Repérer des toponymes dans des titres de cartes topographiques

Catherine Dominguès[1], Iris Eshkol-Taravella[2]
(1) IGN, laboratoire COGIT 73 avenue de Paris 94160 Saint-Mandé
(2) LLL, UMR 7270, 10 Rue de Tours, BP 46527, 45065 ORLEANS cedex 2
catherine.domingues@ign.fr, iris.eshkol@univ-orleans.fr

RÉSUMÉ

Les titres de cartes topographiques personnalisées composent un corpus spécifique caractérisé par des variations orthographiques et un nombre élevé de désignations de lieux. L'article présente le repérage des toponymes dans ces titres. Ce repérage est fondé sur l'utilisation de BDNyme, la base de données de toponymes géoréférencés de l'IGN, et sur une analyse de surface à l'aide de patrons. La méthode proposée élargit la définition du toponyme pour tenir compte de la nature du corpus et des données qu'il contient. Elle se décompose en trois étapes successives qui tirent parti du contexte extralinguistique de géoréférencement des toponymes et du contexte linguistique. Une quatrième étape qui ne retient pas le géoréférencement est aussi étudiée. Le balisage et le typage des toponymes permettent de mettre en avant d'une part la diversité des désignations de lieux et d'autre part leurs variations d'écriture. La méthode est évaluée (rappel, précision, F-mesure) et les erreurs analysées.

ABSTRACT

Localizing toponyms in topographic map titles

The titles of customized topographic maps constitute a specific corpus which is characterized by spelling variations and a very significant number of place names. This paper is about identifying toponyms in these titles. The toponym tracking is based on IGN's toponym data base as well as light parsing according to patterns. The method used broadens the definition of the toponym to include the nature of the corpus and the data in it. It consists of three successive stages where both the extralinguistic context - in this case georeferencing toponyms - and the linguistic context are taken into account. The fourth stage which is without georeferencing is examined too. Toponym tagging and typing allow to highlight toponym naming and spelling variations. The method has been assessed (recall, precision, F-measure) and the results analysed.

MOTS-CLÉS : toponyme, information spatiale, écriture des toponymes, BDNyme, ressource lexicale.

KEYWORDS : toponyme, spatial information, toponyme writing, BDNyme, lexical resource.

1 Introduction/contexte/observation de la réalité

Dans le contexte d'une demande croissante de produits cartographiques adaptés à leurs utilisateurs, de nombreuses entreprises et agences nationales géographiques offrent des services de cartographie qui permettent de concevoir des produits cartographiques plus ou moins personnalisés. En particulier, l'Institut de l'information géographique et forestière (IGN) offre depuis 2007 un service web de *Carte à la carte* qui permet à tout utilisateur d'Internet de définir, à partir des bases de données géographiques de l'institut, une carte topographique personnalisée sur différents aspects (taille, échelle, centre et titre de la carte). L'ensemble de ces demandes constitue une source de renseignements sur les usages de ce service. Une manière d'étudier ces demandes est d'en baliser l'ensemble des titres afin d'identifier les différents types d'informations qu'ils contiennent. Les toponymes étant majoritairement représentés dans les titres, leur traitement constitue une première étape de cette étude ; cet article est consacré à leur identification automatique.

2 Difficultés liées aux noms de lieu[1]

La notion de lieu s'appuie sur des définitions hétérogènes et des règles d'écriture complexes.

2.1 Définitions

Dans le domaine du traitement automatique des langues, les toponymes (localisations, lieux, entités spatiales) font partie des entités nommées. Un état de l'art sur les différents systèmes de reconnaissance des entités nommées est présenté dans (Ehrmann, 2008) et (Nadeau et Sekine, 2009). Selon (Ehrman, 2008) « *on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus* ». Concernant les toponymes, les conventions de la campagne d'évaluation Quaero (2011)[2] distinguent les lieux administratifs des lieux physiques, les voies, les bâtiments et les adresses. (Lesbegueries, 2007) propose la distinction entre « *entité spatiale absolue* » caractérisant les informations propres à une entité nommée (*la ville de Paris*) et « *entité spatiale relative* » caractérisant des indications spatiales associées aux entités nommées (*près de Paris*).

