

Repérage et identification automatiques de noms de lieux avec variations d'écriture dans des corpus

Mathilde Jouvel-Triollet

▶ **To cite this version:**

Mathilde Jouvel-Triollet. Repérage et identification automatiques de noms de lieux avec variations d'écriture dans des corpus. Sciences de l'Homme et Société. 2019. dumas-02302553

HAL Id: dumas-02302553

https://dumas.ccsd.cnrs.fr/dumas-02302553

Submitted on 1 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ Grenoble Alpes

Repérage et identification automatiques de noms de lieux avec variations d'écriture dans des corpus

JOUVEL-TRIOLLET Mathilde

Sous la direction de Monsieur Olivier KRAIF

Tuteurs de stage : Madame Catherine DOMINGUES
et Monsieur Philippe GAMBETTE

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2018-2019

UNIVERSITÉ Grenoble Alpes

Repérage et identification automatiques de noms de lieux avec variations d'écriture dans des corpus

JOUVEL-TRIOLLET Mathilde

Sous la direction de Monsieur Olivier KRAIF

Tuteurs de stage : Madame Catherine DOMINGUES
et Monsieur Philippe GAMBETTE

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2018-2019

UNIVERSITÉ Grenoble Alpes

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : JOUVEL-TRIOLLET

PRENOM : Nathilde

DATE : 31 Août 2019

Sommaire

Introduction **5**

I. Présentation du terrain de stage **7**

1. IGN **9**
1.1. Présentation globale 9
1.2. Champs d'action 9
1.3. Organisation interne 9

2. UPEM **10**
2.1. LIGM 10

II. Objectifs (cahier des charges) **11**

3. Étude du besoin **12**

4. Méthodologie pour répondre à la demande **13**
4.1. Identification de noms de lieux 13
4.1.1. Définition 13
4.1.2. Les variations d'écriture 13
4.1.3. État de l'art 15
4.2. Repérage des noms de lieux 15
4.2.1. Définition 15
4.2.2. État de l'art 15
4.3. Approche envisagée 17

5. Ressources et outils utilisés **19**
5.1. Les corpus 19
5.1.1. La Grande Peur 19
5.1.2. 88milSMS 20
5.1.3. BNF 20
5.1.4. Renumar 21
5.1.5. MATRICIEL 21
5.1.6. Corpus non-utilisés 21
5.2. Les gazetiers 23
5.2.1. BDNyme 23

5.2.2. Geofla . . . 23
5.2.3. GeoNames . . . 24
5.2.4. Prétraitements communs à tous les jeux de données . . . 24
5.3. Les mesures d'évaluation . . . 25
5.3.1. Précision . . . 25
5.3.2. Rappel . . . 25

III. Réponse au cahier des charges **26**

6. Identifier des toponymes **27**
6.1. Introduction . . . 27
6.2. Méthode . . . 27
6.2.1. Prétraitement du gazetier et des toponymes du corpus . . . 27
6.2.2. Match exact . . . 29
6.2.3. Le calcul de similarité . . . 29
6.3. Résultats . . . 31
6.4. Discussion . . . 32
6.5. Variante du processus d'identification . . . 33

7. Repérer des toponymes dans un corpus **34**
7.1. Des mots déclencheurs . . . 34
7.1.1. Introduction . . . 34
7.1.2. Méthode . . . 34
7.1.3. Résultats . . . 35
7.1.4. Discussion . . . 35
7.2. Des mots impossibles . . . 36
7.2.1. Introduction . . . 36
7.2.2. Méthode . . . 37
7.2.3. Résultats . . . 38
7.2.4. Discussion . . . 38

8. Repérage et identification simultanés **39**
8.1. Introduction . . . 39
8.2. Méthode . . . 39
8.2.1. Tokenisation du corpus . . . 40
8.2.2. Création de séquences . . . 40
8.2.3. Prétraitements du gazetier . . . 41
8.2.4. Création de segments . . . 45
8.3. Résultats . . . 48
8.4. Discussion . . . 49

Conclusion **52**

Références **54**

Table des illustrations **55**

Annexes 1 **56**
Répartition des variations par corpus 57
Une variation supplémentaire 59
Quelle mesure de similarité pour quelle variation ? 61

Annexes 2 **63**
Les mots outils 64
Les noms génériques géographiques 65
Un extrait du gazetier de toponymes 68
Toponymes extraits de La Grande Peur 70
Toponymes extraits des SMS 71
Toponymes extraits de la BnF 73
Extrait du corpus MATRICIEL 73
Toponymes extraits de Renumar 75
Extrait du corpus CHOUCAS 75

Introduction

Ce stage de fin d'études a été l'opportunité pour moi d'acquérir une expérience significative dans le domaine du Traitement Automatique du Langage (TAL) et de mettre en pratique mes connaissances acquises au cours du Master Sciences du Langage Parcours Industries de la Langue. Pour une durée de 5 mois (de mars à août), le laboratoire de l'IGN m'a proposé de me confronter aux enjeux du repérage d'entités nommées spatiales dans des corpus et m'a permis de développer des nouvelles compétences à travers différentes missions.

Il s'agit d'un stage qui s'intègre au projet collaboratif ANR 2016 CHOUCAS (Olteanu-Raimond et al., 2017) coordonné par l'équipe COGIT du laboratoire LASTIG de l'IGN et en partenariat avec les équipes STeamer du laboratoire LIG (Université Grenoble Alpes) et MOVIES du laboratoire LIUPPA (Université de Pau et des pays de l'Adour) ainsi que le Peloton de Gendarmerie de Haute Montagne (PGHM) de Grenoble. Le projet CHOUCAS fait suite à une demande exprimée par le PGHM, afin de faciliter le processus de décision lors de la localisation de victimes en montagne. Lors de l'appel téléphonique émis par la victime, le PGHM dispose de ressources textuelles hétérogènes de par leur niveau de langue, leur taille, leur syntaxe, etc. Ces données ne peuvent donc pas être interrogées efficacement.

C'est pourquoi, en vue de réduire le temps de localisation de la victime, les équipes CHOUCAS s'engagent à proposer des méthodes et des outils permettant de constituer et d'enrichir des données géographiques issues de sources hétérogènes, des modèles de raisonnement spatial flou et des environnements de géovisualisation.

Notre stage prend également la suite d'un stage de recherche entrepris par deux étudiants de l'École Nationale des Sciences Géographiques et reposant sur le lien entre les variations linguistiques et les mesures de distance entre deux chaînes de caractères.

Finalement, ce stage vise à apporter, aux secouristes du PGHM, une aide à la recherche des noms de lieux dans le contexte où la graphie recherchée diffère de celle des dictionnaires de noms propres de lieux.

Dans l'état d'avancement actuel de Choucas, il faudrait un outil quiassocie à un toponyme du corpus utilisé par le PGHM (et composé de données de formats et de caractéristiques linguistiques et typographiques différentes) un référent (forme normée) qui constitue une entrée du dictionnaire de noms propres de lieux. Puisque les données sont hétérogènes, la graphie du toponyme peut être identique à celle trouvée dans les dictionnaires de noms propres de lieux, comme elle peut s'en éloigner.

Prenons la phrase suivante : « À Pralognan suivre la route entre l'hôtel (...) ». Grâce à une annotation manuelle antérieure, nous savons que « Pralognan » est un toponyme. Le but est maintenant d'identifier le toponyme, c'est-à-dire d'indiquer qu'il correspond au toponyme référent « Pralognan-la-Vanoise ».

Un autre besoin de Choucas est de proposer, pour un toponyme donné, tous les noms de lieux qui contiennent ce toponyme. Lorsqu'une victime transmet par appel téléphonique sa position au PGHM, elle ne sait pas forcément où elle se situe et elle peut être sous le choc. Dans ces conditions, la position qu'elle communique peut être incomplète. Par exemple, une victime qui se situerait dans la commune Le-Bourg-d'Oisans pourrait indiquer par téléphone qu'elle se trouve dans la région de l'Oisans. Dans ce cas, le sauveteur souhaite disposer d'un outil qui à partir de « Oisans » lui propose tous les toponymes composés de ce mot (comme Le-Freney-d'Oisans, Saint-Christophe-en-Oisans et Le-Bourg-d'Oisans).

Dans cette optique, notre réflexion se place dans un contexte descriptif, et non prescriptif, de la graphie des noms de lieux. L'intérêt du stage ne porte pas sur la recherche des toponymes répondant à des normes de composition et de graphie précises, mais sur tout nom propre désignant un élément géographique.

C'est pourquoi notre réflexion aura pour ligne conductrice la problématique suivante : Identifier des noms de lieux dans des corpus malgré des variations d'écriture.

Le présent mémoire se divise en trois parties. Dans une première partie nous présenterons le terrain de stage. Dans un second temps nous décrirons les missions confiées puis nous exposerons l'approche envisagée ainsi que les outils utilisés pour répondre aux missions. Enfin, nous proposerons des solutions et détaillerons les processus de réflexion.

Première partie

Présentation du terrain de stage

Le présent stage s'est tenu sur deux sites différents, qui sont les lieux de travail des deux encadrants du stage (Catherine Dominguès, chargée de recherche au LASTIG (IGN) et Philippe Gambette, enseignant chercheur à l'Université Paris-Est Marne-la-Vallée). C'est pourquoi nous ferons ici une présentation des terrains de stage.

1. IGN

1.1. Présentation globale

Fondée en 1940, l'IGN est une entreprise publique française basée à Saint-Mandé. Anciennement nommée l'Institut Géographique National, l 'entreprise conserve son siglemais devient l'Institut National de l'Information Géographique et Forestière en 2012, suite à l'intégration de l'inventaire forestier national.

1.2. Champs d'action

L'IGN a pour missions la description du territoire national, le développement de base de données et d'informations géolocalisées pour assurer l'aménagement du territoire, la prévention des risques et la sécurité nationale.

1.3. Organisation interne

Au sein de l'IGN se trouve une unité de recherche, le LaSTIG. Le LaSTIG, Laboratoire en Sciences et technologies de l'information géographique, regroupe le service de recherche en sciences de l'information géographique (SRSIG) de l'IGN ainsi que des enseignants chercheurs de l'Université Paris Est Marne la Vallée et de l'Ecole Nationale des Sciences Géographiques. Il est divisé en quatre équipes qui sont les suivantes :

— Acquisitions et Traitements
— Médiation et enrichissement d'information Géographique (MEIG)
— Structures Spatio-Temporelles pour l'Analyse des Territoires (STRUDEL)
— GéoVisualisation

2. UPEM

L'Université Paris-Est Marne-la-Vallée (UPEM) est un établissement pluridisciplinaire créé en 1991 et localisé principalement à Champs-sur-Marne. L'université compte 15 unités de recherche, dont le LIGM.

2.1. LIGM

Le LIGM (Laboratoire d'Informatique Gaspard-Monge) est un laboratoire d'informatique placé sous la co-tutelle du CNRS, de l'École des Ponts Paristech, de l'ESIEE Paris et de l'UPEM. Le LIGM est divisé en cinq axes de recherche :

1. Algorithmes, architectures, analyse et synthèse d'images
2. Combinatoire algébrique et calcul symbolique
3. Logiciels, réseaux et temps-réel
4. Modèles et algorithmes
5. Signal et communications

Le présent stage est rattaché au thème de recherche "Algorithmique pour la bioinformatique et algorithmique du texte" de l'équipe "Modèles et algorithmes".

Deuxième partie

Objectifs (cahier des charges)

3. Étude du besoin

Les données utilisées par le PGHM sont des corpus de natures diverses (itinéraires de randonnée, guides touristiques, etc.). En fonction du type du corpus, de son niveau de langue, de son lieu ou sa date de création, des phénomènes de variations d'écriture peuvent être observés, notamment sur les noms de lieux. Par exemple, dans un itinéraire de randonnée, lorsqu'une même ville est répétée plusieurs fois, elle peut être amenée à subir une troncature (par exemple, « Pralognan » pour « Pralognan-la-Vanoise ») afin de faciliter la lecture, de rendre le propos plus fluide et moins redondant pour le lecteur.

Comme expliqué plus haut, ce stage a pour objectif principal de fournir un outil contribuant à la résolution du problème de variation des données que le PGHM utilise lors de la localisation des victimes en montagne en identifiant des noms de lieux dans des données hétérogènes.

Le besoin de repérer des toponymes dans un corpus, à la manière des systèmes de reconnaissance d'entités nommées, tout en tenant compte des variations d'écriture présentes sur les toponymes est partagé par différents projets dans lesquels interviennent les encadrants du stage : MATRICIEL [1] (Brando, Dominguès, & Capeyron, 2016), Cité des Dames [2], PARVIS [3].

En somme, il s'agit de créer un outil qui, quel que soit le type du corpus, repère automatiquement chaque toponyme du texte et lui associe son référent (forme normée) dans un dictionnaire de noms de lieux.

1. https ://psigehess.hypotheses.org/peps-matriciel
2. https ://citedesdames.hypotheses.org/42
3. https ://parvis.hypotheses.org/

4. Méthodologie pour répondre à la demande

La réponse à la demande initiale s'organise autour de deux grands axes que nous détaillerons ici.

4.1. Identification de noms de lieux

4.1.1. Définition

L'identification des noms de lieux constitue le premier axe. Identifier un nom de lieu revient à associer, à un toponyme donné, une forme normalisée qui se rapporte à la version « officielle » du toponyme, celle que nous pouvons retrouver dans les dictionnaires de noms propres (le terme « version officielle » est à considérer avec précaution. En effet, d'une part les noms de lieux évoluent dans le temps, d'autre part (Dominguès & Eshkol-Taravella, 2013) souligne que « des règles d'écriture existent, mais elles sont compliquées, subtiles et non homogènes ».

Par ailleurs, dans un contexte descriptif, la notion de variation d'écriture est au coeur du problème d'identification de toponymes.

4.1.2. Les variations d'écriture

Les variations d'écriture sont des phénomènes qui modifient la graphie d'un mot. Un mot est porteur de variation lorsque sa forme diffère de celle trouvée dans la plupart des dictionnaires. En conséquence, la variation peut prendre des formes variées qui dépendent de l'intention du producteur de l'énoncé et/ou dépend du type du corpus dont elle est issue. Par exemple, dans le contexte d'énoncés courts avec un nombre restreint de caractères (comme dans Twitter), le producteur peut choisir d'utiliser des formes abrégées.

Panckhurst établit en 2009 une typologie des variations identifiables dans les SMS, qui complète les typologies déjà constituées par Anis (2004), Fairon et al. (2006),Véronis et Guimier de Neef (2006) et Liénard (2007). Et bien que ces typologies se rapportent à des corpus de SMS et de corpus issus des « nouvelles formes de communication écrite »,elles sont le reflet des variations que nous pouvons retrouver de manière générale dans tout type de corpus et sur toutes les catégories grammaticales de mot, les noms propres et les toponymes notamment.

Dans la démarche d'identification du toponyme, nous tenons compte de la variation que peut porter le toponyme, en procédant à une reconnaissance de cette variation. Les

types de variations étant nombreux et dépendant des corpus,nous avons fait le choix de nous focaliser sur un certain nombre de ces variations, observées dans les corpus étudiés (voir la section sur les corpus p. 19 par les étudiants de l'ENSG lors du stage qui a précédé le présent stage. Ces variations sont la troncature, le squelette consonantique, l'initiale sur le dernier mot et la simplification. Le tableau 4.1 fait la description de chacune d'elles.

Nom de la variation	Description	Illustration
Troncation	Omission d'un ou de plusieurs constituants du toponyme polylexical (l'omission de mot-outil ou de nom générique géographique n'est pas comptée comme troncation)	Beaujeu-Saint-Vallier-Pierrejux-et-Quitteur : *Beaujeu-Saint-Vallier-Pierrejux* ou bien *Pierrejux-et-Quitteur*, ou bien *Beaujeu*, ou bien *Vallier-Pierrejux*, etc.
Initiale	Le dernier constituant du toponyme polylexical est tronqué : seule son initiale est conservée	Marne-la-Vallée : *Marne-la-V*
Squelette consonantique	Suppression des voyelles du toponyme. Si la première lettre est une voyelle, elle est conservée. Les noms génériques géographiques et les mots-outils réduits à leur squelette consonantique ne sont pas pris en compte.	Marne-la-Vallée : *Mrn-la-Vll* Albertville : *Albrtvll*
Simplification	Un mot est simplifié lorsqu'il répond à au moins un des critères suivants : -omission de mot-outil -casse aléatoire (toponyme tout en minuscules, ou bien majuscules au milieu du toponymes, etc.) -omission/substitution/insertion de signe diacritiques -omission/substitution/insertion de signe de ponctuation -omission d'un caractère répété	Marne-la-Vallée : *Marne la Vallée* ou bien *Marne-la-Vale* ou bien *marNe-vAlléE*, etc.

FIGURE 4.1. – Les variations d'écriture

Nous avons fait le choix de ces variations car elles reflètent toutes un phénomène qui est difficile à appréhender, celui de l'abrégement. En effet, tant du point de vue humain que celui de la machine, lorsqu'un mot est abrégé, il est difficile (sans délimitation du contexte et connaissances extra-linguistiques) de détecter d'une part que quelque chose a été supprimé, et de déterminer d'autre part ce qui a été supprimé. Ces variations constituent donc un enjeu dans l'identification de toponymes, et d'autant plus qu'elles concernent aussi bien un abrégement de lettre(s) (squelette consonantique, simplification) que de mot(s).

Par ailleurs, le tableau 4.1 met en évidence le fait que les noms génériques géographiques [1] et les mots outils [2] constitutifs de toponymes ne sont pas pris en compte lors de la détection d'une variation.

4.1.3. État de l'art

À ce jour, il n'existe pas dans la littérature scientifique, d'outils d'identification de toponymes pour le français qui tiennent compte des variations d'écriture. Certaines recherches se sont penchées sur la question, comme (Kogkitsidou, 2018), qui traite les variations de casse par normalisation, tandis que (Dominguès & Eshkol-Taravella, 2013) prennent en compte les variations décrites dans les nouvelles formes de communications écrites et les élargissent : les erreurs de frappe, les abréviations, la suppression ou la substitution des séparateurs, signes diacritiques, prépositions et déterminants, la création lexicale, la transcription phonétique d'un accent régional.

Néanmoins, le nombre de variations abordées reste succinct.

4.2. Repérage des noms de lieux

4.2.1. Définition

À l'instar des systèmes de reconnaissance d'entités nommées, le repérage de toponymes fait partie intégrante du processus d'extraction d'informations dans un texte et se confronte aux mêmes difficultés, à savoir des difficultés propres à l'écrit telles que la segmentation et l'orthographe.

Avant toute chose, le premier défi est d'appréhender la structure du texte et sa construction syntaxique. Extraire des informations depuis un corpus de SMS requiert des méthodes de segmentation du texte différentes de celles employées pour un texte sous format TEI par exemple.

Le second défi est celui imposé par les mots qui, sujets des variations, présentent une graphie non standard.

Nous avons consulté la littérature scientifique afin d'identifier les méthodes existantes qui traitent du repérage d'entités nommées dans le contexte de corpus non-standard.

4.2.2. État de l'art

Le repérage de toponymes et plus largement la reconnaissance d'entités nommées (REN) est l'une des principales missions du TAL. De nombreuses applications ont recours à la REN comme l'analyse de sentiments, la fouille de textes, la traduction automatique, etc.

