Noms de lieux dans le corpus de français parlé : une approche symbolique pour un traitement automatisé

Hélène Flamein, Iris Eshkol-Taravella

Noms de lieux dans le corpus de français parlé : Une approche symbolique pour un traitement automatisé

Hélène FLAMEIN, Iris ESHKOL-TARAVELLA

1. Introduction

Le progrès technologique et Internet ont permis la constitution et la mise à disposition de grandes quantités de données de nature variée. C'est le cas des corpus oraux disponibles en ligne comme ESLO, CLAPI, BNC[1], etc. L'exploitation des outils du TAL rend ces données accessibles, consultables et visualisables.

Le travail présenté propose d'exploiter le corpus ESLO (Enquêtes Sociolinguistique à Orléans) d'une manière nouvelle. Il s'agit d'analyser la perception qu'ont les Orléanais de leur ville à travers la détection et l'analyse des mentions de lieux dans les enregistrements réalisés par les chercheurs orléanais. Le travail présenté est un travail pluridisciplinaire qui s'inscrit dans les domaines de la linguistique de corpus, de la linguistique outillée, du traitement automatique du langage (TAL), de la géomatique et des humanités numériques.

Les objectifs de ce travail sont autant applicatifs que théoriques. Il s'agit d'une part de parvenir au traitement informatique de données jusqu'ici peu exploitées. En effet, la majorité des travaux dédiés à la détection des mentions de lieux ou entités spatiales (Lesbeguerries 2007) reposent sur des données textuelles comme des récits de voyages (Loustau et. al, 2008), des œuvres littéraires (Moncla et al., 2016) ou des documents textuels (journaux, cartes géographiques anciennes, lithographies, cartes postale,…) (Lesbeguerries, 2007). Le travail présenté s'appuie sur des transcriptions de conversations orales. Ce travail exploite donc des données de nature différente présentant de nouveaux défis en ce qui concerne l'extraction d'informations géographiques. Par ailleurs, en chaque maillon de la chaîne de traitement, les spécificités propres de l'oral sont analysées et prises en compte dans l'élaboration des différents modules d'annotation et d'extraction. L'autre enjeu de ce travail est d'étudier les différents procédés lexicaux, syntaxiques et pragmatiques que le locuteur met en œuvre pour désigner un lieu. La dénomination d'un lieu est un processus social réapproprié subjectivement et elle est déterminée par la personnalité, l'histoire du locuteur. La perception d'un lieu peut aussi apparaître à travers sa dénomination.

La notion de lieu est une notion complexe. Sa définition pose question que ce soit du point de vue de la linguistique, du TAL ou même de la géographie. La discussion de cette notion sera présentée dans la section 2. Le corpus traité est décrit dans la section 4. Il est composé de transcriptions de conversations orales spontanées. Le traitement de l'oral diffère de celui de l'écrit et pose certaines difficultés présentées dans la section 3. L'élaboration du système d'annotation automatique des mentions de lieux dans l'oral transcrit guidée par des travaux existants et par l'observation manuelle du corpus est décrite dans les sections 5 et 6. Cette annotation est réalisée selon une approche symbolique associant des règles et la manipulation de ressources lexicales. Enfin, l'analyse des mentions de lieux dans les transcriptions du français parlé est présentée dans la section 7.

2. Notion de lieu

La notion de lieu est l'élément central de ce travail et nécessite d'être explicitée. Le

[1] http://eslo.huma-num.fr/, http://clapi.ish-lyon.cnrs.fr/, https://corpus.byu.edu/bnc/

*Trésor de la Langue Française Informatisé (*TLFi)[2] considère le lieu comme l'espace « déterminé par sa situation dans un ensemble, par la chose qui s'y trouve ou l'événement qui s'y produit » ou « qualifié par un adjectif qui le caractérise dans ses dimensions, son aspect, sa qualité ». Pour le Larousse en ligne[3], le *lieu* est une « situation spatiale de quelque chose, de quelqu'un permettant de le localiser, de déterminer une direction, une trajectoire » ou un « endroit, localité, édifice, local etc., considérés du point de vue de leur affectation ou de ce qui s'y passe ». Le Petit Robert (1993) définit le *lieu* comme une « portion déterminée de l'espace, considérée de façon générale et abstraite ».

Toutes ces définitions relient la notion de lieu avec celle d'espace. Ainsi, le lieu possède cette propriété particulière de l'espace d'occuper une certaine étendue que l'on peut mesurer et situer par rapport à ce qui l'entoure. Cette propriété offre aussi la possibilité d'allouer des coordonnées géographiques au lieu afin de le placer sur un plan, de le localiser dans un espace.

Dans le domaine de la géographie, le concept de lieu est fondamental. Bailly et al. (2016) évoquent la difficulté de poser une définition précise du terme *lieu* et proposent de le considérer comme un point dans l'espace que l'on peut situer sur une carte. Dans le *Dictionnaire de Géographie et de l'espace des sociétés* (Levy, Lussault, 2013), Berque (2013) continue en affirmant que le lieu est « *là où quelque chose se trouve ou/et se passe* ». Lussault (2007) dégage le concept d'« *identité spatiale* ». Le lieu se trouverait doté de caractéristiques particulières, de singularités par les acteurs d'une société donnée. Les signes qui le particularisent permettent de le distinguer des autres objets spatiaux. Enfin, Tuan (1997) offre au lieu une dimension relationnelle : l'homme y est étroitement lié et ses interactions avec l'espace définissent le lieu. Les aspects géographiques et sociaux sont pris en compte de nouveau dans les définitions proposées.

Les linguistes s'intéressent à la notion du lieu majoritairement à travers l'étude des expressions spatiales (Boons 1987, Borillo 1998, Laur 1991, Vandeloise 1986, Le Pesant 2011a, 2012 etc.) ou des noms généraux d'espace (Huyghe 2009). On peut citer le travail de Xavier Gouvert (2008) qui décrit quelques propriétés linguistiques du fonctionnement des noms de lieux dans la phrase et constate qu'ils s'apparentent moins à un « nom propre » qu'à un « adverbe propre », dans la mesure où « les toponymes exercent primordialement et très majoritairement la fonction syntaxique de circonstant ou de second actant à signifié circonstanciel ».

Du point de vue de la lexicologie et plus particulièrement de l'onomastique, les lieux sont étudiés dans le cadre de la toponymie. La science de la toponymie ne fait pas la distinction entre les deux notions *lieu* et *toponyme.* La Commission Nationale de Toponymie (CNT) indique qu'un toponyme dénote un objet géographique déterminé. Pour l'IGN, « un toponyme est un nom de lieu, constitué d'un ou plusieurs mots, en rapport étroit avec un détail géographique localisé et avec le groupe humain qui l'utilise ». La notion de *toponyme* est donc une notion certes géographique mais aussi sociale. Roseline le Squère (2006), au travers d'une étude des toponymes bretons, s'intéresse aux fonctions que peuvent remplir les toponymes. Si « dans leurs premières fonctions, les toponymes guident, informent sur le territoire qu'ils nomment », ils peuvent notamment participer à la réactivation de « la mémoire du lieu et de l'ensemble du territoire auquel il appartient », « faire fructifier le capital culturel et économique du territoire » mais aussi « catégoriser un territoire de façon positive ». Le nom d'un lieu n'est pas une simple étiquette sur un espace : il symbolise son histoire, les enjeux qui lui sont propres.