La Commission Nationale de Toponymie (CNT)[3] indique qu'un toponyme désigne un objet géographique déterminé. Pour l'IGN, « *un toponyme est un nom de lieu, constitué d'un ou plusieurs mots, en rapport étroit avec un détail géographique localisé et avec le groupe humain qui l'utilise* ». La toponymie[4] distingue « *des noms de lieux habités (villes, bourgs, villages, hameaux et écarts) ou non habités (lieux-dits)* […] *les noms liés au relief (oronymes), aux cours d'eau (hydronymes), aux voies de communication (odonymes, ou hodonymes)* », et des microtoponymes « *comme des noms de villas ou d'hôtels* », par exemple.

La notion de toponyme et par conséquent son identification posent différents types de problèmes. D'après les définitions précédentes, la typologie des toponymes est d'ordre référentiel car elle s'appuie sur la nature du référent désigné par le nom de lieu. Il ne peut alors s'agir de limiter cette définition aux seuls noms propres car les noms communs peuvent aussi permettre de désigner des lieux de manière neutre : *le village*, ou personnalisée : *mon village*, ou de faire référence à un lieu imaginaire *le bout de monde*, *mon paradis*, etc. Pour poursuivre cette typologie référentielle, il faut ajouter les déictiques : *ici*, *là*. En outre, un même lieu peut être désigné par plusieurs toponymes et inversement un même toponyme peut désigner des objets géographiques différents.

2.2 Ecriture des toponymes

L'écriture des noms de lieux fait appel à des règles complexes qui s'appuient sur des connaissances linguistiques et extralinguistiques. Le nommage des objets géographiques n'est pas normalisé et provient souvent de la tradition orale. De plus, l'écriture des toponymes diffère selon l'usage ; par exemple, les panneaux indicateurs ou les plaques indicatrices de rue, sont écrits en majuscules. (Bioud, 2006) remarque qu'*« il est de plus en plus courant de trouver des textes où un même mot est orthographié de deux ou trois façons différentes, parfois même plus »*. Sur le Web, l'orthographe change d'un utilisateur à l'autre et les nouvelles formes de communication écrite influencent fortement l'écriture des toponymes.

Cependant, des règles d'écriture existent, mais elles sont compliquées, subtiles et non homogènes d'où des difficultés de mise en application et de compréhension. Deux signes typographiques, en particulier, rendent difficile l'écriture de toponymes composés : la majuscule et le trait d'union.

Pour des règles d'usage de la majuscule (Mathieu-Colas, 1998) souligne que *« chaque auteur présente ses règles sous une forme impérative, on note de l'un à l'autre un certain nombre de divergences qui, dissipant l'illusion d'une norme universelle, ne font que mettre en évidence l'instabilité du système »*.

Les règles concernant le trait d'union sont aussi compliquées, complexes et floues. Les recommandations et observations grammaticales de la CNT, par exemple, juxtaposent des critères sémantiques et syntaxiques : *« Parmi les mots composant en français un toponyme […] sont joints par des traits d'union les mots ayant perdu dans la composition leur sens ou leur syntaxe habituels »*. L'un des sous-exemples de cette affirmation concerne *« les mots appartenant à un groupe de mots ayant une fonction de complément (avec ou sans préposition) au sein du syntagme toponymique et ne se limitant pas à décrire l'objet géographique »* : *le massif du Mont-Blanc*, *le parc des Buttes-Chaumont*. Cependant lorsque ces mots n'ont pas la fonction de complément, cette règle

[1] Dans cet article, les termes *nom de lieu* et *toponyme* seront employés indifféremment.
[2] http://quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf
[3] CNI-CNIG (2010), Recommandations et observations grammaticales.
http://www.cnig.gouv.fr/Front/docs/cms/cnt-grammaire-recommandation_126924688421947500.pdf [consulté le 18/12/2012]
[4] Définition d'un toponyme dans Wikipédia : http:///www.wikipedia.fr [consulté le 10/12/2012].