1. Un nom générique géographique désigne une entité géographique (route, lac, prison, etc.). La liste complète est fournie en annexes 8.4

2. Un mot-outil est un mot grammatical plus ou moins vide de sens (déterminants, conjonctions, prépositions etc.). À cette liste, est ajouté le lemme « saint ». La liste complète est fournie en annexe 8.4

Il existe de multiples systèmes de REN, reposant sur des approches différentes. Kogkitsidou (2018) dénombre trois techniques de REN : les techniques d'apprentissage statistiques, les techniques symboliques à base de règles et les techniques hybrides. La première technique de REN, apparue dans les années 1990, est celle fondée sur les automates, à savoir la technique symbolique. Grâce à l'utilisation de transducteurs, ce système s'avère être un outil précis en terme de qualité de reconnaissance mais un peu moins pour ce qui est de l'exhaustivité. Nous pouvons citer*Exoseme* (Landau, Sillion, & Vichot, 1993) comme système de REN pour le français.

Avec la mise à disposition de grandes quantités de données, les systèmes statistiques émergent et viennent compléter les méthodes symboliques. Il s'agit de modèles d'apprentissage guidés par les données, qui nécessitent une intervention du concepteur sur les données, par le biais d'annotations par exemple (Nouvel, Ehrmann, & Rosset, 2015). *Answer Extraction* (Abney, Collins, & Singhal, 2000) est un système d'apprentissage.

Par la suite, les systèmes hybrides font leur apparition. Ils sont dits hybrides car ils allient connaissances et données : le concepteur agit tant sur les données que sur le modèle. Le système *Nemesis* (Fourour, 2002) en est un exemple.

Toutefois, ces modèles sont difficilement utilisables sur des données différentes de celles sur lesquelles ils ont été entraînés. La plupart des systèmes de REN sont conçus pour être performants sur des corpus qui respectent les règles d'orthographe et de typographie, comme des corpus journalistiques par exemple. Or l'intérêt de ce stage est justement de repérer des noms de lieux dans des données bruitées, instables de par leurs variations.

(Kogkitsidou, 2018) propose une approche de REN pour langage non-standard qui préconise au préalable une normalisation morphosyntaxique du corpus puis un module de normalisation de casse. Elle applique ensuite séparément *OpeNER* (apprentissage automatique), *CasEN* (ingénierie des grammaires) et *mSX* (hybride) sur des corpus bruts puis sur les mêmes corpus mais normalisés. Suite à la comparaison des résultats, elle en déduit que la normalisation permet d'augmenter les performances de REN.

(Dominguès & Eshkol-Taravella, 2013) présentent une façon différente d'aborder le langage non-standard et se penchent plus précisément sur la question des toponymes et leur repérage. La méthode de repérage présentée a recours à une base de données lexicales (BDNyme). Elle est complétée par l'utilisation de patrons qui exploitent plusieurs indices, notamment les noms génériques de lieux et les verbes, noms et prépositions locatives. Les résultats de l'évaluation d'un telle approche laissent penser que le typage des toponymes par des patrons a augmenté le nombre de toponymes correctement détectés.

4.3. Approche envisagée

La première partie de notre travail s'attache à identifier des toponymes dans un corpus, tout en tenant compte des variations que le toponyme est susceptible de porter. Cette partie est indépendante du repérage de noms de lieux, car l'identification se fait à partir d'un corpus déjà annoté en noms de lieux. Dans le processus d'identification, nous distinguons deux missions :

— à partir d'un corpus, faire le lien entre tout segment de toponyme et les toponymes qui le contiennent ;
— associer à un toponyme du corpus, le référent normé le plus similaire au toponyme

Quelle que soit la mission, il est nécessaire de recourir à une ressource qui répertorie des toponymes dans leur forme normée. Cette ressource est un gazetier de toponymes, constitué à partir de plusieurs extraits de bases de données géographiques.

Une fois l'objectif d'identification atteint, nous irons plus loin dans notre réflexion en procédant au repérage de toponymes. Cette fois, le support sera un corpus non-annoté en noms de lieux.

En somme, dans un premier temps, la partie identification et la partie repérage sont traitées indépendamment, l'une ne dépendant pas de l'autre. Elle sont par conséquent évaluées chacune séparément, au moyen des mesures de précision et de rappel. Dans un second temps, la chaîne de traitement combine les deux parties, en identifiant les toponymes repérés automatiquement par le programme. Les toponymes dans ce cas sont repérés et identifiés dans des corpus non-annotés au préalable.

À terme, le traitement suit l'enchaînement montré par la figure 4.2 :

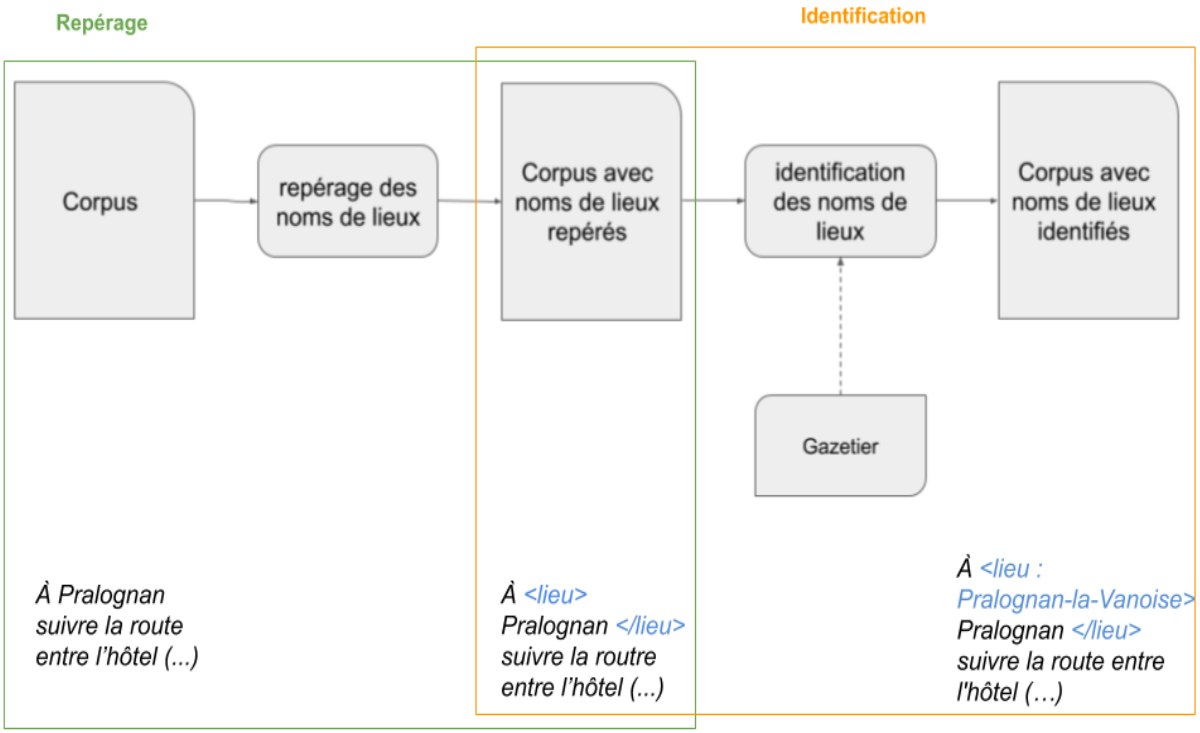


FIGURE 4.2. – Les grandes étapes du traitementdu repérage et de l'identification combinés

5. Ressources et outils utilisés

5.1. Les corpus

Les corpus que nous avons exploités sont des corpus de test. C'est sur les corpus que nous appliquons et évaluons les systèmes d'identification et de repérage de toponymes. L'étape d'identification de toponymes nécessite un corpus de toponymes, c'est-à-dire une liste dans laquelle des toponymes (porteurs de variation ou non) sont associés à leur référent normé. Notre but est de retrouver le bon référent normé de chaque toponyme de la liste. Pour la partie repérage de toponymes, nous avons besoin de textes annotés en toponymes. Nous repérons des mots dans les textes et pour savoir si ces mots sont bien des toponymes, nous les comparons aux toponymes annotés dans les textes. Quant à la dernière partie qui combine repérage et identification, elle implique le recours à un corpus dans lequel les toponymes sont annotés et associés à leur référent normé.

Nous avons étudié chacun des toponymes des corpus pour savoir sous quelle forme les toponymes apparaissent et si des variations d'écriture étudiées dans le cadre du stage sont observables sur ces toponymes (pour connaître le résultat de cette étude, voir l'annexe p. 57).

5.1.1. La Grande Peur

Le corpus de la Grande Peur (Paris, Abadie, & Brando, 2017) [1]est extrait d'un récit (Lefebvre, 2014) portant sur le mouvement de révolte ayant eu lieu en France en 1789. Le texte, au format TEI [2] est annoté en toponymes et tous les toponymes sont associés à leur référent normé. Ils sont encadrés par des balises « <placeName> ». Dans la balise ouvrante se trouve la version normée du nom de lieu balisé ainsi que l'adresse URL DBPedia [3] d'où provient cette forme du nom de lieu normalisé . Voici un exemple de nom de lieu balisé issu du corpus :
<placeName ref="http ://fr.dbpedia.org/resource/Romilly-sur-Seine">Romilly </placeName>

Le corpus La Grande Peur a servi de corpus de test pour la partie de repérage de toponymes et la partie qui combine repérage et identification. L'étude des toponymes contenus dans ce corpus (174 toponymes au total) révèle une faible quantité de variations. Les variations observées sont la simplification (omission de tiret comme Château-

1. un extrait des toponymes issus de ce corpus est disponible à l'annexe 8.4
2. Le TEI, Text Encoding Initiative, est une norme qui énumère des prescriptions et de bonnes pratiques pour standardiser l'encodage. Ces principes sont traduits en XML.
3. DBPedia est une base de données qui met à disposition les données Wikipédia au format du web sémantique.

Renard devenu Châteaurenard ou au contraire ajout de tirer comme Champvallon devenu Champ-vallon) et quelques troncations (Romilly-sur-Seine devenu Romilly) .

5.1.2. 88milSMS

En 2011, une équipe pluridisciplinaire de chercheurs crée un corpus composé de plus de 88 000 SMS en français collectés dans le contexte des nouvelles formes de communication écrite. Ce corpus, intitulé « 88milSMS. A corpus of authentic text messages in French » (ou « 88milSMS ») (Panckhurst et al., 2016), a été constitué dans le cadre du projet *sms4science*, coordonné par le CENTAL (Centre de Traitement Automatique du Langage de l'Université catholique de Louvain, Belgique) et vise à étudier les phénomènes linguistiques véhiculés à travers les SMS.

Les SMS ont été anonymisés semi-automatiquement. Dans le but d'utiliser le corpus pour procéder au repérage seul et au repérage et à l'identification simultanés de toponymes, nous avons extrait manuellement 191 noms de lieux à partir de 2000 SMS. À chaque nom de lieu repéré dans le corpus, nous avons associé une forme normalisée.

À la fin du traitement, un corpus de sms non annoté et une liste de lieux extraits [4] de ces sms sont disponibles.

Parmi les variations étudiées au cours du stage, nous retrouvons dans ce corpus la simplification (comme Montpellier qui perd sa majuscule) et la troncation (comme Carnon-plage tronqué en Carnon).

5.1.3. BNF

La Bibliothèque nationale de France (BnF) possède un catalogue consultable en ligne et qui recense les documents qu'elle conserve tels que des imprimés, des documents sonores, des documents cartographiques, etc. Chaque document recensé est décrit par une notice (notices de références bibliographiques, d'auteurs, de noms communs, de noms géographiques, etc.). Dans une démarche d'ouverture des données, la BnF a regroupé toutes ces informations sur un site web, *data.bnf.fr*. Par le biais de ce site, la BnF facilite ainsi l'échange et la réutilisation des données et métadonnées issues de son catalogue. Le projet data.bnf (Wenz, 2013) utilise les outils du Web Sémantique (RDF) afin de rassembler les données, que nous avons interrogées à l'aide de requêtes SPARQL.

Le corpus [5] constitué est donc un corpus de toponymes associés à leurs propriétés. Il nous a été utile au moment de l'observation des variations présentes sur les toponymes. Lors de cette observation, nous avons constaté que de nombreux toponymes du corpus de la BnF étaient porteurs de variations dont nous ne connaissons pas le type car elles ne font pas partie des variations étudiées dans ce stage. Cela nous a poussé à envisager la possibilité d'appréhender dans notre travail une nouvelle variation, la variation d'une lettre (pour en savoir plus sur la variation d'une lettre, consulter la p. 59)

4. un extrait des toponymes issus de ce corpus est disponible à l'annexe 8.4
5. un extrait des toponymes issus de ce corpus est disponible à l'annexe 8.4

5.1.4. Renumar

Le projet Renumar (ressources numériques pour l'édition d'archives de la renaissance) (Van de Weerd, 2018) traite les données issues de la base de données « De minute en minute » en offrant une structure plus adaptée à la recherche. Cette base de données répertorie des transcriptions et des descriptions d'actes notariaux datant du XVe et XVIe siècles. Ces actes sont annotés en personnes et en lieux.

Pour chaque toponyme, un lien vers Geonames [6] est indiqué et permet d'accéder à la forme normée du toponyme. Nous avons collecté plus de 600 noms de lieux associés à leur forme normée. [7]

Comme pour le corpus BnF, les toponymes issus du corpus Renumar sont nombreux à porter des variations différentes de celles appréhendées dans ce stage. Le corpus a donc lui aussi servi pour l'étude de la variation d'une lettre.

5.1.5. MATRICIEL

Le corpus [8] a été élaboré dans le cadre du projet collaboratif MATRICIEL (Dominguès, Weber, Brando, Jolivet, & Van Damme, 2017). MATRICIEL s'attache à analyser des récits de vie de migrants républicains espagnols exilés en France entre 1936 et 1939. Dans ce but, les récits de vie sont transcrits puis des outils sont conçus afin d'en faciliter l'analyse. L'attention est portée plus particulièrement sur les lieux énoncés dans ces récits, afin de les identifier puis de les cartographier et reconnaître les sentiments associés.

Le corpus des récits de vie construit pour ce projet a été annoté (noms communs de lieux, noms propres de lieux, lexique du sentiment) automatiquement grâce à des plugins intégrés dans une chaîne de traitement GATE [9].

Les annotations permettent d'associer à chaque nom de lieu du texte, des traits d'information extraits des gazetiers comme le pays où est localisé le lieu désigné par ce nom, le continent, les coordonnées du lieu, la langue correspondant à cette forme du toponyme, etc. Néanmoins, la forme normée des toponymes n'apparaît pas. Or au moment de traiter la partie identification des toponymes, le corpus Matriciel était le seul que nous possédions. Un prétraitement sur le corpus s'imposait, afin de disposer de toponymes et de leur forme normée. Nous avons alors mis à jour notre corpus de référence en rajoutant pour chaque nom de lieu, son référent.

Les toponymes du corpus n'ont en majorité pas de variation d'écriture. Seule la troncation a été observée.

5.1.6. Corpus non-utilisés

Les corpus jouent un rôle central dans le TAL. Ils servent à tester la cohérence et à évaluer la performance des outils que nous concevons. Pour qu'il soit efficace, ce test doit être réalisé sur une grande quantité de corpus. C'est pourquoi nous avons tenté de

6. GeoNames est une base de données géographiques consultable sur `https://www.geonames.org/`
7. un extrait des toponymes issus de ce corpus est disponible à l'annexe 8.4
8. un extrait de ce corpus est disponible à l'annexe 8.4
9. GATE est une plateforme mettant à disposition des outils d'exploitation de corpus

réunir un maximum de corpus. Néanmoins, les textes recueillis ne sont pas tout le temps exploitables. C'est le cas des corpus qui vont suivre.

CarteALaCarte

« Carte à la carte » est un service en ligne proposé par l'IGN qui permet aux utilisateurs de créer et de personnaliser leurs propres cartes, le choix du titre constituant une des personnalisations possibles. Le repérage de ces titres est décrit dans (Dominguès & Eshkol-Taravella, 2015). Le sujet des variations d'écriture est abordé dans cette étude et un corpus de titres de cartes est construit. Sont indiqués pour chaque titre de carte le nom de lieu détecté ainsi que la variation d'écriture qu'il porte. Ce corpus n'a pas été retenu car il contient majoritairement des changement de casse et des omissions ou substitutions de signes de ponctuations.

ETAPE

ETAPE (Gravier et al., 2012) est une campagne d'évaluation menée sur divers corpus. Les résultats de ces campagnes sont souvent disponibles sur la plateforme ELRA (European Language Resources Association, plateforme offrant des services en rapport avec les ressources linguistiques). L'échantillon de corpus que nous souhaitions exploiter contient des transcription d'émissions radio. Les fichiers sont annotés en entités nommées. Malheureusement, puisqu'il s'agit de transcriptions d'entretiens oraux nous n'avons pas relevé de variations (mis à part des troncatures des noms de lieu, déjà observées dans d'autres corpus.

ACSYNT

ACSYNT (Delais-Roussarie et al., 2004) est un corpus oral contemporain disponible sur Ortholang. Il comprend de la lecture oralisée de texte, des présentations monologuées et des entretiens guidés. Après analyse du contenu de ces trois types de données orales, nous avons identifié la présence de toponymes dans les textes relatifs aux entretiens. Toutefois, les textes ne sont pas annotés et les noms de lieux n'ont pas de variations d'écriture.

Corpus 14-Praxiling

Le projet « Corpus 14 » (Steuckardt, 2017)soutenu par le laboratoire Praxiling (Université Paul-Valéry Montpellier) a pour objet d'étude les correspondances des soldats lors de la Première Guerre Mondiale. Les correspondances sont accessibles sur la plateforme TXM (plateforme d'analyse textométrique). Cependant, les corpus mis à disposition ne sont pas encore annotés. L'annotation est actuellement en cours.
Ce corpus présente un intérêt pour le stage en terme de variations d'écriture car certains toponymes utilisés dans ces correspondances diffèrent de leur version actuelle par leur construction et leur graphie. Il serait intéressant d'étudier les règles, si elles existent, qui régissent l'évolution graphique des toponymes du début du XXème siècle à aujourd'hui.

CHOUCAS

Le corpus CHOUCAS [10] est composé d'itinéraires de randonnées sous format XML (langage informatique de balisage) et dans lequel des noms de lieux sont annotés. Nous n'avons pas exploité ce corpus pour deux raisons. D'une part, pour chaque nom de lieu annoté, nous ne disposons que du toponyme du corpus, la version normée n'est pas indiquée. Par ailleurs, la nature des lieux annotés est très riche. Sont annotés comme noms de lieux beaucoup de rues, des places, des garages, des chapelles, etc. D'autre part, la quantité de variations d'écriture observées est minime et il ne s'agit pour la plupart que de troncatures.

5.2. Les gazetiers

Un gazetier est une ressource, souvent sous forme de liste, constituée d'entités appartenant à un domaine spécifique (par exemple, un gazetier de toponymes). Ici, ils sont utilisés comme base de données que nous consultons lors de l'étape de l'identification. C'est dans les gazetiers que nous cherchons les référents normés. Par conséquent, la taille du gazetier (et donc le nombre de toponymes qui le constituent) se doit d'être suffisamment grande, afin de permettre l'identification d'un maximum de toponymes. C'est pourquoi nous avons créé un gazetier unique composé de toponymes issus de plusieurs bases de données.