De nombreux travaux en TAL s'intéressent aux lieux et à leurs spécificités dans

---

[2] http://atilf.atilf.fr/dendien/scripts/tlfiv5/visusel.exe?12;s=2924586840;r=1;nat=;sol=1; [consulté le 10 mars 2014]

[3] http://www.larousse.fr/dictionnaires/francais/lieu/47076?q=lieu#47003 [consulté le 10 mars 2014]

l'optique d'un traitement automatique. Les lieux sont le plus souvent identifiés dans le cadre de recherche en TAL autour du concept d'entités nommées « des éléments informationnels pertinents dont on parle et qui jouent un rôle dans la description d'un événement, d'un fait » (Nouvel et al., 2015 :13). Ces travaux abordent les lieux du point de vue de leur détection automatique dans le cadre de la tâche de reconnaissance des entités nommées : lieux, personnes, organisations, etc. (Ehrmann 2008 ; Maurel et al., 2011 ; Nadeau, Sekine, 2009 ; Nouvel et al., 2015).

Lesbeguerries (2007) s'intéresse aussi à la notion de lieu dans le cadre de la Recherche d'Information (RI). Il propose le terme *entité spatiale* qui recouvre la définition du lieu et rend explicite la dimension géolocalisable d'un lieu. Cela suppose que tous les lieux ne sont pas forcément localisables sur une carte. Les expressions : *l'école*, *la grande rue là-bas* ou *dans la ville*, correspondent toutes à des lieux, mais sans éléments plus précis de contexte, il est impossible de déterminer avec précision à quel lieu on se réfère. Ces expressions peuvent donc être considérées comme des lieux mais pas comme des entités spatiales.

En conclusion, la plupart des travaux en linguistique s'intéresse au fonctionnement des noms de lieux dans le discours tandis que la géographie décrit la relation entre ce nom et l'espace qu'il représente. Le courant de la géographie humaniste met en évidence le lien entre l'homme et son environnement. C'est l'homme qui, par son histoire, son vécu, effectue la dénomination de l'espace qui l'entoure. Du point de vue des entités nommées, on se rend compte de la difficulté de la tâche d'annotation de ce type d'information. Le terme *entité spatiale* (Lesbeguerries, 2007) met l'accent sur la dimension géolocalisable et se place directement dans une perspective applicative de production de cartes géographiques.

Dans ce travail, le lieu est considéré comme un espace déterminé auquel l'homme a attribué un ou plusieurs noms, composés de noms propres et/ou de noms communs pour s'y référer. Son objectif est de détecter toutes les mentions de lieux dans le corpus traité. La majorité des travaux dédiés à l'extraction automatique des lieux et des entités spatiales reposent sur des données textuelles comme des tweets (Zesnani et. al, 2016), des récits de voyages (Loustau et. al, 2008), des œuvres littéraires (Moncla et. al, 2016), des écrits du Web (Dominguès, Eshkol-Taravella, 2015) ou des récits de vie (Dominguès, 2018). Contrairement aux travaux cités, le travail présenté est effectué sur le corpus oral transcrit. La démarche proposée prend en compte les caractéristiques propres de l'oral et les intègre dans la chaîne de traitement.

3. La dénomination des lieux à l'oral

Le discours oral diffère de l'écrit. L'écrit se présente au destinataire comme un produit fini tandis que l'oral se construit en temps réel. L'une des caractéristiques de l'oral est la présence de disfluences, des éléments qui rompent le flux de parole comme des marqueurs discursifs (*quoi*, *enfin*, …), des onomatopées ou interjections (*ah*, *euh*, *hum*, …), des répétitions, des reformulations, des reprises ou encore des amorces de mots. Observons l'exemple :

1. y a quoi par-là y a le **cons-** euh **une partie du conservatoire** y a oui si c'est **Sainte-Croix** quoi la **cathédrale** (ESLO2_iti_10_04_C)

où le nom de lieu amorcé *cons-* est suivi par une hésitation *euh* introduisant un lieu *une partie du conservatoire*. Par ailleurs, le marqueur discursif *quoi* apparaît au milieu du nom de lieu *Sainte-Croix la cathédrale*.

Au-delà de la présence des disfluences, l'instantanéité du discours oral a une

influence sur la façon dont un locuteur mentionne un lieu. Dans l'exemple [1], le lieu la *cathédrale Sainte-Croix* est mentionné par le locuteur dans l'ordre inverse *Sainte Croix la cathédrale* ce qui ne se produira jamais dans le discours écrit.

Par ailleurs, la dimension relationnelle qui existe entre homme et lieu rend possible l'appropriation, la personnalisation d'un lieu par un individu (Dominguès, Eshkol-Taravella, 2015) comme dans l'exemple [2].

2. on regardait **notre cathédrale** tous les deux (ESLO2_iti_07_01_C)

Les noms de lieux peuvent être tronqués, abrégés, voire remplacés par un surnom. Ces formes divergentes du nommage traditionnel sont à la fois porteuses d'indices sur la perception qu'a un locuteur d'un lieu et source d'erreurs au moment de la détection automatique des différentes mentions. Par exemple, dans :

3. bon puis les **bords de Loire** sont magnifiques maintenant (ESLO2_ENT_1070)

le lieu *bords de Loire* ne correspond pas au nom officiel auquel il fait référence : *Quai du Roi* ou *Quai du Châtelet*. Le locuteur préfère un nom plus générique à la convention pour se référer à son environnement.

Le locuteur peut se référer à un lieu à travers les déictiques (*cette place*, *ici*, *là*) ou en préférant des appellations plus génériques comme dans l'exemple [4] où le locuteur ZL473 évoque *Orléans* en utilisant le nom commun *ville*.

4. WT075: « et tu trouves que **Orléans** c'est une bonne **ville** ? en général ? »
ZL473: « euh non je pense que c'est pas une **ville** très étudiante »

L'emploi du nom de lieu dans le discours est une représentation de l'espace construite par le locuteur. La référence à un lieu donné est déterminée par le système toponymique du locuteur.

La détection automatique de lieux dans les transcriptions est une tâche complexe. La méthode proposée est guidée par la nature des données traitées et tient compte de leurs spécificités. Ainsi, les noms de lieux tronqués, généralisés, personnalisés, une partie des coréférences font partie de la chaîne de traitement proposée.

4. Corpus ESLO

Le corpus étudié est le corpus ESLO (Enquête Sociolinguistique à Orléans) et, plus précisément le corpus ESLO2, la suite d'une enquête réalisée au début des années 70 par des chercheurs britanniques (ESLO1). Quarante ans plus tard, dans les années 2010, une nouvelle enquête est lancée par le laboratoire LLL (Laboratoire Ligérien de Linguistique) pour constituer un corpus comparable à ESLO1. Le corpus ainsi constitué est appelé ESLO2[4]. Il comprend 18 modules représentant chacun une situation d'enregistrement différente : interviews d'habitants et de personnalités de la ville, des paroles captées dans la rue, les transports publics, les commerces, les lieux de travail, etc.

