Etude de la perception d'une ville : Repérage automatique, analyse et visualisation

Hélène Flamein

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE HUMANITE ET LANGUES

LABORATOIRE LIGERIEN DE LINGUISTIQUE

THÈSE présentée par :

Hélène FLAMEIN

soutenue le : **10 décembre 2019**

pour obtenir le grade de : **Docteur de l'université d'Orléans**

Discipline/ Spécialité : Sciences du Langage

Etude de la perception d'une ville

Repérage automatique, analyse et visualisation

THÈSE dirigée par :

Iris ESHKOL-TARAVELLA	Professeur, Université Paris-Nanterre
Gabriel BERGOUNIOUX	Professeur, Université d'Orléans

RAPPORTEURS :

Thierry POIBEAU	Directeur de recherche CNRS
Mathieu ROCHE	Chercheur HDR, TETIS, Cirad

JURY :

Olivier BAUDE	Professeur, Université Paris-Nanterre....Président du Jury
Gabriel BERGOUNIOUX	Professeur, Université d'Orléans.....Directeur
Iris ESHKOL-TARAVELLA	Professeur, Université Paris-Nanterre.....Directrice
Thierry POIBEAU	Directeur de recherche CNRS-LATTICE .....Rapporteur
Mathieu ROCHE	Chercheur HDR, TETIS, Cirad.....Rapporteur

REMERCIEMENTS

Je tiens tout d'abord à remercier Thierry Poibeau et Mathieu Roche d'avoir accepté d'être mes rapporteurs et pour le temps qu'ils ont consacré à mon travail. Je remercie également Olivier Baude d'avoir accepté d'évaluer mon travail en faisant partie de mon jury.

Pour tout ce qu'ils ont pu m'apporter, mes remerciements les plus sincères vont à mes deux directeurs :

A Gabriel Bergounioux, qui m'a accordé sa confiance et qui m'a fait découvrir différents aspects de la recherche et de la vie d'une université. Je n'imagine pas ce qu'aurait été cette thèse sans ces expériences passionnantes.

A Iris Eshkol-Taravella, et son soutien de tous les instants. Merci pour votre gentillesse, votre bienveillance ou encore votre générosité sans lesquelles je n'aurais pas pu arriver aussi loin.

Je remercie également l'ensemble du Laboratoire Ligérien de Linguistique pour m'avoir accueillie et guidée ces dernières années. Que ce soit pour mon travail de recherche ou pour les cours assurés dans le département, j'aimerais notamment remercier Caroline, Céline, Flora, Anne-Lyse, et Anaïs pour leur disponibilité et leur aide. Merci aussi à Layal et toute l'équipe de l'IUT d'Informatique d'Orléans qui m'a chaleureusement accueillie et accompagnée durant la fin de cette thèse.

Bien sûr une grande part de ces remerciements s'adresse à tous les doctorants que j'ai pu croiser de près comme de loin. Que ce soit lors de conférences, avec l'ADDOSHS, ou dans toute l'université, ces années ont été riches de belles rencontres. Merci tout particulièrement à mes copines linguistes : Hyun Jung, Camille, Jennifer, Fatma et Yossra pour leur soutien et leur amitié. Merci à mes chers voisins géographes : Romain, Tarek, Quentin, Margaux et Cathy. Grâce à vous, la géographie n'aura plus jamais la même saveur. Merci à ma Britannique préférée, Alice (*cette année c'est la nôtre !*). Aujourd'hui je vous compte parmi mes plus chers amis et je mesure la chance que j'ai eue de vous rencontrer.

Evidemment je remercie mes amis de toujours : Manon, Sandy, Marie (Pavel, Pacôme et Quentin), Aurélie, Florian, François, Thomas, Eloi. Vous me connaissez suffisamment pour savoir l'importance que vous avez dans ma vie. Je remercie aussi ma famille. Je pense à mes grands-mères Marie-Louise, Yvette et Edith et à mon grand-père Roger. Une pensée pour mes cousins ardéchois Mathieu, Morgane et Estelle. Et même si les mercis ne suffisent plus, merci à mes sœurs Annabelle et Caroline pour leur soutien à tous les niveaux et à qui je souhaite le meilleur et plus encore. Merci papa et merci maman, pour tout, merci.

Et finalement je remercie ma ville : Orléans, ses quais, ses rues pavés, ses habitants pas si froids et sa cathédrale. J'espère ne pas avoir travesti son authenticité et suscité l'envie d'aller s'y promener.

SOMMAIRE

INTRODUCTION

Contexte de la thèse

A l'heure où le numérique est omniprésent, de plus en plus de corpus et de données sont accessibles. Articles de presse, productions littéraires, modes d'emploi et désormais commentaires d'utilisateurs sur le web, conversations par messageries instantanées, SMS, tweets ou encore contenus vidéoludiques ne sont que des exemples de la variété de données disponibles aujourd'hui. Le traitement de ces données multimodales renferme différents enjeux dont les principaux seraient de gérer la quantité incommensurable de données disponibles tout en étant capable de s'adapter à leur diversité de contenu et de forme. Le développement des outils informatiques contribue à l'accroissement de la masse de données disponibles mais participe surtout à l'offre d'outils nécessaires à leur traitement et consultation. Pour répondre à ces enjeux, il est nécessaire d'avoir une approche pluridisciplinaire. Le domaine des humanités numériques illustre bien cette problématique en se positionnant au croisement de l'informatique et des sciences humaines et sociales.

