Typage de noms toponymiques à des fins d'indexation géographique

Mauro GAIO* — **Christian SALLABERRY*** — **Van Tien NGUYEN***

* *Laboratoire LIUPPA*
BP-1155, 64013 PAU Université Cedex
prénom.nom@univ-pau.fr

RÉSUMÉ. *Cet article présente une annotation automatique d'expressions spatiales géolocalisables permettant de contribuer à l'amélioration du processus d'indexation de l'information géographique non structurée. La méthode proposée combine l'exploitation de relations spécifiques intraphrastiques et de ressources externes. Elle repose sur la reconnaissance d'entités nommées spatiales étendues afin, d'une part, d'en déduire une représentation symbolique exprimée en termes de traits sémantiques puis, d'autre part, de leur attribuer au moins une représentation numérique. Notre méthode totalement implantée et a été expérimentée sur un fonds documentaire réel de type « récit de voyage ».*

ABSTRACT. *This paper presents an automatic annotation of spatial expressions enabling the improvement of the indexing process of non-structured geographic information. The proposed method combines the use of specific intra-sentential relationships and external resources. It relies on expanded spatial named entities recognition to deduce a symbolic representation expressed in terms of semantic features, on the one hand, and, to give them a numerical representation, on the other hand. Our method completely implemented and was tested on a corpus of travelogues.*

MOTS-CLÉS : *extraction de lexique, modèle spatial, index géographique, patrons linguistiques, enrichissement d'ontologie.*

KEYWORDS: *lexicon extraction, spatial model, geographic index, linguistic pattern, ontology enrichment.*

1. Introduction

L'analyse automatique d'expressions spatiales dans la langue connaît un regain d'intérêt notamment dans des problématiques d'extraction d'information géographique à partir du contenu textuel de documents. Les enjeux scientifiques se situent dans deux thématiques distinctes. D'une part, dans la création ou l'enrichissement d'ontologies dans le domaine de la géographie, comme cela a été mis en avant dans les travaux de Uitermark (2001) ou ceux de Brodeur (2004). D'autre part, dans la recherche d'information géographique (*Geographical Information Retrieval*) telle qu'évoquée pour la première fois par Larson (1996) et précisée par Purves et Jones (2004).

Les objectifs du projet GéOnto[1] couvrent ces deux thématiques. Une première ontologie (qualifiée de *noyau* ou encore *initiale*) spécifique à un domaine géographique (la topographie) a tout d'abord été créée. Un ensemble de réponses a ensuite été proposé pour la mise en place d'une méthodologie permettant de transformer des termes, extraits automatiquement de textes, en concepts potentiels pour enrichir cette ontologie noyau. L'ontologie ainsi obtenue a été exploitée comme ressource dans deux scénarios d'usage. Le premier s'est appuyé sur celle-ci pour assurer l'interopérabilité de différentes bases de données géographiques tandis que le second l'a intégrée dans une chaîne d'indexation spatiale de fonds documentaires territorialisés pour augmenter l'adéquation des représentations géométriques utilisées comme index.

La méthode présentée dans cet article a été conçue et testée dans le cadre de ce projet. Elle repose sur la reconnaissance « d'entités nommées spatiales étendues » afin d'en déduire une meilleure qualification (typage d'un point de vue BD) des noms de lieux. Cette qualification permettant d'une part, de produire une représentation symbolique plus cohérente de ces lieux puis, d'autre part, de leur attribuer une ou plusieurs représentations numériques plus appropriées. La représentation symbolique est exprimée en termes de traits sémantiques dans un esprit proche de celui proposé par (Charnois et Enjalbert, 2005) et (Bilhaut *et al.*, 2007). Cette méthode est intégralement opérationnelle et a été expérimentée sur un corpus réel : une collection de récits de voyage.

1.1. *Problématique*

Le problème central ici est la constitution d'un lexique de termes topographiques. Ce lexique doit être obtenu à partir de l'extraction des syntagmes nominaux employés pour leur dénotation topographique (par exemple, *territoire aride, au sud de l'étroite vallée*, etc.) dans un contexte documentaire donné. Notre cadre expérimental est constitué d'un fonds documentaire comportant plusieurs centaines de récits de voyage dans les Pyrénées. Afin d'opérer automatiquement cette extraction de manière ciblée, notre première contribution consiste à proposer des patrons linguistiques per-

1. GéOnto est un projet ANR (ANR-07-MDCO-005-04) `http://geonto.lri.fr/`.

mettant de différencier les syntagmes à dénotation topographique parmi tous ceux contenus dans le fonds documentaire cible.

La figure 1 exprime graphiquement les processus de filtrage successifs devant être appliqués à l'ensemble A des termes candidats, pour construire progressivement un sous-ensemble A' de termes dont la relation à une ontologie du domaine[2] sera très fortement présumée. Ce processus de filtrage n'a de sens que si l'on dispose *a priori* d'une ontologie noyau du domaine souhaité. Dans le contexte du projet GéOnto, cette ontologie noyau a été produite à partir des documents techniques de spécification des bases de données géographiques de l'IGN.

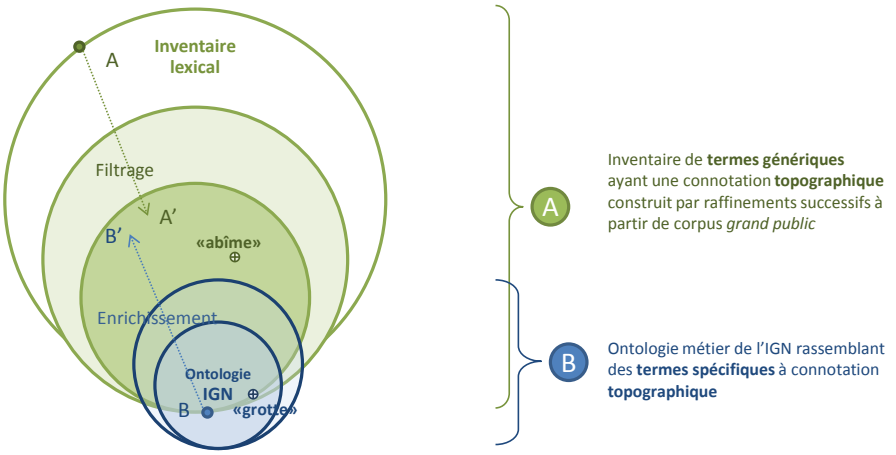

Figure 1. *Dans le cadre de notre expérimentation, l'inventaire (A') de termes topographiques extrait du corpus « récits de voyage » est en intersection avec celui (B) décrit dans l'ontologie noyau bâtie à partir des spécifications de bases de données géographiques de l'IGN.*

Elle décrit notamment un ensemble B de termes spécifiques usités par des spécialistes du domaine (dans notre cas la topographie). Comme le montre la figure 1, les ensembles A' et B ont une intersection non vide. Les termes communs décrits par l'ensemble $A' \cap B$ permettent de préciser la nature du nom de lieu, c'est-à-dire son type dans des ressources externes adaptées pour une recherche plus efficace des géométries correspondantes. L'ensemble $A' - B$, quant à lui, correspond à des termes ne labélisant aucun concept dans l'ontologie et qui, par conséquent, ne permettent pas ce typage. La seconde contribution concernera l'enrichissement de l'ontologie B afin de tendre vers un ensemble B' de termes représentés dont l'intersection avec A' sera plus

2. L'ontologie de domaine a été définie conjointement par Gruber (1995) comme une spécification explicite d'une conceptualisation et par Guarino et Giaretta (1995) qui en soulignent le caractère formel : une ontologie est essentiellement une représentation explicite de la sémantique d'un domaine, telle qu'elle est perçue par une communauté donnée.

importante. Cet enrichissement consiste à relier des termes de A' aux structures sémantiques décrites dans B. Nous avons choisi d'utiliser une ressource terminologique générique de type thésaurus pour obtenir des micro-arborescences de termes qui sont considérés dans ce thésaurus comme ayant un sens proche. Ces micro-arborescences nous permettent ensuite de mettre en œuvre des méthodes d'alignement pour que l'enrichissement puisse être proposé sur l'ontologie noyau. Ainsi, par exemple, le terme *abîme* pourra être proposé, comme candidat à l'enrichissement de la structure ontologique contenant initialement le terme *grotte* dans B', à condition que les structures arborescentes correspondantes aient en commun un certain nombre de propriétés.

1.2. *Quelques observations sur le corpus*

Comme nous pouvons le constater dans l'extrait (1) 3., les syntagmes nominaux (SN) à annoter sont très souvent associés à des noms de lieux (*la charpente altière des Monts-Maudits ; les roches calcaires de la Pèna-Blanca ;* etc.) ayant tous comme propriété intrinsèque une géolocalisation.

(1) *1.* Depuis quelques temps une vive curiosité avait porté mes regards vers la Maladetta [...]

2. Je parlai de mes intentions à plusieurs guides de Luchon [...]

3. Après avoir contemplé, avec une admiration mêlée d'effroi, **la charpente altière** des Monts-Maudits, nous songeâmes bientôt à descendre sur le **territoire aride** au sud de la **région** d'Aragon. Le temps était menaçant : de légers brouillards parcouraient les **hauteurs**, et précédaient des nuages d'une teinte grisâtre, qui roulaient vers nous, venant de l'ouest des Pyrénées, un orage s'amoncelait : il ne tarda pas à éclater. Ayant renvoyé nos chevaux et payé le tribut accoutumé à la complaisance des carabinieros (douaniers) espagnols, nos guides chargèrent nos provisions sur leurs épaules, et nous descendîmes, assez lestement, vers le **pied** de la Maladetta, laissant à notre droite les **roches calcaires** de la Pèna-Blanca. Arrivés au fond de **la vallée** du Plan-des-Etangs, qui est plus élevée que sa voisine, la **vallée latérale** de l'**hospice** de Bagnères, de 446 mètres, nous laissâmes derrière nous une **cabane** habitée pendant l'été par des bergers espagnols, pour remonter, par un **plan rocailleux**, jusqu'au **gouffre** de Tourmon, qui absorbe les eaux d'un **torrent rapide**, descendant de la partie orientale du **glacier** de la Maladetta [...]

Cette observation est corroborée par les travaux de Vandeloise (1986) sur le couple [cible, site] et de Borillo (1998) sur le couple [entité concrète, repère spatial] qui correspondent au couple (terme, nom de lieu). Toutefois, si cette observation est intéressante, elle reste incomplète car des SN considérés comme n'ayant aucune dénotation ni connotation géographique peuvent également être associés à des noms de lieux comme dans les extraits (1) 1. et (1) 2. La question qui se pose ici est de savoir comment filtrer ces SN qui, compte tenu de nos objectifs, doivent être ignorés car considérés comme étant du bruit. L'étude de notre corpus a permis d'observer que très fréquemment ce couple se trouve en relation, au sein d'une même phrase, avec des verbes de déplacement (*descendre vers le pied de la Maladetta ; remonter par un plan rocailleux jusqu'au gouffre de Tourmon*) ou des verbes de perception (*contempler la charpente altière des Monts-Maudits*) ou encore des verbes permettant de décrire les aspects topographiques (*dominent ce village, s'étendent entre les différents mamelons*). Dans plusieurs cas, la construction de ce couple fait appel à des relations spatiales, telles que modélisées, pour certaines, par Egenhofer et Franzosa (1991) et, pour d'autres, par Ligozat (1998), afin de faire référence à un lieu complexe (*descendre sur le territoire aride au sud de la région d'Aragon ; arrivés au fond de la vallée du Plan-des-Etangs*).