5.2.1. BDNyme

Produite par l'IGN, la BD NYME est un extrait de la base de données BD TOPO® (description vectorielle 3D des éléments d'un territoire tels que adresses postales, hydrographie, unités administratives, réseau routier, etc.), et couvre l'ensemble du territoire français métropolitain. Chaque entrée contient un toponyme et les coordonnées du lieu désigné. Cette ressource nous a été utile pour les chefs-lieux qu'elle contient. Néanmoins, après plusieurs tests sur différents corpus, il s'est avéré que la quantité de chefs-lieux BDNYME dont nous disposions n'était suffisante. Nous avons donc remplacé ces chefs-lieux par ceux provenant du jeu de données GEOFLA.

5.2.2. Geofla

Geofla est une base de donnée produite par l'IGN qui fournit des descriptions d'unités administratives de France métropolitaine et DOM à des échelles régionales, nationales, etc. Les données sont disponibles au format shapefile, en projection Lambert-93. C'est sur les chefs-lieux que nous avons porté notre intérêt. Notre jeu de données était initialement un shapefile, composé d'une représentation géographique des chefs-lieux du territoire français. Par conséquent, un logiciel était nécessaire pour exploiter ce fichier et en extraire des données sous format textuel. Nous avons donc eu recours à QGIS desktop, un logiciel consacré aux systèmes d'informations géographiques permettant notamment

10. un extrait de ce corpus est disponible à l'annexe 8.4

de visualiser et d'analyser des données spatiales. Le fichier de sortie obtenu à partir du shapefile de chefs-lieux après passage dans QGIS [11] est un fichier csv qui pour chaque chef-lieu indique les propriétés suivantes :

— des coordonnées géographiques x et y
— un identifiant
— le nom du chef-lieux
— son statut (arrondissement ou commune)
— son INSEE_COM

5.2.3. GeoNames

GeoNames est une base de données géographiques consultable sur internet et qui comprend plus de 25 millions de toponymes dont 11 millions de noms de lieux géoréférencés. Les toponymes proviennent de diverses sources, dont l'IGN pour le territoire français. GeoNames est aussi enrichie par ses utilisateurs, qui peuvent ajouter ou corriger des toponymes. Les toponymes géoréférencés sont associés à diverses données telles que les coordonnées géographiques du lieu en projection WGS84 [12], son identifiant GeoNames, sa population, son code postal, etc. Des formes alternatives du toponyme, correspondant à différentes langues, sont parfois proposées.

5.2.4. Prétraitements communs à tous les jeux de données

En vue d'obtenir un gazetier unique [13], les gazetiers GeoNames et GEOFLA ont été combinés. Pour chaque entrée du gazetier résultant, nous obtenons les propriétés suivantes :

— le nom du toponyme ;
— son identifiant, unique pour chaque toponyme dans la base de données source ;
— sa source (GeoNames, GEOFLA) ;
— ses coordonnées géographiques

Une entrée du gazetier ainsi formé se présente de la façon suivante :
Toponyme [toponyme : identifiant : source dont est issu le toponyme (coordonnées géographiques)]
Il est possible de trouver des toponymes identiques dans la même base, mais qui correspondent à des lieux différents. Par exemple :
Saint-Nazaire [**Saint-Nazaire** : PAIHABIT0000000038521159 : BDNYme (2.9906597,42.666954), **Saint-Nazaire** : PAIHABIT0000000095087142 : BDNYme (4.6223545,44.198536), **Saint-Nazaire** : PAIHABIT0000000028857311 : BDNYme (-2.2239106,47.27336)]

11. https ://www.qgis.org/fr/site/
12. « Système de Référence Terrestre réalisé par géodésie spatiale par le département de la défense américain qui est inhérent au Global Positioning System qui l'utilise pour ses éphémérides radiodiffusées. Il existe plusieurs réalisations du WGS84 avec des différences décimétriques. Il est à présent très proche des réalisations ITRF. » https://geodesie.ign.fr/?p=72page=glossaire
13. un extrait du gazetier est disponible à l'annexe 8.4

5.3. Les mesures d'évaluation

5.3.1. Précision

La mesure d'évaluation pour laquelle nous avons opté est la précision. La précision calcule le nombre d'éléments pertinents trouvés (les vrais positifs) par rapport à l'ensemble des éléments sélectionnés.

$$\frac{\text{vrais positifs}}{\text{éléments sélectionnés}}$$

Il s'agit là de mesurer la qualité des données trouvées sur une échelle de 0 à 1. Une précision de 0.9 est une précision presque parfaite (avec une marge d'erreurs de 10%).

Dans le cas de l'évaluation de l'identification, la précision permet de quantifier, la proportion de toponymes du corpus auxquels est associé le bon référent normé.

Initialement, nous appelions "bon référent normé" un candidat dont l'identifiant était le même que celui du toponyme du corpus. De cette manière, nous incrémentions de un le nombre de vrais positifs dès lors qu'un toponyme du corpus avait le même identifiant que l'un des candidats qui lui étaient associés. Mais penser de cette façon revenait à procéder à une désambiguïsation géographique. En effet, chaque toponyme possède un identifiant unique. Paris au Texas n'aura pas le même identifiant que la capitale française. Il faudrait donc être en mesure de déterminer que pour un toponyme situé à un endroit précis du corpus, il s'agit de Paris en France. Afin d'apporter une solution à cette problématique nous pourrions prendre en compte le contexte du corpus, en considérant la sémantique du texte, et recourir à un gazetier qui couvre la même zone géographique que le corpus. Toutefois, nous avons fait le choix de ne pas nous confronter à la désambiguïsation géographique car, bien qu'intéressant, c'est un domaine trop éloigné du sujet du stage.

Pour l'évaluation de la partie repérage, la précision indique, parmi tous les mots repérés comme toponymes, la part de mots qui sont réellement des toponymes.

5.3.2. Rappel

Le rappel est le révélateur de l'exhaustivité des données trouvées. Il calcule le nombre d'éléments pertinents trouvés par rapport à l'ensemble des éléments pertinents.

$$\frac{\text{vrais positifs}}{\text{éléments pertinents}}$$

Pour l'identification, il n'est pas calculé car un référent normé est attribué à tous les toponymes. En revanche, le rappel du repérage révèle la quantité de toponymes du corpus qui auraient dû être repérés mais qui ont été manqués.

Troisième partie

Réponse au cahier des charges

6. Identifier des toponymes

6.1. Introduction

Nous nous sommes dans un premier temps concentrés sur l'identification de toponymes éventuellement porteurs de variations d'écriture. Pour cette partie, le corpus d'entrée est constitué de séquences qui sont considérées comme des noms de lieu qui diffèrent éventuellement de la forme normée, figurant dans le gazetier. Le premier prototype vise à attribuer un référent normé à chaque séquence. L'identification de la variation est aussi fournie pour chaque nom de lieu reconnu. Par exemple, pour la séquence du corpus « Marne la V », le programme désignera le toponyme « Marne-la-Vallée » comme possible référent normé en indiquant la présence de la variation « initiale » sur le nom de lieu du corpus.

6.2. Méthode

Pour mener à bien cette tâche, nous avons recours au gazetier de toponymes formé auparavant. Chaque séquence est comparée à toutes les entrées du gazetier, le but étant d'obtenir un match exact [1] entre la séquence du corpus et l'entrée du gazetier. En revanche, de cette façon, seuls les noms de lieux normés trouveront un référent normé, puisque le gazetier est composé de toponymes normalisés. C'est pourquoi nous avons prétraité le gazetier.

6.2.1. Prétraitement du gazetier et des toponymes du corpus

Pour rappel, notre gazetier est un ensemble de toponymes dont les sources sont diverses (GEOFLA et GeoNames). Pour chaque toponyme, l'identifiant, la source et les coordonnées géographiques sont mentionnés.

Faire le lien entre un toponyme du corpus avec variation et un nom de lieu du gazetier sous sa forme normée implique quelques modifications du gazetier en particulier. Pour qu'un nom de lieu avec variation coïncide avec une entrée du gazetier, nous avons adapté le gazetier aux variations en créant un gazetier prétraité pour chaque type de variations : la simplification, la troncature, le squelette consonantique, et l'initiale. Chaque entrée du gazetier prétraité est obtenue en appliquant la transformation correspondante, au toponyme du gazetier initial. Le tableau 6.1 présente les transformations appliquées selon le type de variation.

1. il y a match exact entre deux mots lorsqu'ils sont identiques

Nom de la transformation	Processus de transformation	Illustration
Troncation	1) mise en minuscules 2) suppression de la ponctuation et des espaces 2) suppression des mots outils et des noms génériques géographiques 3) suppression des diacritiques 4) suppression des caractères répétés 5) suppression d'un ou plusieurs mots	Pralognan-la-Vanoise : *pralognan* ou *vanoise* Beaujeu-Saint-Vallier-Pierrejux-et-Quitteur : *beaujeuvalierpierejux* ou bien *pierejuxquiteur*, ou bien *beaujeu*, ou bien *valierpierejux*, etc.
Initiale (seulement sur les toponymes polylexicaux)	1) mise en minuscule 2) suppression de la ponctuation et des espaces 3) suppression des mots outils et noms génériques géographiques 4) suppression des caractères répétés 6) suppression des diacritiques 5) pour le dernier token, préservation uniquement de l'initiale	Noisy-le-Grand : *noisyg*
Squelette consonantique	1) mise en minuscules 2) suppression des mots outils, des noms génériques géo, de la ponctuation, des espaces, des diacritiques et caractères répétés 3) suppression des voyelles (si le mot commence par une voyelle, on ne la supprime pas)	Paris : *prs* Noisy-le-Grand : *nsgrnd*
Simplification	1) mise en minuscules 2) suppression des mots outils, des noms génériques géo, de la ponctuation, des espaces, des diacritiques et caractères répétés	Pralognan-la-Vanoise : *pralognanvanoise*

FIGURE 6.1. – Transformations appliquées selon le type de variation

Les étapes de transformations relatives à la variation de simplification se retrouvent dans toutes les autres transformations, peu importe le type de variation. À la fin du prétraitement, nous disposons donc d'un gazetier prétraité avec la variation squelette consonantique, un pour la simplification, un pour la variation initiale, un pour la troncation de niveau un, et enfin un dernier gazetier prétraité avec la troncation de niveau supérieur à un. À cette liste se rajoute une version du gazetier où les toponymes ne sont pas transformés. Nous parlerons alors du gazetier « identité ».

Les entrées d'un gazetier prétraité prennent la forme suivante : toponyme prétraité avec variation [nom de lieux original : identifiant : source (coordonnées géographiques)]

Plusieurs toponymes peuvent partager la même forme prétraitée. Par exemple, en appliquant la troncature sur les entrées de notre gazetier original, les toponymes « Noisy-le-Grand », « Noisy-le-Sec » et « Noisy-Rudignon » obtiendront tous « noisy » comme

forme prétraitée. Par conséquent, dans le gazetier prétraité avec la variation troncature, l'entrée « noisy » se présentera de cette manière :
noisy [**Noisy-le-Grand** : PAIHABIT0000000002597894 : BDNYme (2.5599918,48.839775), **Noisy-le-Sec** : PAIHABIT0000000002597792 : BDNYme (2.4578044,48.89007), **Noisy-Rudignon** : PAIHABIT0000000001794082 : BDNYme (2.932189,48.336254)]

Il se peut qu'après transformation, il ne reste plus rien du toponyme. Alors, ce toponyme n'apparaîtra pas dans cette version prétraitée du gazetier. Par exemple, comme il est impossible d'appliquer la variation initiale sur des toponymes non-polylexicaux, « Paris » ne figurera pas dans le gazetier prétraité avec la variation initiale. Cela réduit la taille des gazetiers prétraités, et donc les temps de traitement.

La gestion des variations d'écriture présentes sur les toponymes peut donc se faire en intervenant sur le gazetier, mais aussi sur le corpus. Chaque nom de lieu du corpus est simplifié. Dans le cadre notre travail, les omissions/insertions/substitutions de signes de ponctuation, la casse aléatoire, les espaces en trop etc. seuls ne constituent pas une variation à proprement dit mais peuvent empêcher le match exact avec le gazetier. En simplifiant les mots du corpus, nous procédant en quelques sortes à une normalisation.

6.2.2. Match exact

L'étape suivante est celle du match exact entre le toponyme du corpus et les gazetiers prétraités. En vue de diminuer au possible le temps de traitement, le toponyme du corpus, non simplifié, est d'abord comparé aux entrées du gazetier identité. S'il trouve un nom de lieu du gazetier qui lui est identique, nul besoin de poursuivre la recherche dans les autres gazetiers prétraités. Dans le cas contraire, il est simplifié puis comparé aux entrées des autres gazetiers prétraités.

6.2.3. Le calcul de similarité

Lors du match exact, il est possible que le toponyme transformé du gazetier soit associé à plusieurs toponymes normés (comme nous l'avons vu avec l'exemple de la troncature « noisy »). Il devient donc nécessaire de faire un choix entre les référents normés proposés, en optant pour le plus similaire au toponyme du corpus (non simplifié). La similarité est calculée au moyen de mesures de similarité entre chaînes de caractères.

Les mesures de similarités choisies dans notre étude sont celles qui ont été analysées lors du stage de recherche entrepris par deux étudiants de l'ENSG. Il s'agit des mesures suivantes : le coefficient de Dice(Dice, 1945), la mesure de Jaccard(Hadjieleftheriou & Srivastava, 2010), de Levenshtein(Levenshtein, 1966), de Jaro-Winkler(Cohen, Ravikumar, Fienberg, et al., 2003), de Jaro(Jaro, 1989) et de Needleman-Wunsch(Needleman & Wunsch, 1970).

L'article (Nguyen, Sallaberry, & Gaio, 2013) divise les mesures de similarités en trois

catégories : les méthodes basées sur des caractères (Jaro, Jaro-Winkler, Levenshtein, Needleman-Wunsch), celles sur des tokens (Jaccard et Dice) et les hybrides. Il définit la première méthode de la manière suivante :

> (...) la similarité entre deux chaînes est déterminée par des caractères communs et la position de ces caractères dans les chaînes (Jaro (Jaro, 1989), Jaro-Winkler (Winkler, 1999)) ou par le nombre d'opérations(suppression ; insertion ; remplacement) nécessaires pour construire une chaîne à partir de l'autre (Levenshtein (Levenshtein, 1966), Needleman-wunsch (Needleman et Wunsch, 1970), Smith-Waterman (Smith et Waterman, 1981)). L'inconvénient principal des ces métriques est la non distinction des mots lorsqu'une chaîne en comporte plusieurs.

Levenshtein a un second point faible : son temps de calcul long dû à une complexité de $O(n)$.

Quant aux méthodes basées sur des tokens, il les décrit ainsi :

> considèrent une chaîne comme un ensemble de tokens. Un token est une sous-chaîne de caractères délimitée par des caractères spécifiques (espaces, tirets, . . .).

Les mesures de Dice et Jaccard ne tiennent pas compte de l'ordre des caractères. En outre Jaccard pénalise davantage lorsqu'il y a peu de mots en commun.

Dans le calcul de similarité entre le toponyme du corpus et celui du gazetier, nous utilisons une mesure de similarité par défaut.

Le schéma 6.2 reprend les grandes étapes d'identification.

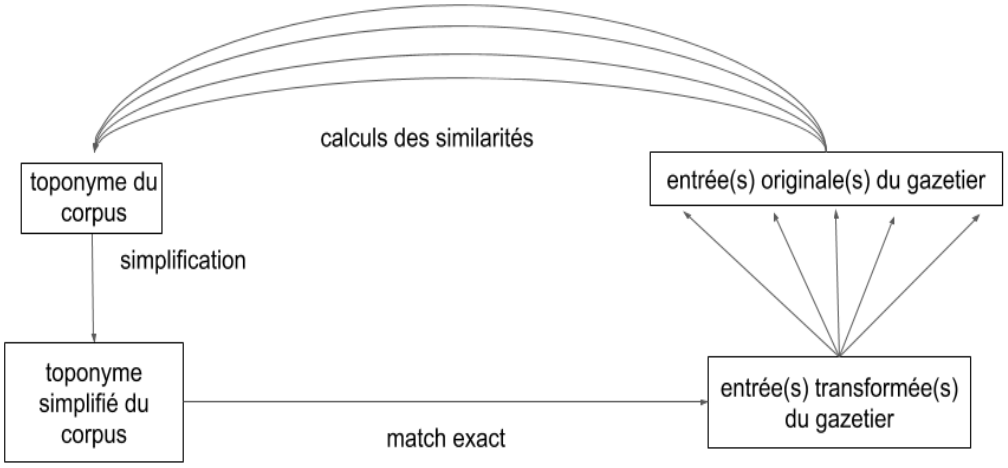


FIGURE 6.2. – Étapes d'identification

6.3. Résultats

Pour évaluer la performance et la pertinence du processus d'identification, nous avons eu recours à un corpus de référence, Matriciel.

Le corpus de référence est annoté en toponymes pour lesquels sont indiqués à chaque fois l'identifiant, le nom, les coordonnées géographiques et la source.

Notre but étant d'associer les bons référents normés au plus grand nombre de toponymes, l'efficacité de notre programme est mesuré par la précision.

Nous cherchons à savoir si un toponyme du corpus a pour nom référent celui d'un de ses candidats.

Le corpus sur lequel nous avons calculé la précision est un extrait du corpus Matriciel, comprenant un total de 351 noms de lieux. Quatre mesures de similarité ont servi à comparer les toponymes du corpus aux entrées des gazetiers prétraités et à fournir le meilleur candidat à chaque token du corpus. Les mesures de similarité sélectionnées sont Jaro-Winkler, Dice, Levenshtein et Jaccard. La précision obtenue est de 0.85 pour les trois premières mesures de similarité. Elle est un peu plus faible pour Jaccard avec 0.79, ce qui reste tout de même correct. La précision est donc bonne : sur les 351 toponymes, 300 toponymes ont été associés au bon candidat. Le rappel n'a pas été évalué car chaque

toponyme du corpus sans exception se voit attribuer un candidat. Le rappel est donc automatiquement de 1.

Mesure de similarité	Précision
Jaro-Winkler	0.85
Dice	0.85
Levenshtein	0.85
Jaccard	0.79

FIGURE 6.3. – Précision obtenue selon la mesure de similarité

6.4. Discussion

Si la précision est aussi bonne, c'est à cause de la nature même du corpus Matriciel. Pour rappel, Matriciel est une transcription de l'oral. Par conséquent, les seules variations linguistiques susceptibles d'être rencontrées dans le corpus sont soit des variations propres à l'oral, telle que la troncature (par exemple, « Argelès » à défaut de « Argelès-sur-Mer »), soit des erreurs propres à la transcription comme des fautes de frappes ou des omissions de diacritiques et ponctuation.

D'autre part, dans le corpus Matriciel, certains toponymes étiquetés contiennent des noms génériques géographiques comme « camp de Mérignac » ou des locutions prépositives spatiales telles que « en direction de Perpignan ». Ceci explique pourquoi 51 toponymes ne sont pas identiques à leur référent associé.

Par ailleurs, la manière de choisir le meilleur référent normé parmi plusieurs référents présente une limite. Prenons pour exemple « noisy » comme toponyme du corpus. Il matche avec l'entrée du gazetier transformé par troncation « noisy ». Parmi les référents normés sont proposés « Noisy-le-Sec » et « Noisy-le-Grand ». Le référent normé qui est associé au toponyme du corpus est celui qui est le plus similaire. Noisy-le-Sec l'emporte automatiquement car comme il contient moins de caractères que Noisy-le-Grand,

il se rapproche davantage de « noisy ». Pourtant, dans le corpus, « noisy » peut être la troncature de Noisy-le-Grand. En l'occurrence, il s'agit d'un cas de figure qui ne peut être résolu à moins d'avoir des connaissances complémentaires sur le corpus, le contexte sémantique, la couverture géographique, etc.

