidats sont traités pour validation. Ainsi, «Ossau», «Eaux-Chaudes», «Echelles», «Marsous», «Pau» sont validés

et «Visos» ne l'est pas. Ainsi, seuls les cas a), b) c) d) et e) comprennent des toponymes considérés pour la suite du traitement.

cas	Toponyme candidat	Nom_toponymique validé	Source du syntagme nominal qualifiant
a	vallée d'Ossau	Oui	Ontologie
b	route des Eaux-Chaudes	Oui	Ontologie
c	pas des Echelles	Oui	Rameau
d	provisions à Marsous	Oui	Rameau
e	gave de Pau	Oui	Inconnu
f	fontaine de Visos	Non	Ontologie

Les cas a) et b) ne nous poseront aucun problème pour la suite car ce sont des toponymes validés dont le syntagme nominal est connu dans notre ontologie initiale. Ces deux cas ne seront pas considérés dans les deux prochaines étapes.

(iii) Identification et filtre des syntagmes nominaux liés aux EN grâce au patron (V,P ?,_,E)
Nous allons, dans cette étape, nous baser sur le patron (V,P ?,_,E) pour un dernier filtre. Dans les phrases c), d) et e), nous allons considérer les verbes de déplacement afin de ne garder que les toponymes de ce type de construction syntaxique.

Dans la phrase c), nous avons bien une structure (V,P ?,_,E) <marcher, jusqu'au, pas, des Echelles>. De plus, le terme *pas* est connu du thésaurus Rameau. Ce terme pourra donc être conservé comme un candidat potentiel à l'enrichissement de l'ontologie. Dans la phrase d), nous avons le terme *provisions* connu du thésaurus Rameau mais aucunement une structure (V,P ?,_,E). Ce terme ne sera donc pas conservé pour être un candidat à l'enrichissement de l'ontologie. Dans la phrase e), nous avons bien une structure (V,P ?,_,E) <traverser, gave, de Pau>. Le terme *gave* est de source inconnue et ce terme pourra donc être conservé comme un candidat potentiel à l'enrichissement de l'ontologie.

(iv) Enrichissement de l'ontologie avec les termes candidats
Sur les six cas du départ, grâce à la mise en œuvre de notre processus, seuls deux termes seront considérés dans cette étape : «pas» et «gave». Le terme «pas» appartient au thésaurus Rameau dans un sens géographique (Figure 2). Dans ce cas, notre approche permet d'exploiter de façon automatisée ce type de termes, en vérifiant d'éventuelles approximations de sens avec les concepts de l'ontologie. Pour le terme « gave », cette étape est plus délicate. Nous devons faire appel à d'autres ressources externes, du type Dictionnaire Larousse (Figure 3). Cependant, le terme « gave » doit être validé. Le processus automatisé trouvera ici ses limites.

Figure ***2. Extrait du thésaurus Rameau*** *Figure 3*. ***Extrait du Dictionnaire Larousse***

Abordons la chaine de traitement mise en œuvre dans nos travaux.

4. Une chaine de traitement du TAL

Le contexte dans lequel nous nous plaçons du point de vue du TAL est celui de l'extraction d'informations ciblées (l'information géographique) dans un corpus relativement homogène.

4.1. Analyse des EN et des termes associés

La chaine de traitement que nous proposons a été implémentée grâce à l'utilisation de LinguaStream[3] (figure 4). Elle s'appuie sur une démarche de « tokenisation » classique (1).