Deux modules favorisant la présence de noms de lieux sont sélectionnés pour ce travail : Entretiens (84 transcriptions, 1 167 000 mots, discussions en face à face entre un chercheur et un locuteur témoin à partir d'une trame d'entretien semi-directive) et Itinéraire (91 transcriptions, 70000 mots, enregistrements plus courts en micro-trottoir

[4] http://eslo.huma-num.fr

dans lesquels les enquêteurs demandent leur chemin au locuteur témoin) (cf. Tableau 1).

	Entretiens	***Itinéraires***
Description	Discussion en face à face entre un chercheur et un locuteur témoin à partir d'une trame d'entretien	Enregistrements dans la rue de demandes d'itinéraire (Comment aller à la mairie d'Orléans ?)
Protocole	Prise de contact préalable : présentation du projet et prise de rendez-vous. Enregistrements réalisés chez les locuteurs témoins	Demandes d'itinéraire dans la rue. Le début est à micro discret ; la suite à micro montré donne lieu à une reformulation de l'itinéraire. La collecte a été effectuée dans divers endroits de la ville afin de couvrir des quartiers représentatifs de la diversité sociologique de la ville.
Nombre de transcriptions	84	91
Nombre de mots total	1 166 654	69 328

Tableau 1 : Description des modules sélectionnés dans ESLO2[5]

Chaque enregistrement est transcrit orthographiquement avec une distinction entre les tours de parole. La convention de transcription préconise de transcrire sans signes de ponctuation et sans majuscules au début des énoncés. Seules exceptions, les points d'interrogation qui différencient les questions et les majuscules pour les noms propres.

5. Constitution du corpus de référence ou annotation manuelle

Pour pouvoir repérer et analyser les noms de lieux, nous avons procédé à leur annotation manuelle. Le corpus annoté manuellement est un corpus de référence pour le traitement automatique ainsi que pour l'analyse quantitative des lieux annotés.

5.1. Conventions d'annotation

Pour annoter l'ensemble des mentions de lieux présentes dans le corpus, nous utilisons la balise XML : <loc> à laquelle sont ajoutés trois attributs indiquant la nature des lieux identifiés :

- le type du lieu : naturel, ville, pays, oro (voies), monuments, etc.; - la zone géographique : 0 (lieux en dehors de l'agglomération orléanaise), 1 (lieux de l'agglomération) et 2 (lieux d'Orléans), afin de distinguer les lieux à géoréférencer des

[5] Les informations du tableau sont issues du site officiel du corpus : http://eslo.huma-num.fr/

autres ;
- le label ou le nom conventionnel du lieu annoté pour la manipulation des bases de données géographiques et l'association du lieu avec les coordonnées GPS correspondantes.

5.1.1 Typologie des lieux

En vue de l'analyse de l'image d'une ville, il est nécessaire de conserver une typologie fine des lieux urbains.

L'annotation des types de lieux s'inspire de la typologie des campagnes d'évaluation ESTER2[6] et ETAPE[7] (Rosset, Grouin et Zweigenbaum, 2011)[8]. La typologie établie comprend onze étiquettes (Tableau 2).

<loc type=" ">	
Villes	type="ville"
Orléans, Paris, La Ferté-St-Aubain, Dunois, La Source…	
Région	type="region"
Loiret, région Centre Val-de-Loire, Beauce, Gâtinais…	
Pays	type="pays"
France, Espagne, Royaume-Uni, Chine…	
Supranational	type="supra"
Europe, Asie du Sud Est, le Nord, la Flandre…	
Rues, avenues, ponts…	type ="voie"
rue de la République, Pont Royal…	
Lieux physiques naturels	type ="naturel"
Forêt d'Orléans, Loire, Canal de Briare…	
Lieux à dimension historique, touristique	type ="monument"
Cathédrale Sainte Croix, Hôtel Groslot…	

6 http://www.afcp-parole.org/camp_eval_systemes_transcription/

7 http://www.afcp-parole.org/etape.html

8 La campagne ETAPE a pour objectif d'évaluer les performances des technologies vocales appliquées à l'analyse de flux télévisés en langue française tandis qu'ESTER2 est un projet antérieur à ETAPE avec des objectifs similaires de mesure de performances de systèmes de transcriptions d'émissions radiophoniques.

Lieux à fonction administrative	type ="admin"
Mairie d'Orléans, Office du Tourisme, CAF…	
Lieux à fonction éducative	type ="educatif"
Lycée Pothier, Université d'Orléans…	
Lieux à fonction commerciale	type ="commerce"
Carrefour, H&M, Menphis Coffee…	
Lieux à fonction non commerciale	type ="ncommerce"
Hôpital de la Source, Secours Populaire,…	

Tableau 2 : Typologie des lieux

On peut observer les résultats de cette annotation dans l'exemple [5] :

5. comme je vais à la à l'**<loc type="educatif">université</loc>** à **<loc type="ville">La Source</loc>** ça me fait quand même de la route (ESLO2_ENTJEUN_03_C)

Suite à la typologie proposée, le nom de lieu *université* est annoté comme un lieu à fonction éducative, alors que *La Source* est une ville. Que le locuteur se réfère à l'endroit physique ou à l'institution *université d'Orléans*, il fait toujours référence à une entité ayant une portée éducative située à Orléans, qui fait partie intégrante de la ville et de son histoire.

5.1.2 Zone géographique

Outre l'information du type de lieu identifié, l'annotation des lieux doit présenter certaines informations relatives à leur localisation géographique. Avant de chercher à attribuer des coordonnées précises à chacune des mentions de lieux identifiées, trois zones géographiques ont été déterminées pour l'annotation. Les conventions d'annotation différencient ainsi les lieux situés à Orléans, les lieux hors Orléans mais situés dans son agglomération et les lieux en dehors de l'agglomération. Le découpage de ces trois zones correspond aux découpages administratifs de la ville d'Orléans et de son agglomération (Tableau 3) :

<loc type=" " zone=" ">	
zone ="0"	lieux hors agglomération orléanaise
Paris, Tours, Indre, Bretagne, Rhône, Seine …	
zone ="1"	lieux hors Orléans mais inclus dans l'agglomération

Saint Jean de la Ruelle, Saran, Auchan…	
zone ="2"	lieux situé à Orléans
Orléans, rue de Bourgogne, Key-West…	

Tableau 3 : Zones géographiques

L'information de la zone géographique permet un traitement différencié des annotations. Par exemple, un lieu considéré hors agglomération orléanaise n'est pas géoréférencé sur la carte finale. Ainsi, dans l'exemple [6], *Paris* ne sera pas géolocalisé sur la carte contrairement à *Orléans*.