n'est plus valable (*le mont Blanc*) et l'initiale du nom générique reste minuscule. Le *Bon Usage* (Grevisse et Goosse, 1993) parlant des signes typographiques dont le trait d'union fait partie et sans distinguer le cas des toponymes, indique la présence de ce signe *« à la suite d'un changement de signification »* (*la rue Saint-Pierre*, *la ville de Saint-Etienne*). Comment le scripteur pourrait-il mémoriser toutes ces nuances typographiques qui présupposent des connaissances linguistiques profondes et qui s'appuient sur une analyse syntaxique et sémantique préalable du syntagme écrit ?

Il n'existe donc pas de consensus réel entre les divers auteurs pour trancher de l'usage de la majuscule ou du trait d'union dans la plupart de leurs emplois ; ce constat milite pour l'écriture libre des toponymes : *« Qu'on opte pour une "harmonisation orthographique" [...] ou pour une tolérance bien tempérée, il convient de se libérer des "délires" de l'orthographe [...]. Rien ne serait pire pour le traitement automatique que de vouloir s'accrocher à des normes aussi pointilleuses qu'arbitraires »*. (Mathieu-Colas 1998 :12). En conséquence, le point de vue adopté ici ne se veut pas normatif mais tente de tenir compte de cette liberté orthographique en n'imposant aucune règle d'écriture du toponyme et en acceptant de nombreuses variations.

3 Description du corpus d'étude

Le corpus de travail est composé de lignes, chacune contient un titre formé de phrases ou de groupes de mots servant à dénommer la carte[5] et de spécifications techniques permettant de dessiner la carte :

15000;	*portrait;*	*704074;7059443;*	*LILLE-VILLE - PROMENONS NOUS Carte spéciale pour Victoire*.
échelle	orientation	coordonnées du centre	titre

Le corpus est constitué par des internautes variés utilisant des règles d'écriture disparates. Les internautes sont guidés par des objectifs très différents dans la création et donc dans la dénomination de leur carte : il peut s'agir d'un souvenir des vacances, de la préparation d'un événement partagé par une très petite communauté, ou son rappel, d'un cadeau à un proche… Le langage y est libre, les règles typographiques ne sont pas appliquées ou interprétées de manière individuelle et non homogène. Par conséquent, le texte n'est donc pas normalisé et présente différents types de variations orthographiques :

- l'absence ou la présence des majuscules, séparateurs, signes diacritiques, prépositions, déterminants : *Fontenay sous bois* (au lieu de *Fontenay-sous-Bois*) ;
- l'utilisation d'abréviations plus ou moins normalisées : *⁄* au lieu de *sous* ;
- les erreurs de frappe : *LA MONTAGHE* (au lieu de *LA MONTAGNE*) ;
- la transcription phonétique d'un accent régional : *NOT'BARAQUE ché par ichi* ;
- de nouvelles formes de communication écrite : *LADOUJEVIENS OUPETIGEAITAI* ;
- la création lexicale *: AZZAZ et ses environz*, *Tamalou-Land*.

Tous les titres ne sont pas rédigés en langue française. Les internautes composent des titres en langues étrangères (anglais, allemand, etc.) ou régionales (corse, basque, provençal, etc.) qu'ils peuvent mélanger au français dans un même titre.

Les internautes peuvent aussi appartenir à des communautés dont les activités s'appuient sur l'utilisation de cartes topographiques et utiliser le langage de ces communautés dans le titre de la carte, comme par exemple dans : *BLEAU TOP30 où Bleau* est *« l'appellation familière de la forêt de Fontainebleau dans les milieux sportifs, notamment le Groupe de Bleau »*[6].