Il faut enfin constater que dans certains cas des SN faisant référence à un concept géographique sont présents dans nos textes sans être directement associés à un nom propre de lieu (*cabane ; plan rocailleux ;* etc.). Ils ne sont donc pas directement repérables par ce biais.

Compte tenu de toutes ces observations notre problématique générale concerne la modélisation des relations entre SN candidats au typage de toponymes pour la mise en œuvre d'un processus d'annotation et d'extraction automatique.

L'article est structuré comme suit. La section 2 relate des travaux dédiés à la reconnaissance d'entités nommées, de repères spatiaux et à l'évocation de déplacements dans la langue. La section 3 est consacrée à la description des structures syntaxico-sémantiques employées dans les textes étudiés et des patrons lexico-syntaxiques mis au point pour leur annotation. Ces patrons servent de support à l'extraction de termes et noms propres de lieux (nous verrons en section 3 pourquoi nous n'utilisons pas délibérément ici le terme de *toponyme*) et mènent à leur inventaire. La section 4 concerne l'utilisation d'un tel inventaire terminologique pour enrichir une ontologie topographique en connectant les termes relevés au thésaurus RAMEAU pour bâtir des hiérarchies locales. Ces hiérarchies permettent de proposer ces termes en ajout dans l'ontologie grâce au principe de l'alignement. En section 5, des expérimentations évaluent cette méthode d'extraction et l'indexation spatiale de documents à l'aide de l'ontologie ainsi obtenue. Ces expérimentations montrent comment le typage de noms de lieux est amélioré par nos propositions. Enfin, la dernière section conclut l'article.

2. Travaux connexes et repères bibliographiques

2.1. *Annotation des entités nommées et gazetiers*

De manière générale l'annotation des entités nommées (personnes, entreprises, lieux, etc.) est une problématique reconnue comme jouant un rôle important dans de nombreux traitements automatiques de la langue (Sagot et Boullier, 2008) et notamment dans le cas de l'extraction automatique d'information (Poibeau, 2003). Elle est encore considérée actuellement comme une tâche difficile.

La reconnaissance d'entités nommées (NER pour *Named Entity Recognition* en anglais) consiste à traiter un flux de mots issus d'une analyse lexicale préalable. Un détecteur d'entités nommées utilise généralement un système d'apprentissage ou une base de règles *ad hoc* pour détecter les entités nommées et les catégoriser (Poibeau, 2003). La reconnaissance d'entités nommées par apprentissage se fait à partir de textes étiquetés à la main par des experts : des méthodes d'analyse statistique (les textes sont considérés comme un flux de caractères) permettent de bâtir des patrons génériques utilisables sur un corpus plus large. L'approche *ad hoc*, quant à elle, s'appuie sur des patrons lexicaux bâtis manuellement avec l'aide d'experts comme : « un nom propre précédé par la préposition *à*, est potentiellement un lieu » qui est un exemple de règle lexicale. Ces règles sont ensuite appliquées à un corpus.

Nous nous intéressons exclusivement aux entités nommées spatiales étendues comme définies dans la section 3. Plusieurs techniques, ayant fait leurs preuves pour mettre en œuvre cette annotation, comme par exemple dans les travaux de (Rocío et Erick, 2010), ou au sein de notre équipe (Loustau, 2008), sont élaborées à l'aide de ressources externes nommées « gazetiers » (ou *gazetteers* en anglais). Un gazetier [3] est un dictionnaire ou répertoire géographique dont les entrées sont des noms de lieux. À chaque entrée du dictionnaire peuvent être associées des informations comme l'appartenance à une ou plusieurs structures administratives (commune, région, pays, etc.), la caractéristique physique (montagne, rivière, route, etc.), des données statistiques, une représentation géométrique exprimée dans un référentiel géographique. Il existe plusieurs gazetiers accessibles par Internet tels que : Geonames[4], BDNyme[5], Word Gazetteer[6], GEOnet Names Serve (GNS[7]), etc.

3. http://en.wikipedia.org/wiki/Gazetteer
4. http://geonames.org
5. http://www.ign.fr
6. http://www.world-gazetteer.com/
7. http://earth-info.nga.mil/gns/html/

Il existe également de nombreux outils de reconnaissance automatique d'entités nommées. *GATE ANNIE*[8], *LingPipe*[9], *OpenCalais*[10], *Stanford NER*[11] et *OpenNLP*[12] sont qualifiés de généralistes et visent le marquage de plusieurs catégories d'entités nommées, *MetaCarta*[13] et *Yahoo !Placemarker*[14] ciblent les entités nommées de type lieux tandis que *GuTime*[15] et *HeidelTime*[16] sont dédiés aux entités nommées de type dates.

En ce qui concerne les entités nommées spatiales, nous pouvons citer le projet CasEN (Maurel *et al.*, 2011) qui propose différentes typologies de noms de lieux utilisées dans des cascades de transducteurs pour la reconnaissance d'entités nommées. D'autres travaux (Buscaldi et Rosso, 2008) (Bouamor, 2009) procèdent à la catégorisation de noms de lieux après leur identification. La démarche proposée dans (Bouamor, 2009) exploite notamment la structure des documents : par exemple, dans l'encyclopédie collaborative Wikipédia, l'identification des entités nommées se fait dans le *titre* et leur catégorisation s'appuie sur l'analyse de la première phrase de la partie *description* ou encore de la partie *catégorie* en fin d'article. La démarche proposée dans (Buscaldi et Rosso, 2008) (Buscaldi, 2009), quant à elle, vise plus particulièrement la désambiguïsation des noms de lieux reconnus.

Nous proposons une approche hybride qui, comme Buscaldi et Rosso (2008) et Bouamor (2009), repère d'abord des noms de lieux mais qui recherche également, comme proposé par Maurel *et al.* (2011), des termes présents dans des ressources (ici ontologiques) pour repérer des termes associés, potentiellement géographiques.

2.2. *Repères spatiaux dans la langue et relations spatiales*

Selon Borillo (1998), un lieu est une portion de l'espace matériel dans lequel nous nous situons et nous évoluons. On peut alors considérer que dans l'expression *la région d'Aragon* l'entité concrète est dénotée par le terme *région*, et que le repère spatial peut être déduit du nom toponymique *Aragon*. Il en est de même pour l'expression plus étendue *la partie orientale du glacier de la Maladetta*, le repère spatial peut être déduit, de la même manière que dans l'exemple précédent, du nom toponymique *Maladetta* et l'entité concrète incarne ici l'expression *partie orientale du glacier*. Cette dernière expression contient donc une précision de localisation au sein de l'entité concrète, cette précision étant exprimée par une relation spatiale. L'inter-

8. http://gate.ac.uk/ie/annie.html
9. http://alias-i.com/lingpipe
10. http://www.opencalais.com
11. http://nlp.stanford.edu/software/CRF-NER.shtml
12. http://opennlp.sourceforge.net
13. http://www.metacarta.com
14. http://developer.yahoo.com/geo/placemaker/
15. http://www.timeml.org/site/tarsqi/modules/gutime/index.html
16. http://dbs.ifi.uni-heidelberg.de/index.php?id=106

prétation de cette relation nécessite le déclenchement d'un raisonnement spatial. Clementini (2009) propose une classification des relations spatiales en trois catégories : topologiques, projectives, et métriques. Ces classes sont respectivement fondées sur les propriétés de l'espace topologique, projectif, et euclidien. Les relations topologiques ont été de loin les plus étudiées, et parmi les premiers modèles proposés le RCC-8 de Randell *et al.* (1992) est devenu la base de nombreuses autres propositions. Dans l'ouvrage dirigé par Aiello *et al.* (2007) une tentative de synthèse est proposée autour de ces modèles. Les deux autres catégories ont été moins explorées. Les relations projectives, dont l'intérêt premier est de pouvoir être décrites par des propriétés projectives sans avoir recours aux propriétés métriques (Billen et Clementini, 2004). Les relations projectives tentent de formaliser des relations exprimées en langue naturelle par des expressions comme : *à droite de, en avant de, entre, le long de, à la banlieue de, au nord de*, etc. Bien que des modèles spécifiques aient été développés pour certaines de ces relations, comme les relations d'orientation (Freksa, 1992 ; Hernández, 1993) et les directions cardinales (Frank, 1992 ; Ligozat, 1998) il manque un modèle capable de représenter toutes les variantes de cette classe de relations. Comme les relations topologiques, les relations projectives peuvent être considérées comme étant de nature qualitative, car elles n'ont pas besoin de s'appuyer sur une représentation euclidienne lorsque impliquées dans l'implantation d'un raisonneur sur l'espace. Ce problème de la représentation logique et du traitement algorithmique de l'espace est parfaitement exposé par Balbiani et Muller (2000). Les relations métriques, telles que la distance entre deux points, quant à elles, sont généralement considérées comme étant de nature quantitative, comme par exemple dans l'approche proposée par Berretti *et al.* (2003)

Afin d'obtenir automatiquement une représentation approchant un lieu, nous prenons en compte la présence éventuelle de relations spatiales grâce à une approche hybride (Gaio *et al.*, 2008) combinant les relations topologiques comme décrites par Egenhofer et Franzosa (1991) (par exemple, *dans, à l'intersection*, etc.), les relations directionnelles telles que formalisées par Ligozat (1998) (par exemple, *au sud de*, etc.) et les relations de distance exprimées de manière métrique (par exemple, *à 10 km de*, etc.).

2.3. *Expression de déplacement*

Selon Talmy (2000), dans les langues latines comme le français, le mouvement est caractérisé par le verbe. Dans notre corpus, d'après une première étude réalisée dans notre équipe par Loustau (2008), l'expression du déplacement est essentielle dans un récit de voyage. Plusieurs travaux linguistiques comme ceux de Boons (1987), de Laur (1991) et de Sarda (2000) ont été réalisés afin d'étudier le rôle des verbes de déplacement dans la langue. Ces auteurs ont proposé une catégorisation des verbes de déplacement par leur polarité. De manière un peu réductrice, mais suffisante, nous dirons que, dans notre approche, les polarités sont initiale (*quitter*), médiane (*visiter*), ou finale (*arriver*).

D'autre part, d'après les observations faites au cours de ce travail, les verbes de perception (par exemple, *voir*, etc.) nous sont apparus comme ayant également une importance dans certain contexte d'évocation, en particulier, lorsque le narrateur au cours des déplacements de l'acteur souhaite rendre compte de certaines situations ou sensations. De même nous avons étudié l'intérêt de traiter des verbes que nous appelons topographiques (par exemple, *entourer*, etc.).

2.4. *Outils TAL et grammaire hors contexte*

De plus en plus d'outils permettant des traitements linguistiques plus ou moins profonds deviennent exploitables, mais leurs performances s'amenuisent rapidement dès que les formulations se complexifient. Pour pouvoir utiliser notre méthode sur des corpus réels, il nous a donc semblé préférable de nous appuyer sur des outils plus robustes car se cantonnant à réaliser des analyses de faible profondeur telles que les analyses morphosyntaxiques proposées par TreeTagger (Schmidt, 1994) ou par Melt (Denis et Sagot, 2009)[17].

Les grammaires hors contexte[18] sont depuis longtemps utilisées en TAL, citons parmi les premiers travaux ceux de Hearst (1992), pour la mise au point d'analyses fondées sur des patrons. Ces grammaires se composent d'un ensemble de règles qui permettent de remplacer une séquence d'expressions (nom, adjectif, verbe, etc.) par un nouvel identifiant unique d'un niveau d'abstraction plus élevé (syntagme nominal, syntagme verbal, etc.). Dans le cas de ce travail, la grammaire hors contexte est utilisée pour marquer non seulement des informations à un niveau d'abstraction syntaxique plus élevé (groupes de noms propres, groupes de nom communs) mais également à un niveau sémantique (verbe de déplacement, nom toponymique, etc.) grâce à l'utilisation combinée de diverses ressources contenant des informations soit syntaxiques, soit sémantiques, soit les deux en même temps.