Finalement, ce premier prototype permet de mettre en lumière les rouages du processus d'identification de toponymes ainsi que ces contraintes. Il faut néanmoins approfondir la démarche, en concevant un programme capable d'opérer sur divers corpus et donc d'autres variations.

6.5. Variante du processus d'identification

Une des variantes de l'identification que nous avons envisagée se focalise sur le gazetier de troncatures et répond à un des besoins exprimés par Choucas. En effet, le programme actuel doit être en mesure d'associer à un nom de lieu tous les toponymes du gazetier qui contiennent ce nom de lieu (des candidats).

Pour ce faire, le toponyme du corpus est tronqué à chaque espace ou signe de ponctuation. Les troncatures obtenues sont comparées séparément au gazetier transformé par troncation. Dès qu'il y a match exact avec un toponyme transformé du gazetier, les toponymes normés qui lui sont associés deviennent des candidats du nom de lieu du corpus.

Pour aller plus loin, il serait intéressant de travailler sur les coordonnées géographiques de façon à ne proposer pour un nom de lieu que des toponymes proches géographiquement parlant. En effet, puisque le but de cette variante d'identification est la conception d'un outil facilitant la localisation des victimes par les sauveteurs du PGHM, il n'est pas envisageable que les toponymes proposés soient localisés sur une zone géographique très large. Si la victime déclare se trouver à Lans-en-Vercors (Isère), il est plus pertinent que l'outil d'identification propose Corrençon-en-Vercors (Isère) plutôt que Vassieux-en-Vercors (Drôme).

7. Repérer des toponymes dans un corpus

Pour repérer des toponymes dans un corpus ou bien plus largement pour procéder à l'analyse lexicale d'un texte, il faut avant tout segmenter ce dernier et définir l'unité linguistique minimale sur laquelle portera l'analyse, c'est-à-dire définir le token. En d'autres termes, (Palmer, 2000) définit la tokenisation de la manière suivante :

> La tokenisation est le processus de segmentation d'une séquence de caractères dans un texte en localisant les limites de mots, les points où un mot se termine et un autre commence. Pour les besoins de la linguistique informatique les mots ainsi identiés sont souvent appelés tokens. Dans les langues écrites où aucune limite de mots n'est explicitement marquée dans le système d'écriture, la tokenisation est également connue comme la segmentation de mots.

Dans ce chapitre nous testons deux façons de segmenter le texte, l'une avec des mots déclencheurs, l'autre avec une liste de mots dits impossibles.

7.1. Des mots déclencheurs

7.1.1. Introduction

Nous avons pensé, dans un premier temps, recourir à un système de mots déclencheurs. Nous avons fait l'hypothèse que chaque mot déclencheur est suivi d'un toponyme. Ainsi, un toponyme est un mot introduit par un mot déclencheur.

7.1.2. Méthode

La première étape est celle de la constitution de la liste des mots déclencheurs. Celle-ci est composée des noms génériques géographiques, de prépositions spatiales (« sur », « vers », etc.) et de locutions prépositives de lieu (« à côté de » par exemple).

Une fois la liste établie vient la tokenisation du corpus.

Dans cette étape de segmentation sont considérés comme segmenteurs les retours à la ligne et les signes de ponctuation (« ., :;!?"()#/[] »). Le tiret et l'apostrophe ne sont pas considérés comme des segmenteurs car il s'agit de séparateurs couramment observés dans les noms de lieu. De même que l'espace qui est aussi un séparateur observé dans les toponymes polylexicaux mais il a été considéré dans ce travail comme un segmenteur car notre système de tokenisation doit pouvoir être appliqué à toutes sortes de corpus. Et dans le cas de corpus SMS, la diminution ou l'absence de ponctuation (Rachel Panckhurst, 2009) ne laisse pas d'autres choix que de prendre l'espace comme segmenteur.

Enfin, nous parcourons le texte de token en token. Si le token courant est un mot déclencheur, alors le mot suivant est un toponyme.

7.1.3. Résultats

Nous avons testé notre méthode sur deux corpus qui sont La Grande Peur et celui de SMS. Le choix de ces corpus est stratégique. En effet, ces deux corpus présentent des caractéristiques bien distinctes.

Le premier est un corpus qui respecte les règles de graphie du français mais qui se démarque de par l'époque du sujet qu'il traite : 1789. Les toponymes présents dans ce corpus sont pour la plupart ceux de l'époque, et leur graphie aussi.

Quant aux SMS, c'est une forme de communication écrite empreinte de néologismes et néographie. Les variations sont diverses.

Pour chacun des deux corpus, les toponymes ont été extraits préalablement sous forme de liste. Pour chaque corpus, la précision et le rappel sont évalués et prennent comme corpus de référence la liste de toponymes associée au corpus évalué.
Nous cherchons ici à quantifier la part de toponymes correctement repérés (vrais positifs) par rapport au nombre total de mots repérés comme toponymes (rappel) mais aussi par rapport au nombre total de toponymes du corpus (précision). En outre l'étude des toponymes non-détectés et celle des non-toponymes détectés comme toponymes permet de cibler les points faibles de l'approche par mots déclencheurs.

Sur le corpus *La Grande Peur*, l'approche par mots déclencheurs obtient une précision de 0.362 et un rappel identique. Parmi les 174 noms de lieux du corpus, seuls 63 ont été repérés. Le nombre de non-toponymes détectés comme toponyme s'élève à 111.

Le corpus de SMS donne quant à lui un meilleur rappel (0.5119) mais une précision beaucoup plus faible (0.181). Cette faible précision trouve son origine dans la quantité considérable de mots détectés comme toponymes : 237 mots repérés comme toponymes dont 43 vrais positifs pour un corpus constitué de 84 toponymes.

7.1.4. Discussion

L'étude des faux positifs (non-toponymes détectés) du corpus de SMS permet de comprendre pourquoi la précision est faible. En observant le contexte des faux positifs dans le corpus, un détail surprend : le contexte gauche. Tous ces faux positifs sont bel-et-bien introduits par des mots déclencheurs, mais dans le contexte du corpus, leur catégorie grammaticale n'est pas celle recherchée. Ce ne sont pas des prépositions spatiales, mais des prépositions qui combinées au mot suivant (le faux positif) forment des locutions prépositionnelles (sous prétexte), adverbiales (en amont, en fait), ou conjonctive (en sorte).Ce phénomène est très récurrent dans le corpus de SMS, un peu moins dans

l'autre corpus. Nous notons aussi que « en » est particulièrement ambigu car il peut être employé comme indicateur de moyen (en voiture) et également pour la construction de gérondifs.

Par ailleurs, la nature même du corpus SMS impliquant une néographie des mots influe sur la qualité et l'exhaustivité des toponymes détectés. Certains mots « mal orthographiés » sont reconnus à tort comme mots déclencheurs augmentant ainsi le nombre de mots détectés comme toponymes. C'est notamment le cas de « sûr » devenu « sur ». Mais l'effet inverse est également observable. Certains mots déclencheurs subissent un changement de graphie et ne sont donc plus reconnus (« a » au lieu de « à » et « sr » à défaut de « sur »). Par conséquent, les toponymes qu'ils introduisent ne sont pas repérés. L'énumération de toponymes avec l'emploi de la virgule ou de la conjonction « et » rend les toponymes invisibles lors du repérage (dans « Paris, Montpellier et Marseille », nous comptons trois toponymes mais aucun mot déclencheur).

Pour contrer l'ambiguïté des mots déclencheurs, il serait intéressant d'utiliser des patrons, plutôt que de simples mots. Il s'agirait de définir des constructions morpho-syntaxiques qui introduisent des toponymes en partant des observations établies à partir de corpus. Les patrons pourraient être construits à partir de verbes de localisation statiques (être, se trouver, se situer, etc.) ou dynamiques (aller, arriver, partir, venir) suivis de prépositions (aller à , partir de, se trouver vers, etc.).

Outre sa précision et son rappel peu concluants, cette démarche est problématique en ce qui concerne le repérage de noms de lieux polylexicaux. Le corpus de SMS compte 13 toponymes polylexicaux, pourtant aucun d'eux n'a été repéré. Le corpus La Grande Peur n'en dénombre pas moins de 80 (45% des toponymes du corpus sont polylexicaux), et seuls 11 ont été repérés. Après analyse de ces toponymes, nous distinguons que la grande majorité d'entre eux ont pour séparateur l'espace (c'est le cas de tous les toponymes polylexicaux des SMS comme « saint jean de vedas » par exemple et de plus de la moitié de ceux de la Grande Peur comme « la Puisaye » ou « le Berry »par exemple). Le corpus étant tokenisé notamment aux espaces, les toponymes polylexicaux sont divisés en plusieurs tokens. En plus, le repérage par mots déclencheurs n'admet comme toponyme qu'un unique token.

En conclusion, l'ambiguïté des mots déclencheurs et la méthode de tokenisation empêchant la détection de certains toponymes polylexicaux nous conduisent à changer de stratégie, de façon à maximiser la précision et le rappel de notre programme.

7.2. Des mots impossibles

7.2.1. Introduction

Notre priorité est donc à présent les noms de lieux polylexicaux que nous souhaitons pouvoir repérer dans leur intégralité. À cette fin, nous avons revu la liste des mots déclencheurs.

7.2.2. Méthode

Nous avons remplacé notre liste de mots déclencheurs par une liste de mots dits « impossibles ». Nous considérons comme mot impossible tout mot qui ne peut être un constituant de toponyme. Pour former cette liste de mots impossibles, nous avons eu recours au logiciel *Unitex*(Paumier, 2011) [1].

Avec Unitex, nous cherchons à connaître les catégories grammaticales des mots qui ne sont pas ou peu présents dans les toponymes. Nous avons donc appliqué les dictionnaires français d'Unitex de mots simples et de mots composés sur notre gazetier de toponymes, soit sur près de 80 000 toponymes.

Après application des dictionnaires, Unitex fournit en sortie la liste des mots simples et composés qu'il a retrouvé parmi les toponymes fournis en entrée. La liste des mots composés ne nous semblait pas pertinente car elle se compose principalement de noms propres et de combinaisons « nom commun + 'de' » . Nous ne l'avons donc pas traitée. La seconde liste regroupe 64340 mots simples. Comme nous pouvons voir dans le tableau 7.1, près de 94% de ces mots simples sont des noms. Les deux autres catégories grammaticales les plus récurrentes sont les verbes et les adjectifs, avec une occurrence respective de 3.5% et 1.9%, ce qui reste assez faible.

POS	Occurrence	%
noms	**60285**	**93,714%**
Verbes	**2293**	**3,564%**
adjectifs	**1274**	**1,980%**
adverbes	85	0,132%
déterminants	51	0,079%
pronoms	32	0,050%
prépositions	29	0,045%
Conjonction de coordination	5	0,008%
Conjonction de subordination	1	0,002%
Autre	274	0,426%
Nombre total de mots simples	64329	100,00 %

FIGURE 7.1. – Les catégories grammaticales retrouvées dans les toponymes du gazetier

Les mots dont la catégorie grammaticale peut être « nom », « verbe » ou bien « adjectif » sont à exclure de la liste de mots impossibles.

Par le biais d'Unitex, nous voulons également valider ou non l'hypothèse selon laquelle un pronom ne peut être un constituant de toponyme. D'après les résultats, 32 pronoms ont été retrouvés dans les toponymes du gazetier, soit 0.05%. En regardant en contexte ces pronoms, nous avons constaté que les pronoms détectés n'en étaient pas.

1. Suite logicielle fondée sur des grammaires pour l'analyse de corpus

En fait, le gazetier de toponymes fourni en entrée d'Unitex englobe plusieurs langues, ce qui implique plusieurs alphabet. Or, lors de l'application des dictionnaires Unitex sur notre gazetier, c'est un alphabet français que nous avons sélectionné. Par conséquent, les caractères issus d'autres alphabets n'étaient pas reconnus et donc comptés comme séparateurs. Ainsi, « Quiñones » est devenu « qui ones », où « qui » est un pronom.

Au terme de notre étude Unitex, nous avons finalement retenu les mots étiquetés avec les catégories grammaticales suivantes : les pronoms personnels, les pronoms relatifs, les conjonctions de coordinations (hormis « ou » et « et » car souvent observables dans les toponymes), les conjonctions de subordination et les locutions conjonctives de subordination.

Nous avons ensuite tokenisé notre texte de la même façon que pour l'approche précédente, soit à chaque ponctuation. Néanmoins, puisque nous travaillons désormais avec une liste de mots impossibles, un nom de lieu repéré n'est plus un unique token, mais un token ou groupe de tokens compris entre deux mots impossibles (la fin du texte agit comme un mot impossible). Le token ou groupe de tokens constituent une séquence. De cette façon, nous maximisons nos chances de repérer les toponymes polylexicaux pour lesquels l'espace s'avère être un séparateur.

7.2.3. Résultats

Nous avons appliqué cette stratégie aux corpus utilisés lors de l'approche précédente. Cependant, la précision et le rappel ne seront pas évalués ici. L'objet de notre étude pour cette approche est la longueur des séquences formées. Nous cherchons à savoir si la granularité des mots impossibles est suffisante pour créer des séquences de toponymes. Dans l'idéal, une séquence correspond à un toponyme simple ou polylexical.

Le corpus La Grande Peur est constitué de 1591 tokens. Le nombre de séquences repérées est de 120. La longueur moyenne des séquences repérées est de 8 tokens, ce qui est bien au-dessus de la longueur moyenne des toponymes présents dans le corpus (2 tokens). Le constat est le même pour le corpus de SMS. Nous comptons un total de 398 séquences constituées et dont la longueur moyenne est de 7 tokens. Sachant que la taille de toponymes du corpus n'excède pas 4 tokens et que la moyenne est de 1 token, il est clair que les séquences détectées sont beaucoup trop grandes. Toutefois, quel que soit le corpus, tous les toponymes ont été repérés.

7.2.4. Discussion

Au cours de cette approche, l'idéal n'a pas été atteint. La taille des séquences est bien trop importante par rapport à la taille moyenne des toponymes. Il paraît donc évident que tous les toponymes, polylexicaux ou non, ont été repérés. Cependant, ils sont noyés au milieu d'autres tokens qui ne sont pas des toponymes.

Les mots impossibles seuls ne constituent donc pas des séquenceurs suffisants. Malgré tout, cette approche met en lumière la nécessité de créer des séquences afin de repérer les toponymes polylexicaux et le besoin d'affiner la taille des séquences. Pour cela, il est possible d'intervenir sur la liste des mots impossibles, en l'étoffant.

8. Repérage et identification simultanés

8.1. Introduction

Dans l'état actuel de notre travail, nous avons identifié deux besoins fondamentaux qui sont l'identification de toponymes et le repérage de toponymes, le tout à partir de corpus hétérogènes. Les deux besoins ont été traités séparément et des solutions sont proposées pour chacun des deux problèmes. Les résultats obtenus lors de l'identification sont satisfaisants. Bien que cela ne soit pas le cas pour le repérage, ils permettent d'aller plus loin dans la réflexion et de proposer de nouvelles stratégies de repérage.

Jusqu'à présent, la stratégie de repérage se concentrait sur la segmentation du corpus, d'abord par la tokénisation du texte, puis par la création d'unités plus vastes que le token par le biais de séquenceurs, permettant ainsi de regrouper tous les constituants d'un toponyme polylexical au sein de la même unité. Les résultats de la segmentation montrent néanmoins la nécessité d'affiner la taille des séquences. Une fois les séquences suffisamment affinées, il faut s'assurer qu'elles contiennent bien des toponymes, puis définir les limites du toponyme dans la séquence. Le recours à une base de données de toponymes (un gazetier) devient indispensable à cette étape du repérage. Mais puisque les toponymes peuvent être sujets aux variations, la base de données doit subir des prétraitements pour que le lien puisse être établi entre un toponyme du corpus et les toponymes du gazetier.

Le recours au gazetier revient tout compte fait à déterminer si un mot du corpus peut être un toponyme, et par la même occasion à attribuer un référent au toponyme repéré. Le repérage et l'identification ne sont ici plus perçus comme deux axes indépendants, mais comme deux étapes qui se complètent.

8.2. Méthode

Cette stratégie repose sur trois niveaux de segmentation du corpus qui correspondent aux trois paliers de la sémantique (texte, phrase, mot) (Bilhaut & Enjalbert, 2005). Le corpus sera découpé en tokens, puis ces tokens seront assemblés pour former des séquences et enfin des segments (potentiellement des toponymes) seront créés au sein même des séquences. Cette nouvelle segmentation voit l'introduction des notions de séquenceurs et de mots non-signifiants.

8.2.1. Tokenisation du corpus

Le premier niveau de segmentation du corpus se fait par tokenisation aux ponctuations. Cette fois-ci, l'objectif est d'atteindre une granularité très fine. De ce fait, tous les symboles de ponctuation, y compris le tiret et l'apostrophe, séquencent le texte.

Une fois le texte tokenisé, il s'agit de regrouper des tokens contigus dans le but de créer de nouvelles unités, que nous appelons séquences. C'est au moyen de séquenceurs que sont formées les « séquences ».

8.2.2. Création de séquences

Les mots séquenceurs

Dans cette nouvelle méthode, les mots impossibles prennent le nom de « séquenceurs ». Leur nature reste identique (pronoms personnels, pronoms relatifs, conjonctions de coordinations, conjonctions de subordination et locutions conjonctives de subordination). Il s'agit de mots qui sont très peu, voire jamais, constitutifs de toponymes. Ainsi, une séquence est un ensemble de tokens compris entre deux séquenceurs.

Les mots non-signifiants

Outre les séquenceurs, nous avons construit une liste de mots dits « non-signifiants ». En linguistique, les mots non-signifiants sont des mots outils. Ils appartiennent à des classes fermées et s'opposent aux mots lexicaux dont les possibles catégories grammaticales sont nom, adjectif, verbe et adverbe. Dans le cadre de notre stage et plus précisément de la présente méthode, la définition que nous avons attribuée aux non-signifiants est plus complexe. Un non-signifiant est ici un mot que nous pouvons parfois retrouver dans un nom de lieu mais qui est ambigu. Nous comptons parmi les non-signifiants les adverbes de temps, de lieux, de négation, les prépositions spatiales, les déterminants, les articles définis et indéfinis, les points cardinaux, les noms de lieux géographiques et les mots outils définis plus haut. Tous ces mots sont ambigus car considérer comme séquenceur la préposition « devant » (par exemple) reviendrait à la déclarer comme mot qui ne peut pas constituer un toponyme et donc faire l'impasse sur de nombreux toponymes comme « Savonnières-devant-Bar » ou « Fléville-devant-Nancy ». Pourtant, nous ne pouvons pas non plus déclarer systématiquement la préposition « devant » comme constituant de toponymes car cela créerait beaucoup trop de bruit.