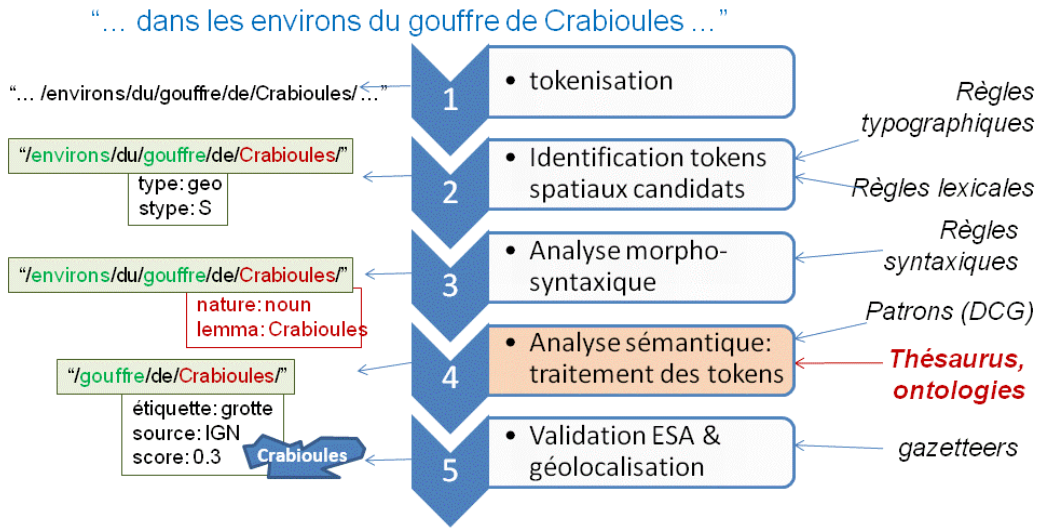

Figure 4. ***Chaine de traitement***

Nous marquons rapidement des EN spatiales candidates puis nous appliquons les étapes suivantes de l'analyse à ces EN uniquement. Un marqueur de *token* spatial candidat (2) utilise

[3] http://www.linguastream.org/

des règles lexicales (lexiques d'introducteurs d'information spatiales) et typographiques (majuscule en début de token par exemple). Puis, un analyseur morpho-syntaxique associe un lemme et une nature à chaque *token* spatial candidat (i.e. *Crabioules*, nom). Un analyseur sémantique (4) associé à des règles de grammaire DCG (Definite Clause Grammar) exprimées en Prolog qualifie des entités spatiales absolues (ESA) et les entités spatiales avec indirections que nous appelons entités spatiales relatives (ESR). Rappelons que dans le cadre du processus ici présenté, seule la partie ENs des ESR est exploitée, la partie marquée comme étant une indirection est écartée. La fin du processus donnera donc des ESA accompagnées d'un terme : *(mont, Crabioules)*, *(ville, Pau)*, (*vallée, Ossau)*, *(roche carbonatée, Montagne Pelée),* etc. Les ESA sont validées et géolocalisées (5) automatiquement à l'aide de ressources externes ou de *gazetteers* contributives internes.

La liste des termes associés à des Noms_toponymiques est marquée grâce à l'ontologie géographique initiale et à l'aide de ressources externes de type thésaurus (ici Rameau). Nous obtenons ainsi un ensemble de termes associés à des ENs. Nous avons montré dans (Kergosien et al, 2009) que sur un ensemble de 14 récits, nous obtenons des termes présents dans l'ontologie géographique initiale ou pas. Parmi ces autres termes, grâce à la ressource externe Rameau, nous savons que certains ont un caractère géographique et pourraient enrichir la taxonomie (comme « gave » trouvé dans « le gave de Pau » qui pourrait aider à identifier une représentation spatiale adéquate à « Pau »), contrairement à d'autres (comme « maire » trouvé dans « le maire de Pau »).

4.2. Analyse des verbes

Le sous-ensemble des verbes de déplacement auxquels nous nous intéressons ici sont ceux qui entrent dans une construction [`verbe de déplacement, préposition ?, syntagme ?, toponyme`], que nous notons (V,P ?,_,E) (Loustau et al, 2007). Cette construction permet dans une certaine mesure de lever une grande partie des problèmes d'ambiguïté que l'on peut trouver dans des propositions comme « quitter son mari », « traverser une mauvaise période », etc. De plus, nous nous basons sur les principes de la polarité aspectuelle d'un verbe telle que définie par (Boons, 1987) : ce critère de classification des verbes de déplacement est basé sur la phase temporelle à laquelle le verbe de déplacement fait intrinsèquement référence.