À partir de ces différents éléments bibliographiques nous pouvons proposer notre méthode qui permet, grâce à la formalisation d'un nombre maîtrisable de relations syntaxico-sémantiques entre des éléments de diverses catégories de lexiques, de réaliser un inventaire de syntagmes nominaux à connotation géographique.

3. Annotation des « toponymes » grâce aux structures VT

Une entité nommée spatiale étendue telle que nous l'entendons et que nous nommerons par convenance « toponyme », est composée *a minima* d'un nom propre at-

17. Pour la version standard de notre chaîne de traitement, nous utilisons TreeTagger. Toutefois, cet analyseur produit dans certains cas des erreurs bien connues. Une version parallèle est donc en cours d'implantation afin d'intégrer Melt, dans l'espoir de réduire certaines de ces erreurs.

18. Formellement, un langage est hors contexte si et seulement si il existe un automate à pile qui le reconnaît (Hopcroft *et al.*, 2001).

tribué à un lieu ou « nom toponymique », éventuellement associé à un ou plusieurs concepts de nature ontologique (que nous nommerons « type ») et à un ou plusieurs concepts relatifs à l'expression de la localisation dans la langue, nommés « indirections ». Les concepts de nature ontologique sont représentés la plupart du temps par des syntagmes nominaux SN (*villes, régions, fleuves,* etc.) qui ont tout à la fois des propriétés de synonymie (plusieurs SN pour un même concept) et sont porteurs d'ambiguïtés (plusieurs concepts pour un même SN). Les indirections, quant à elles, font intervenir des relations spatiales (*dans, proche de, à l'ouest de, au sommet de, entre,* etc.). Par exemple, dans le cadre de notre expérimentation, le toponyme *au sud du gave de Pau* a pour nom toponymique *Pau* dont le type, ou nature ontologique en topographie, obtenu par une relation de synonymie, est *hydronyme* et auquel est associée la relation spatiale d'orientation *sud*. Ce typage permet de lever un certain nombre d'ambiguïtés et, par exemple ici, de distinguer le toponyme *gave de Pau* du toponyme *ville de Pau*. Le raisonnement sur les relations spatiales associées aux noms toponymiques permet, quant à lui, de déterminer une représentation numérique, c'est-à-dire, par exemple, une liste de coordonnées géographiques qui sera considérée comme une approximation numérique de la zone géographique évoquée dans le texte.

3.1. *Les structures* VT

Ces structures intègrent les travaux relatifs à l'expression spatiale dans la langue (Borillo, 1998), aux relations spatiales (Egenhofer et Franzosa, 1991 ; Ligozat, 1998) et aux déplacements (Boons, 1987 ; Laur, 1991).

De manière formelle, soit V, I, T, G respectivement des ensembles de verbes, d'indirections, de termes topographiques et de noms toponymiques. $VTo = (v, t)$ avec $v \subset V$ et l'ensemble t étant défini de la manière suivante : $t = (te, i, nt|t)$ tel que $te \subset T$, $i \subset I$ et $nt \subset G$. Le symbole $nt|t$ indique que le troisième ensemble t peut être constitué d'un t (récursivité) ou bien d'un nt.

Nous présentons d'abord les deux cas les plus courants VTo et VTr, puis les patrons linguistiques permettant de repérer les instances de ces structures VT dans le texte.

3.1.1. *La structure* VTo

Cette structure illustrée dans la figure 2 est décrite par la structure VTo qui se compose d'au moins un **v**erbe particulier et d'au moins un **to**ponyme.

La structure VTo de la figure 2 est la suivante : $VTo = (v, t)$ avec : $v =$ « descendre », $t =$ « sur le territoire aride au sud de la région d'Aragon ». Nous pouvons constater que le toponyme t est défini récursivement $t_1 = (te_1, i_1, nt_1|t_2)$ tel que : $te_1 =$ « territoire aride », $i_1 =$ « sur », $t_2 = (te_2, i_2, nt_2|t_3)$ tel que : $te_2 = \emptyset$, $i_2 =$ « au sud », $t_3 = (te_3, i_3, nt_3|t_4)$ tel que : $te_3 =$ « région », $i_3 = \emptyset$, $nt_3 =$ « Aragon ».

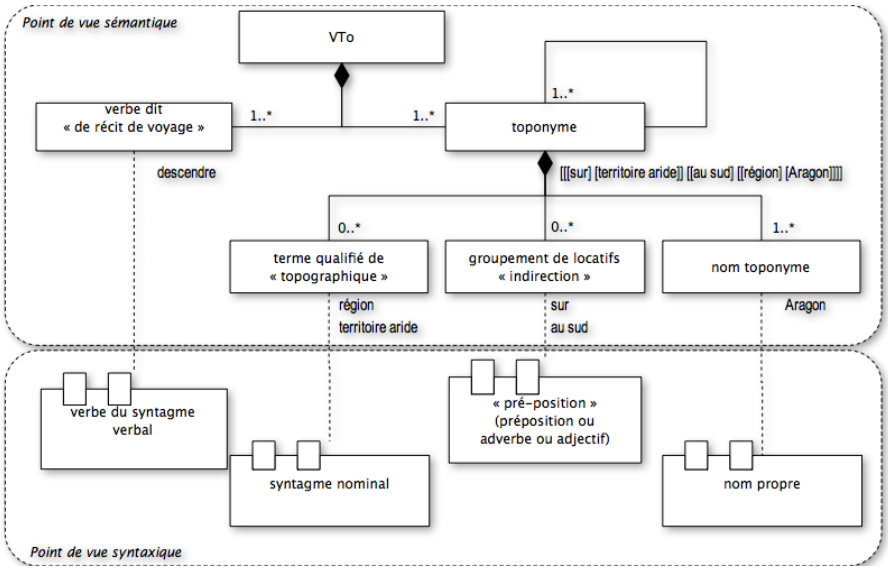

Figure 2. *Diagramme UML de la composition d'une structure VTo, illustré par un exemple*

Voici quelques exemples relatifs à la structure VTo (le 4 est illustré dans la figure 2) :

(2) *1.* quitter *Bayonne* ;

2. remonter vers *Gavarnie* ;

3. arriver au *fond de la vallée du Plan-des-Etangs* ;

4. descendre sur le *territoire aride au sud de la région d'Aragon* ;

5. contempler *la charpente altière des Monts-Maudits* ;

6. voir le *cirque mystérieux de Barrosa* ;

7. découvrir nettement *la chaîne des montagnes d'Aspe* ;

8. arroser le *Gave* ;

9. entourer les divers *versants du Néthou* ;

10. dominer l'*hospice de Bénasque* ;

11. séparer les *montagnes d'Arras* et *Sireix*.

Grammaticalement, les indirections appartiennent à une catégorie composite que nous avons nommée « pré-position » (à cause de leur position en français) appartenant soit à la catégorie des prépositions, soit à celle des adverbes, soit à celle des adjectifs. Les noms toponymiques correspondent à des noms propres et les termes dits topographiques correspondent à des syntagmes nominaux. Le verbe dit « verbe de récit de voyage » correspond au verbe du syntagme verbal.

Du point du vue sémantique, les « pré-positions » sont des locatifs, par exemple, dans la catégorie des prépositions nous aurons *sur, sous, dehors,* etc., dans celle des adverbes *à proximité, autour,* etc. et, enfin, dans celle des adjectifs *au sud, au nord,* etc. Elles permettent d'exprimer la direction d'un mouvement lors d'une composition avec des verbes de déplacement. Dans le cas de leur association à des verbes de déplacement on se retrouve dans les situations décrites par Boons (1987) selon le principe de polarité des verbes et par Laur (1991) selon la polarité aspectuelle des prépositions. Par exemple, dans la phrase *Je quitte Pau pour Paris*, la polarité du verbe *quitter* et celle de la préposition *pour* nous permettent de déduire que la source et la destination du mouvement ici sont respectivement *Pau* et *Paris*.

Le syntagme verbal sera donc, *a minima*, un verbe de déplacement, mais dans certains contextes il peut appartenir à d'autres catégories, comme, par exemple, celle dite de perception, ou à celle permettant d'avoir la capacité de décrire des propriétés topographiques. Les verbes de déplacement permettent de relater des actions de l'acteur (par exemple, les phrases (2) 1., (2) 2., (2) 3., (2) 4.) tandis que les verbes de perception servent à décrire ses observations (par exemple, les phrases (2) 5., (2) 6.). Les verbes topographiques sont utilisés dans les cas où le narrateur donne une description du lieu en s'appuyant sur les spécificités topographiques de ce dernier (par exemple, les phrases (2) 8., (2) 9., (2) 10., (2) 11.). Retenons que l'ensemble de ces verbes constituera un lexique spécialisé. Celui-ci a été construit de manière *ad hoc*, nous l'appellerons par convenance « le lexique verbal ». C'est cet ensemble de verbes que nous avons appelé « verbes de récit de voyage ».

Les cardinalités dans la figure 2 représentent le nombre potentiel d'instances modélisées dans la composition de la structure VTo. Par exemple, l'indirection, de cardinanité $[0..*]$, peut figurer, zéro fois (comme dans les exemples (2) 1., (2) 5., (2) 6., (2) 7., (2) 8., (2) 9., (2) 10., (2) 11.), une fois (comme dans (2) 2., (2) 3.), deux fois (comme dans (2) 4.), voire plus, dans la structure VTo.

Comme indiqué ci-dessus, le toponyme tel que nous l'entendons ici se construit de manière récursive à partir d'un nom toponymique mais, par extension, il peut également être obtenu par résolution d'une référence à un nom toponymique. En effet, on peut distinguer deux principaux cas de figure : le nom est explicite ou alors il est formulé de manière anaphorique (le nom toponyme peut être considéré comme implicite).
Dans le premier cas, il peut s'agir d'un nom toponymique ou d'un couple (terme, nom toponymique) dans lequel le terme précise la nature topographique du nom toponymique, par exemple *Bayonne, Gavarnie, vallée du Plan-des-Etangs, région d'Aragon*. La géoréférence peut alors être recherchée directement dans les ressources de type ga-

zetier. Il peut s'agir également d'un toponyme construit à partir d'un ou de plusieurs noms toponymiques de manière récursive englobant les indirections et des termes topographiques qui précisent, pour chaque sous-ensemble, la nature du lieu nommée, comme dans : *au sud de Pau, la ville au sud de Pau, au fond de la vallée du Plan-des-Etangs, le territoire aride au sud de la région d'Aragon*. Considérons le toponyme *la ville au sud de Pau*, le nom toponymique *Pau* est le repère spatial qui va servir de référence pour déterminer l'entité concrète évoquée. Cette détermination pouvant être opérée en grande partie par l'utilisation d'opérateurs spatiaux comme ceux implantés dans la composante logicielle d'un SIG[19]. Ici, il pourra être déduit que le lieu en question peut être l'une des villes se trouvant *au sud de Pau* comme par exemple *Gelos*.

Le rôle du nom toponymique dans la structure VTo est de permettre de déterminer la géoréférence noyau du toponyme. Grâce à cette géoréférence, on pourra classifier un nom toponymique en fonction de sa forme géométrique (ligne, point, surface). Un des rôles de l'indirection est de décrire les relations spatiales associées aux noms toponymiques.

Dans le second cas et afin de traiter l'ensemble des situations correspondant à des formulations anaphoriques, nous proposons la structure VTr ci-après qui se compose d'au moins un syntagme nominal faisant référence à un nom toponymique.