6. c'est pas ça pose pas de problème donc euh ce qui manque à <loc type="ville" **zone="2" >Orléans**</loc> je dirais tu peux l'avoir à <loc type="ville" **zone="0">Paris**</loc> donc c'est vrai que euh (ESLO2_ENT_1008_C)

5.1.1. Label officiel

La nature orale des données étudiées influence l'emploi des noms de lieux. Il est rare qu'un lieu soit cité sous sa forme administrative et présente dans les bases de données institutionnelles. Lorsqu'un locuteur mentionne un lieu, il peut se l'approprier, le personnifier en opérant des modifications sur son nom (troncations, utilisation de surnom, etc.), alors que le nom officiel du lieu correspond à sa forme complète, sans aucune modification du locuteur. Cette différence de forme est marquée par une étiquette *label* correspondant au nom officiel du lieu identifié :

7. ah ben si tu peux redescendre tu prends la tu prends la rue qui est là et tu vas tout au bout jusqu'à la <loc type="voie" zone="2" **label="rue de la République">rue de la Rép-**</loc> tu vois où elle est ? la <loc type="voie" label=" rue de la République">rue de la République</loc> ? (ESLO2_iti_06_11_C)
8. je passais pas <loc type="ville" zone="0" **label="La Ferté-Saint Aubin">La Ferté**</loc> ça faisait loin hein ça me faisait cinquante kilomètres (ESLO2_ENT_1023_C)

Dans les exemples [7], [8], les lieux *rue de la République* et *La Ferté-Saint-Aubin* sont tronqués par le locuteur (*rue de la Rép-* et *La Ferté*). L'attribut *label* prend pour valeur la forme complète du nom de ces lieux.

L'intérêt de faire figurer le nom officiel dans la balise du lieu se justifie également par l'un des objectifs du travail mené qui vise le géoréférencement des lieux identifiés sur la carte de la ville. Cette information sert à rechercher dans une base de données les coordonnées géographiques du lieu pour le placer sur la carte.

5.2. Processus d'annotation manuelle du corpus

Pour constituer un corpus de référence, un nouvel échantillon de transcriptions, distinct du précédent, est sélectionné. Cet échantillon (quatre enregistrements) est annoté manuellement par deux annotatrices, selon les conventions d'annotations établies. L'accord inter-annotateur est évalué avec la mesure du Kappa de Cohen (1960). Il obtient (0.81), un score jugé excellent selon la grille de lecture proposée par Landis & Koch (1977). Ce bon score montre que les conventions d'annotation

proposées sont pertinentes et homogènes. Les erreurs de désaccord observées proviennent de l'ambiguïté de certains lieux. Par exemple, dans :

9. A1 : « la mairie la <loc **type="admin"** zone="2" label="mairie d'Orléans">mairie d'Orléans</loc> elle est belle »
   A2 : « la mairie la <loc **type="monument"** zone="2" label="mairie d'Orléans ">mairie d'Orléans</loc> elle est belle »

les deux annotatrices A1 et A2, résidentes d'Orléans, ne sont pas d'accord sur la façon de catégoriser le lieu *mairie d'Orléans*, un ancien hôtel particulier du XVIe siècle. Malgré l'intérêt touristique du bâtiment qui induit en erreur l'annotatrice A2, c'est la fonction principale du lieu, le type *"admin"*, qui est privilégiée. Après avoir effectué une version de consensus, les deux annotatrices ont annoté six transcriptions supplémentaires afin de compléter le corpus de référence.

6. Traitement automatique pour l'annotation des lieux

Un lieu peut être mentionné dans le discours de manières différentes : un nom propre « Paris », un groupe nominal contenant un nom propre « université de Nanterre », un groupe nominal sans nom propre « cette place ». Un endroit peut faire référence à un lieu imaginaire ou métaphorique « le bout de monde », « mon paradis »
etc. Les déictiques « ici », « là » désignent aussi le lieu d'une manière référentielle. Les noms de lieu peuvent être composés d'un ou de plusieurs mots, être liés ou non par un trait d'union. La tâche de leur repérage automatique est donc complexe mais elle se heurte par ailleurs aux difficultés supplémentaires liées aux corpus oraux.

6.1. Méthodologie

L'annotation automatique des lieux dans l'oral transcrit est réalisée en suivant trois grandes étapes :

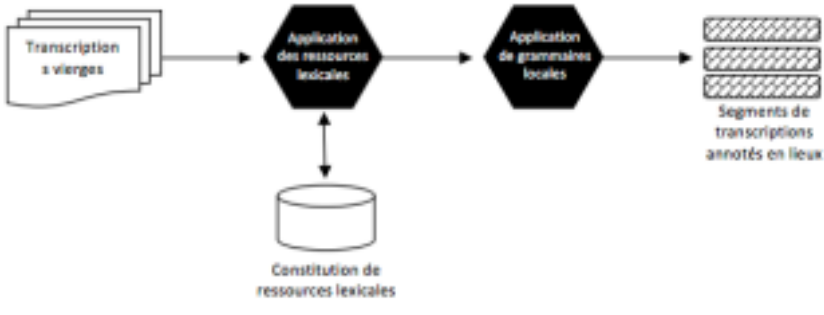

Figure 1 : Schéma du traitement automatique mis en place

- constitution du lexique de noms de lieux à partir des bases de données libres Géonames et GEOFLA[9] et son enrichissement par la génération d'abréviations observées dans le corpus,
- application de grammaires locales, c'est-à-dire de règles d'extraction fondées sur la description du contexte d'apparition du nom de lieu dans le discours ; - sauvegarde des

[9] http://www.geonames.org/, http://professionnels.ign.fr/geofla. Ces bases présentent l'intérêt principal d'associer chaque lieu aux coordonnées GPS lui correspondant et de permettre plus tard son géoréférencement.

lieux identifiés et de leurs attributs dans une base de données, ce qui permet de résoudre en partie les problèmes de coréférence et d'effectuer l'analyse de variations du nommage d'un même lieu.

Pour détecter les noms de lieux, nous utilisons les méthodes symboliques fondées sur les règles d'extraction décrivant le contexte d'emploi de noms de lieux. Le système déclenche ces règles (appelées *patrons*) lorsqu'apparaissent les indicateurs désignant les types de lieux mentionnés (*boulangerie*, *école*, *maison*, *rue*, etc.). Ces noms ont été répertoriés dans le lexique construit préalablement. Dans l'exemple suivant,

10. **rue Royale** euh **l'avenue de la cathédrale** aussi euh voilà (ESLO2_iti_03_01)

les deux lieux *rue Royale* et *avenue de la cathédrale* sont reconnus grâce à une règle montrée dans la Figure 2. Si le mot suivant le mot déclencheur commence par une majuscule, ou s'il correspond à une préposition (PREP) et/ou un déterminant (DET) suivi d'un nom, on suppose que ces termes font partie du nom du lieu. Ainsi dans l'exemple 10, le déclencheur *avenue* est suivi par la préposition *de*, elle-même suivie par le déterminant *la* et le nom *cathédrale.* Le déclencheur *rue* est suivi directement par le nom propre *Royale.* Ces deux contextes décrits par la règle sont reconnus et annotés par le système en tant que lieu.