Ainsi, conformément à cette notion de polarité aspectuelle, les verbes de déplacement qui interviennent dans l'évocation d'un déplacement seront selon (Borillo, 1998), (Garcia-Debanc et al, 2009):

- Des « verbes initiaux » ou de polarité initiale, comme *quitter, partir, sortir, s'échapper, s'éloigner*, etc. : « le déplacement exprimé par le verbe prend implicitement le site comme lieu d'origine de la cible » ;
- Des « verbes finaux » ou de polarité finale, comme *arriver à, atteindre, entrer dans, regagner*, etc. « L'emplacement du site représente la destination vers laquelle la cible se déplace ou est déplacée » ;
- Des « verbes médians » ou de polarité médiane, comme *traverser*, *franchir*, *parcourir, passer par, se déplacer dans, etc.* représentent les cas dans lesquels « la zone du site représente le lieu parcouru ou traversé par la cible durant son déplacement ».

Dans les travaux linguistiques, nous trouvons une classification des verbes de déplacement (Vandeloise, 1987), (Laur, 1991), (Borillo, 1998), (Aurnague, 2008). Ces divers travaux nous aident ici à caractériser les verbes de déplacement dans l'analyse de notre corpus.

Dans notre processus, le mode opératoire pour le repérage et l'identification des syntagmes verbaux de type déplacement est mis en œuvre grâce au principe des transducteurs. Nous rappelons que les transducteurs sont basés sur des machines à états finis mettant en correspondance deux langages réguliers. Compte tenu des observations faites sur notre corpus, la construction des syntagmes verbaux de déplacement peut être apparentée à un langage régulier. Cette modélisation du déplacement sous forme de transducteurs est générique à tout déplacement exprimé principalement sous forme verbale.

Les transducteurs sont traduits en règles de grammaire dans lesquelles nous retrouvons les principaux objets du modèle : le verbe, la préposition et l'ES. Cette analyse à base de règles s'appuie sur les résultats d'analyses plus en amont. La première est l'analyse morphosyntaxique de chaque unité lexicale. Elle permet, notamment, de s'abstraire des formes fléchies des mots, comme par exemple celles des verbes conjugués. La deuxième analyse est l'extraction des ESA et des ESR.

5. Catégorisation des termes candidats à l'enrichissement

Le graphique suivant (figure 5) donne diverses statistiques sur notre corpus vis-à-vis de l'ontologie initiale, du thésaurus Rameau et de source inconnue.

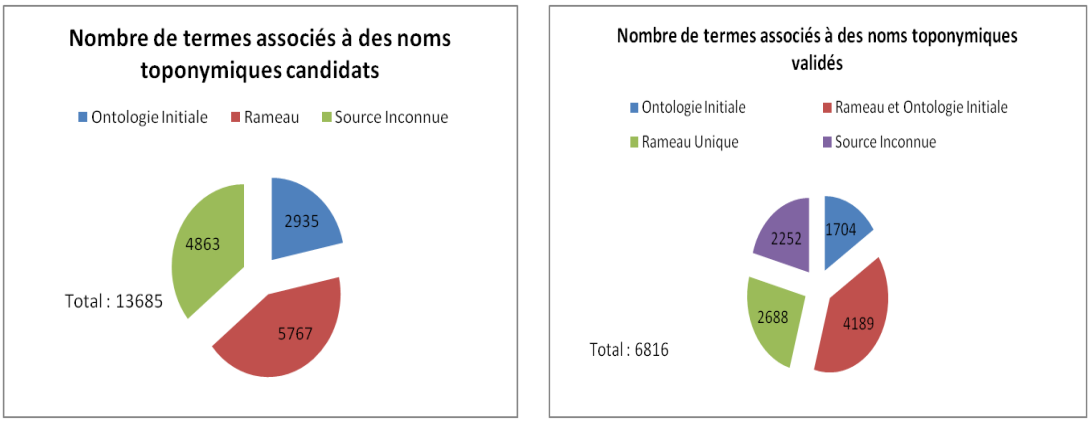

Figure 5. ***Une première catégorisation des termes***

Ainsi, la validation des EN permet quasiment de diviser par 2 la liste initiale de toponymes candidats : sur les 13685 occurrences de départ, nous avons 6816 occurrences de termes associées à des noms toponymiques validés (Figure 5).