3.1.2. *La structure VTr*

Rappelons que l'un des objectifs est de repérer les syntagmes nominaux pouvant porter un concept géographique. Afin d'atteindre cet objectif et afin d'éviter toute résolution anaphorique souvent coûteuse et parfois difficile, nous avons conçu la structure VTr.

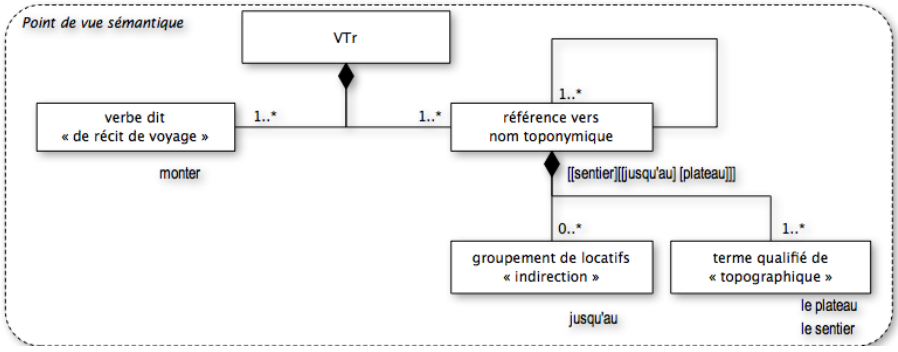

Figure 3. *Diagramme UML pour la structure VTr*

19. Système d'information géographique

Cette structure (figure 3) se compose d'un verbe de récit de voyage, éventuellement d'une ou plusieurs prépositions, et d'au moins un syntagme nominal. Voici quelques exemples : *gagne la montagne* ou *montons sur le plateau, se précipitent les gaves, voit deux ou trois pics,* etc. La structure VTr est donc construite simplement à partir de la structure VTo dans laquelle le nom toponymique est remplacé par un syntagme nominal dénotant un concept topographique.

Comme l'illustrent les phrases suivantes nous nous trouvons en présence de diverses situations anaphoriques :

(3) *1.* Nous quittons Gèdre à 6 heures et montons **le sentier** jusqu'au **plateau**.

2. Le chasseur descend au fond du **vallon** pour remonter sur une **saillie** opposée.

3. Nous descendîmes dans une **vallée** longitunale où le **lac** se décharge.

4. La route passe un **petit pont** qui domine un autre **pont grisâtre**.

5. Ils montent des **escaliers** et descendent des **échelles**.

6. Le choc tomba sur la **barrière noire crénelée** qu'on aperçoit vers Gavarnie.

7. Les filets d'eau tombent par douze **ruisseaux** qui glissent de la dernière **assise**.

8. Nous tournons à un second **pont**, et nous entrons dans la campagne de Gèdres.

Les termes ci-dessus ne sont pas associés directement aux noms de lieux mais pourront l'être après résolution anaphorique. Une indication pour que les termes puissent être repérés est que, dans certains cas, il existe des relations sémantiques entre des termes dans des contextes précis tels que :

– la description du déplacement de l'acteur : les termes sont associés à la description d'un mouvement de l'acteur d'un lieu vers un autre (par exemple, la phrase (3) 1.). Les termes peuvent également être associés à des actions consécutives de l'acteur (par exemple, les phrases (3) 2., (3) 5. et (3) 8.) ;

– la description des caractéristiques topographiques des lieux : dans un récit de voyage, le narrateur décrit très souvent les caractéristiques topographiques du lieu traversé ou décrit (par exemple la phrase (3) 3.). Dans certains cas, les lieux sont mis en relations les uns avec les autres sans évoquer les actions de l'acteur (les phrases (3) 4., (3) 7.).

Symbole	Signification	Exemples
V	verbe dont le lemme appartient à notre lexique verbal	[remonte] [remonté] [remonter]
TE	terme candidat	[vallon] [chemin de fer] [tableau]
NC	nom commun	[ville] [vallée] [pic]
A	adjectif	[calcaire] [rocheuse]
$(X\|Y)$	soit X, soit Y (disjonction)	(de | d') : soit on met « de », soit on met « d' »
NT	nom toponymique candidat	[Toulouse] [Mont-de-Marsan] [Nicolas]
NP	nom propre	[Pau] [Nicolas]
TO	toponyme	[ville de Pau] [région de l'Aragon]
I	indirection	[sud] [nord] [fond] [à] [au] [dans]
VT	la structure VT candidate	[visiter la ville de Pau] [visiter la ville]
#	de 1 à n mots quelconques	[à pied] [en conséquence]
mVT	relation entre plusieurs structures VT	[quitter Gèdre] à 6 heures et [atteindre le plateau]
$teVT$	relation entre un terme et une structure VT	Dans cette montagne on trouve un [chemin qui conduire en Espagne]
$\{X\}n$	X n fois	$\{VT\}n$: la structure VT apparaît n fois dans le texte

Tableau 1. *Les symboles utilisés dans les patrons linguistiques*

3.1.3. *Patrons linguistiques*

Les instances des structures VT dans le texte sont très variées. En effet, nous définissons 16 patrons différents dans les tableaux suivants, pour un total de 240 cas de structures VT. Le tableau 1 présente la signification de la symbolique utilisée pour la définition de ces patrons.

Comme montré dans le tableau 2, les termes candidats doivent être des groupes de noms communs définis récursivement. En effet, un terme peut être un mot simple comme *vallée* (patron TE_1) ou composé de plusieurs mots comme par exemple *chemin de fer* (patrons TE_2, TE_3, TE_4). Dans certains cas, il peut comporter des adjectifs comme par exemple *massif calcaire* (patrons TE_2, TE_3).

Comme indiqué dans le tableau 3, les noms toponymiques peuvent être formés par un mot simple (*Pau*) (patron NT_1) ou par un groupe de noms propres (*Marie Blanque*) (patrons NT_2, NT_3). Dans plusieurs cas, des noms de lieux français comportent des prépositions (*de, sur*), des articles (*la, l'*) ou le caractère tiret (-) pour connecter les noms propres (*Chambre d'Amour, Pic-d'Ossau, Mont-de-Marsan*) (patron NT_3).

Patron	Définition	Exemples
TE_1	$TE = NC$	[vallée] [ville]
TE_2	$TE = TE\ A$	[massif calcaire] [crête rocheuse]
TE_3	$TE = A\ TE$	[hautes montagnes] [hautes cimes neigeuses
TE_4	$TE = TE$ (de | d') TE	[chemin de fer] [infini panorama des cimes]

Tableau 2. *Patrons pour les termes candidats*

Patron	Définition	Exemples
NT_1	$NT = NP$	[Pau] [Bordeaux]
NT_2	$NT = NT\ NT$	[Marie Blanque] [Saint Sever-de-Rustan]
NT_3	$NT = NT$ (de | de la | du | des | d'| de l' | - | de- | des- | sur- | sur-l' | la- | les- | à-) NT	[Mont-de-Marsan] [Pic du Midi] [Côte de la Fontaine] [Chambre d'Amour] [Puigmal-de-Cerdagne] [Plan-des-Etangs] [Pic-d'Ossau]

Tableau 3. *Patrons pour les noms toponymiques*

Patron	Définition	Exemples
TO_1	$TO = NT$	[Pau] [Bordeaux]
TO_2	$TO = TE$ (de | d'| du| de l'| des | de la) TO	[granges de Portalieu] [vallon du Neiss] [port d'Azun]
TO_3	$TO = I$ (de | d'| du| de l'| des | de la) TO	[fond de la vallée du Plan-des-Etangs] [à côté de Pau]
TO_4	$TO = TE\ I$ (de | d'| du| de l'| des | de la) TO	[vallée au sud de Pau] [territoire aride au sud de la région de l'Aragon]
TO_5	$TO = TE\ \{TO\}n$	[route Pau Bordeaux] [chemin Lescar Oloron-Sainte-Marie Saint-Jean-Pied-de-Port]

Tableau 4. *Patrons pour les toponymes*

Comme nous pouvons le constater dans le tableau 4, ces patrons sont construits à partir des patrons de termes candidats TE et des patrons de noms toponymiques NT. Nous définissons ici les toponymes comme étant un nom toponymique seul (*Pau*) (patron TO_1), ou des termes liés à des noms toponymiques avec ou sans des indirections I (*vallon du Neiss, vallée au sud de Pau*) (patrons TO_2, TO_3, TO_4). Le ta-

bleau 4 montre que les termes peuvent être attachés à un nom toponymique (patrons TO_2, TO_3), à plusieurs (TO_5) ou non (TO_4).

La position des indirections dans les toponymes est très variée. Les indirections peuvent être attachées directement (*à côté de Pau*) ou indirectement au nom toponymique. Dans certains cas, elles peuvent également se trouver entre les termes et les noms toponymiques (*vallée au sud de Pau*), ou entre deux structures qui comportent des termes (*territoire aride au sud de la région de l'Aragon*). Nous ne faisons volontairement référence à aucune relation grammaticale car dans la méthode proposée ici nous ne les utilisons pas.

Patron	Définition	Exemples
VT_1	$VT = V\ (TO\|TE\|I\ \ TE)$	[quitter Gèdre] [visiter le vallon au sud de Pau] [contempler la charpente altière des Monts-Maudits] [dominer ce village] [gagner la montagne]
VT_2	$VT = V\ \#\ (TO\|TE\|I\ \ TE)$	[remonter à pied la vallée d'Ossau] [aller en conséquence à Bagnères de Luchon] [monter à cheval jusqu'au cirque de Gavarnie]

Tableau 5. *Patrons pour les structures VT*

Les patrons de VT, dans le tableau 5, sont définis à partir des patrons de toponymes (pour les VTo) ou ceux de termes candidats (pour les VTr). Comme montré par le tableau 5, les indirections peuvent être présentes dans des structures VT ou absentes. Dans certains cas, les expressions qui désignent des moyens de transport sont employées après les verbes, par exemple *J'ai remonté à pied la vallée d'Ossau*. Le patron VT_2 traite ces cas.

3.2. *Relations entre structures ou entre termes et structures*

En ce qui concerne les relations entre structures : soit $vt_1, vt_2, ...vt_m$ des structures VT au sein de la même phrase, la relation entre celles-ci est définie par $mVT = (vt_1, vt_2, ...vt_m)$.

Comme indiqué dans le tableau 6, les patrons des relations mVT sont définis à partir des patrons pour les structures VT. Les quatre premiers patrons ($mVT_1, mVT_2, mVT_3, mVT_4$) ont pour but de repérer des structures VT qui partagent le même verbe, par exemple *Nous avons remonté la vallée d'Ossau jusqu'à Laruns* (patron mVT_1), tandis que le dernier patron (mVT_5) permet de déterminer les structures VT dont les verbes sont différents, par exemple les VT *quittons Gèdre* et *atteignons le plateau* dans la phrase *Nous quittons Gèdre à 6 heures et atteignons*

le plateau à 7 h 30. Le caractère (#) dans le patron mVT_5 signifie qu'il existe des séquences de mots quelconques entres les structures VT.

Patron	Définition	Exemples
mVT_1	$mVT =$ $VT_1\ (TO\|TE\|I\ \ TE)$	[remonter la vallée d'Aspe (remonter) jusqu'à Lescun]
mVT_2	$mVT = VT_2$ $(TO\|TE\|I\ \ TE)$	[emprunter à pied la vallée d'Ossau (remonter) jusqu'à Laruns] [passer à cheval sur le versant de Cauterets (passer) par la brèche de Courouaou de Bouc]
mVT_3	$mVT = VT_2$ (et) TO	[visiter avec intérêt le château de Pau et (visiter) le cirque de Gavarnie]
mVT_4	$mVT = VT_1\ \ TO$ $\{,TO\}n$ (et) TO	[passer par le Tourmalet, (passer par) le village d'Hourquette d'Arreau et (passer par) la belle vallée de Louron]
mVT_5	$mVT = VT$ # $\{VT\}n$ # (. | ? | !)	[quitter Gèdre à 6 heures et atteindre le plateau à 7 h 30 !] [descendre au fond du vallon pour remonter sur une saillie opposée.]