Repérer les toponymes dans un tel corpus est une tâche difficile. Le plus souvent, les systèmes de détection des entités nommées utilisent soit une approche symbolique fondée sur des grammaires locales (Bontcheva *et al.*, 2002), (Friburger, 2002), (Poibeau, 2003), soit une approche statistique à base d'apprentissage automatique, soit des systèmes hybrides comme (Béchet *et al.*, 2011). Notre démarche est guidée par les caractéristiques du corpus décrits ci-dessus : information spatiale, variations orthographiques[7] et d'emploi des toponymes. Elle utilise donc à la fois une ressource lexicale qui associe les noms propres de lieu et leur géoréférencement, et des patrons qui détectent les noms de lieu formés sur des noms communs.

4 Ressource pour identifier les toponymes : BDNyme

L'IGN propose une ressource lexicale spécifique recensant les toponymes de France métropolitaine et leur localisation, BDNyme. Le principe du recueil des toponymes est fondé sur une enquête terrain et a pour objectif de demeurer aussi proche que possible de l'usage local actuel. Des critères de qualité sont aussi garantis ; par exemple, tous les toponymes, après avoir été soumis à un représentant de l'usage local, sont validés par le bureau de toponymie de l'IGN. Le nombre de toponymes retenus dans la base répond à des critères cartographiques, ce qui revient généralement à un critère de densité. Enfin sa couverture est de plus de 1,7 million d'entrées. Les toponymes se distinguent par leur type (cette segmentation s'appuie à la fois sur des critères géographiques et administratifs) : chef-lieu, lieu-dit habité, lieu-dit non habité, hydronyme, oronyme, toponyme de communication, toponyme ferré et toponyme divers. D'autres ressources existent, comme GeoNames[8], mais elles ne fournissent pas la même couverture ; par exemple, GeoNames propose 1277 occurrences du terme *Ardèche*, alors que BDNyme retrouve 19 804 toponymes situés dans le département de l'Ardèche et répartis selon leurs catégories.

[5] La longueur est limitée à 55 caractères.

[6] Wikipédia : http://fr.wikipedia.org/wiki/Groupe_de_Bleau [consulté le 12/12/2012]

[7] Toutes les variations n'ont pas été résolues dans le travail présenté ici.

[8] GeoNames propose une base de données de toponymes gratuite et accessible par Internet sous une licence Creative Commons. http://www.geonames.org/ [consulté le 22/03/2013]

Les toponymes de BDNyme sont écrits en lettres minuscules accentuées (codage utf-8), simples ou composés et dont le séparateur est, selon le cas, l'espace, le trait d'union ou l'apostrophe. Chaque toponyme est suivi de ses coordonnées géographiques :

lille; *705009.20;7059266.70*
toponyme coordonnées du toponyme

Conformément à ses spécifications, BDNyme contient les noms d'objets dont l'implantation est ponctuelle (le point culminant d'une montagne : *pic carlit*) ou peut être ramenée à un point géographique (une ville dont l'implantation est ramenée à son centroïde : *paris*). En conséquence, les objets dont l'implantation est linéaire (comme un fleuve : *la seine*), ou surfacique (par exemple les entités administratives comme les régions ou les départements) et pour lesquels la notion de centroïde n'est pas pertinente ne sont pas contenus dans BDNyme. Ces entités étant largement présentes dans les titres de cartes, des listes de départements, régions administratives, montagnes, régions naturelles et pays, fleuves et rivières, parcs naturels ont été constituées manuellement et utilisées.

5 Méthode employée pour le repérage des toponymes dans les titres

Le repérage des toponymes se décompose en quatre étapes successives. A chaque étape, les toponymes reconnus dans le corpus sont balisés et typés[9]. Le résultat de ces transformations constitue le corpus d'entrée de l'étape suivante. Seules les trois premières étapes s'appuient sur le contexte. Ce recours au contexte recouvre deux aspects : le géoréférencement de la carte (l'emprise exacte : étape 1 et l'emprise élargie : étape 2) et une analyse de surface du corpus (étape 3) à l'aide de grammaires locales en utilisant la plateforme Unitex (Paumier, 2003). La dernière étape (étape 4) recherche les toponymes sans tenir compte du contexte de géolocalisation (cf. figure 1).