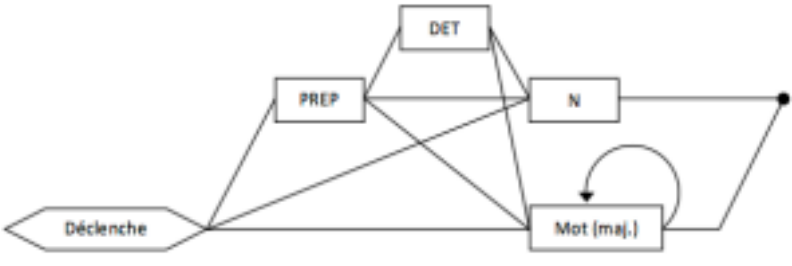

Figure 2 : Règle de reconnaissance d'un lieu fondé sur un déclencheur

Une phase importante de l'annotation des lieux mentionnés dans le corpus consiste en l'application de ressources lexicales provenant des bases de données géographiques. Cependant, ces lexiques ne suffisent pas pour l'annotation exhaustive des lieux contenus dans le corpus car seuls sont recensés les noms officiels, conventionnels des lieux. Un lieu peut être mentionné différemment de la norme établie.

6.2. Traitement des abréviations

Les noms de lieux divergeant de la norme sont nombreux à l'oral. L'abréviation est un phénomène récurrent dans leur désignation. L'abréviation raccourcit un syntagme par la suppression d'un ou plusieurs mots. Les lieux plurilexicaux peuvent être abrégés à l'oral (par exemple : *La Ferté* pour *La Ferté-Saint-Aubin*).

Lorsque le locuteur abrège le nom d'une voie, il ne conserve que le dernier mot (*rue Gauguin* au lieu de la *rue Paul Gauguin*, ou *rue Madeleine* au lieu de *rue Porte Madeleine*) ou groupe prépositionnel (*rue de Sonis* au lieu de la *rue du Général de Sonis*) composant le nom officiel ainsi que le mot caractérisant le type de voie. Dans l'exemple :

11. on voit l'état de la **la place De Gaulle** aussi euh (ESLO2_ENT_1031_C)

le nom de lieu normalisé *Place du Général de Gaulle* est abrégé en *place De Gaulle*. Le terme *place* caractérisant le type de voie a été conservé tout comme le nom de famille *De Gaulle*. L'expression *du Général* par contre a été supprimée. Cette information a été jugée facultative car elle n'empêche pas d'être compris.

Par ailleurs, les locuteurs peuvent omettre les mots grammaticaux comme les déterminants ou les prépositions (*place Cheval Rouge* au lieu de *place du Cheval Rouge*, *rue Porte Dunoise* au lieu de *rue de la Porte Dunoise*).

À partir de ces constatations, la liste de toutes les suppressions possibles entre le mot type et le dernier mot du nom de la voie est générée automatiquement, ce qui permet de reconnaître les noms de lieu abrégés *place De Gaulle, place Cheval Rouge, rue Porte Dunoise* et de les lier à leur label officiel : *Place du Général de Gaulle, place du Cheval Rouge, rue de la Porte Dunoise.*

Il existe une exception : celle des rues comportant le mot *faubourg*. Un faubourg est une ancienne dénomination pour désigner des quartiers entourant une ville qui reste dans des noms de rues comme *la rue du Faubourg Saint-Antoine* ou la *rue du Faubourg Saint-Honoré* à Paris, ou bien la *rue du Faubourg Saint-Jean* ou la *rue du Faubourg Madeleine* à Orléans. Si l'on suit la règle précédente pour l'abréviation de ces noms de rues, on obtient respectivement les *rues Saint-Jean* ou *Madeleine*. Pourtant, il existe une autre façon d'abréger ces noms de rues. La plupart du temps, les Orléanais se réfèrent à ces rues sous le nom de *faubourg Saint-Jean* et *faubourg Madeleine*. Cette fois, c'est le terme *rue* caractérisant le type de voie qui est supprimé au profit de l'ancien terme *faubourg*. Ce cas particulier est pris en compte pour l'enrichissement du lexique et des variantes sont générées en conséquence.

Les noms de villes, composés de trois mots ou plus peuvent aussi être abrégés. De la même façon que pour les noms de voies, des variantes sont générées pour compléter les ressources lexicales. Cependant, le fonctionnement de l'abréviation d'un nom de ville ne fonctionne pas de la même manière qu'une voie. Pour les noms de ville, c'est le premier terme qui est conservé et les derniers termes sont supprimés. On dit *Rio* pour *Rio de Janeiro* ou *Aix* pour *Aix-en-Provence.* Ainsi, les villes *La Ferté Saint-Aubin*, *Saint-Jean-de-la-Ruelle*, ou *Fleury les Aubrais* auront respectivement pour variantes : *La Ferté*, *Saint-Jean* et *Fleury*.

6.3. Traitement des noms de lieux tronqués

Fréquemment, les locuteurs tronquent les noms de lieux comme *rue de la rép* pour *rue de la République* ou peuvent s'interrompre et ne produire que les amorces comme *Orl-* à la place d'*Orléans*. Dans les conventions de transcription du corpus ESLO, les mots non finalisés sont marqués par un tiret. Grâce à cet indice, le système peut prévoir un traitement spécifique pour ce type de lieux amorcés et tronqués. Au delà de l'intérêt qu'il y a à repérer de nouvelles mentions dans le corpus, l'enjeu principal autour de la détection des noms de lieux tronqués est de relier la forme tronquée à sa forme normalisée. Faire ce lien est primordial pour permettre l'élaboration de la carte finale présentant la perception de la ville d'Orléans.

Pour identifier les noms de lieux tronqués, des patrons sont construits. Les patrons sont appliqués sur le tour de parole à partir du moment où un tiret est identifié. Une fenêtre d'observation est alors établie. Celle-ci s'étend depuis un déclencheur éventuellement identifié dans le tour de parole (une majuscule, un nom commun générique comme *rue, place, etc.)* jusqu'au tiret. Dans l'exemple

12. ah ben si tu peux redescendre tu prends la tu prends la rue qui est là et tu vas tout au bout jusqu'à **la rue de la Rép-** tu vois où elle est ? **la rue de la République** ? (ESLO2_iti_06_11)

la mention *rue de la Rép-* est la version tronquée de la *rue de la République*. L'étape d'application des ressources lexicales permet l'identification uniquement de la version normalisée de la *rue de la République*. Pour détecter cette mention de lieu non normalisée, le système établit une fenêtre d'observation qui s'étend de la deuxième occurrence du mot *rue* jusqu'au mot tronqué *Rép-*, soit : *rue de la Rép-*. Pour retrouver le nom normalisé d'origine, les ressources lexicales sont artificiellement tronquées (voir Figure 3).

Figure 3 : Réduction de la fenêtre d'observation pour la détection d'une voie

Dans cet exemple, l'entrée du lexique *rue de la Rape* est devenue *rue de la Rap* après troncation, ce qui ne correspond pas à la séquence candidate *rue de la République*. Par contre, l'entrée du lexique *rue de la République* est devenue *rue de la Rép-* après troncation. Le système fait donc le lien entre cette entrée et la séquence candidate *rue de la Rép-*.

6.4. Traitement des coréférences

On parle de coréférence lorsque deux ou plusieurs termes ou expressions ont le même référent. La gestion des coréférences est l'une des problématiques actuelles du TAL. Plusieurs projets comme ANCOR[10] ou DEMOCRAT[11] étudient les chaînes de coréférences et travaillent à l'élaboration de systèmes automatisés dédiés à leur identification.