Tableau 6. *Patrons pour les relations entre plusieurs structures VT : mVT*

En ce qui concerne les relations entre termes et structures : soit un terme et une structure VT, la relation entre eux est définie par $teVT = (TO, VT)$. Les patrons linguistiques relatifs à cette relation sont présentés dans le tableau 7.

Patron	Définition	Exemples
$teVT_1$	$teVT =$ TO # $(VT\|mVT)$	[Ce sentier conduit au village de Ger] [un pont de bois traverse la Garonne au bas de Montréjeau]
$teVT_2$	$teVT = TO$ # (qui | par lequel | où | que | sous lequel) # $(VT\|mVT)$	[c'est un pont de glace sous lequel s'échappe le gave] [mais je ne vois pas de route encore qui puisse nous conduire sur la montagne.]

Tableau 7. *Patrons pour les relations entre terme et structure VT : $teVT$*

3.3. *Opérationnalisation de l'inventaire lexical à partir des structures* VT

3.3.1. *Principe de raisonnement*

Le principe général de raisonnement est le suivant : la présence d'une relation géographique, appartenant à l'ensemble des relations géographiques RG, entre syntagmes au sein d'une même phrase qualifie, dans ces syntagmes, les groupes de noms

communs de termes géographiques et les groupes de noms propres de noms toponymiques.

Soit le lexique des verbes de récits de voyage V = {descendre, ...}, l'ensemble des indirections I = {sur, ...}, l'ensemble des syntagmes nominaux SN du corpus, les ensembles de termes et de noms toponymiques candidats, respectivement $TE \subset SN$ et $TO \subset SN$, le gazetier G = {Pau, Aragon, Bordeaux, ...} et l'ensemble des concepts de l'ontologie O.
Soit, d'autre part, la relation géographique candidate :
$R = (\{v_1, \ldots, v_m\}, \{te_1, \ldots, te_p\}, \{i, \ldots, i_n\}, \{to_1, \ldots, to_q\})$, où $v \subset V$, $i \subset I$, te et $to \subset SN$. Les v_i, te_i, i_i et to_i sont des instances reconnues par un patron de relations de l'un des trois types VT, mVT et $teVT$.
Soit les quatre règles de raisonnement suivantes :

règle 1 : Si $\exists to_i \in G$, alors $R \in RG$
une relation candidate est une relation géographique si au moins un des noms toponymiques de la relation est défini dans un gazetier ;

règle 2 : Si $\exists te_i \in O$, alors $R \in RG$
une relation candidate est une relation géographique si au moins un des termes de la relation est l'un des labels d'un concept dans une ontologie ;

règle 3 : Si $R \in RG$, alors $\forall te_i,\ te_i$ proposé pour $\in O$
tous les groupes de noms communs de la relation sont des termes candidats a être label d'un concept topographique dans l'ontologie O si la relation est géographique ;

règle 4 : Si $R \in RG$, alors $\forall to_i,\ to_i \in G$
tous les groupes de noms propres de la relation sont considérés comme des noms toponymiques si la relation est géographique.

Enfin, la combinaison des règles 1 ou 2 avec la règle 3 formalise l'opérationnalisation de notre méthode de repérage des SN pour établir l'inventaire lexical. De plus, la combinaison des règles 1 ou 2 avec la règle 4 permet d'opérationnaliser une méthode de reconnaissance de noms toponymiques.

3.3.2. *Exemples de repérage des SN*

Par la structure VT. Reprenons l'exemple *Nous descendons sur le territoire aride au sud de la région d'Aragon*. A l'aide des patrons linguistiques, la structure VT candidate R = (descendons, sur, territoire aride, région, sud, Aragon) est repérée. Dans ce cas v ={descendons}, i = {sur, sud}, te = {territoire aride, région}, to = {Aragon}. On a $to_1 \in G$, la règle 1 est satisfaite, R est considérée comme une relation géographique. Selon la règle 3, les éléments de te sont repérés en tant que termes géographiques.

Par les relations mVT. Voyons maintenant comment les règles de raisonnement sont appliquées sur des relations mVT.

> Exemple 1 : dans la phrase : *Nous quittons Gèdre à 6 heures et atteignons le plateau à 7 h 30*, la relation candidate repérée par des patrons linguistiques correspond à $mVT = (vt_1, vt_2) =$ (quittons Gèdre, atteignons le plateau) avec $v =$ {quittons, atteignons}, $i =$ { }, $te =$ {plateau}, $to =$ {Gèdre}. On note que la règle 1 est satisfaite car ($to_1 \in G$). Par conséquent, selon la règle 3, te_1 doit être un terme géographique. Dans ce cas, le modèle proposé permet de repérer un terme qui n'est directement attaché à aucun nom toponymique.

> Exemple 2 : dans la phrase : *Le chasseur descend au fond du vallon pour remonter sur une saillie opposée*, la relation candidate repérée par des patrons linguistiques correspond à $mVT = (vt_1, vt_2) =$ (descend au fond du vallon, remonter sur une saillie opposée) avec $v =$ {descend, remonter}, $i =$ {au, sur, fond}, $te =$ {vallon, saillie opposée}, $to =$ { }. On note que la règle 2 est satisfaite ($te_1 \in O$). En conséquence, selon la règle 3, te_2 sera repéré comme terme géographique. Il s'agit ici d'un cas dans lequel l'ontologie est utilisée pour détecter de nouveaux termes géographiques.

Par les relations $teVT$. Dans la phrase : *C'est un pont de glace sous lequel s'échappe la rivière*, la relation géographique repérée est la relation $teVT$ (pont de glace, s'échappe la rivière). Le syntagme nominal *rivière* est un terme présent dans l'ontologie, par conséquent, selon les règles 2 et 3, le syntagme *pont de glace* est un terme géographique à repérer. Il s'agit ici aussi d'un cas dans lequel l'ontologie est utilisée pour détecter de nouveaux termes géographiques mais cet exemple permet également d'illustrer la situation dans laquelle le terme à extraire n'est associé à aucun nom toponymique.

4. Enrichissement de l'ontologie topographique

Une fois ces traitements effectués, nous disposons d'un inventaire lexical issu d'un corpus. Le corpus « récits de voyage », retenu pour notre expérimentation, a la particularité de refléter un vocabulaire plus grand public que celui des spécifications techniques de bases de données topographiques qui ont guidé la construction de l'ontologie noyau. Afin de combiner cette nouvelle ressource avec la première ontologie, il est nécessaire d'identifier les concepts communs, ce qui est le rôle des techniques d'alignement d'ontologies, pour ensuite les fusionner en une nouvelle ontologie topographique étendue.

L'enrichissement de l'ontologie topographique initiale se déroule en plusieurs étapes. Une première étape a pour but d'éliminer du lexique tous les termes déjà pré-

sents dans l'ontologie. La deuxième étape, quant à elle, s'appuie sur une ressource externe pour un nouveau filtrage dont l'objectif est de ne retenir que des termes *a priori* topographiques. Cette ressource permet également la construction d'arborescences de termes associés à chaque élément du lexique. Enfin, la troisième et dernière étape consiste à proposer l'enrichissement de l'ontologie topographique initiale par l'ensemble des micro-arborescences ainsi construites[20].

4.1. *Filtrage des termes du lexique déjà présents dans l'ontologie : calcul de similarité entre termes*

Nous proposons une méthode qui permet de vérifier si un terme (mot ou groupe de mots) extrait du lexique existe déjà dans l'ontologie cible. À cet effet, chaque terme est comparé avec les labels des concepts définis dans l'ontologie. Le problème clé ici est donc celui de la comparaison de deux chaînes de caractères constituées par la paire [terme, label]. La caractéristique principale de ces chaînes de caractères est qu'elles sont souvent composées de plusieurs mots. Considérons quelques exemples de paires [concept, terme] : (*chemin de fer touristique, voie ferrée touristique, centre de formation professionnelle des adultes, centre de formation des adultes*), (*haras national, nation*), (*bureau de poste, poste de radio*), etc.

Les métriques de similarité entre termes peuvent être classées en trois catégories principales : fondées sur les caractères, fondées sur les tokens et hybrides.

Les métriques fondées sur des caractères considèrent les chaînes comme une séquence de caractères. En conséquence, la similarité entre deux chaînes est déterminée par des caractères communs et la position de ces caractères dans les chaînes (Jaro (1989), Jaro-Winkler (1999), I_Sub (Stoilos *et al.*, 2005)) ou par le nombre d'opérations (suppression, insertion, remplacement) nécessaires pour construire une chaîne à partir de l'autre (Levenshtein (1966), Needleman et Wunsch (1970), Smith et Waterman (1981)). L'inconvénient principal des métriques de ce type est de ne pas distinguer les mots lorsqu'une chaîne en comporte plusieurs.

Les méthodes fondées sur des tokens, comme TFIDF (Cohen *et al.*, 2003 ; Jaccard, 1901), considèrent une chaîne comme un ensemble de tokens. Un token est une sous-chaîne de caractères délimitée par des caractères spécifiques tels que des espaces ou des tirets, par exemple. Ces méthodes s'appuient sur le nombre de tokens communs dans les chaînes à comparer mais ne mettent pas en œuvre de métriques fondées sur des caractères pour déterminer des tokens similaires. Par exemple, pour les chaînes *chemin de fer touristique* et *voie ferrée touristique*, les tokens *fer* et *ferrée* sont considérés comme différents.

20. Le lecteur pourra accéder à plus de détails concernant les deux dernières étapes dans (Kergosien *et al.*, 2009 ; Mustière *et al.*, 2011). Elles sont toutes trois décrites succinctement ci-après.

Les méthodes hybrides (SoftTFIDF (Cohen *et al.*, 2003) , Monge-Elkan (Monge et Elkan, 1996), TagLink (Camacho et Salhi, 2006)) tentent de combiner ces deux types d'approches. Elles utilisent une métrique fondée sur les caractères pour évaluer le degré de similarité de paires de tokens.

La caractéristique commune aux méthodes hybrides et aux méthodes fondées sur les tokens est qu'elles ne prennent pas en compte l'ordre des tokens dans les chaînes (par exemple, le score de similarité calculé par Jaccard, TFIDF, SoftTFIDF, Monge-Elkan ou TagLink pour les chaînes *piste de ski* et *ski de piste* est égal à 1).

C'est la raison pour laquelle nous avons mis au point une nouvelle méthode hybride. Nous considérons les chaînes à comparer comme des séquences de tokens et proposons de traiter les tokens de la même manière que les méthodes classiques comparent les caractères.

Soit S l'ensemble des chaînes de caractères. Notre métrique est définie comme une fonction $\mu : S \times S \rightarrow \mathbf{R}$ telle que :

$$\begin{aligned} 0 \leq \mu(S_1, S_2) \leq 1, \forall S_1, S_2 \in S \\ \mu(S_1, S_1) = 1, \forall S_1 \in S \end{aligned} \quad [1]$$

La valeur de la fonction $\mu(S_1, S_2)$ dépend non seulement des tokens communs (ou presque similaires) à deux chaînes mais encore d'autres caractéristiques de leurs tokens (par exemple, la position (l'ordre) des tokens dans les chaînes de caractères ; le nombre d'opérations de suppression, d'insertion ou de remplacement de tokens nécessaire à la construction d'une des deux chaînes à partir de l'autre, etc.). Ainsi, la valeur de la fonction $\mu(S_1, S_2)$ est calculée en deux étapes :

1) transformation des tokens en symboles : une métrique (μ_1) opérant sur les caractères associe le même symbole[21] à deux tokens similaires : soit $\alpha\beta\gamma$ pour *piste de ski* et $\gamma\beta\alpha$ pour *ski de piste* ;

2) mesure de la similarité des chaînes de symboles : une métrique (μ_2) opérant également sur les caractères calcule la similarité entre ces séquences de symboles.