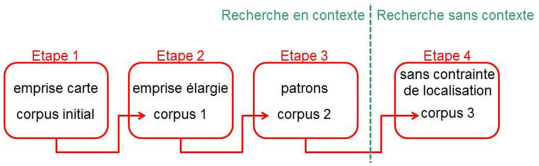

FIGURE 1 – Etapes de traitement du corpus des titres de cartes

L'identification des toponymes repose sur la comparaison des chaînes de caractères de BDNyme et celles contenues dans les titres. L'écriture des toponymes dans les titres présente de nombreuses variations orthographiques. En conséquence, certaines variations ont été prises en compte dans la reconnaissance d'une chaîne de BDNyme dans un titre :

- l'absence ou la présence de majuscules : *LILLE* , *Lille*, *lille* ;
- l'absence ou la présence de signes diacritiques : *FRESNES-LES-MONTAUBAN* et *FRESNES-LÈS-MONTAUBAN* ;
- le séparateur peut être le blanc, le trait d'union ou l'apostrophe ;
- les abréviations *st* et *ste* sont respectivement identifiées à *saint* et *sainte ;*
- les mots vides (déterminants, prépositions) peuvent être omis ;
- un toponyme composé de plusieurs mots peut être abrégé et reconnu sous cette forme à condition que les mots du toponyme ne soient pas un mot vide, l'adjectif *saint(e)* ou ses abréviations, un générique de noms de lieux[10], ni un prénom[11]. Par exemple, dans le titre : *AUTOUR DE BOUC Attention ça grimpe*, *BOUC* est reconnu comme le nom abrégé de *Bouc-Bel-Air*.

5.1 Recherche des toponymes à l'aide du contexte

5.1.1 Le contexte de géoréférencement : étapes 1 et 2

L'identification des toponymes dans le corpus des titres est fondée en premier lieu sur le contexte extralinguistique : la géolocalisation. Dans l'étape 1, ne sont examinés que les toponymes qui se situent dans l'emprise de la carte (rectangle a_1xb_1 de la figure 2) ; dans l'étape 2, l'emprise considérée est élargie (rectangle a_2xb_2). Les zones prises en compte sont représentées dans la figure 2. Dans l'exemple suivant : *{CHOLET,.<ChefLieuE2>} {Forêt de Nuaillé,.<ToponymeDiversE1>}*, *Forêt de Nuaillé* est reconnu comme un toponyme dans l'étape 1 et *Cholet* comme un chef- lieu dans l'étape 2.

Dans les cas d'ambiguïté où des analyses différentes peuvent être avancées pour une même chaîne de caractères, seule la balise correspondant à la chaîne la plus longue est posée. Dans l'exemple : *La Sainte Baume* où *la sainte* et *sainte-baume* sont des lieux-dits non habités, c'est la séquence la plus longue *Sainte-Baume* qui sera balisée : *La {Sainte Baume,.LieuDitNonHabiteE1}.*

[9] Au total, quatorze types de toponymes sont différenciés par les ressources lexicales (neuf avec BDNyme et six avec celles constituées manuellement) et neuf par les grammaires locales.

[10] Une liste de génériques pour les noms de lieux a été constituée et compte 291 entrées.

[11] Une liste de prénoms a été constituée et compte 1 641 entrées.

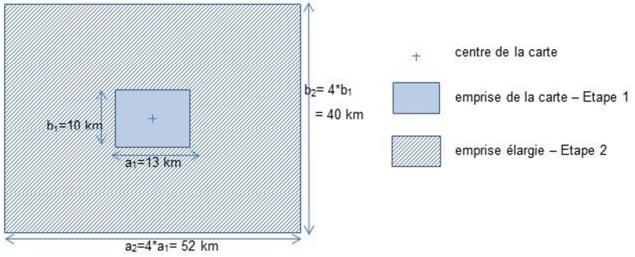

FIGURE 2 – Représentation de l'emprise de la carte (Etape 1) et de l'emprise élargie (Etape 2) pour une carte de format *NORMAL* : 91,5 cm x 69 cm - échelle 1/15 000 - orientation *paysage*)