Le module développé ne détecte pas toutes les coréférences. Ainsi, les cas concernant des pronoms, souvent sujets à ambiguïtés, ne sont pas traités par notre module. Dans l'exemple

13. euh mais euh mais même **cette rue-là** euh disons qu'y a allez dix quinze ans ouais voir quinze ans je pense euh **elle** était malfamée

le pronom *elle* ne sera pas reconnue par le système. Ajoutons qu'à l'oral, la gestuelle peut être utilisée pour montrer un lieu mais cette information n'apparaît pas dans la transcription.

Cependant, comme il a été mentionné dans la section précédente, le lieu peut aussi être répété dans le discours sous une forme abrégée, tronquée, etc. Grâce aux points communs existant entre les termes se faisant référence, le système développé parvient à traiter ce type de coréférence : c'est-à-dire à établir le lien entre deux mentions comme *rue de la Rép-* et *rue de la République* ou comme *La Ferté* et *La Ferté Saint Aubin*.

Certaines coréférences peuvent être reconnues grâce aux informations collectées au

[10] http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html

[11] http://www.lattice.cnrs.fr/Projet-ANR-DEMOCRAT

moment de l'annotation automatique des lieux. Les conversations analysées sont celles de locuteurs qui ne se connaissent pas ou peu. Nous faisons l'hypothèse que le locuteur mentionne le nom du lieu d'une manière complète la première fois qu'il s'y réfère. Dès lors, chaque lieu identifié et ses attributs sont stockés en parallèle de l'annotation principale. Lorsque le système identifie un nom de lieu qui diffère de la convention, il consulte la liste des détections précédentes pour établir un lien de coréférence. Dans l'exemple [4], *Orléans* est annoté comme une ville et cette caractéristique permet de faire le lien avec la reprise *ville* présente dans le tour suivant.

Le traitement réalisé permet aussi de résoudre des cas d'homonymie provoqué par l'abréviation d'un nom de ville. Le locuteur a la volonté d'être compris par son interlocuteur. Il va désambiguïser son discours en donnant la forme conventionnelle du lieu avant d'utiliser une abréviation comme dans l'exemple suivant :

14. et euh bon je sais pas trop par où il m'a fait passer puis à un moment j'ai j'ai cru être perdue parce que j'étais à **Saint-Jean-de-la-Ruelle** alors j'ai dit mince et en fait Ingré c'est après **Saint-Jean** donc j'ai continué et puis finalement

Pour résoudre ces cas de coréférences, le système détecte, dans la transcription, un nom de lieu non-normalisé (*Saint-Jean*) pour lequel il existe plusieurs correspondances (*Saint-Jean-de-la-Ruelle*, *Saint-Jean-de-Braye* et *Saint-Jean-le Blanc*), il va consulter la liste des lieux déjà identifiés dans l'enregistrement transcrit pour retrouver sa première mention complète et établir un lien de coréférence. Pour cela, le système utilise la distance de Levenshtein (1965) pour évaluer la similarité entre deux chaînes de caractères.

6.5. Évaluation

Les mesures de Rappel, Précision et F-Mesure sont le plus souvent utilisées pour déterminer les performances de l'outil développé. Ces mesures s'appuient sur la comparaison de corpus de référence et l'annotation automatique réalisée par le système à évaluer. Plus la version annotée automatiquement est similaire à celle annotée manuellement, plus le module est considéré comme efficace.

La mesure du Rappel montre la part de détections pertinentes par rapport à l'ensemble des détections réalisées par le système. La Précision représente la part des détections pertinentes par rapport à la totalité des détections que le système est censé effectuer. Ces deux mesures prennent en compte le nombre d'éléments correctement (Vrais positifs) et incorrectement (Faux positifs) identifiés ainsi que le nombre d'éléments qui n'ont pas été identifiés (Faux négatifs).

Le module d'annotation automatique des lieux obtient une F-mesure (une moyenne) de 0,91, un rappel de 0,90, une précision de 0,93. Le Rappel (0,90) du module est considéré comme satisfaisant. Ce score démontre que la plupart des détections attendues ont été opérées, comme dans l'exemple [15] dans lequel trois lieux sont cités (*Orléans*, la *rue de la République* et la *rue Royale*).

15. mais celle d'**<loc type="ville" zone="0" label="Orléans">Orléans</loc>** non j'ai toujours un mal fou entre la **<loc type="voie" zone="0" label="Rue de la République">rue de la République</loc>** la **<loc type="voie" zone="0" label="Rue Royale">rue Royale</loc>** vous voyez c'est des (ESLO2_iti_03_01)

Néanmoins, certaines détections manquent comme le bateau restaurant *l'Inexplosible* mentionné dans l'exemple [16]. Les absences de détection s'expliquent principalement par l'absence d'exhaustivité des ressources lexicales employées.

16. oui y a le **l'Inexplosible** là un un bateau oui qui fais- qui faisait <loc type="commerce" zone="2" label="bar">bar</loc> à tapas au début puis maintenant il fait <loc type="commerce" zone="2" label="restaurant">restaurant</loc> euh (ESLO2_ENT_1042)

La Précision, qui représente le degré de pertinence de l'annotation, est aussi jugée satisfaisante (0,93). Le module produit une annotation de qualité dans laquelle la grande majorité des détections sont pertinentes. Une forte Précision est une caractéristique des systèmes fondés sur des méthodes symboliques, c'est-à-dire sur des règles d'extraction qui reconnaissent les entités selon leur contexte.

Cependant, certaines détections sont erronées. Dans l'exemple [17], si le système n'a pas identifié le *Campo Santo*, il a annoté *Campo* comme une ville en renvoyant à la commune corse Campo qui est référencée dans la base de données GEOFLA.

17. oui c'est vrai elle doit pas passer sous vos fenêtres elle passe plutôt <loc type="voie" zone="2" label="Rue de Bourgogne">rue de Bourgogne</loc> vous êtes pas allée voir au **<loc type="ville" zone="0" label="Campo">Campo</loc> Santo** (ESLO2_ENT_1042)

D'une manière générale, le module présente de bonnes performances pour la tâche de détection des désignations de lieux dans l'oral transcrit. Les annotations attendues sont effectuées de manière congruente comme en témoigne la F-Mesure de 0,91.

7. Analyse des données annotées

L'analyse quantitative et qualitative des lieux mentionnés est faite sur l'échantillon du corpus annoté manuellement composé de 15 entretiens et de 5 itinéraires, pour un total de 6622 tours de parole, parmi lesquels 1223 contiennent une mention de lieu, soit 18% du total des tours de parole observés. Parmi ces transcriptions, 2292 mentions de lieux ont été annotées. Certains tours de parole peuvent contenir plusieurs lieux :

18. quoi ça ça ça disparaît pas Maingourd maintenant est sur **La Chapelle** est venu dans la périphérie d'**Orléans** et puis bon à cette époque-là **Orléans** n'avait pas l- toute **la région de La Source La Source** a été achetée par la **ville d'Orléans** à à à **Olivet** et les communes qui étaient là-bas c'est y avait p- **Orléans** s'arrêtait du côté de **Saint-Marceau** quoi a- al- elle arrivait jusqu'à **Olivet** mais La rs- **La Source** a été c'était la plaine (ESLO2_ENT_1059_C)

Les transcriptions extraites du module Itinéraire comportent beaucoup moins de noms de lieux que celles extraites du module Entretien. Cette différence s'explique par le fait que les enregistrements du module Itinéraire durent en moyenne une dizaine de minutes alors que ceux du module Entretien peuvent dépasser une heure et demie.