Grâce à la chaine de traitement, nous avons donc mis en évidence dans le texte les constructions [`verbe de déplacement, préposition (facultative), syntagme nominal, entité spatiale`] . Parmi les 6816 occurrences de termes identifiés qualifiant un nom toponymique validé, 4189 sont présentes dans Rameau dont 2688 ne sont pas partagés avec l'ontologie initiale, 2252 occurrences de termes sont de source inconnue. Elles sont donc candidates à son enrichissement. Nous devrions donc a priori les traiter. Grâce à l'application du patron `(V,P ?, _,E)` nous allons éliminer un ensemble conséquent de séquences pour lesquelles des termes de source Rameau ou de source inconnue sont associés à des noms toponymiques. Ainsi, nous réduisons notre champ d'analyse puisque, nous aurons au final 628 occurrences de termes identifiées participant à un patron `(V,P ?, _,E)` et au final 300 occurrences de termes identifiées ne seront pas présentes dans l'ontologie initiale (plus de 200 termes distincts).

Les verbes de déplacement de notre corpus pour les 628 occurrences sont classés selon les polarités suivantes (figure 6.1), dont 67% de polarité finale. Parmi ces verbes, 3 catégories

peuvent être repérées (Figure 6. 2): des verbes très utilisés dans le corpus (cf Cat 1), des verbes très peu utilisés (cf Cat. 2) et des verbes liés exclusivement à des syntagmes nominaux de l'ontologie initiale (cf Cat. 3).

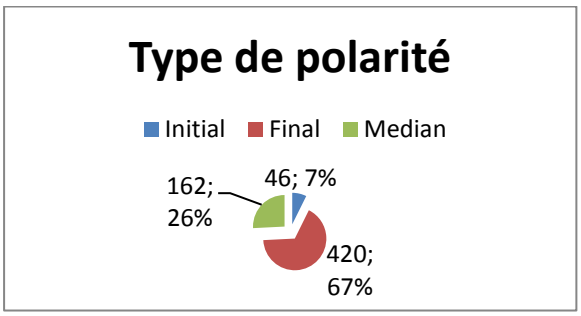

Figure 6.1 : ***type de polarité***

Cat. 1	Cat. 2	Cat. 3
aller	border	engager
arriver	dépasser	remonter
atteindre	grimper	sortir
conduire	marcher	
descendre	précipiter	
entrer	retourner	
monter	rôder	
partir		
passer		
rendre		
revenir		
suivre		
traverser		
venir		

Figure 6.2 : ***Un échantillon de catégories de verbes***

L'usage des relations verbales tel que décrit dans cet article nous permet de mieux cibler la proposition des termes candidats à l'enrichissement.

6. Limites et perspectives

L'hypothèse que nous avons émise sur la présence de verbes de déplacement comme indicateur quant à la connotation géographique des syntagmes antéposés aux toponymes est vérifiée. La méthodologie proposée nous permet d'extraire de notre corpus de récits de voyage un lexique d'étiquettes sémantiques, élément de précision lors de la récupération de la géométrie ou candidat à l'enrichissement de l'ontologie.

Cependant, dans notre corpus, nous avons également des constructions `[verbe de déplacement, préposition (facultative), syntagme nominal, entité spatiale ]` qui posent problème et qui méritent donc d'autres analyses.

(a) *« j'ai marché jusqu'au maire d'Accous »*
(b) *« Je pus pénétrer dans les solitudes des Landes »*
(c) *« Je suis parti à cheval pour Gavarnie »*

Le cas a) est retenu dans notre analyse via le patron `(V,P ?, _,E)` avec un terme de source Rameau. Mais ce cas s'avère erroné du fait de la signification du terme « maire » qui n'est en aucun cas géographique. Le cas b) est retenu dans notre analyse via le patron `(V,P ?, _,E)` avec un terme de source inconnue. Mais ce cas s'avère erroné du fait de la signification du terme *solitudes* qui n'est en aucun cas géographique. Le cas (c) pourra être correctement analysé : le syntagme nominal *à cheval*, grâce à un nouveau patron lexico-syntaxique identifiant ces syntagmes comme une modalité de déplacement, va permettre ainsi de lever l'ambigüité.