5.1.2 Le contexte linguistique : étape 3

La troisième étape procède au repérage et au typage des toponymes à l'aide de patrons[12] [13] :

- les patrons repèrent les noms génériques de lieux, seuls[14] : *lac*, *plaine*, *hôtel*, *mont*, *maison*, *gîte*, etc. ou accompagnés d'un complément : *forêt de St Cucufa*, *maison de campagne (*étiquette : *LieuGenerique)* ;
- pour repérer les lieux, les patrons se fondent aussi sur les verbes, noms et prépositions locatifs : *vivre à (étiquette : LieuStatif)*, *partir de*, *départ de (étiquette : LieuDepart)*, *arrivée à (étiquette : LieuArrivee)*, *à côté de*, *autour de*, *alentour de*, *les environs de (*étiquette : *LieuApproximatif)*, *chez (étiquette : LieuChez) ;*
- l'étiquette *LieuSubj* repère des lieux appropriés et personnalisés par l'utilisateur : *mon paradis*, *far east* ou des lieux imaginaires comme *: Tamalou-Land ;*
- les déictiques : *ici*, *là*, *là où* sont balisés par une étiquette *LieuDeict* ;
- enfin, l'étiquette *LieuAdresse* identifie les adresses : *20 r de la Mollanche*.

La ligne suivante est un exemple de titre balisé à l'issue de l'étape 3 :
{TILLY,.ChefLieuE1} Les Millerus {Chez Olive et Sand,.LieuChez}

5.2 Identification sans contexte : étape 4

Dans cette dernière étape, le contexte que constitue le géoréférencement de la carte et des toponymes n'est pas pris en compte, c'est-à-dire que les toponymes de BDNymes sont recherchés sans tenir compte de la localisation du toponyme par rapport à l'emprise de la carte ou à l'emprise élargie. Les objets dont l'implantation est surfacique ou linéaire ne peuvent être localisés et donc recherchés dans les étapes précédentes ; ils sont intégrés à la recherche de toponymes dans cette étape.

La ligne suivante donne un exemple de titre balisé à l'issue de l'étape 4 :

{LE ROURET,.LieuDitHabiteE4} Escapades Jeep 2007.

6 Evaluation

Le corpus a été séparé en deux parties : un corpus de travail qui a permis d'identifier les ressources lexicales complémentaires nécessaires et de construire les patrons, un corpus de test sur lequel la méthode a été évaluée. Le corpus de test annoté manuellement constitue le corpus de référence. L'évaluation repose sur la comparaison des balises obtenues automatiquement dans le corpus de test et celles posées manuellement dans ce même corpus ; des mesures de rappel, précision et F-mesure permettent d'objectiver les résultats obtenus.

6.1 Corpus de référence

La constitution du corpus de référence s'est heurtée à des difficultés, liées à ses spécificités :

- l'absence de contexte entraîne de nombreux cas d'ambiguïté, par exemple : *STE AGA*THE peut être un nom de personne, un surnom, un nom de lieu (*Ste Agathe en Donzy*) ou une fête. Ces séquences ont été repérées mais non balisées en tant que lieu ;

[12] Les patrons regroupent onze graphes principaux.
[13] A l'opposé de la nature géographique des étiquettes proposées par BDNyme, la typologie des étiquettes est ici de nature sémantique et rend compte de l'emploi du lieu dans le corpus.
[14] Certains noms génériques étant contenus dans le toponyme de BDNyme, par exemple : *pic carlit*, ils ont été repérés dans les étapes précédentes si la localisation est dans l'emprise prise en compte.

- certains toponymes ne sont pas utilisés pour désigner un lieu et ne doivent donc pas être étiquetés :
  * nom de lieu qui contribue à définir l'auteur ou le destinataire de la carte : *LES COUSINES XXX*[15] *de Pietrosella* ;
  * nom de lieu qui contribue à définir une autre entité, ici un club cycliste : *TEAM U NANTES ATLANTIQUE* ;
  * nom générique ambigu qui peut désigner un lieu mais aussi une activité : *LA MONTAGNE ça vous gagne*.