Parcours des séquences

Ensuite, nous parcourons le contenu des séquences, en examinant chaque token successivement, hormis les tokens non-signifiants, car ils ne sont pas révélateurs de toponymes. Nous cherchons à savoir si un token de la séquence peut être un toponyme ou bien un morceau de toponyme. Chaque token repéré comme potentiel toponyme forme un segment de la séquence courante (la création de segments constitue le troisième niveau de segmentation du corpus). Si le token s'avère être une partie de toponyme, notre but sera alors de repérer tous ses constituants. Pour mener à bien cette tâche et puisque

notre approche repose sur l'exploitation de bases de données lexicales, il nous fallait un gazetier de toponymes auquel se référer.

8.2.3. Prétraitements du gazetier

Nous avons utilisé le gazetier de toponymes constitué en début de stage. Chaque token est comparé à toutes les entrées du gazetier. Nous considérons que le token est un toponyme ou un morceau de toponyme lorsqu'il y a match exact entre le token et le toponyme du gazetier (c'est-à-dire que les deux formes sont identiques). En conservant notre gazetier en l'état, il ne peut y avoir match exact qu'entre un token du corpus sans variation et un toponyme constitué d'un seul mot du gazetier.



FIGURE 8.1. – Extrait de séquences de corpus et du gazetier de toponymes

Dans la figure 8.1, nous disposons de trois séquences du corpus tokenisées et d'un extrait du gazetier de toponymes. L'extrait du gazetier est composé de trois entrées avec pour clés les termes en bleu. Pour chaque séquence du corpus, nous cherchons à savoir si les tokens (en vert) sont des toponymes ou constituants de toponymes. Aussi, nous comparons chaque token en vert avec les clés de chaque entrée du gazetier. Le token "Paris" de la séquence 1) matche avec l'une des clés du gazetier. Il est donc repéré comme toponyme simple. Dans la séquence 2), "pariiiiis" ne trouve pas de match dans le gazetier. Pourtant, il pourrait s'agir du toponyme "Paris" avec une répétition de la lettre "i". Enfin, dans la dernière séquence, "Noisy" semble être une troncature des deux dernières clés du gazetier, mais comme il ne matche pas exactement avec une clé du gazetier, il n'est pas repéré comme toponyme.

En somme, pour faire le lien entre un toponyme polylexical issu d'un corpus avec

variations d'écriture et une entrée du gazetier, deux points sont à prendre en considération :

— l'éventuelle présence de variations sur le toponyme du corpus

— les toponymes polylexicaux du corpus sont divisés en plusieurs tokens

Ces deux points nous ont conduit à revoir notre système de prétraitement des gazetiers. Désormais, tous les toponymes polylexicaux du gazetier sont tronqués. La troncation se fait aux signes de ponctuation (y compris l'espace) et les mots-outils ou noms génériques géographiques ne sont pas retenus comme troncatures possibles.

La figure 8.2 donne un aperçu du gazetier avant et après troncation des clés.

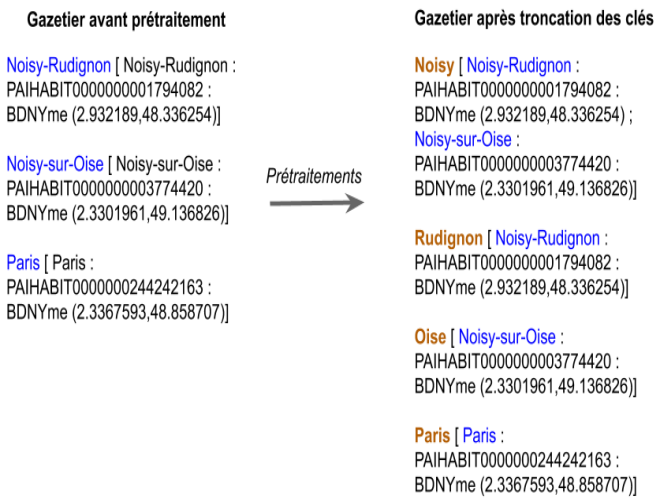


FIGURE 8.2. – Extrait du gazetier avant et après segmentation des clés

L'entrée « Noisy-sur-Oise » est divisée en deux tokens, « sur » étant un mot outil et n'étant donc pas pris en compte lors de la procédure de troncation (comme indiqué plus haut). Ainsi, si notre corpus contient le token « Noisy », il matchera avec l'entrée du gazetier « Noisy », qui correspond au toponyme « Noisy-sur-Oise », mais aussi à « Noisy-Rudignon ».

Pour ce qui est des variations, nous avons suivi le même raisonnement de prétraitement des gazetiers que celui adopté précédemment, mais en l'adaptant aux nouvelles contraintes.

Pour rappel, nous disposons d'un gazetier sur lequel nous appliquons plusieurs prétraitements successivement. Un prétraitement appliqué au gazetier dépend d'une transformation particulière (qui elle-même dépend d'une variation) et il en résulte une nouvelle version du gazetier. En d'autres termes, la quantité de gazetiers prétraités dépend du nombre de variations étudiées.

À présent, le nombre de prétraitements est réduit. Sachant que les clés du gazetier

sont déjà tronquées, la troncation ne fait plus partie des transformations appliquées. Les variations prises en charge sont donc la simplification, le squelette consonantique, la variation « identité » et la variation « initiale ». Le tableau 8.3 confronte les processus de transformations antérieurs aux actuels.

FIGURE 8.3. – Tableau comparatif des processus de transformations du gazetier

Nom de la variation	Processus de transformation antérieur	Processus de transformation actuel	Illustration avec un cas concret
initiale (ne s'applique que sur un toponyme polylexical)	1) mise en minuscule 2) suppression de la ponctuation et des espaces 3) suppression des mots outils et noms génériques géographiques 4) suppression des caractères répétés 6) suppression des diacritiques 5) pour le dernier token, préservation uniquement de l'initiale	Seule l'initiale du dernier token du toponyme est préservée	Toponyme :Marseille-en-Beauvaisis Avant Entrée du gazetier obtenue après transformation : marseileb [Marseille-en-Beauvaisis:id : source : coordonnées géo] Maintenant Entrée du gazetier obtenue après transformation : b [Marseille-en-Beauvaisis : id : source : coordonnées géo]
simplification	1) mise en minuscules 2) suppression des mots outils, des noms génériques géo, de la ponctuation, des espaces, des diacritiques et caractères répétés	pas de changements	Toponyme : Marseille-en-Beauvaisis → marseilebeauvaisis[Marseille-en-Beauvaisis : id : source : coordonnées géo]
squelette consonantique	1) mise en minuscules 2) suppression des mots outils, des noms génériques géo, de la ponctuation, des espaces, des diacritiques et caractères répétés 3) suppression des voyelles (si le mot commence par une voyelle, on ne la supprime pas)	pas de changements	Toponyme : Marseille-en-Beauvaisis → mrslbvss [Marseille-en-Beauvais : id : source : coordonnées géo]
identité	pas de transformation	mise en minuscules	Toponyme : Paris Avant Paris [Paris : id : source : coordonnées géo] Maintenant paris [Paris : id : source : coordonnées géo]

Nous relevons peu de changements entre les prétraitements antérieurs et les actuels. Le changement majeur est celui concernant la variation « initiale ». Le gazetier prétraité avec la transformation « initiale » n'a pour clés que des lettres uniques (ou des chiffres). En transformant un toponyme polylexical comme « Noisy Rudignon » avec la transformation « initiale », nous obtenons d'un côté le token « noisy » et de l'autre « r ». Conserver « noisy » dans le gazetier prétraité avec la transformation « initiale » n'aurait pas de sens, puisqu'il s'agit d'une forme qui se trouve déjà dans nos gazetiers.

La figure 8.4 donne un aperçu des étapes de prétraitements du gazetier.

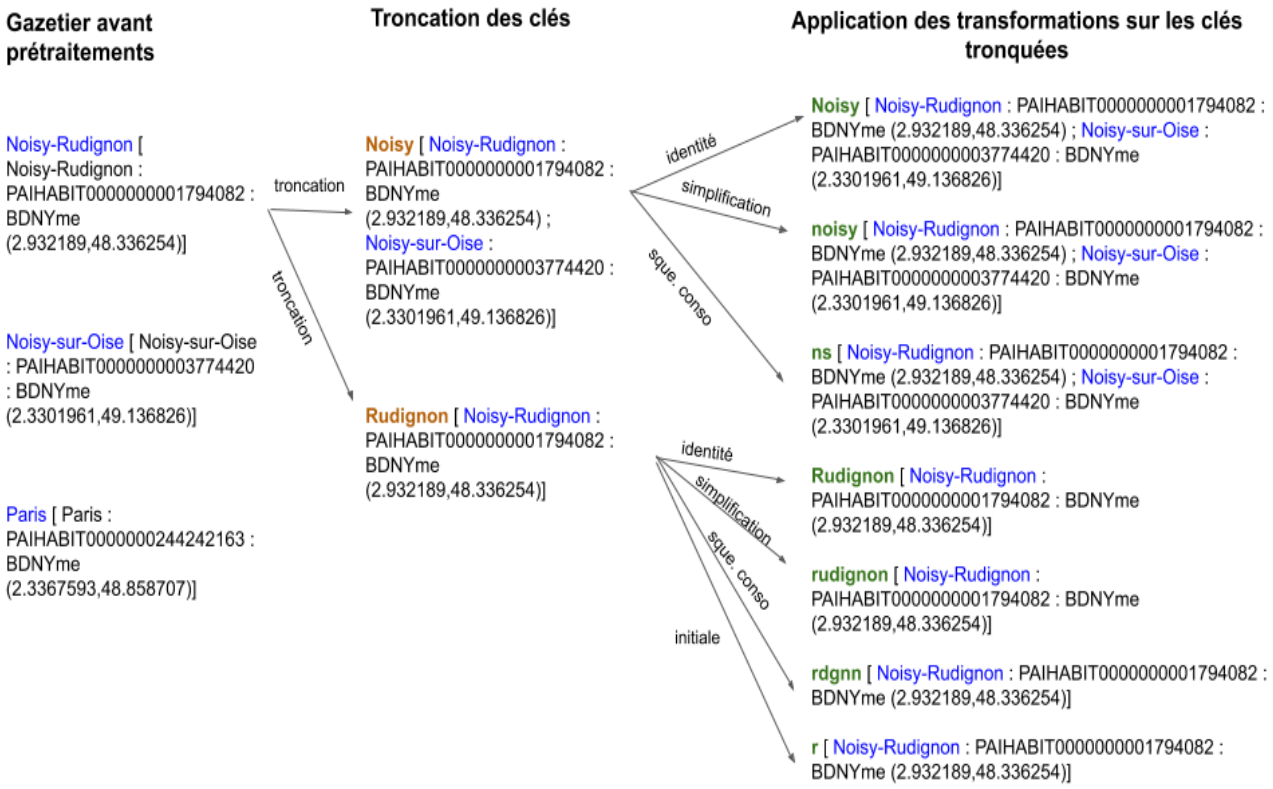


FIGURE 8.4. – Étapes de prétraitements du gazetier

Toutefois, le prétraitement ne s'arrête pas là. Puisque les clés des gazetiers prétraités peuvent être des troncatures de toponymes (sauf en cas de toponyme simple), il nous fallait garder en mémoire la position du morceau de toponyme par rapport à sa version originale (nous appelons cette position *posGazetier*). Par exemple, pour l'entrée du gazetier suivante : grnd [Noisy-le-Grand : PAIHABIT0000000002597894 : BDNYme (2.5599918,48.839775)], « grnd » est le squelette consonantique de « Grand » qui se situe à la place 2 du toponyme original (les indices de position débutent à partir de 0). Nous avons donc rajouté cette information à tous les gazetiers prétraités et ce pour les toponymes simples comme pour les toponymes polylexicaux. Cela donne :

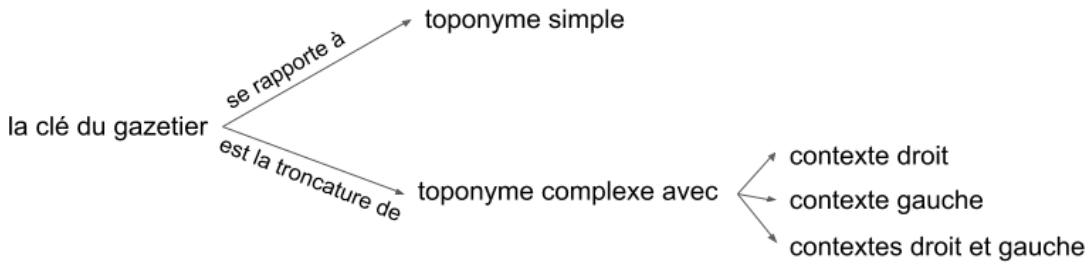
grnd { [Noisy-le-Grand : PAIHABIT0000000002597894 : BDNYme (2.5599918,48.839775), *posGazetier*=2],
[Saint-Gérand-de-Vaux : PAIHABIT0000000105730546 : BDNYme (3.3985417,46.380157), *posGazetier*=1],
[Grand : PAIHABIT0000000043275697 : BDNYme (5.48725,48.384888), *posGazetier*=0
}

(dans cet exemple, nous constatons que le squelette consonantique grnd se rapporte à plusieurs toponymes).

8.2.4. Création de segments

Une fois le texte tokenisé, les séquences créées et le gazetier prétraité vient l'étape de création des segments, soit parcourir les tokens du corpus et les chercher dans les gazetiers prétraités. Le parcours des séquences se fait de token en token. Si le token courant est un mot non-signifiant comme défini plus haut, nous l'ignorons. En effet, les mots non-signifiants peuvent être des constituants de toponymes, mais le plus souvent ils sont adverbes, déterminants, etc. Les ignorer lors du parcours des tokens nous permet de réduire le nombre de segments repérés qui ne sont pas des toponymes. Avant de comparer le token courant aux entrées des gazetiers, nous transformons le token. Lorsque nous recherchons le token dans le gazetier « identité », le toponyme est mis en minuscules (car les entrées de ce gazetier sont en minuscules). Pour les autres gazetiers, le token subi la transformation de simplification (car les transformations squelette consonantique, initiale, et bien sûr simplification contiennent les étapes de simplification).

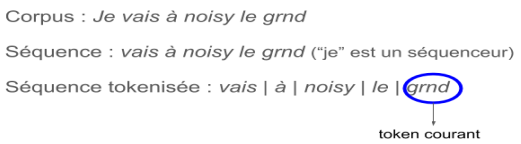
Lorsqu'un match exact a lieu entre le token courant du corpus et une clé d'un gazetier (c'est-à-dire que les deux mots sont identiques), nous souhaitons savoir si la clé du gazetier est la troncature d'un toponyme du gazetier. Si oui, nous voulons savoir si le token du corpus peut lui aussi être la troncature de ce toponyme. À ce stade, quatre cas de figure sont possibles :



Nous savons que si la position de la clé du gazetier par rapport au toponyme original (*posGazetier*) est supérieure à 0, cela signifie que la clé du gazetier est un constituant d'un toponyme polylexical du gazetier. Il faut donc définir les limites du toponyme au sein de la séquence. Par conséquent, nous observons dans le toponyme polylexical du gazetier le nombre de mots qu'il y a avant ce constituant (n) et après ce constituant (n'). Puis, dans le corpus, à partir du token courant, nous nous déplaçons de n tokens vers la gauche et de n' tokens vers la droite : nous obtenons ainsi un segment. Attention, quand nous identifions les contextes droit et gauche du token courant, nous ne sortons jamais de la séquence courante. En effet, chaque séquence est encadrée par des mots séquenceurs, autrement dit par des mots qui ne peuvent pas être constituants de toponyme.

En revanche, si la *posGazetier* est égale à 0, alors soit la clé du gazetier correspond à un toponyme simple du gazetier, soit la clé du gazetier est le premier constituant d'un toponyme polylexical. De ce fait, il nous faut observer la taille du toponyme du gazetier. Pour une taille supérieure à un, nous avons affaire à un toponyme polylexical. Dans le cas d'un toponyme simple, le token constitue un segment à lui seul.

Prenons l'exemple suivant :



En comparant le token courant avec les entrées des gazetiers prétraités, nous obtenons un match exact avec l'entrée suivante :

grnd {[Noisy-le-Grand : PAIHABIT0000000002597894 : BDNYme (2.5599918,48.839775), *posGazetier*=2],
[Saint-Gérand-de-Vaux : PAIHABIT0000000105730546 : BDNYme (3.3985417,46.380157), *posGazetier*=1],
[Grand : PAIHABIT0000000043275697 : BDNYme (5.48725,48.384888), *posGazetier*=0
}

Cette entrée est issue du gazetier prétraité par squelette consonantique. Ainsi, chaque toponyme associé à la clé « grnd » est un potentiel référent normé du token. Maintenant, pour chaque potentiel référent normé, nous observons *posGazetier*, puis en fonction de *posGazetier* nous examinons les contextes gauche et droit du token dans le corpus.

- Dans « Noisy-le-Grand », la clé « grnd » est à la place numéro 2 et ne possède pas de contexte droit. En reportant cela sur le corpus, cela donne le segment « noisy le grnd ».
- le token « grnd » est précédé d'un mot et suivi de deux dans « saint-Gérand-de-Vaux ». Pas de segment créé, car dans le corpus, le token courant n'a pas de contexte droit. Donc le token courant n'est pas un morceau du toponyme « saint-Gérand-de-Vaux »
- Enfin, « grnd » correspond aussi au toponyme simple « Grand ». Le token courant devient lui-même un segment.

Les segments repérés sont les suivants :

vais | *à* | [*noisy* | *le* | [*grnd*]]

[*noisy le grnd*] serait le squelette consonantique de [Noisy-le-Grand : PAIHABIT0000000002597894 : BDNYme (2.5599918,48.839775), posGazetier=2]

[*grnd*] serait le squelette consonantique de [Grand : PAIHABIT0000000043275697 : BDNYme (5.48725,48.384888), posGazetier=0]

Dans cet exemple, nous constatons que les deux segments repérés s'intersectent, ce qui n'est pas envisageable dans le contexte de reconnaissance de toponymes. Il faut donc trier les segments, de façon à supprimer tout chevauchement. Pour cela, nous transformons les potentiels référents normés avec la transformation qui nous a permis de les atteindre (dans l'exemple, c'est le squelette consonantique).

$$\text{Noisy-le-Grand : nsgrnd} \; -- \; \text{Grand : grnd}$$

En outre, nous simplifions les segments.

$$\text{noisy le grnd : noisygrnd} \; -- \; \text{grnd : grnd}$$

Puis nous calculons la similarité entre les segments simplifiés et leur référent transformé (la mesure de similarité dépend de la transformation subie par le référent).

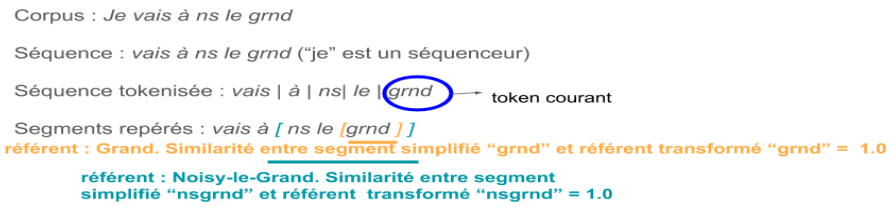
$$\text{référent : nsgrnd, segment : noisygrnd, similarité} = 0.7$$

$$\text{référent : grnd, segment : grnd, similarité} = 1.0$$

Dans cet exemple, le segment « grnd » simplifié et son référent transformé sont identiques, tandis que le segment « noisy le grnd » est moins similaire à son référent transformé. Nous en déduisons que c'est le segment « grnd » qui a le plus de chance d'être un toponyme et donc nous ne conservons pas le second segment.