En fait, notre méthode utilise deux métriques de base qui sont paramétrables : μ_1 pour comparer une paire de tokens, μ_2 pour comparer deux séquences de symboles. μ_1, μ_2 peuvent être une même métrique ou bien deux métriques distinctes. Ainsi, chaque combinaison de métriques opérant sur des caractères produit une nouvelle métrique hybride.

Notre proposition correspond à une métaméthode permettant de générer autant de méthodes distinctes que de combinaisons possibles. Après une première expérimentation, nous avons retenu JaroWinkler à la fois pour μ_1 et μ_2 avec $\varepsilon = 0,84$ comme

21. Un symbole est représenté par un caractère unique, ce qui permet ensuite de mettre en œuvre des formules de calcul de similarité entre chaînes de caractères pour déterminer la similarité entre chaînes de symboles.

seuil de similarité entre deux tokens pour μ_1. Nous avons ensuite appliqué cette métrique à l'ensemble du lexique topographique en vue de l'élimination des termes déjà présents dans l'ontologie cible.

4.2. *Préparation des termes du lexique candidats à l'enrichissement à l'aide de ressources externes*

Parmi les termes conservés dans le lexique, certains ont un caractère topographique et pourraient enrichir l'ontologie (comme *gave* dans textitle gave de Pau), contrairement à d'autres (comme *maire* dans *le maire de Pau*) qu'il s'agit d'écarter.

Le thésaurus RAMEAU[22] (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) est apparu comme un choix pertinent de ressource externe pour l'analyse des termes candidats. Rappelons que RAMEAU est « le langage d'indexation matière » utilisé, en France, par la BNF[23], les bibliothèques universitaires, de nombreuses bibliothèques de lecture publique ou de recherche ainsi que plusieurs organismes privés. Il se compose d'un vocabulaire de termes sémantiquement reliés entre eux et d'une syntaxe indiquant les règles d'utilisation des « vedettes-matière » (concepts) pour indexer un fonds documentaire.

Chacun des termes du lexique est donc confronté à la ressource RAMEAU. La composante de RAMEAU relative à une entrée du lexique est extraite puis traduite en OWL. Notre première tâche est ainsi d'évaluer si les concepts intervenant dans une composante particulière relèvent ou non du domaine de la topographie. Notre deuxième tâche consiste ensuite à identifier dans la composante, les termes pertinents pour l'enrichissement.

4.2.1. *Exemple de composante dans le thésaurus RAMEAU*

Dans RAMEAU, les éléments terminologiques s'organisent autour d'un terme dit « vedette ». Une telle composante regroupe des termes génériques et spécifiques relatifs à cette « vedette-matière », ainsi qu'un ensemble de termes étiquetés « employé pour », c'est-à-dire des termes qui doivent être remplacés par la vedette-matière, quand on fait de l'indexation.

Si on prend l'exemple du terme *abîme* dans le lexique, ce dernier n'appartient pas à l'ontologie topographique noyau mais figure bien dans le thésaurus RAMEAU. Il a pour vedette-matière *grottes* et, d'autre part, *grottes* aura comme employé-pour *abîmes, antres, avens, cavernes, gouffres,* etc..

4.2.2. *Extraction des micro-arborescences à partir du thésaurus RAMEAU*

Des mots ou groupes de mots du lexique sont confrontés au thésaurus RAMEAU. Les composantes RAMEAU obtenues deviennent des entrées « candidates » à l'enri-

22. `http://rameau.bnf.fr/informations/rameauenbref.htm`
23. Bibliothèque nationale de France

chissement de l'ontologie. Ces entrées sont représentées dans un format OWL dans lequel les termes « génériques » et « spécifiques » de la vedette sont respectivement représentés comme généralisant et spécialisant et les termes employé-pour sont représentés comme des spécialisations de la vedette.

Tous les termes de la composante sont considérés comme, *a priori*, potentiellement intéressants pour l'enrichissement. Ainsi, la partie extraite de RAMEAU correspond à une micro-arborescence devant être alignée avec l'ontologie topographique noyau.

4.3. *Enrichissement de l'ontologie topographique noyau*

La proposition consiste à vérifier que la « vedette-matière » de la composante RAMEAU extraite à partir d'un terme du lexique appartient à l'ontologie cible et que ce même terme figure dans le champ « employé-pour » de la « vedette-matière ». Auquel cas, ce terme peut être proposé comme un nouveau concept spécialisant le concept de l'ontologie correspondant à la « vedette-matière ». Dans la figure 4, *grotte* appartient à la fois à une composante RAMEAU et à l'ontologie cible, par conséquent, *abîme*, élément du lexique géographique, peut être candidat à l'enrichissement dans l'ontologie. Il s'agit toutefois d'écarter le risque éventuel de fausse homonymie entre un concept de RAMEAU et un concept de l'ontologie cible qui auraient le même label sans avoir le même sens.

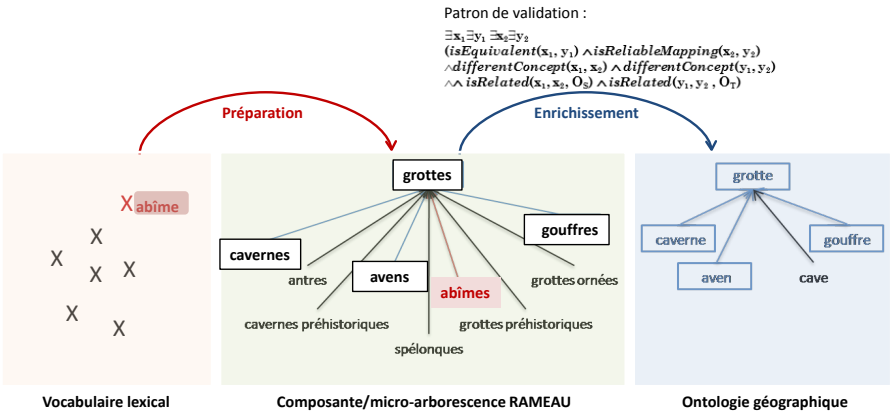

Figure 4. *Exemple de composante RAMEAU validée pour l'enrichissement*

Pour cela, nous nous appuyons sur les travaux de Safar et Reynaud (2009) et de Hamdi *et al.* (2010) qui définissent la notion d'alignement d'ontologies comme la recherche de *mappings*, l'appariement, ou encore, la mise en correspondance d'arborescences de concepts. Le test consiste à vérifier l'existence d'au moins deux correspondances entre deux ontologies : une première correspondance ou *mapping* d'équivalence entre deux concepts et une autre correspondance ou *mapping* jugé fiable entre

deux autres concepts. Une correspondance d'équivalence entre concepts (C_{Si} d'une ontologie source est équivalent à C_{Cj} d'une ontologie cible) est proposée lorsque la similarité d'un des labels de C_{Si} avec un des labels de C_{Cj} est supérieure ou égale à un certain seuil. Il s'agit ensuite de déterminer une autre correspondance qualifiée de fiable entre un autre concept C_{Sk} avec un concept C_{Cl} distinct de C_{Cj} mais qui lui soit relié. On qualifie de « reliés » les concepts d'une même ontologie qui sont reliés par une relation « père_fils » ou « frère_de ».

La procédure de mise en correspondance que nous appliquons en vue de l'enrichissement de l'ontologie cible a été proposée par Mustière *et al.* (2011). Il s'agit tout simplement de construire une première arborescence de termes à partir d'un élément de notre lexique géographique et du thésaurus RAMEAU (phase de préparation, figure 4). Cette micro-arborescence est ensuite mise en correspondance avec l'ontologie cible en vue de son enrichissement (phase d'enrichissement, figure 4). Le test impose que les concepts de la cible soient reliés.

Ainsi, nous distinguons trois principaux cas de figure :

– les différentes conditions explicitées plus haut sont réunies : le domaine de la source est qualifié de valide ;

– moins de deux correspondances sont détectées : la proposition d'enrichissement est rejetée ;

– plusieurs correspondances d'équivalence sont détectées, mais les différents concepts correspondants de l'ontologie cible ne sont pas reliés : le domaine de la source doit être validé par un expert.

La figure 4 illustre, dans les encadrés, des concepts identifiés comme équivalents lors de l'appariement. Il existe plus de deux correspondances fiables telles que les concepts considérés dans les appariements soient en relation deux à deux (ici, quatre appariements d'équivalence relatifs à quatre concepts tous reliés entre eux). Le domaine de la composante extraite de RAMEAU a donc été jugé compatible avec celui de l'ontologie à enrichir. La mise en correspondance produit des résultats qui vont ensuite faire l'objet de validations et de traitements complémentaires spécifiques, parfois en interaction avec l'expert. Ces traitements ont été implémentés dans l'environnement TaxoMap Framework (Hamdi *et al.*, 2010) à l'aide de patrons (Hamdi *et al.*, 2011) dans le cadre du projet GéOnto.

5. Expérimentations

Nous présentons une première expérimentation visant l'évaluation de la contribution de nos différentes propositions (lexiques verbaux, grammaires, autres ressources) à l'extraction d'un ensemble de termes géographiques contenus dans un échantillon de documents de type récit de voyage. Comme nous l'avons montré, les termes extraits contribuent à l'enrichissement de l'ontologie géographique du projet GéOnto. Enfin,

une seconde expérimentation a pour but de mesurer l'apport d'une telle ontologie topographique dans une chaîne d'indexation spatiale de documents textuels.

5.1. *Expérimentation de la méthode d'extraction des termes topographiques*

5.1.1. *Évaluation quantitative*

Nous avons mené des expérimentations sur douze livres (récits de voyage) fournis par la médiathèque de Pau (MIDR), ce qui fait un total de 2 400 pages. En prenant cet échantillon de documents, nous évaluons notre méthode selon trois scénarios présentés dans le tableau 8 et la figure 5. Ainsi, nous évaluons l'apport de chaque type de verbe de récit de voyage (scénario 1), des différents patrons linguistiques (scénario 2) et d'autres ressources (scénario 3) pour le raisonnement. Les résultats présentent trois principaux indicateurs : le nombre de termes extraits, le nombre de termes validés par des experts et la précision. Notons que chaque colonne du tableau 8 est indépendante et qu'il n'y a pas de cumul de valeurs. Les nouveaux termes sont déduits des termes valides distincts et représentent les termes, sans correspondance dans l'ontologie, susceptibles d'être candidats pour l'enrichissement.