Finalement, le corpus de référence constitué est composé de 1 457 lignes (soit 6 388 mots) et contient 1 576 désignations de lieux.

6.2 Résultats de l'évaluation

Afin de mesurer l'apport de chaque étape, quatre évaluations ont été mises en place. Le tableau 2 présente ces évaluations en termes de rappel, précision et F-mesure.

	A l'aide du contexte			Sans contexte
	Etape 1	Etape 2	**Etape 3**	Etape 4
Rappel	0,37	0,44	0,72	**0,80**
Précision	0,82	0,79	**0,82**	0,50
F-mesure	0,50	0,57	**0,76**	0,61

TABLE 1 – Evaluation des quatre étapes de la recherche des toponymes

Le gain entre les étapes 1 et 2 correspond à l'élargissement de l'emprise. Ceci signifie que les noms de lieux dans les titres ne figurent pas nécessairement dans la carte. Dans l'exemple :

{CHOLET,.ChefLieuE2} {Forêt de Nuaillé,.ToponymeDiversE1}

Forêt de Nuaillé est reconnu dès la première étape parce qu'il figure dans l'emprise de la carte, ce qui n'est pas le cas de *Cholet*. L'hypothèse correspondante serait que, pour désigner un lieu qu'il considère peu connu, l'utilisateur ajoute le nom d'un lieu plus populaire ou qui a un statut administratif plus important (chef-lieu ou lieu-dit).

L'étape 3 présente les meilleurs résultats en termes de précision et F-mesure ; le gain à cette étape est dû à l'utilisation des patrons qui tiennent compte des différents types de désignations de lieux et exploitent des indices linguistiques, absents dans BDNyme. Dans l'étape 4, le rappel est amélioré parce que la contrainte de géolocalisation n'est pas appliquée. En contrepartie, le relâchement de cette contrainte entraîne de nombreuses erreurs. Dans l'exemple : *{ERSTEIN,.ChefLieuE1} _ VELO {Carte,.LieuDitHabiteE3} des Reibel*, le mot *Carte* a été reconnu en tant que lieu-dit habité *Carté*. Ces cas fréquents d'ambiguïté dégradent la précision de l'étape.

6.3 Analyse des erreurs

Les erreurs de balisage ont été analysées. Certaines proviennent des cas mentionnés dans le paragraphe 6.1 et conduisent à identifier des séquences qui ne désignent pas des lieux ou qui sont ambiguës et donc non balisées dans le corpus de référence en tant que lieu (ce qui affecte la précision).

D'autres toponymes ne sont pas reconnus (le rappel est donc moins bon) parce :

- certaines variations orthographiques présentes dans les titres sont difficilement prévisibles :
  * les abréviations peu courantes comme *ch* pour *chemin* (*Ch-St Hilaire* pour *chemin St-Hilaire*, ou *L'* pour *longue* (*Saint-Germain de L'Chaume* pour *Saint-Germain de Longue Chaume*) ,
  * l'absence de séparateur : *VillardBonnot* au lieu de *Villard-Bonnot* ou *CCBEAUNOIS* au lieu de *CC BEAUNOIS> ;*
  * les erreurs de frappe : *St Aygul* au lieu de *St Aygulf, Pointeuils-et-Brésis* au lieu de *Ponteuils-et-Brésis ;*
- certains lieux ne peuvent être repérés automatiquement, par exemple dans : *ARAMOUN Autour de chez Jérôme et Yoann* où *Aramoun*, qui est un village au Liban, est un nom donné métaphoriquement à un lieu en France.

7 Perspectives et conclusions

Ce travail ouvre de nombreuses perspectives. Bien que BDNyme soit une base de données pérenne et homogène qui constitue une ressource de référence, d'autres types de lieux, en particulier les microtoponymes et les objets d'implantation surfacique pourraient être trouvées sur le Web. Wikipedia est une des ressources les plus utilisées. Une perspective serait d'ajouter une étape de balisage fondée sur des résultats de repérage automatique de toponymes dans Wikipedia[16].