Finalement, dans le corpus « Je vais à noisy le grnd », à partir du token « grnd », le toponyme « grnd » a été repéré et associé au référent normé « Grand : PAIHABIT0000000043275697 : BDNYme (5.48725,48.384888)] ». Du point de vue de la machine, ce résultat est tout à fait correct. Pourtant, si nous avions annoté manuellement en noms de lieux le corpus, c'est « noisy le grnd » qui serait annoté comme nom de lieu, avec pour référent normé « Noisy-le-Grand ». Cet écart entre le résultat attendu et le résultat obtenu trouve son explication dans le type de variations prises en compte dans ce stage. Le toponyme manuellement repéré n'est pas un squelette consonantique tel que nous l'envisageons car seul le dernier mot a perdu ses voyelles.

Dans le cas suivant, « ns le grnd » a été annoté manuellement comme toponyme. Il est le squelette consonantique de « Noisy-le-Grand ». En passant notre chaîne de traitement sur ce corpus, voici ce que nous obtenons :



Cette fois-ci, le segment « ns le grnd » est un squelette consonantique qui répond aux critères du squelette consonantique que nous avons imaginé dans le cadre de notre stage. D'autre part, les deux segments repérés s'intersectent et ils ont la même similarité. Parce que nous voulons maximiser nos chances de repérer des toponymes polylexicaux dans leur intégralité, le segment le plus long sera retenu, et tous les autres segments intersectant avec lui seront supprimés.

8.3. Résultats

Ce qui différencie cette approche des précédentes, c'est qu'elle combine le repérage et l'identification. Le principe de l'identification reste le même : à partir d'un corpus, le toponyme repéré est transformé puis comparé aux entrées des gazetiers prétraités. S'il y a match exact entre le toponyme transformé et un entrée des gazetiers, alors l'entrée du gazetier est considérée comme potentiel référent du nom de lieu du corpus. Par contre, le processus de repérage est différent. C'est donc ce qui sera évalué. Nous quantifions ici la part de toponymes du corpus correctement trouvés par rapport au nombre total de toponymes du corpus puis par rapport au nombre total de mots détectés (rappel et précision). Nous indiquons également le nombre de toponymes du corpus repérés auxquels le bon référent normé a été attribué.

Les corpus évalués sont celui de SMS, Matriciel et La Grande Peur. Il s'agit des corpus pour lesquels nous disposons à la fois du corpus vierge (non-annoté en noms de lieux) et de la liste des toponymes du corpus associés à leur référent (qui a été obtenue par annotation manuelle).

Le corpus Matriciel est celui qui obtient le meilleur rappel. Sur un total de 700 lieux, 621 ont été repérés (soit un rappel de 0.887). Et parmi les noms de lieux du corpus repérés, 84% se sont vus attribuer le bon référent normé. Néanmoins le nombre de segments repérés est exponentiel (4672) et fait chuter la précision, qui tombe à seulement 0.133.

Le corpus de SMS suit plus ou moins la même tendance, mais avec un rappel plus faible (0.548). 46 toponymes du corpus sur 84 ont été détectés et seuls 39% sont associés au bon référent normé. Encore une fois, la précision est faible (0.116) à cause du trop grand nombre de mots repérés comme toponymes (395 segments).

Enfin, le corpus de La Grande Peur donne des résultats moyens, avec un rappel de

0.517, une précision de 0.484 et un pourcentage de toponymes repérés associés au bon référent normé qui s'élève à 41%. Le nombre de segments repérés (186) est presque similaire à la quantité de toponymes que comporte le corpus (174) mais il comprend seulement 90 vrais positifs.

8.4. Discussion

Le premier élément à étudier est le nombre de segments repérés. De manière générale, il est trop important. Nous en avons relevé deux causes. La principale cause est l'ambiguïté des toponymes. La plupart des segments détectés comme toponymes sont en réalité des verbes et des noms. Ce phénomène est particulièrement observable avec le corpus Matriciel, dans lequel nous relevons une occurrence du participe passé « eu » de 132. Pourtant, il est repéré comme étant le toponyme Eu (Normandie) à chaque fois, de même que « parce » reconnu 378 fois comme simplification du toponyme Parcé, à tort. Dans le même cas nous retrouvons entre autres « grand », « armes », « arrive », « douze », « cours », « aller », « rives », « long », etc.

La seconde raison, moins marquante, est l'ambiguïté des squelettes consonantiques cette fois-ci. D'une part, certains toponymes du gazetier prennent la forme de mots courants lorsqu'ils sont réduits à leur squelette consonantique, comme Oñate qui devient « ont ». Par conséquent, le participe passé « ont » du corpus sera repéré comme étant le squelette consonantique de Oñate. D'autre part, dans le cas de langage informel comme dans les SMS, certains mots récurrents sont réduits à leur squelette consonantique. De cette façon, « pour » devient « pr » repéré comme squelette consonantique du toponyme Pori, « désolé » devient « dsl », squelette consonantique de la commune Dasle. En outre, dans le registre de langue des SMS, certains mots courts ou abrégés matchent avec des squelettes consonantiques du gazetier de toponymes. (« ok » avec Oikia et « rdv » avec Riodeva). De par ces ambiguïtés, 69 squelettes consonantiques ont été repérés comme toponymes du corpus de SMS, en comparant avec les annotations de référence, pas un seul ne s'avère correspondre à un nom de lieu du corpus.

Du côté des noms de lieux non repérés, nous dénombrons deux raisons qui relèvent du type de variations appréhendées (« tlse » pour Toulouse est une forme de squelette consonantique différente de celle que nous traitons car elle conserve la dernière voyelle), et du type de gazetier utilisé selon le corpus. Parfois la couverture du gazetier n'est pas suffisante par rapport au corpus (certains toponymes des corpus n'y sont pas répertoriés). D'autres fois, la façon dont sont annotés les noms de lieux du corpus ne concorde pas avec le type de lieux recensés dans le gazetier (par exemple, un grand nombre de toponymes des corpus Matriciel et La Grande Peur sont introduits par des noms génériques géographiques du type « la vallée de »).

Pour conclure, cette approche ne peut pas être performante pour un contexte général. Il est impératif d'adapter le gazetier au type de corpus étudié. Et pour savoir quel

gazetier choisir selon le corpus, il faut connaître la zone géographique couverte par le corpus et le type de langue et donc les variations d'écriture susceptibles d'apparaître.

Conclusion

La ligne conductrice de ce stage était l'identification de toponymes avec variations d'écriture à partir de données hétérogènes. Deux processus d'identification étaient attendus :

— attribuer à un toponyme un référent normé
— proposer pour un toponyme donné tous les noms de lieux qui contiennent ce toponyme

L'objectif principal étant atteint, nous avons poussé la réflexion plus loin en nous penchant sur la question du repérage de toponymes avec variations. L'évaluation du processus de repérage a permis de mettre en lumière les difficultés engendrées par les toponymes polylexicaux et nous a permis de revoir notre stratégie. Au lieu de diviser les étapes de traitement en deux grands axes, nous les avons combinés. Nous avons d'abord réfléchi à comment segmenter le corpus, puis nous avons segmenté le gazetier. Les résultats obtenus après évaluation du système ne sont pas très bons. Pour une meilleure performance, il faudrait adapter le gazetier au corpus. Or il est illusoire de vouloir appréhender tout type de corpus et toute couverture de gazetier. Toutefois, le programme est adaptable.

D'autre part, nous avions illustré notre problématique par une phrase, « À Pralognan suivre la route entre l'hôtel (...) », dans laquelle « Pralognan » devait être repéré comme troncature de Pralognan-la-Vanoise. Dans l'état actuel de notre programme, cet objectif ne peut être atteint qu'en dissociant les étapes de repérage et d'identification. En appliquant les deux méthodes simultanément, « Pralognan » sera repéré comme toponyme, Pralognan-la-Vanoise lui sera associé, mais la variation troncation ne sera pas indiquée. Certains post-traitements s'imposent.

Dans une perspective d'amélioration des travaux réalisés, nous pensons qu'il serait judicieux d'augmenter les possibilités de notre chaîne de traitement en étoffant la liste des variations abordées. Un plus grand nombre de variations permettrait à notre projet de s'adapter à de nouveaux types de corpus. Une variation phonétique par exemple, qui donnerait la transcription phonétique de chaque toponyme permettrait au programme de faire face à d'autres cas de figures, comme celui de la substitution de phonèmes par d'autres phonèmes.

Par ailleurs, il nous paraît important de pouvoir varier la granularité des données sur lesquelles nous travaillons. Quelle est l'échelle des toponymes étudiés ? Des chefs-lieux ? Des pays ? Cette question de granularité des toponymes va de paire avec le choix du gazetier. Quelle surface du territoire couvrir ? Un gazetier qui couvre un large territoire entraînera en contrepartie un temps de traitement plus long. Mais cibler la zone géographique couverte par les gazetiers en fonction du corpus requiert une connaissance préalable des toponymes compris dans le corpus. Or il arrive parfois que même des connaissances humaines ne suffisent pas à localiser un toponyme, en particulier lorsqu'il est porteur d'une variation d'écriture. Les variations d'écriture sont en effet souvent propres au producteur et ne suivent aucune règle. Il est difficile de reconnaître Saint-Laurent-d'Onay (Rhône-Alpes) au lieu de Saint-Laurent-d'Oingt (Rhône-Alpes) à partir

de « Saint-Laurent-d'O ». La désambiguisation géographique peut parfois permettre de faire un choix entre des potentiels toponymes en regardant la localisation de ces derniers. Mais dans le cas de notre exemple, les deux toponymes sont proches en terme de graphie mais aussi au niveau de leur localisation. Les départager s'avère délicat.

Le choix des ressources pose aussi la question de la langue. La langue du corpus doit être la même que celle des bases de données lexicales. D'autre part, la langue ne définit pas le contexte géographique. Un texte en français ne signifie pas qu'il traite des toponymes du territoire français.

D'une manière plus globale, le stage fait l'état de certains verrous qui au-delà de notre sujet, sont propres au domaine du TAL. Toute tâche de TAL requiert une quantité de données importantes sur lesquelles travailler. Bien que de nombreux acteurs du TAL s'inscrivent dans une démarche d'ouverture des données, l'accès aux corpus est parfois restreint. Dans le cadre de ce stage, le manque de corpus annotés en noms de lieux s'est fait ressentir. L'annotation de corpus demande beaucoup de temps de travail. C'est pourquoi il est peu envisageable de s'y atteler le temps d'un stage.

Références

Abney, S., Collins, M., & Singhal, A. (2000). Answer extraction. In *Proceedings of the sixth conference on applied natural language processing* (pp. 296–301).

Bilhaut, F., & Enjalbert, P. (2005). Sémantique et traitement automatique du langage naturel, chapter 10, recherche d'information géographique. *Hermes, Lavoisier*, 371–406.

Brando, C., Dominguès, C., & Capeyron, M. (2016). Evaluation of ner systems for the recognition of place mentions in french thematic corpora. In *Proceedings of the 10th workshop on geographic information retrieval.*

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Cohen, W. W., Ravikumar, P., Fienberg, S. E., et al. (2003). A comparison of string distance metrics for name-matching tasks. In *Iiweb* (Vol. 2003, pp. 73–78).

Delais-Roussarie, E., Bourigault, D., Choi-Jonin, I., Fabre, C., Molinu, L., Rouquier, M., & Tarrier, J.-M. (2004). *Acsynt, un corpus oral du français contemporain.*

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Dominguès, C., & Eshkol-Taravella, I. (2013). Repérer des toponymes dans les titres de cartes topographiques..

Dominguès, C., & Eshkol-Taravella, I. (2015). Toponym recognition in custom-made map titles. *International Journal of Cartography*, *1*(1), 109–120.

Dominguès, C., Weber, S., Brando, C., Jolivet, L., & Van Damme, M.-D. (2017). Analyse et cartographie des sentiments dans des récits de vie de migrants..

Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In *Actes, neuvieme conférence nationale sur le traitement automatique des langues naturelles (taln 2002)* (Vol. 1, pp. 265–274).

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., & Galibert, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language..

Hadjieleftheriou, M., & Srivastava, D. (2010). Weighted set-based string similarity. *IEEE Data Eng. Bull.*, *33*(1), 25–36.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, *84*(406), 414–420.

Kogkitsidou, E. (2018). *Communiquer par sms : Analyse automatique du langage et extraction de l'information véhiculée* (Thèse de doctorat non publiée). Grenoble Alpes.

Landau, M., Sillion, F., & Vichot, F. (1993). Exoseme : A document filtering system based on conceptual graphs. In *International conference on conceptual structures,*

iccs (Vol. 93).

Lefebvre, G. (2014). *La grande peur de 1789 : suivi de les foules révolutionnaires*. Armand Colin.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453.

Nguyen, V. T., Sallaberry, C., & Gaio, M. (2013). Mesure de la similarit\'e entre termes et labels de concepts ontologiques. *arXiv preprint arXiv :1307.6422*.

Nouvel, D., Ehrmann, M., & Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE Group.

Olteanu-Raimond, A.-M., Davoine, P.-A., Gaio, M., Gouarderes, E., Van Damme, M.-D., Villanova-Oliver, M., ... others (2017). Projet choucas : Intégration de données hétérogènes et raisonnement spatial pour l'aide à la localisation des victimes en montagne..

Panckhurst, R. (2009). Short Message Service (SMS) : typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin* (p. 33-52). Université Paul-Valéry Montpellier 3. Consulté sur `https://hal.archives-ouvertes.fr/hal-00443014`

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., & Verine, B. (2016). 88milsms. a corpus of authentic text messages in french. *Banque de corpus CoMeRe. Chanier T.(éd)-Ortolang : Nancy*.

Paris, P.-H., Abadie, N., & Brando, C. (2017). Linking spatial named entities to the web of data for geographical analysis of historical texts. *Journal of Map & Geography Libraries*, *13*(1), 82–110.

Paumier, S. (2011). Unitex-manuel d'utilisation.

Steuckardt, A. (2017). La ponctuation choisie des peu-lettrés, d'après «corpus 14». *Linx. Revue des linguistes de l'université Paris X Nanterre*(75), 145–160.

Van de Weerd, J. (2018). Vers les origines sémantiques du conditionnel épistémique. étude d'un genre juridique en français classique (xvie-xviiie siècles). *Langue française*(4), 77–89.

Wenz, R. (2013). Linked open data for new library services : the example of data. bnf. fr. *JLIS. it*, *4*(1), 403.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann.

Table des figures

4.1. Les variations d'écriture 14
4.2. Les grandes étapes du traitementdu repérage et de l'identification combinés 18

6.1. Transformations appliquées selon le type de variation 28
6.2. Étapes d'identification 31
6.3. Précision obtenue selon la mesure de similarité 32

7.1. Les catégories grammaticales retrouvées dans les toponymes du gazetier . 37

8.1. Extrait de séquences de corpus et du gazetier de toponymes 41
8.2. Extrait du gazetier avant et après segmentation des clés 42
8.3. Tableau comparatif des processus de transformations du gazetier 43
8.4. Étapes de prétraitements du gazetier 44
.5. Répartition des variations dans le corpus Renumar 57
.6. Répartition des variations dans le corpus SMS 57
.7. Répartition des variations dans le corpus BnF 58
.8. Répartition des variations dans le corpus Matriciel 58
.9. Pourcentage d'occurrences des variations d'une lettre, deux lettres ou trois lettres 59
.10. Qualité des mesures de similarité selon la variation 62
.11. Jeu de données pour apprentissage supervisé 63

Annexes 1

Répartition des variations par corpus

Pour chaque corpus, les diagrammes suivants révèlent la répartition des variations observées sur les toponymes.

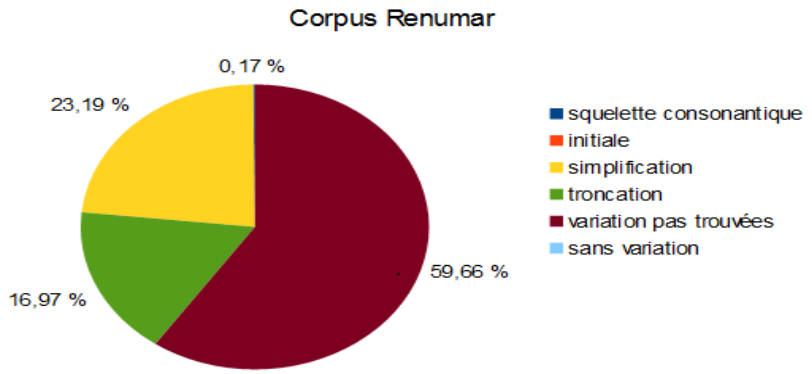


FIGURE .5. – Répartition des variations dans le corpus Renumar

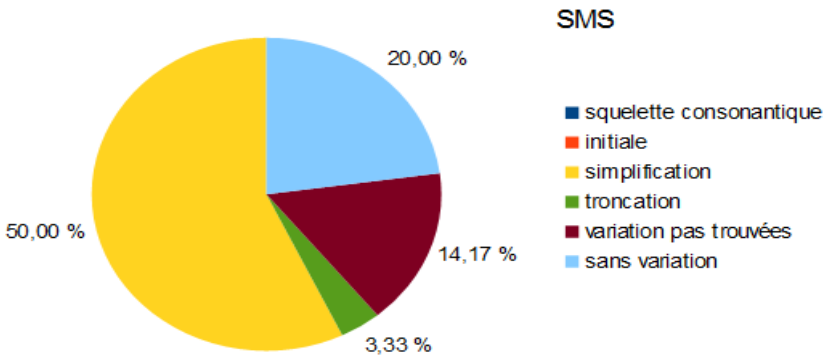


FIGURE .6. – Répartition des variations dans le corpus SMS

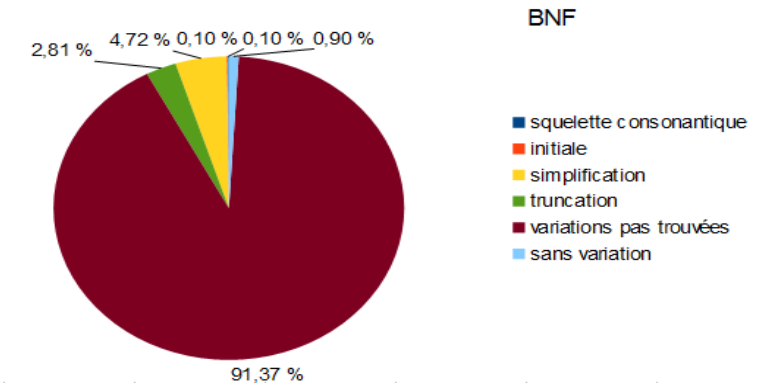


FIGURE .7. – Répartition des variations dans le corpus BnF

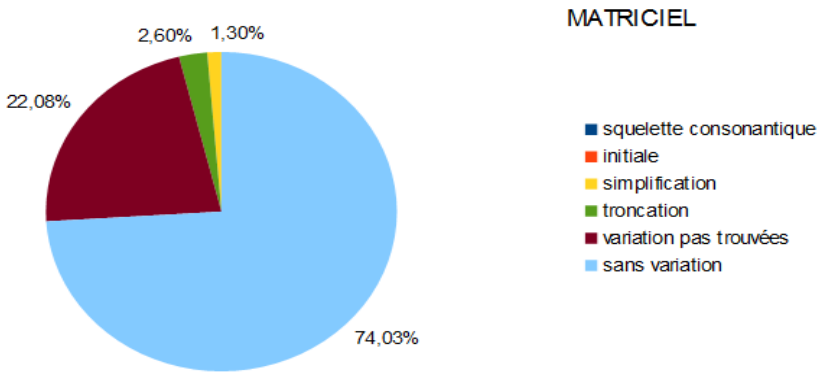


FIGURE .8. – Répartition des variations dans le corpus Matriciel

Une variation supplémentaire

Après l'étude de la typologie des variations selon le type de corpus, force est de constater que la part de variations non identifiées est importante dans les corpus Renumar et BnF. L'observation des toponymes issus de ces corpus fait émerger la possibilité d'appréhender dans notre travail une nouvelle variation : la variation « insertion/substitution/délétion d'une lettre ».