		Scénario 1			Scénario 2			Scénario 3		Tous
		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	
Verbe	Déplacement	X			X	X	X	X	X	X
	Perception		X		X	X	X	X	X	X
	Topographique			X	X	X	X	X	X	X
Relations	VT	X	X	X	X			X	X	X
	mVT	X	X	X		X		X	X	X
	$teVT$	X	X	X			X	X	X	X
Ressources	Gazetiers	X	X	X	X	X	X	X		X
	Ontologies	X	X	X	X	X	X		X	X
Évaluation	Nb de termes	2 523	452	1 219	2 089	2 215	2 381	2 430	4 082	5 600
	Nb. de termes valides	1 779	281	863	1 512	1 545	1 598	1 609	2 932	3 841
	Nb. de termes distincts	1 112	299	639	919	1 023	1 138	1 186	1 567	**2 173**
	Nb. de termes valides distincts	630	154	400	543	588	620	617	917	**1191**
	Nb. de nouveaux termes	503	95	300	427	464	480	514	752	**1016**
	Précision	0,57	0,52	0,63	0,59	0,57	0,54	0,52	0,59	**0,55**

Tableau 8. *Résultats d'expérimentation*

Le premier scénario a pour but d'évaluer le lexique verbal associé aux relations VT, mVT et $teVT$. Initialement, notre lexique verbal (cas 1.1 du tableau 8) comporte 75 verbes de déplacement. Comme montré par le tableau, notre chaîne de traitement a extrait 1 112 termes distincts dont 630 validés par des experts, soit une précision de $0,57$. Afin d'étendre la capacité de repérage de ce lexique verbal nous avons également proposé un lexique de 29 verbes de perception (cas 1.2 du tableau 8) et un autre de 59 verbes topographiques (cas 1.3 du tableau 8). Pour chaque cas du scénario 1, nous prenons les trois relations VT, mVT, $teVT$ et combinons les deux types de ressources, gazetier et ontologie. Le meilleur paramétrage est le 1.1 qui retourne

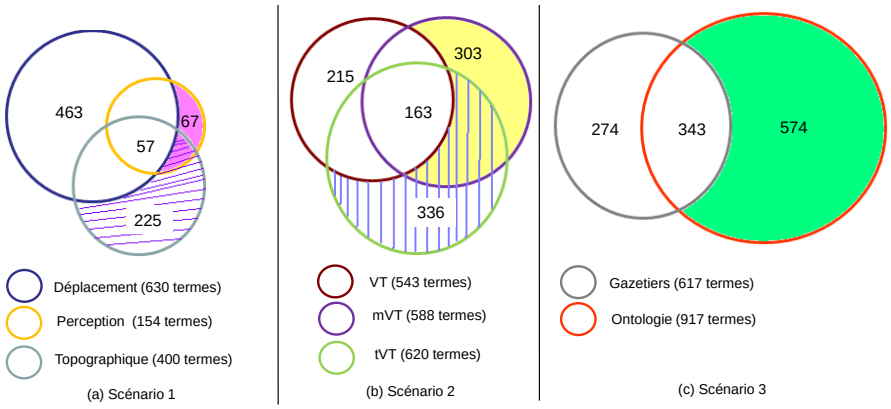

Figure 5. *Complémentarité des scénarios d'expérimentation*

le maximum de termes valides et une précision de $0,57$. La figure 5(a) nous permet d'observer la complémentarité de ces scénarios et montre notamment que le scénario 1.2 (verbe de perception) permet le marquage de 67 nouveaux termes par rapport au scénario 1.1. De même le scénario 1.3 (verbe topographique) permet le marquage de 225 nouveaux termes par rapport au scénario 1.1.

Le deuxième scénario a pour but d'évaluer l'apport des relations par rapport aux patrons VT, mVT et $teVT$. Comme indiqué dans le tableau 8, chaque paramétrage dans ce scénario prend toutes les catégories de verbes et toutes les ressources disponibles. Le meilleur paramétrage en fonction de la précision est le 2.1 (relation VT) mais le meilleur paramétrage en ce qui concerne le nombre de termes distincts valides et le nombre de nouveaux termes est le 2.3 (relation $teVT$). Selon l'expérimentation, le nombre de termes en commun pour les trois cas du scénario 2 est de 163 termes distincts valides. Ainsi, la figure 5(b) montre que le scénario 2.2 permet le marquage de 303 nouveaux termes par rapport au scénario 2.1. Le scénario 2.3, quant à lui, permet le marquage de 336 nouveaux termes par rapport au scénario 2.1.

Dans le dernier scénario, nous évaluons notre méthode en fonction de ressources externes utilisées pour le raisonnement. En effet, notre méthode permet d'utiliser l'ontologie seule ou le gazetier seul pour valider les relations géographiques candidates et extraire des termes. L'expérimentation a montré que l'utilisation des ontologies pour le repérage des termes donne un meilleur résultat que l'utilisation des gazetiers : le nombre de termes extraits et la qualité des termes extraits avec les ontologies sont considérablement plus élevés que ceux obtenus avec les gazetiers. En effet, la figure 5(c) montre que ces scénarios sont complémentaires : le scénario 3.2 (ontologie) permet d'extraire 574 nouveaux termes par rapport au scénario 3.1 (gazetiers). Ces résultats mesurent deux phénomènes linguistiques différents : l'association de termes à des concepts de l'ontologie, d'une part, et l'association de termes à des noms toponymiques de gazetiers, d'autre part.

Enfin, l'expérimentation a également montré que l'utilisation des relations VT, mVT et $teVT$, qui donnent 1 191 termes distincts valides, extraits avec une précision de $0,55$, est plus efficace que celle des couples [termes, nom toponymique], qui donnent 733 termes distincts valides extraits avec une précision de $0,38$. Ces derniers résultats ont été obtenus par l'analyse basique de tous les toponymes, indépendamment des scénarios étudiés ici. Cette dernière approche correspond à la manière dont fonctionnent la plupart des moteurs génériques de reconnaissance d'entités nommées.

5.1.2. *Évaluation qualitative : traitement de phrases complexes*

L'expérimentation a montré que notre approche est robuste car elle permet de traiter des phrases très complexes du corpus. En voici quelques-unes :

(4) *1.* Je partis en conséquence pour Bagnères de Luchon une seconde fois en passant par le Tourmalet, un beau village près d'Hourquette d'Arreau et un autre au sud de la belle vallée de Louron.

2. De cet océan de glaces et de neiges, partent comme autant de mers, des prolongements qui s'étendent entre les différents mamelons des montagnes circonvoisines, et parviennent jusqu'aux sommets des vallées qui aboutissent à Vignemale, pour fournir à l'écoulement de divers gaves ou torrents : le plus considérable de ces prolongements paraît être celui qui descend dans la vallée d'Ossoue.

Dans le cas de la phrase (4) 1., les règles étiquetant les textes de manière incrémentale permettent de marquer deux structures VT dont la seconde est composée d'une liste de toponymes. Les termes *village* et *vallée* sont extraits. Il s'agit ici de termes associés directement aux noms toponymiques.

Dans le cas de la phrase (4) 2., notre chaîne a repéré une relation mVT et trois relations $teVT$: mVT (*partent comme autant de mers, étendent entre les différents mamelons des montagnes circonvoisines, parviennent jusqu'aux sommets des vallées, aboutissent à Vignemale, descend dans la vallée d'Ossoue*), $teVT$ (*prolongements qui s'étendent entre les différents mamelons des montagnes circonvoisines*), $teVT$ (*sommets des vallées qui aboutissent à Vignemale*), $teVT$ (*celui qui descend dans la vallée d'Ossoue*). À l'aide de ces relations, les termes suivants sont extraits : *mers, mamelons des montagnes circonvoisines, sommets des vallées, vallée*. Cet exemple montre encore un fois que notre méthode est capable d'extraire des termes composés et des termes qui ne sont pas associés aux noms toponymiques (*mamelons des montagnes circonvoisines, sommets des vallées*).

5.2. *Exploitation d'une ontologie topographique dans une chaîne d'indexation spatiale de documents textuels*

Une chaîne d'indexation spatiale bénéficiant de l'ontologie géographique pour le typage des noms toponymiques détectés dans des textes peut, par exemple, désambiguïser *pic d'Argelès, vallée d'Argelès et ville d'Argelès* qui font référence à des zones géographiques distinctes. Nous allons, dans un premier temps décrire succinctement la chaîne de traitement PIV (Pyrénées Itinéraires Virtuels). Nous allons ensuite présenter les résultats d'une expérimentation qui nous a permis d'évaluer l'apport de l'ontologie dans un tel contexte.

5.2.1. *La chaîne d'indexation spatiale de documents textuels*

La chaîne d'indexation spatiale PIV (Gaio *et al.*, 2008) est composée de trois principales étapes de traitement dont les deux premières s'appuient sur les règles et ressources de marquage exposées en section 3.

L'étape 1 consiste à repérer les toponymes avec le processus décrit plus haut.

L'étape 2 détermine une représentation symbolique sous la forme de traits sémantiques pour chacun des toponymes détectés comme précédemment. Il s'agit de la réutilisation de l'étape de marquage des indirections décrite figure 2. Ainsi, lorsque des relations spatiales sont détectées au sein d'un toponyme, chaque relation est analysée et des étiquettes (telles qu'une relation d'adjacence, d'inclusion, d'orientation, etc.) sont associées au toponyme. De même, pour chaque toponyme comportant un SN, un processus de typage permet d'apposer une deuxième catégorie d'étiquettes (oronyme, hydronyme, voie de communication, commune, lieu-dit, etc.) qualifiant ces entités.

L'étape 3 interprète les représentations symboliques ainsi obtenues et calcule des représentations numériques. L'interprétation est supportée par des algorithmes d'approximation qui associent des géométries aux toponymes (Gaio *et al.*, 2008), à l'aide des opérateurs spatiaux d'un SIG (le logiciel *PostGIS* en la circonstance). Concernant l'approximation spatiale, nous utilisons des ressources de type gazetier pour valider et géolocaliser chaque nom toponymique. Les relations spatiales (par exemple, $orientation\,(nord)$) sont ensuite interprétées et leurs représentations calculées ; des opérateurs SIG (par exemple, $translation$, $intersection$) sont appliqués à la géométrie correspondant au nom toponymique de référence.

5.2.2. *Apport d'une ontologie topographique dans une chaîne d'indexation spatiale de documents textuel*

Expérimentation. Les ressources mobilisées pour l'expérimentation correspondent à l'ontologie géographique GéOnto et au corpus décrit à la section 5.1. Un premier traitement a permis d'extraire 15 572 toponymes (TO).

Le protocole d'expérimentation vise le traitement de ce corpus selon trois versions spécifiques de la chaîne de traitement PIV :

– PIV_a : la chaîne ne dispose pas de ressource pour le typage des noms toponymiques ;

– PIV_b : la chaîne dispose d'un lexique *ad hoc* construit manuellement, à partir des *catégories/natures* de lieux répertoriées dans les gazetiers Geonames[24] et BDNYME[25], pour le typage des noms toponymiques ;

– PIV_c : la chaîne dispose de la ressource ontologique du projet GéOnto pour le typage des noms toponymiques.

Dans chaque configuration, nous procédons à différentes évaluations quantitatives et qualitatives. L'évaluation quantitative détermine le nombre de toponymes (TO) identifiés à l'étape 1 du processus, le nombre de TO typés à l'étape 2 et le nombre de TO géolocalisés à l'étape 3. L'évaluation qualitative, quant à elle, mesure, pour un échantillon de 100 TO vérifiés manuellement, le nombre des TO correctement typés et géolocalisés : il s'agit de mesurer la précision du typage et de la géolocalisation dans la chaîne de traitement PIV.

Résultats. Du point de vue quantitatif, si l'on considère les 15 572 TO candidats, les résultats de l'expérimentation, décrits dans le tableau 9, montrent que le prototype PIV_c, avec l'usage de l'ontologie, permet le typage de beaucoup plus de TO candidats. Le nombre de toponymes typés est d'environ 1/3 de l'ensemble des toponymes candidats. Ceci contribue à une meilleure précision de la géolocalisation des TO.