La répartition des lieux mentionnés en fonction de la typologie définie est affichée dans la Figure 4. On observe que c'est l'étiquette ville 44% qui est la plus représentée, ce qui reflète la nature des données traitées. Les voies représentent 13% des lieux mentionnés, les commerces 10%, les monuments 9%, les lieux naturels 8%, et les régions, les lieux éducatifs et les pays représentent chacun 5%. Les lieux administratifs sont rares et n'apparaissent que vingt-quatre fois. Une part de l'explication tiendrait à l'ambiguïté de ces noms qui peuvent avoir d'autres interprétations comme celle de monument. C'est le cas par exemple de la mairie d'Orléans : les locuteurs auront

tendance à se référer au bâtiment pour son attrait touristique plutôt qu'à sa fonction administrative. Enfin, seulement dix lieux à fonction non commerciale et un seul lieu supranational sont présents dans le corpus de référence.

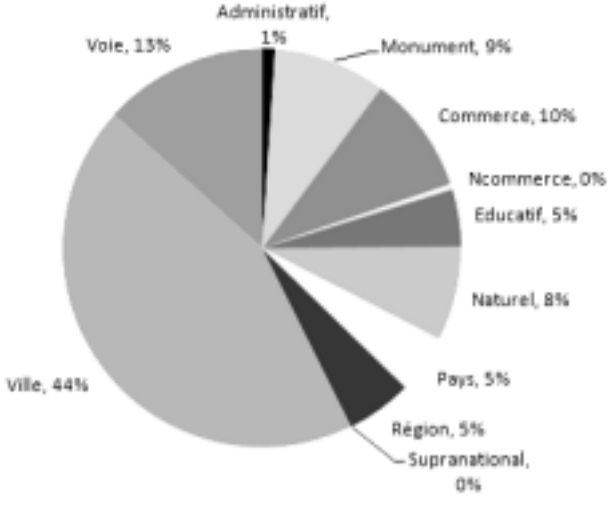

Figure 4 : Répartition de lieux selon leurs types

Cette répartition est conditionnée par le contenu du corpus : les entretiens portent sur la ville d'Orléans appréhendée à partir de la vie de ses habitants. C'est la raison pour laquelle les lieux de type voies, monuments, commerces, éducation, comptent parmi les mentions les plus fréquentes après les noms de villes. On peut considérer que ces observations reflètent la perception que les habitants ont de leur cité, une vision qui se manifeste à travers la mention de lieux dont les Orléanais ont eu envie de parler, dont ils se sentent proches tant sur un plan géographique que sentimental.

La proximité sentimentale peut se retrouver dans la façon dont les locuteurs désignent les lieux auxquels ils font référence. Le fait de modifier, tronquer, abréger, etc. le nom d'un lieu est déjà un indice des modalités de perception de la ville. Ainsi, les mentions de lieux ayant subi des modifications de la part des locuteurs représentent 36% des lieux annotés.

Par ailleurs, un lieu peut être nommé par un surnom ou un nom connu et figé au sein de la communauté. Dans l'exemple suivant :

19. donc c- je suppose que c'est la génération d'après et euh mais y avait quand même plein d'oies sur ce ce bout de d'île euh qui est euh pas très loin du pont euh euh du **pont George Cinq** oui c'est **le pont Royal** (ESLO2_ENT_1034_C)

le locuteur mentionne deux fois le même pont. Ce dernier a changé plusieurs fois de noms au cours de son histoire. *George Cinq* est le nom officiel actuel tandis que *Royal* était le nom qu'on lui avait attribué au moment de son inauguration en 1763. Malgré l'ancienneté de ce dernier nom, il est très régulier que les Orléanais se réfèrent à ce pont en utilisant son ancien nom. Dans l'exemple [20] :

20. en gros euh sous **les Arcades** (ESLO2_ENTJEUN_04_C)

Les Arcades correspondent à un surnom donné à la *rue Royale*, une rue centrale à

Orléans bordée sur toute sa longueur par des galeries à arcades. On observe ainsi une véritable réappropriation du nom d'un lieu. « Faire allusion à une entité en utilisant un surnom est un cas de la personnalisation, de l'appropriation d'un lieu par un locuteur ». (Eshkol-Taravella, Flamein, 2017).

Un lieu n'est pas toujours mentionné en utilisant un nom propre. Dans l'exemple :

21. enfin je sais pas c'est mais à Orléans je le trouve moche le tram ils veulent le faire passer devant **la cathédrale** devant **ma cathédrale** (ESLO2_ENT_1003_C)

le locuteur se réfère au lieu d'abord de façon neutre, *la cathédrale,* pour le reprendre en remplaçant l'article défini *la* par le déterminant possessif *ma.* On assiste ainsi au processus de l'appropriation du lieu par le locuteur. Par ailleurs, le locuteur n'a pas besoin de préciser le lieu *cathédrale d'Orléans,* c'est-à-dire d'utiliser un nom propre. L'acte de désignation dépend fortement du contexte de l'énonciation. Dans cet exemple, le nom commun *la cathédrale* désigne directement un référent (Orléans n'a qu'une cathédrale). Les noms communs représentent 34,73% des noms de lieu annotés dans le corpus étudié.

8. Conclusions et perspectives

L'article présente le système de détection des lieux dans les transcriptions du corpus ESLO2. La nature orale du corpus favorise les variations dans les noms de lieux qui peuvent ne pas être mentionnés de façon conventionnelle (*La Ferté* pour *La Ferté Saint-Aubin* ; *rue de la Rép-* pour *rue de la République, la cathédrale* pour *la cathédrale d'Orléans, etc.*).

Le système développé d'annotation automatique de lieux tient compte de ces variations et des caractéristiques du corpus oral. Il obtient 93% de précision, 90% de rappel, 91% de F-mesure. Les mentions de lieux identifiées sont stockées dans une base de données, une ressource lexicale riche pour les linguistes, géographes et chercheurs de la toponymie qui s'intéressent à la façon dont un locuteur français mentionne un lieu.

Le travail présenté est un travail pluridisciplinaire qui s'inscrit dans les domaines de la linguistique de corpus, de la linguistique outillée, du traitement automatique du langage, de la géomatique et plus particulièrement dans un de ses axes consacrés à l'extraction d'information géographique et analyse spatiale à partir de textes ainsi que dans le domaine des humanités numériques.