Cette nouvel objet de recherche mettrait en lumière les cas pour lesquels une différence d'une lettre est observable entre un nom de lieu avec variation et la forme normée de ce même nom de lieu.

Nous cherchons à connaître, pour un corpus donné, le nombre de fois pour lesquelles il y a n lettres qui varient entre un nom de lieu avec variation et sa forme normée. Il peut s'agir de lettres supprimées, rajoutées, ou remplacées. Dans notre cas, nous nous contentons de rechercher le nombre de fois pour lesquelles il y a 1, 2 ou 3 lettres qui varient. Pour ce faire, nous calculons la similarité entre le nom de lieux avec variation et sa forme normée (parfois, le nom de lieu avec variation est associé à plusieurs formes normées ; toutes les formes normées sont prises en compte). La métrique utilisée est la mesure de similarité Levenshtein, sous sa forme «absolue» (c'est-à-dire que la valeur obtenue n'est pas comprise entre 0 et 1 mais correspond au nombre de lettres variantes). Par exemple, en calculant la similarité absolue de Levenshtein entre «Dieppe» et «Dieppa» on obtient $s = 1$.

L'étude porte sur les corpus de SMS, Renumar, BnF et Matriciel. Ces corpus rassemblés comprennent un total de 1176 paires, chacune formée d'un nom de lieu avec variation associé à sa forme normée.

Nous avons donc commencé par étudier le pourcentage de noms de lieux pour lesquels une lettre, deux lettres ou trois lettres varient par rapport à la forme normée. Sur l'ensemble des corpus, cette variation de lettres est observable sur moins de la moitié des noms de lieux(45,27%).

		%
une lettre	335	18,86 %
deux lettres	266	14,98 %
trois lettres	203	11,43 %
autre	972	54,73 %
nombre total de noms de lieux	1776	100,00 %

FIGURE .9. – Pourcentage d'occurrences des variations d'une lettre, deux lettres ou trois lettres

La variation d'une lettre étant la plus fréquente, elle devient l'objet de notre étude.

La substitution de lettre correspond à 63% des cas de variations d'une lettre. Nous appelons substitution d'une lettre lorsqu'un nom de lieu ne varie que d'une lettre par

rapport à sa forme normée. Nous avons distingué deux phénomènes de substitution d'une lettre observable sur des toponymes : l'un identifiable en milieu de mot et l'autre en fin de mot. Les permutations portant uniquement sur la casse ou les signes diacritiques ne sont pas comptabilisées dans cette étude. La permutation du suffixe est le phénomène le plus rare puisqu'il n'est observable que dans un cinquième des cas. Nous constatons néanmoins que la permutation du suffixe regroupe, dans notre corpus, une vingtaine de combinaisons possibles. Les plus récurrentes sont le passage du « d » final sur le nom de lieu normé au « t », du « s » au « z » et du « a » au « e ». Il s'agit là de variations propres aux corpus Renumar et BnF, des corpus portant sur des récits d'une époque antérieure. Les variations sont donc significatives de l'évolution diachronique des toponymes. La permutation d'une lettre en milieu de mot dénombre quant à elle un total de 37 combinaisons possibles. L'une de ces combinaisons se démarque particulièrement : le changement du « i » du toponyme normé en « y », avec une occurrence de 70%. Un remplacement du « e » par le « a » (et inversement) est également identifiable. Ici encore, ces variations proviennent majoritairement des corpus Renumar et BnF.

Pour ce qui est de l'ajout ou la suppression de lettres finales, le « s » est récurrent.

Quelle mesure de similarité pour quelle variation ?

Au cours du processus d'identification, nous avons indiqué utiliser une mesure de similarité par défaut. Nous nous sommes par la suite demandé si le choix de la mesure de similarité pouvait influencer la quantité de toponymes correctement identifiés en fonction de leurs variations. L'objectif est alors de définir, si elle existe, la mesure de similarité adaptée à chaque variation. Pour cela, il faut calculer la qualité de chaque similarité pour une variation donnée, évaluée comme le nombre de faux positifs qu'elle permet d'écarter. Nous avons tenté de répondre à la question en utilisant deux approches différentes : l'une par apprentissage supervisé guidé par l'outil *Weka*[1] (Witten, Frank, Hall, & Pal, 2016) et l'autre par apprentissage statistique.

Statistiques

Ici, la qualité des mesures de similarité est évaluée pour chaque variation de la façon suivante.

Pour un gazetier $g1$, une mesure de similarité $simi1$ est utilisée pour comparer un toponyme de $g1$ transformé par une variation $v1$ à son référent normé. La similarité obtenue constitue le seuil de référence ref.

Prenons l'exemple d'un corpus dans lequel les toponymes ont été annotés et associés à leur référent normé. Dans ce corpus,« noisy »est un toponyme associé au référent normé Noisy-le-Grand. La similarité Jaro entre la troncature noisy et son référent normé Noisy-le-Grand est de 0.6 et constitue le seuil de référence (ref).

Ensuite, le mot transformé par $v1$ est comparé à tous les toponymes du gazetier sans variation $g1$. À chaque fois, si la similarité obtenue est inférieure au seuil de référence ref, la quantité de mots moins similaires augmente.

noisy est ainsi comparé à Noisy-le-Sec. Dans les faits, bien que Noisy-le-Grand soit le bon référent normé, il est moins similaire à noisy que Noisy-le-Sec car il possède plus de caractères. En revanche, nous souhaitons que le seuil de référence ref reste supérieur aux autres similarités, c'est-à-dire que la similarité entre noisy et Noisy-le-Grand reste meilleure que celle entre noisy et Noisy-le-Sec et que donc le nombre de mots moins similaires soit maximal. L'opération est répétée avec tous les mots transformés du gazetier prétraité avec la variation $v1$. Si la part de mots moins similaires est élevée (proche de 100 %), alors nous considérons que pour la variation $v1$, la mesure de similarité $simi1$ est de bonne qualité.

Nous observons ensuite la qualité de toutes les mesures de similarités, pour toutes les variations et ce pour différents gazetiers.

Nous obtenons le tableau .10.

1. Weka regroupe un ensemble d'algorithmes pour la tâche de fouille de données, et propose entre autres des outils de classification et de partitionnement de données (clustering).

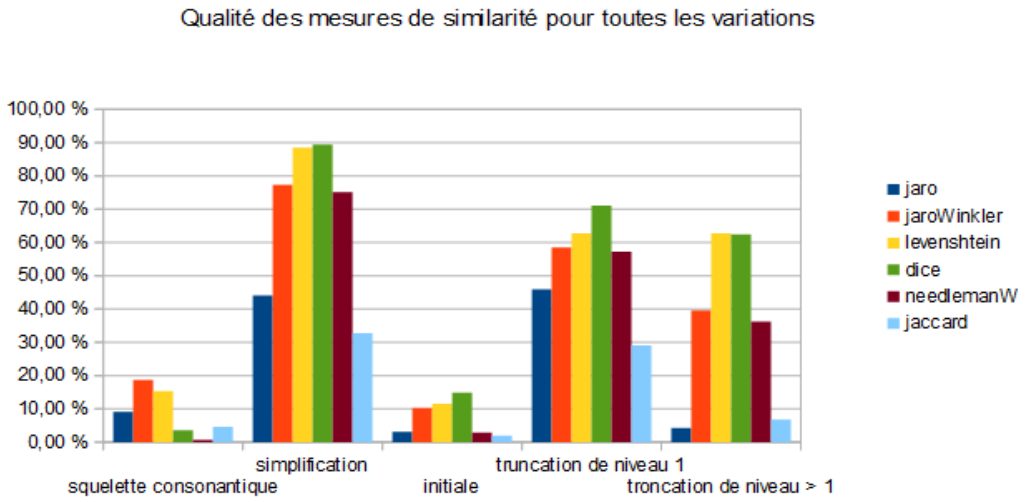


FIGURE .10. – Qualité des mesures de similarité selon la variation

À partir de ce tableau, plusieurs point sont été dégagés :

— La similarité à utiliser pour le squelette consonantique semble être Jaro Winkler, tandis que pour les autres variations c'est Dice

— La simplification et la troncation à un niveau sont les variations les mieux repérées

— la taille du gazetier a une influence sur la qualité de la similarité Jaro Winkler par rapport à la variation squelette consonantique : plus le gazetier a d'entrées et plus la qualité de Jaro Winkler est mauvaise. La similarité Jaro Winkler n'est donc pas la mesure la plus adaptée au squelette consonantique

— Pour la variation initiale, plus un toponyme polylexical a de constituants, moins il a de chance d'être associé à plusieurs toponymes normés. La part de mots moins similaires a donc plus de chances d'atteindre les 100%. La qualité de la mesure de similarité est donc automatiquement améliorée. Finalement, il n'est pas certain que Dice soit la mesure la plus adaptée à la variation initiale

En somme, certains facteurs influencent la qualité des mesures de similarité et remettent ainsi en cause le classement des meilleures mesure à utiliser selon la variation.

Weka

La seconde approche est une méthode par apprentissage supervisé.

Il faut dans un premier temps constituer un jeu de données. Le jeu de données utilisé ici a été conçu de la façon suivante :

— calcul de similarité entre un toponyme avec variation $t1$ et tous les toponymes transformés issus d'un gazetier prétraité avec une variation $v1$

— seule la similarité la plus élevée est conservée

— si le référent normé de $t1$ correspond à la forme normée du toponyme avec la plus haute similarité , alors nous indiquons C1

— les étapes précédentes sont réitérées avec un toponyme $t2$
— puis avec une autre mesure de similarité
— quand toutes les mesures de similarité ont été testées avec $v1$, nous passons à la variation $v2$ et nous répétons le processus, ainsi de suite.

	Variation 1			Variation 2			
	mesure de simi 1	mesure de simi 2	mesure de simi 3	mesure de simi 1	mesure de simi 2	mesure de simi 3	c1 / c0
toponyme a) avec variation	sa11	sa12	sa13	sa21	sa22	sa23	c0
toponyme b) avec variation	sb11	sb12	sb13	sb21	sb22	sb23	c1

FIGURE .11. – Jeu de données pour apprentissage supervisé

Le jeu de données présenté comme dans la figure .11 passe ensuite par Weka. L'algorithme utilisé est **Random forests** [2] (Breiman, 2001). Par défaut, le corpus de test est découpé en 10 et à chaque fois l'algorithme apprend sur 90% et teste sur les 10% restants.

À partir des données apprises, l'algorithme va tenter d'indiquer pour chaque ligne de similarités s'il s'agit d'un C0 ou d'un C1. Notre intérêt porte néanmoins sur l'arbre décisionnel généré par l'algorithme. Nous voulons savoir si une variation associée à une certaine mesure de similarité permet de trancher entre C0 et C1.

Malheureusement, la tâche d'apprentissage supervisé comme celle par statistiques ne permet pas d'établir un lien entre une variation spécifique et une mesure de similarité.

Nous pourrions forcer la donne et construire un jeu de données constitué uniquement de toponymes réduits à leur squelette consonantique par exemple et associés à leur référent normé auquel nous appliquons toutes les variations. Ensuite, toutes les mesures de similarités seraient calculées. La démarche serait jugée performante si d'une part le squelette consonantique devient la variation la plus détectée, et d'autre part si une similarité se démarque.

2. Random forests est un algorithme pour les arbres de décisions utilisant la classification ou la régression

Annexes 2

Les mots outils

saint, st, ste, sainte, stes, saintes, sn, san, sta, santa, santo, le, la, las, les, lès, l', l, el, los, the, dans, ds, aux, à, du, d, de, of, des, del, et, vers, hors, entre, en, sur, s/, sous, s/s

Les noms génériques géographiques

abattoir, abbatiale, abbaye, abreuvoir, abri de montagne, adret, aérodrome, aérogare, aéroport, aérotrain, agglomération, aiguille, aire d'autoroute, aire de garage, aire de péage, aire de repos, aire de service, aire de stationnement, aire de triage, allée, allée forestière, allée piétonne, alpage, altiport, amer, amer, amphithéâtre romain, amphithéâtre romain, amphithéâtre romain, antécime, antenne, antre, appontement, appontement, aquarium, aqueduc, aqueduc, arbre, arc de triomphe, arcade, arène, arène, arête, arrêt ferroviaire, arrondissement, ascenseur à bateaux, ascenseur à bateaux, atelier, auberge, autodrome, autoroute, auvent, aven, bac, balcon, balisage, balise, balise, ballastière, banc, baraquement, barrage, barrage, barrage hydroélectrique, barrage hydroélectrique, barrage réservoir, barrage réservoir, barrage voûte, barrage voûte, barranco, barre rocheuse, barrière, base, base nautique, basilique, bassin, bassin d'écluse, bassin d'écluse, bassin d'épuration, bassin d'épuration, bassin de compensation, bassin de compensation, bassin de décantation, bassin de décantation, bassin de filtrage, bassin de filtrage, bassin de lagunage, bassin de lagunage, bassin de rétention, bassin de rétention, bassin de rétention, bassin fermé, bassin lacustre, bastille, bastille, bâtiment administratif, bâtiment commercial, bâtiment culturel, bâtiment d'habitation, bâtiment et installation, bâtiment et installation agricoles, bâtiment et installation industriels, bâtiment et installation sport et loisirs, bâtiment militaire, bâtiment religieux, bâtiment remarquable, bâtiment remarquable, beffroi, beffroi, belvédère, berge, bergerie, bergerie, biaou, biaou, bief, bief, blockhaus, blockhaus, bois, borne frontière, borne géodésique, borne kilométrique, bosquet, bouée, bouée, bourg, brèche, bretelle, bretelle d'accès, bretelle d'échangeur, bretelle de sortie, bric, broussaille, bungalow, bureau de poste, cabane, câble transporteur, caillou, cairn, cale, cale, cale sèche, cale sèche, calotte, calotte, calvaire, camp, camp de vacances, camp militaire, campanile, camping, canal, canal, canal d'alimentation, canal d'alimentation, canal d'irrigation, canal d'irrigation, canal de décharge, canal de décharge, canal de dérivation, canal de dérivation, canal de navigation, canal de navigation, canal de restitution, canal de restitution, canalisation, canalisation, canon à neige, canton, canyon, captage, captage, carrefour dénivelé, carrière, cascade, casemate, casemate, caserne, caserne de pompiers, castel, castel, cathédrale, cave, caverne, cavité, cénote, centrale électrique, centrale hydro-electrique, centrale hydro-électrique, centrale nucléaire, centrale thermique, centre, centre culturel, centre de soin, chaîne de montagnes, chalet, champ, champ de courses équestres, champ de tir, champignonnière, chaos, chapeau, chapelet de lacs, chapelle, château, château, château d'eau, château d'eau, château fort, château fort, chaumière, chaumière, chaumière, chemin, chemin d'exploitation, chemin de fer, chemin de halage, chemin de petite randonnée, chemin muletier, cheminée, choroum, chute d'eau, cime, cimetière, cimetière communal, cimetière islamique, cimetière israélite, cimetière militaire, cimetière pour animaux, circuit auto-moto, circuit automobile, circuit de VTT, cirque, citadelle, citadelle, cité, citerne, citerne, clairière, clinique, clocher, clôture, cluse, col, collège, collégiale, colline, colonie de vacances, combe, commerce, communauté de communes, commune, complexe sportif, conduite hydrocarbure, confluence, contrefort, contrefort, contrefort, coopérative, corniche, corniche, corniche, côte, coteau, couloir, cours d'eau, cours d'eau artificiel, cours

d'eau artificiel, court de tennis, couvent, crassier, cratère, crèche, cressonnière, crête, crevasse, crevasse, croix, croupe, croupe, cuve à vin, cuvette, cynodrome, dalle de protection, déblai, déclivité, déclivité, delta, demeure, demoiselle coiffée, dent, département, dépression, descente, descente, déversoir, déversoir, diffluence, digue, digue, dispensaire, district, division territoriale administrative, doline, dolmen, donjon, donjon, drawbridge, éboulis, écart, échangeur, écluse, écluse, école, école maternelle, école primaire, écrêteur de crues, écrêteur de crues, écurie, écurie, église, église protestante, égout, égout, électrique, élément du patrimoine, élévation, embarcadère, embarcadère, embouchure, embranchement, embranchement de voie ferrée, embut, enceinte, entaille, entraille, entraille, entrepôt, éolienne, épaulement, éperon, éperon, équipement agricole, équipement de protection, équipement de protection hydrographique, équipement de protection hydrographique, équipement de sports d'hiver, équipement géodésique, équipement scientifique et technique, équipement sportif, escalier, escarpement, espace vert, estacade, estacade, estuaire, établissement d'enseignement, établissement de santé, établissement thermal, étang, état, étier, étier, excavation, exploitation maraîchère, exsurgence, exutoire, fabrique, face, faille, faîte, falaise, faubourg, ferme, ferme-auberge, feu, feu, fissure, flanc, fleuve, foire, fontaine, fontaine, forêt, fort, fort, forteresse, forteresse, fortification, fortification, fossé, fossé, franchissement, fronton de pelote basque, funérarium, funiculaire, galerie d'amenée d'eau, galerie d'amenée d'eau, garage (individuel), gare, gare de fret, gare téléphérique, gare téléphérique, gare voyageurs, garrigue, gave, gazoduc, gazon, gendarmerie, gentilhommière, gîte, gîte d'étape, glacier, gorge, gouffre, gour, GR, gradin, grande école, grange, gravière, grève, grotte, gué, gymnase, habitation de loisir, habitation troglodytique, habitation troglodytique, haie, haie, halle, hameau, hangar, hangar industriel, haras, haras, haut fourneau, havre, havre, héliport, herbage, hôpital, hospice, hôtel, hôtel de montagne, hourquette, hutte, hutte, hutte, hydrographie, hypermarché, igue, île, îlot, impasse, infrastructure, infrastructure hertzienne, infrastructure de captage d'eau douce, infrastructure de captage d'eau douce, infrastructure de déplacement aérien, infrastructure de déplacement maritime ou fluvial, infrastructure de déplacement maritime ou fluvial, infrastructure de déplacement terrestre, infrastructure de gestion de l'eau, infrastructure de gestion de l'eau, infrastructure de transport d'eau, infrastructure de transport d'eau, infrastructure de transport ferré, infrastructure de transport par câble, infrastructure de transport routier, infrastructure déplacement, infrastructure eau, infrastructure eau énergie communication, infrastructure électricité, infrastructure hydrocarbure, installation agricole, installation d'aquaculture, installation d'élevage, installation loisir, installation minière, installation service, intersection, intersection, itinéraire, itinéraire à ski, itinéraire de randonnée pédestre, itinéraire vélo, jardin public, jardinerie, jetée, jetée, karst, lac, lac d'altitude, laie forestière, lande, lapiaz, lavoir, lavoir, levée, levée de terre, levée de terre, lézarde, lieu-dit, ligne de faîte, ligne électrique, limite matérialisée, lit de rivière, localité, lotissement, lycée, magasin, mairie, maison, maison à thème, maison de repos, maison de retraite, maison du parc, maison forestière, maison médicale, maisonnette, mamelon, manoir, manoir, manoir, manteau, manteau, manufacture, maquis, marais, marbre, marché, mare, marécage, marquage, massif boisé, menhir, mer de glace, métro aérien, métropolitain, minaret, mine, minoterie, mont, montagne, monument, monument aux morts, monument religieux, moraine, mosquée, moulin, mou-