Nous travaillons actuellement sur ces points en étudiant d'autres ressources (EuroWordNet, Larrousse) qui pourraient nous permettre, de la même façon que Rameau, d'améliorer encore plus l'inventaire du lexique à connotation géographique. Nous travaillons également sur une analyse plus fine des prépositions accompagnant les verbes intransitifs afin de savoir si nous

pouvons repérer des patrons [verbe, préposition, complément] où le complément serait nécessairement à connotation géographique.

Remerciements

Cette recherche a été réalisée dans le cadre du projet GeOnto "Constitution, alignement et exploitation d'ontologies géographiques » (http://geonto.lri.fr/), en partie financé par l'Agence Nationale de la Recherche (ANR-O7-MDCO-005).

Références

AURNAGUE M. (2008), Qu'est-ce qu'un verbe de déplacement ? : Critères spatiaux pour une classification des verbes de déplacement intransitifs du français, CONGRES MONDIAL DE LINGUISTIQUE FRANÇAISE, PARIS, FRANCE, 2008 DOI: 10.1051/CMLF08041.

BOONS, J.-P. (1987). `La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs'. LANGUE FRANÇAISE 76, 5 – 40.

BUNESCU R., PAŞCA M. (2006), Using Encyclopedic Knowledge for Named entity Disambiguation, in EACL The Association for Computer Linguistics (2006), Dernier accès Web le 10/02/2010 http://www.aclweb.org/anthology/E/E06/E06-1002.pdf

EHRMANN M., JACQUET G. (2006), Vers une double annotation des Entités Nommées, Traitement Automatique des Langues 2006 Volume 47 Numéro 3

FORT K., EHRMANN M., NAZARENKO A., (2009) Vers une méthodologie d'annotation des entités nommées en corpus ?, TALN 2009, Dernier accès Web le 05/01/2010 http://hal.archives-ouvertes.fr/docs/00/40/23/21/PDF/taln09_right.pdf

GARCIA-DEBANC C., DUVIGNAU K., DUTRAIT C., GANGNEUX M. (2009). «Enseignement du lexique et production écrite. Un travail sur les verbes de déplacement à la fin de l'école primaire ». Pratiques 141-142, Juin 2009 (Masseron & Lecolle, coord.), pp. 208-232.

KERKOSIEN E, KAMEL M, SALLABERRY C, BESSAGNET MN, AUSSENAC N., GAIO M, (2009), Construction automatique d'ontologie et enrichissement à partir de ressources externes, JFO 2009, ISBN 978 1 60558 842 1, pp11-20

LAUR D. (1991). Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple. *PhD thesis, Université de Toulouse II.*

LEIDNER JOCHEN L.(2004). Toponym resolution in text: "which Sheffield is it?". In Proceedings of the the 27th, Annual International ACM SIGIR Conference (SIGIR 2004), pages 602–606, Sheffield, UK. ACM Press

LOUSTAU P. , GAIO M., NODENOT T. (2007), Des déplacements à l'itinéraire, du syntagme au discours. Extraction d'itinéraires d'un corpus de récits de voyages. *SAGEO 2007. Clermont-Ferrand, France. Juin 2007. CD-ROM. ISBN : 978-2-85710-078-2,* Dernier accès Web le 13/12/2009 http://www.emse.fr/site/SAGEO2007/CDROM/p36.pdf

PAŞCA M., 2004. Acquisition of categorized named entities for web search, In Proc. of CIKM, 2004.

SALLABERRY C., GAIO M., LESBEGUERIES J., LOUSTAU P. (2007) *A Semantic Approach for Geospatial Information Extraction from Unstructured Documents*. Chapitre du livre The Geospatial Web book, publié par Springer dans The Advanced Information and Knowledge Processing Series. ISBN 1-84628-826-6. pp. 93-105. 2007