	TO typés	TO géolocalisés	Temps de traitement en secondes
PIV_a	0	9 840	13 451
PIV_b	2 736	9 408	13 144
PIV_c	4 993	9 421	11 977

Tableau 9. *Évaluation quantitative du typage des toponymes*

Du point de vue qualitatif, nous focalisons notre analyse sur un échantillon de 100 TO dont 96 correspondent à des toponymes étendus (par exemple, *pic de Torte, vallée d'Argelès*) et 4 correspondent à des noms toponymiques (par exemple, *Pau*).

24. http ://www.geonames.org/
25. http ://professionnels.ign.fr/bdnyme

Si l'on considère les résultats de l'expérimentation décrits tableau 10, pour la chaîne PIV_c exploitant l'ontologie, le nombre de TO de cet échantillon correctement typés atteint 92. Ce nombre est de 56 pour les TO correctement géolocalisés. L'amélioration

	TO typés correctement (%)	TO géolocalisés correctement (%)	Temps de traitement d'une phrase en secondes	Temps de traitement d'un TO en secondes
PIV_a	0	35	5,30	0,90
PIV_b	42	36	4,01	0,81
PIV_c	92	56	3,78	0,14

Tableau 10. *Évaluation qualitative du typage des toponymes*

de la qualité d'interprétation (géolocalisation) liée au typage des toponymes est donc de 85 %. Ceci s'explique naturellement par l'exemple du toponyme *gave de Cauterets* qui, dans la chaîne standard, est géolocalisé par la géométrie de la *commune de Cauterets* alors que dans la chaîne exploitant le typage *hydronyme* (cours d'eau), sa géolocalisation est bien celle du *gave de Cauterets* et la géométrie obtenue dans la table *hydronyme* des ressources gazetier de l'IGN.

6. Conclusion

La méthode proposée vise la construction d'un inventaire lexical, à partir d'un fonds documentaire donné, en repérant, au sein d'une même phrase, les GN en relation avec un nom toponymique ou bien avec d'autres SN déjà qualifiés de topographiques. Il s'agit ensuite d'exploiter un faisceau d'indices linguistiques permettant progressivement de conserver l'ensemble le plus restreint possible de GN candidats à être rattachés à un concept de l'ontologie noyau ou à constituer un nouveau concept de cette ontologie afin de l'enrichir. La méthode combine l'exploitation de relations spécifiques intraphrastiques et des ressources externes : lexiques de verbes spécialisés, thésaurus généraux, ontologie géographique et gazetiers.

Intégré dans un processus plus large, le résultat d'un tel traitement permet de déduire, d'une part, une représentation symbolique, exprimée en traits sémantiques, des toponymes annotés et, d'autre part, de leur attribuer au moins une représentation numérique, l'objectif de ce processus étant de proposer un système cohérent pour la recherche d'information géographique dans des bases documentaires territorialisées.

Une première phase d'expérimentations sur un corpus réel a montré que la méthode permet d'élargir de manière significative le champ lexical utilisé pour la définition des concepts d'une ontologie de domaine. Une seconde phase d'expérimentations a montré que la méthode permet également d'améliorer la pertinence des index spatiaux issus de traitements similaires appliqués au contenu d'un corpus textuel. L'amélioration provient en particulier du typage des toponymes qui diminue le nombre

d'ambiguïtés et améliore la précision des index géocodés (représentations numériques des lieux) associés aux toponymes ainsi annotés.

7. Bibliographie

Aiello M., Pratt-Hartmann I., Van Benthem J. F., (eds), « Handbook of spatial logics », edn, Springer, 2007.

Balbiani P., Muller P., « Le raisonnement spatial », edn, Cepadues Editions, 2000.

Berretti S., Del Bimbo A., Enrico V., « Weighted walkthroughs between extended entities for retrieval by spatial arrangement », *IEEE Transactions on Multimedia*, vol. 5, n° 1, p. 52-70, 2003.

Bilhaut F., Dumoncel F., Enjalbert P., Hernandez N., « Indexation sémantique et recherche d'information interactive », *CORIA, Saint-Etienne*, p. 65-76, 2007.

Billen R., Clementini E., « Étude des caractéristiques projectives des objets spatiaux et de leurs relations », *Revue Internationale de Géomatique*, vol. 14, n° 2, p. 145-165, 2004.

Boons J.-P., « La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs », *Langue Française*, n° 76, p. 5-40, 1987.

Borillo A., « L'espace et son expression en français, L'essentiel », edn, Orphrys, 1998.

Bouamor H., « Extraction des connaissances à partir du Web pour la recherche des images géoréférencées », *CORIA*, p. 519-526, 2009.

Brodeur J., Interopérabilité des données géospatiales : élaboration du concept de proximité géosémantique., PhD thesis, Université Laval, Québec, 2004.

Buscaldi D., « Toponym ambiguity in geographical information retrieval », *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, ACM, New York, NY, USA, p. 847-847, 2009.

Buscaldi D., Rosso P., « Using GeoWordNet for Geographical Information Retrieval », *CLEF*, p. 863-866, 2008.

Camacho H., Salhi A., « A string metric based on a one to-one greedy matching algorithm », *Research in Computing Science*, 2006.

Charnois T., Enjalbert P., *Sémantique et traitement automatique du langage naturel*, edn, P. Enjalbert, Paris, France, chapter Compréhension automatique, p. 267-308, 2005.

Clementini E., A Conceptual Framework for Modelling Spatial Relations, PhD thesis, Institut National des Sciences Appliquées de Lyon, Lyon, Juin, 2009.

Cohen W. W., Ravikumar P., Fienberg S., « A Comparison of String Distance Metrics for Name-Matching Tasks », *IJCAI-03 Workshop on Information Integration*, p. 73-78, 2003.

Denis P., Sagot B., « Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort », *Proceedings of PACLIC 2009*, Hong Kong, China, 2009.

Egenhofer M., Franzosa R., « Point-set topological spatial relations », *International journal for Geographical Information Systems*, vol. 5, n° 2, p. 161-174, 1991.

Frank A. U., « Qualitative Reasoning about Distances and Directions in Geographic Space », *Journal of Visual Languages and Computing*, vol. 3, n° 4, p. 343-371, 1992.

Freksa C., « Using orientation information for qualitative spatial reasoning », *in* A. U. Frank, I. Campari, U. Formentini (eds), *Theories and methods of spatio-temporal reasoning in geographic space*, vol. 639 of *LNCS*, Springer, Berlin, p. 162-178, 1992.

Gaio M., Sallaberry C., Etcheverry P., Marquesuzaà C., Lesbegueries J., « A global Process to Access Documents' Contents from a Geographical Point of View. », *Journal of Visual Languages & Computing*, vol. 19, n° 1, p. 03-23, 2008.

Gruber T. R., « Toward principles for the design of ontologies used for knowledge sharing », *International Journal of Human-Computer Studies*, vol. 43, n° 4–5, p. 907-928, 1995.

Guarino N., Giaretta P. edn, IOS Press, Amsterdam, NL., chapter Ontologies and knowledge bases : Towards a terminological clarification, p. 25-32, 1995.

Hamdi F., Reynaud C., Safar B., « Pattern-based mapping refinement », *Proceedings of the 17th international conference on Knowledge Engineering and Knowledge Management by the masses*, EKAW'10, Springer-Verlag, Berlin, Heidelberg, p. 1-15, 2010.

Hamdi F., Safar B., Reynaud C., « Utiliser des résultats d'alignement pour enrichir une ontologie », *in* A. Khenchaf, P. Poncelet (eds), *EGC*, vol. RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, Hermann-Éditions, p. 407-412, 2011.

Hearst M., « Automatic Acquisition of Hyponyms from Large Text Corpora », *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, p. 539-545, Aout, 1992.

Hernández D., « Maintaining Qualitative Spatial Knowledge », *in* A. U. Frank, I. Campari (eds), *COSIT'93*, vol. 761 of *LNCS*, Springer-Verlag, p. 33-53, 1993.

Hopcroft J. E., Motwani R., Ullman J. D., « Introduction to automata theory, languages, and computation, 2nd edition », edn, Addison-Wesley, 2001.

Jaccard P., « Étude comparative de la distribution florale dans une portion des Alpes et des Jura », *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, p. 547-579, 1901.

Jaro M. A., « Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida », *Journal of the American Statistical Association*p. 414-420, 1989.

Kergosien E., Kamel M., Sallaberry C., Bessagnet M.-N., Aussenac-Gilles N., Gaio M., « Construction et enrichissement automatique d'ontologie à partir de ressources externes », *JFO'09*, Poitiers, France, p. 11-20, December, 2009.

Larson R. R., « Geographic Information Retrieval and Spatial Browsing », *GIS and Libraries : Patrons, Maps and Spatial Information*p. 81-124, April, 1996.

Laur D., 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, 1991.

Levenshtein V., « Binary codes capable of correcting deletions, insertions, and reversals », *Soviet Physics Doklady*, 1966.

Ligozat G., « Reasoning about Cardinal Directions », *J. Vis. Lang. Comput.*, vol. 9, n° 1, p. 23-44, 1998.

Loustau P., Interprétation automatique d'itinéraires dans des recits de voyage, PhD thesis, Université de Pau et des Pays de l'Adour, 2008.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D., « Cascades de transducteurs autour de la reconnaissance des entités nommées », *TAL*, vol. 52, n° 1, p. 69-96, 2011.

Monge A., Elkan C., « The field-matching problem : algorithm and applications », *Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Mustière S., Abadie N., Aussenac-Gilles N., Bessagnet M.-N., Kamel M., Kergosien E., Reynaud C., Safar B., Sallaberry C., « Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique », *(RIG) Revue internationale de géomatique*, vol. 21, n° 2, p. 155-179, 2011.

Needleman S. B., Wunsch C. D., « A general method applicable to the search for similarities in the amino acid sequence of two proteins », *Journal of Molecular Biology*, 1970.

Poibeau T., « Extraction automatique d'information : du texte brut au web sémantique », edn, Hermès Lavoisier, 2003.

Purves R., Jones C. (eds), *Workshop on geographic information retrieval*, vol. 2, SIGIR 2004, SIGIR Forum, 2004.

Randell D. A., Cui Z., Cohn A. G., « A spatial logic based on regions and connection », *3rd Int. Conf. on Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, p. 165-176, 1992.

Rocío A.-M., Erick L.-O., « Geo information extraction and processing from travel narratives », *Transforming the Nature of Communication, 14th International Conference on Electronic*, Helsinki, Finland, p. 363-373, 2010.

Safar B., Reynaud C., « Alignement d'ontologies basé sur des ressources complémentaires Illustration sur le système TaxoMap », *Technique et Science Informatiques*, vol. 28, n° 10, p. 1211-1232, 2009.

Sagot B., Boullier P., « SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts », *Traitement Automatique des Langues*, vol. 49, p. 155-188, 2008.

Sarda L., « L'expression du déplacement dans la construction transitive directe », *Syntaxe et Sémantique,* p. 121-137, 2000.

Schmidt H., « Probabilistic part-of-speech tagging using decision trees », *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Smith T. F., Waterman M. S., « Identification of Common Molecular Subsequences », 1981. Academic Press Inc. (London) Ltd.

Stoilos G., Stamou G. B., Kollias S. D., « A String Metric for Ontology Alignment », *International Semantic Web Conference*, p. 624-637, 2005.

Talmy L., « Toward a Cognitive Semantics », edn, The MIT Press, chapter How language structures space, 2000.

Uitermark H., Ontology-Based Geographic Data Set Integration, PhD thesis, Universiteit Twente, the Netherlands, 2001.

Vandeloise C., « L'espace en français », edn, Seuil, Paris, France, 1986.

Winkler W. E., « The state of record linkage and current research problems », 1999. Statistics of Income Division, Internal Revenue Service Publication R99/04.