lin, moulin à eau, moulin à eau, moulin à vent, moulin à vent, mur, mur, mur anti-bruit, mur d'escalade artificiel, mur de fortification, mur de fortification, mur de soutènement, musée, nant, nappe d'eau, nécropole, névé, noeud du réseau hydrographique, nœud du réseau hydrographique, obélisque, observatoire à incendie, observatoire astronomique, observatoire de tir, office de tourisme, oléoduc, oppidum, oratoire, orographie artificielle, ossuaire, ouverture, ouvrage fortifié, ouvrage fortifié, pacage, pancarte, panneau, panorama, parc aquatique, parc d'activité, parc d'attraction, parc de loisir, parc des expositions, parc marin, parc marin, parc national, parc naturel regional, parc zoologique, pare feu, pare-avalanche, parking, paroi, paroi, pas, passage, passage à bétail, passe, passerelle, pature, pâture, pays (région), pédiment, pelouse, pente, pépinière, perte, pertuis, phare, phare, phare, phare, pic, pied, pierrier, pigeonnier, pigeonnier, pilier, pipe-line, piscine, piste cyclable, piste d'athlétisme, piste d'aviation, piste de bobsleigh, piste de cross, piste de ski, piste de sport, piste équestre, piton, place, plage, plaine, plan d'eau, plantation, plaque de pierre, plateau, point de desserte, point de vue, pointe, ponceau, pont, pont mobile, pont suspendu, pont transbordeur, pont-canal, ponton, ponton, port, port, port de plaisance, port de plaisance, port de plaisance, porte de ville, portique, poste de douane, poste de police, poste de transformation, poste électrique, prairie, pré, précipice, prefecture, préfecture, prieuré, prise d'eau, prise d'eau, prison, promontoire, puisard, puits, puits, pylône, pyramide, pyramide, quai, quai, quartier, radar, rade, rade, radier, radoub, radoub, raffinerie, raillère, rails, rangée d'arbres, ravin, ravine, reculée, refuge, région, régional, relais hertzien, relief, remblai, remonte pente, remonte pente, rempart, rempart, repère de nivellement, replat, réseau express régional, réseau hydrographique naturel, réserve nationale de chasse, réserve naturelle, réservoir, réservoir, réservoir agricole, réservoir industriel, ressaut, ressaut, ressaut, ressaut, résurgence, retenue collinaire, retenue collinaire, retenue d'eau, retenue d'eau, retenue sur cours d'eau, retenue sur cours d'eau, ride, rigole, rigole, rimaye, rimaye, riou, rivage, rive, rivière, roche, rocher d'escalade, rond point, roubine, roubine, route, route à chaussées séparées et carrefours dénivelés, route départementale, route nationale, route touristique, route vicinale, ru, rue, ruelle, ruine, ruisseau, sablière, salle de spectacle, salle de sport, sanctuaire, saut, scierie, sente, sérac, serre, silo, siphon, site d'escalade, site de fouilles, site de vol libre, socle, sommet, source, square, stade, station balnéaire, station d'épuration, station d'épuration, station de filtrage, station de filtrage, station de lagunage, station de lagunage, station de pompage, station de pompage, station de relèvement, station de relèvement, station de sports d'hiver, station de voyageurs, station météorologique, station scientifique, station thermale, statue, statue de la vierge, statue religieuse, stèle, superette, supermarché, surface d'eau, surplomb, synagogue, table d'orientation, taillis, talus, talweg, tapis roulant industriel, télébenne, télébenne, télécabine, télécabine, téléphérique, téléphérique, télésiège, télésiège, téléski, téléski, temple, temple bouddhiste, temple hindouiste, terrain d'aviation, terrain de football, terrain de golf, terrain de manoeuvre, terrain de manœuvre, terrain de rugby, terrain de sport, terrasse, terril, théâtre, théatre antique, théatre antique, théâtre antique, théâtre de plein air, théâtre de plein air, toboggan, toit, tombe, torchère, torrent, tour, tour, tour de contrôle, tour de télécommunication, tourbière, tourelle, tourelle, tourment, tramway, tranchée, transformateur, transport urbain, tribunal, tribune, trou, trouée, tumulus, tunnel, ubac, université, usine,

usine de traitement des eaux, usine de traitement des eaux, val, vallée, vallon, vanne, vanne, vasque, végétation, végétation arborée, végétation basse, vélodrome, verger, versant, vestige, viaduc, vigne, village, village de vacances, ville, voie, voie antique, voie de communication non carrossable, voie de garage, voie de service, voie de triage, voie ferrée à crémaillère, voie ferrée de transit, voie ferrée industrielle, voie ferrée principale, voie ferrée touristique, voie ferrée urbaine, voie rapide, voie TGV, voie-mère d'embranchement, volcan, watergang, watergang, zonage, zone arbustive, zone artisanale, zone bâtie, zone commerciale, zone d'activité, zone humide, zone industrielle, zone militaire, zone naturelle protégée

Un extrait du gazetier de toponymes

Villers-le-Tourneur [Villers-le-Tourneur : PAIHABIT0000000026840588 : BDNYme (4.5690517,49.627083)]

Cubzac-les-Ponts [Cubzac-les-Ponts : PAIHABIT0000000050812855 : BDNYme (-0.44946605,44.971577)]

La Campanuca [La Campanuca : 3119905 : Geonames (-3.79473,43.41833)]

Landorthe [Landorthe : PAIHABIT0000000073101374 : BDNYme (0.77292335,43.13401)]

Andicona [Andicona : 3130075 : Geonames (-2.58333,43.18333)]

le Verdier [le Verdier : PAIHABIT0000000051869639 : BDNYme (1.8411552,43.988384)]

Museo Carmen Thyssen [Museo Carmen Thyssen : 10280675 : Geonames (-4.42283,36.72148)]

Gron [Gron : PAIHABIT0000000016748443 : BDNYme (2.74365,47.11991), Gron : PAIHABIT0000000041799613 : BDNYme (3.261832,48.158695)]

Pescadoires [Pescadoires : PAIHABIT0000000015054891 : BDNYme (1.1580505,44.50545)]

Saint-Loup-du-Dorat [Saint-Loup-du-Dorat : PAIHABIT0000000026332559 : BDNYme (-0.41758454,47.890194)]

Gros [Gros : 6325023 : Geonames (-1.9726,43.32485)]

Mayence [Mayence : 2874225 : Geonames (8.2791,49.98419)]

Villarejo-Sobrehuerta [Villarejo-Sobrehuerta : 3104978 : Geonames (-2.48681,40.01907)]

na Berenguera [na Berenguera : 6690252 : Geonames (1.51019,39.11456)]

Fuente de las Donas [Fuente de las Donas : 2518821 : Geonames (-0.75,39.25)]

Revilla [Revilla : 3111921 : Geonames (-3.86616,43.40642), Revilla : 3111922 : Geonames (0.14683,42.59888), Revilla : 3111923 : Geonames (-3.76522,41.15891), Revilla : 3111924 : Geonames (-5.29098,41.07865)]

Grou [Grou : 3121080 : Geonames (-8.06667,41.95), Grou : 3121081 : Geonames (-8.03333,41.95), Grou : 3121082 : Geonames (-8.05,41.93333)]

Aubière [Aubière : PAIHABIT0000000119957356 : BDNYme (3.1240816,45.752373)]

Montmain [Montmain : PAIHABIT0000000052809910 : BDNYme (5.0640655,47.028984), Montmain : PAIHABIT0000000059071294 : BDNYme (1.244253,49.407837)]

Sierra del Calvario [Sierra del Calvario : 2520494 : Geonames (-5.2,38.36667)]

Gamarte [Gamarte : 3016734 : Geonames (-1.14243,43.20114)]

Sanry-sur-Nied [Sanry-sur-Nied : PAIHABIT0000000059959655 : BDNYme (6.3441143,49.05172)]

Toponymes extraits de La Grande Peur

En raison de la licence du corpus La Grande Peur n'autorisant pas son partage, voici un extrait des toponymes issus du corpus :

Forme normée ; Forme avec variation

Champvallon ; Champvallon
Champvallon ; Champ-vallon
Gâtinais ; Gâtinais
Château-Renard ; Châteaurenard
Châtillon-Coligny ; Chatillon-sur-Loing
Saint-Fargeau ; Saint-Fargeau
Aillant-sur-Tholon ; Aillant
Villiers-Saint-Benoît ; Yilliers-sous-Benoît
Puisaye ; la Puisaye
Thury ; Thury
Entrains-sur-Nohain ; Entrains
Loire ; le val de Loire
Briare ; Briare
Sancerre ; Sancerre
La Charité-sur-Loire ; La Charité
Nevers ; Nevers
La Charité-sur-Loire ; La Charité
Nevers ; Nevers
Yonne ; la vallée de l' Yonne
Auxerre ; Auxerre
Champs-sur-Yonne ; Champs
Cure ; La vallée de la Cure
Avallon ; Avallon
Vézelay ; Vézelay
Clamecy ; Clamecy
Tannay ; Tannay
Lormes ; Lormes
Corbigny ; Corbigny
Montsauche-les-Settons ; Montsauche
Saulieu ; Saulieu
Yonne ; l' Yonne
Château-Chinon ; Château-Chinon
Autun ; Autun
Decize ; Decize
Loire ; la Loire
Arroux ; l' Arroux
Bourbon-Lancy ; Bourbon-Lancy

Toponymes extraits des SMS

En raison de la licence du corpus de SMS n'autorisant pas son partage, voici un extrait des toponymes issus du corpus :

Forme normée ; Forme avec variation

Saint-Éloi ; st eloi
Boutonnet ; boutonnet
Saint-Éloi ; st eloi
Saint-Éloi ; st eloi
Saint-Éloi ; st eloi
Nîmes ; Nimes
Montpellier ; Montpel
Marseille ; Marseille
Montpellier ; montpel
Bédarieux ; bédarieux
Montpellier ; montpel
Collioure ; collioure
Collioure ; collioure
Montpellier ; montpellier
Montpellier ; montpellier
Marguerittes ; Marguerittes
Toulouse ; Tlse
Toulouse ; tlse
Toulouse ; Tlse
Peaugres ; peaugres
Sète ; Sète
Le Vigan ; au vigan
Montpellier ; mtpl
Montpellier ; mtpl
Bougnol ; bougnol
Béziers ; Bezier
Carnon-Plage ; Carnon
Béziers ; Bezier
Nîmes ; Nimes
Nîmes ; Nimes
Collioure ; collioure
Collioure ; collioure
Montpellier ; mtpl
Saint Jean de Védas ; Saint Jean de Védas
Peyrens ; peyrens
Millau ; millau
Aix-en-Provence ; AIX

Montpellier ; mtpl
Montpellier ; mtpl
Perpignan ; perpignan
Montpellier ; mtpl
Montpellier ; montpellier
Montpellier ; Mtp
Montpellier ; mtp
Nîmes ; Nimes
La Grande-Motte ; la grande motte
Laborde ; Laborde
Aigues Mortes ; aigue morte
Orléans ; orléan
Carcassonne ; Carcassonne

Toponymes extraits de la BnF

Voici un extrait des toponymes issus de la base de données data.bnf :

Forme normée ; Forme avec variation

Amareins ; Amarains
Amareins ; Amarins
Ambérieu-en-Bugey ; Amberieux
Ambérieu-en-Bugey ; Amberiacus
Ambérieu-en-Bugey ; Ambereu
Ambérieu-en-Bugey ; Ambeiriacus
Ambérieu-en-Bugey ; De Ambayreu
Ambérieux-en-Dombes ; Ambayreu
Ambérieux-en-Dombes ; Ambeyrieux
Ambérieux-en-Dombes ; Ambrei
Ambérieux-en-Dombes ; Amberiacus
Ambronay ; Ambrognay
Ambronay ; Ambronais
Ambronay ; Ambournay en Bugey
Ambronay ; Anbronnay
Andert-et-Condon ; Andert-Condon
Andert-et-Condon ; Villa d'Anderno
Anglefort ; Inflafol
Aranc ; Aran
Aranc ; Arenc
Arandas ; Arandas en Bugey
Arandas ; Arandaz
Arbent ; Albeins
Arbent ; Arbens
Arbent ; Arban
Arbigny ; Albinies
Arbigny ; Arbignia
Argis ; Argit
Argis ; Argil
Arlod ; Arlos
Armix ; Armieis
Armix ; Hermis
Armix ; Armex
Ars-sur-Formans ; Arz
Ars-sur-Formans ; Art
Artemare ; Arthamaraz
Asnieres-sur-Saône ; Anires
Asnieres-sur-Saône ; Anieres

Extrait du corpus MATRICIEL

```
<?xml version='1.0' encoding='utf-8'?>
<GateDocument version="3">

<!-- The document content area with serialized nodes -->

<TextWithNodes>p<Node id="1"/>JV17&#xd;
<Node id="463"/>JV17 : Bien mes mes parents Ã©taient tous les deux de <Node
id="516"/>Salamanca<Node id="525"/>(...)
</TextWithNodes>


<!-- The default annotation set -->
<AnnotationSet>
<Annotation Id="283" Type="Geonames" StartNode="516" EndNode="525">
<Feature>
  <Name className="java.lang.String">62-country</Name>
  <Value className="java.lang.String">Espagne</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">language</Name>
  <Value className="java.lang.String">fr</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">67-geometry</Name>
  <Value className="java.lang.String">NR</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">minorType</Name>
  <Value className="java.lang.String">min</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">5-id</Name>
  <Value className="java.lang.String">3111107</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">52-subcategory</Name>
  <Value className="java.lang.String">ADM2</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">majorType</Name>
  <Value className="java.lang.String">maj</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">65-x</Name>
  <Value className="java.lang.String">-6</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">3-theme</Name>
  <Value className="java.lang.String">political</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">1-source</Name>
  <Value className="java.lang.String">geonames</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">66-y</Name>
  <Value className="java.lang.String">40.83333</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">2-entityType</Name>
  <Value className="java.lang.String">loc</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">42-sentiment</Name>
```

Toponymes extraits de Renumar

Forme normée ; Forme avec variation

Château de Loches ; Loches
Ballan-Miré ; Ballan
Saint-Cyr-sur-Loire ; Saint-Cyr
Joué-lès-Tours ; Joué
Châteaudun ; Chasteaudun
Bourré ; Bourray
Le Mans ; Mans
Blois ; Bloys
Saint-Aignan ; Sainct Aignen
Orléans ; Orleans
Chambray-lès-Tours ; Chambray
Véretz ; Veretz
Berchères-les-Pierres ; Bercheres l'Evesque
Parçay-Meslay ; Parçay
Lussault-sur-Loire ; Lussault
Chamblay ; Chamblé
Notre-Dame-d'Oé ; Notre-Dame-d'Oé
Mainvilliers ; Mainvillier
Bailleau-l'Évêque ; Baillau l'Evesque
Saint-Pierre-des-Corps ; Saint Pierre des Corps
Sénarmont ; Senermont
Lussault-sur-Loire ; Lusault
Cinq-Mars-la-Pile ; Cinq-Mars
Beaulieu-lès-Loches ; Beaulieu
Amiens ; Amyens
Chanceaux-sur-Choisille ; Chanceaux
Vendomois ; Vendômois
Sainte-Maure-de-Touraine ; Sainte-Maure
Saint-Branchs ; Saint Branch
Berchères-les-Pierres ; Bercheres
Bailleau-l'Évêque ; Bailleau l'Evesque
Ver-lès-Chartres ; Ver
Vendôme ; Vendosme
Gellainville ; Gelainville
Ferrara ; Ferrare
Bailleau-l'Évêque ; Baillau
Rouziers-de-Touraine ; Rouziers
Nogent-le-Phaye ; Nogent
La Ville-aux-Dames ; Ville-aux-Dames
Rosay-au-Val ; Rosay

Extrait du corpus CHOUCAS

```
      <w lemma="suivre" type="V" subtype="motion_median">Suivre</w>
      <geogFeat>
        <w lemma="le" type="DET">la</w>
        <w lemma="route" type="N">route</w>
      </geogFeat>
      <offset type="direction" subtype="final">
        <w lemma="jusque" type="PREP">jusqu'</w>
        <w lemma="à" type="PREP">à</w>
      </offset>
      <placeName n="1" xml:id="ene.18">
        <geogName type="S" subtype="CH">
          <w lemma="le" type="DET">la</w>
          <geogFeat>
            <w lemma="chapelle" type="N">chapelle</w>
          </geogFeat>
          <placeName n="0" xml:id="ene.19">
            <name>
              <w type="NPr" lemma="">Saint-Roch</w>
            </name>
          </placeName>
        </geogName>
      </placeName>
    </phr> .</s>
  <s>
    <phr type="verb_phrase" subtype="motion" xml:id="phr.17">
      <w lemma="traverser" type="V" subtype="motion_median">Traverser</w>
      <placeName n="0" xml:id="ene.20">
        <name>
          <w type="NPr" lemma="">Monbonnet</w>
        </name>
      </placeName>
    </phr>
    <w lemma="pour" type="CONJC">pour</w>
    <phr type="verb_phrase" subtype="motion" xml:id="phr.18">
      <w lemma="rejoindre" type="V" subtype="motion_final">rejoindre</w>
      <w lemma="le" type="DET">la</w>
      <rs type="place">
        <name subtype="roadName">
```

Résumé

MOTS-CLÉS : toponymes, variations d'écriture, néographie, corpus hétérogènes, reconnaissance d'entités nommées, mesures de similarité

Les toponymes sont parfois amenés à subir des variations d'écriture et voient leur graphie s'éloigner de celle que nous trouvons habituellement dans les dictionnaires de noms propres. Ces variations d'écriture peuvent dépendre du type de corpus dont les toponymes sont issus, du registre, du temps ou du langage que couvre le corpus. Nous proposons une méthode pour identifier un toponyme, c'est-à-dire faire le lien entre un toponyme avec variations d'écriture et sa forme normée. Le présent mémoire se divise en trois parties. Dans une première partie nous présenterons le terrain de stage. Dans un second temps nous décrirons les missions confiées puis nous exposerons l'approche envisagée ainsi que les outils utilisés pour répondre aux missions. Enfin, nous proposerons des solutions afin de répondre au mieux au besoin d'identification de toponymes avec variations d'écriture.

Abstract

KEY WORDS : Toponyms, written variations, new written form, heterogeneous corpus, named entity recognition, string metrics

Toponyms can sometimes be the target of written variations and experience a different written form from what we usually find in proper names dictionnaries. These written variations can depend on the type of the corpus the toponyms come from, the register, the time or the language that the corpus covers. We suggest a method to identify a toponym, which means making the link between a toponym with written variations and its normalised form. This Master's thesis is divided into three parts. In the first part we will present the context of the internship. Then we will describe the given tasks and we will expound the contemplated approachs and the tools that we used. Finally, we will suggest solutions to meet at best the need of toponyms identification with written variations.