Plus largement, l'enjeu applicatif principal de ce travail est de proposer de nouveaux moyens pour le traitement d'un nouveau type de données. Le corpus ESLO est une ressource riche tant sur le plan qualitatif que quantitatif. L'un des enjeux de ce travail est d'offrir aux utilisateurs une nouvelle manière d'accéder au corpus, et plus particulièrement à la perception de leur ville par les Orléanais. C'est la réalisation de la carte finale qui rend explicite cette information, une visualisation de cette information qui répond directement aux objectifs d'ESLO qui se présente comme le portrait sonore d'Orléans. La matérialisation de ce portrait de la ville d'Orléans restitue d'une part la dimension patrimoniale et anthropologique du corpus, d'autre part, sa valeur de témoignage.

Hélène Flamein, Iris Eshkol-Taravella
LLL UMR 7270, Université d'Orléans, MoDyCo UMR7114, Université Paris Nanterre
helene.flamein@univ-orleans.fr ; iris.eshkoltaravella@parisnanterre.fr

Bibliographie

BAILLY, Antoine, BEGUIN, Hubert, SCARIATI, Renato (2016), *Introduction à la géographie humaine*, Paris, Armand Colin.

BOONS, Jean-Paul (1987). « La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs ». *Langue Française*, n° 76, pp. 5-40.

BORILLO, Andrée (1998), *L'Espace et son expression en français, Paris,* Ophrys.

COHEN, Jacob (1960), « A coefficient of agreement for nominal scales », *Educational and psychological measurement*, 20(1), pp. 37-46.

DOMINGUES, Catherine (2018), « Lieu et nom de lieu, du texte vers sa représentation cartographique », 25e conférence sur le Traitement Automatique des Langues Naturelles.

DOMINGUES, Catherine, ESHKOL-TARAVELLA, Iris (2015), « Toponym recognition in custom-made map titles », International Journal of Cartography, Taylor & Francis, 2015, pp. 109-120. DOI : 10.1080/23729333.2015.1055935. hal-01174721.

EHRMANN, Maud (2008), *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation* (Doctoral dissertation, Paris Diderot University).

ESHKOL-TARAVELLA, Iris, FLAMEIN, Hélène (2017), « Dis-moi Orléans ». Repérage et analyse de la perception d'un lieu dans l'oral transcrit. *Echo des études romanes*, vol. XIII, n1, 61-72.

GOUVERT, Xavier (2008), *Problèmes et méthodes en toponymie française, Essais de linguistique historique sur les noms de lieux du Roannais,* Thèse de doctorat, Université de Paris Sorbonne (Paris IV).

HUYGHE, Richard (2009), *Les Noms généraux d'espace en français. Enquête linguistique sur la notion de lieu,* De Boeck, Duculot.

LAUR, Dany (1991), *Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple*, Thèse de doctorat, Université de Toulouse I.

LANDIS, J. Richard, KOCH, Gary G., (1977). « The measurement of observer agreement for categorical data », *Biometrics*, pp. 159-174.

LE PESANT, Denis (2011), « Problèmes de morphologie, de syntaxe et de classification sémantique dans le domaine des prépositions locatives », *in* F. Neveu, P. Blumenthal et N. Le Querler (dir.), *Au commencement était le verbe. Syntaxe, Sémantique et Cognition, Mélanges en l'honneur du Professeur Jacques François*, Peter Lang, Bern, Berlin, pp. 349-372.

LE PESANT, Denis (2012), « Essai de classification des prépositions de localisation », Actes du *CMLF 2012*, pp. 921-936.

LE SQUERE, Roseline (2006), « Analyse des perceptions, usages et fonctions des toponymes actuels des territoires ruraux et urbains de Bretagne », *Cahiers de sociolinguistique,* 11, Presses universitaires de Rennes, pp. 81-99.

LESBEGUERRIES, Julien (2007), *Plate-forme pour l'indexation spatiale multi niveaux d'un corpus territorialisé* (Doctoral dissertation, Université de Pau et des Pays de l'Adour).

LEVENSHTEIN, Vladimir (1965), *Levenshtein distance*.

LEVY, Jacques, LUSSAULT, Michel (2013), *Dictionnaire de géographie et de l'espace des sociétés* (pp. 1137-p). La Documentation Française.

LOUSTAU, Pierre, GAIO, Mauro, NODENOT, Thierry (2008), « Interprétation automatique d'itinéraires à partir d'un corpus de récits de voyages pilotée par un usage pédagogique », *Revue des Nouvelles Technologies de l'Information*, (13), pp.

177-206.
LUSSAULT, Michel (2007), *L'homme spatial. La construction sociale de l'espace humain,* Paris, Seuil.
MAUREL, Denis, FRIBURGER, Nathalie, ANTOINE, Jean-Yves, ESHKOL TARAVELLA, Iris, NOUVEL, Damien (2011), « Cascades de transducteurs autour de la reconnaissance des entités nommées », *Traitement automatique des langues*, *52*(1), pp. 69-96.
MONCLA, Ludovic, GAIO, Mauro, NOGUERAS-ISO, Javier, MUSTIERE, Sébastien (2016), « Reconstruction of itineraries from annotated text with an informed spanning tree algorithm », *International Journal of Geographical Information Science*, *30*(6), pp. 1137-1160.
NADEAU, David, SEKINE, Satoshi (2007), « A survey of named entity recognition and classification », *Lingvisticae Investigationes*, *30*(1), 3-26.
NOUVEL, Damien, EHRMANN, Maud, ROSSET, Sophie (2015), *Les entités nommées pour le traitement automatique des langues*. ISTE éditions. ROSSET, Sophie,
GROUIN, Cyril, ZWEIGENBAUM, Pierre (2011), *Entités nommées structurées : guide d'annotation Quaero*, LIMSI-Centre national de la recherche scientifique.
TUAN, Yi Fu (1977), *Space and place : The perspective of experience*, U of Minnesota Press.
VANDELOISE, Claude (1986), *L'Espace en français,* Paris, Seuil. ZENASNI, Sarah, KERGOSIEN, Eric, ROCHE, Mathieu, TEISSEIRE, Maguelonne (2016, november). « Extracting new spatial entities and relations from short messages », *in Proceedings of the 8th International Conference on Management of Digital EcoSystems* (pp. 189-196).

Résumé
Le travail proposé exploite d'une manière nouvelle les données orales recueillies par des chercheurs dans le cadre d'une enquête sociolinguistique urbaine. Son objectif est de détecter automatiquement les mentions de lieux dans le français parlé pour permettre d'une part l'analyse linguistique de variations dans le nommage de lieux et pour, d'autre part, permettre la visualisation de la ville telle qu'elle est perçue par ses habitants. L'article propose une méthode pour la détection des noms de lieux dans les enregistrements oraux prenant en compte les particularités liées à la nature des données traitées.

Mots-clés : Noms de lieux, Variation, Traitement Automatique du Langage, Visualisation de l'information, ESLO, Corpus oral.

Abstract
The proposed work makes new use of oral data collected by researchers as part of an urban sociolinguistic survey. Its objective is to automatically detect mentions of places in spoken French in order to allow, on the one hand, the linguistic analysis of variations in the naming of places and, on the other hand, to allow the visualization of the city as it is perceived by its inhabitants. The article proposes a method for the detection of place names in oral records that takes into account the specific characteristics related to the nature of the processed data.

Keywords
Place-names, Variation, Natural Language Processing, Information visualization, ESLO, Oral corpus.