La méthode développée ici est guidée par le respect et la prise en compte de la nature du corpus de titres. Elle s'appuie sur une

[15] Tous les exemples du corpus sont anonymisés.

[16] Des travaux de thèse de (Brando-Escobar, 2013) ont exploré une méthode d'extraction automatique des relations spatiales à partir des articles de Wikipédia ; ces travaux pourraient être adaptés à notre tâche. Celle-ci serait complémentaire à l'utilisation de BDNyme mais ne pourrait s'y substituer car Wikipedia ne contient pas systématiquement les coordonnées géographiques des toponymes.

définition étendue des toponymes et une typologie[17] adaptée à la fois à la ressource BDNyme et au corpus. Elle élargit la notion de contexte à des informations extralinguistiques de géoréfé-rencement. Enfin, le balisage et le typage des toponymes à l'aide des patrons a permis d'augmenter signi-ficativement le rappel et d'ajouter d'autres types d'information sur la nature des noms désignant des lieux.

Le repérage des toponymes constitue une étape préliminaire et nécessaire pour un étiquetage complet du corpus des titres qui permettrait de mieux cerner la demande de cartes des usagers de ce service : les destinataires (*la carte pour papi*), les encouragements (*en avant*), les éléments temporels (*été 2007*), les événements (*20 ans de mariage*). Un des objectifs serait alors d'adapter les typographies, les légendes de cartes, les illustrations de couverture, etc. aux besoins des usagers des nombreux services cartographiques disponibles sur le Web. Cette perspective s'inscrit dans le cadre plus large de la recherche et l'exploitation d'information spatiale contenue dans du texte, par exemple (Loustau *et al.*, 2008).

Références

BÉCHET, F., SAGOT, B.et STERN, R. (2011), « Coopération de méthodes statistiques et symboliques pour l'adaptation non supervisée d'un système d'étiquetage en entités nommées ». *TALN 2011*.

BIOUD, M. (2006). *Une normalisation de l'emploi de la majuscule et sa représentation formelle pour un système de vérification automatique des majuscules dans un texte*, Thèse de doctorat, Université de Franche-Comté.

BONTCHEVA, K., DIMITROV, M., MAYNARD, D., TABLAN, V. et CUNNINGHAM, H. (2002), « Shallow Methods for Named Entity Coreference Resolution ». *TALN 2002*.

BRANDO-ESCOBAR, C. (2013). *Coalla : Un modèle pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence*, Thèse de doctorat, Université de Marne-la-Vallée.

EHRMANN, M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguisation*, Thèse de doctorat, Université Paris 7 - Denis Diderot.

FRIBURGER, N. (2002), *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*, thèse de doctorat, Université François-Rabelais de Tours.

GREVISSE, M. et GOOSSE, A. (1993). *Le Bon Usage*. Duculot. Paris, Louvain-la-Neuve.

LESBEGUERIES, J. (2007). *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé*, Thèse de doctorat, Université de Pau et des Pays de l'Adour.

LOUSTAU, P., GAIO, M. et NODENOT, T. (2008). Interprétation automatique d'itinéraires à partir d'un corpus de récits de voyages pilotée par un usage pédagogique. RNTI, 2008, E (13), pages 177-206.

MATHIEU-COLAS, M. (1998). La majuscule flottante. Remarques sur l'orthographe des noms propres composés (type *N Adj*). *BULAG* n° 23, Centre Lucien Tesnière, Université de Franche-Comté, Besançon, 1998, pages 123-144.

NADEAU, N. et SEKINE, S. (2009). *A survey of named entity recognition and classification*. Satoshi Sekine and Elisabete Ranchhod, ed. John Benjamins publishing company, pages 3-28.

PAUMIER, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Université de Marne-la-Vallée.

POIBEAU, T. (2003). *Extraction automatique d'information, du texte brut au web sémantique*, Lavoisier.

[17] Cette typologie n'a pas, pour le moment, été exploitée mais permettrait d'affiner le traitement de la commande en proposant une symbologie adaptée aux thèmes géographiques dominants déduits du ou des toponymes figurant dans le titre.