En formalisant les théories linguistiques grâce à l'informatique, les technologies du Traitement Automatique des Langues (TAL) constituent des solutions pour répondre à ces enjeux. Le TAL est l'ensemble des méthodes permettant de traiter automatiquement les données exprimées dans une langue. C'est un domaine de recherche ayant « quatre principaux pôles disciplinaires autour duquel il gravite : la linguistique ; l'informatique ; les mathématiques [….] ; l'intelligence artificielle ». Selon Fuchs & Habert (2004 : 1) :

> L'objectif du traitement automatique des langues est la conception de logiciels capables de traiter de façon automatique des données exprimées dans une langue (dite « naturelle », par opposition aux langages formels de la logique mathématique).

L'un des axes principaux du TAL est l'extraction d'information dans lequel on retrouve par exemple les tâches de Reconnaissance d'Entités Nommées (REN), la catégorisation de documents ou encore l'analyse de sentiments. La captation et le traitement de nos ressentis décident d'enjeux majeur. Toutes nos émotions, décodées, révélées, déchiffrées sont aujourd'hui autant de trésors pour les industriels et les utilisateurs de leurs produits.

Notre thèse analyse la perception qu'ont les habitants de leur ville. Pour atteindre cet objectif nous nous positionnons dans un cadre pluridisciplinaire associant principalement la linguistique, le TAL et la géographie. Plus précisément, il s'agit d'extraire cette perception d'un corpus de conversations orales spontanées tiré d'ESLO (Enquête SocioLinguistique à Orléans). L'objectif est de déterminer comment les habitants perçoivent et parlent de leur ville. Pour avoir accès à ces informations, il est nécessaire de modéliser la perception, de l'extraire, de l'analyser pour finalement la visualiser. Ce type de données est à notre connaissance peu exploité dans les travaux existants en TAL. Nous proposons donc de répondre à ce manque en présentant une méthodologie pour leur traitement et la visualisation de leur contenu afin d'en faciliter l'exploitation. Avant de préciser les différentes étapes mises en œuvre dans ce projet, on présente le corpus analysé.

ESLO : Enquête SocioLinguistique à Orléans

L'étude est fondée sur le corpus ESLO (Enquête SocioLinguistique à Orléans) (Abouda & Baude, 2006 ; Eshkol-Taravella *et al.*, 2011 ; Baude & Dugua, 2011), programme du Laboratoire Ligérien de Linguistique (UMR7270) de l'université d'Orléans, qui met au cœur de son investigation les pratiques langagières dans la ville d'Orléans.

Présentation du corpus

Au début des années 70, des universitaires britanniques ont initié le projet ESLO en enregistrant plusieurs centaines d'Orléanais dans leur vie de tous les jours. Cette collecte avait une visée didactique et devait permettre l'élaboration d'une méthode d'apprentissage du français à partir de l'étude d'enregistrements oraux. L'exploitation de ce projet se

poursuivit au sein du Laboratoire Ligérien de Linguistique avec l'idée de rendre disponibles les données collectées tout en suivant un cahier des charges scientifique et juridique.

Quarante ans plus tard, une nouvelle enquête ESLO2 est lancée afin de constituer un corpus comparable au premier et cumuler près de 700 heures d'enregistrements.

Les corpus des ESLO forment un témoignage sur la ville et sur le français et les langues parlées quotidiennement dans toutes leur variété et leurs diversités. D'un point de vue tant quantitatif que qualitatif, ces collections de paroles spontanées, anonymisées et informatisées constituent une ressource très riche pour des chercheurs de tous domaines : historiens, sociologues, linguistes, etc. Cette richesse se retrouve dans la diversité des problématiques soulevées dans différents projets et travaux s'appuyant sur ESLO, comme par exemple :

- **[ESLO-MD] :** ESLO Micro-Diachronie[1]. Sous-corpus oral équilibré entre les deux parties du corpus ESLO pour une étude micro-diachronique de l'alternance du futur simple et du futur périphrastique en français contemporain. (Abouda & Skrovec, 2017 ; 2018)
- **[MODAL] :** Modèle de l'annotation de la modalité à l'oral[2]. Corpus multilingue (anglais, français et italien) dédié à l'étude des constructions épistémiques caractérisant les interactions orales. (Nissim & Pietrandrea, 2017 ; Pietrandrea, 2018)
- **[ANCOR] :** Anaphore et Coréférence dans les Corpus Oraux[3]. Projet pluridisplinaire (TAL, typologie, sémantique) pour l'étude de toutes les formes de reprises anaphoriques et de coréférence à l'oral. ANCOR est le premier corpus français d'envergure répertoriant les relations de coréférence et relations anaphorique. (Muzerelle *et al.*, 2014 ; Desoyer *et al.*, 2015 ; Grobol et. al, 2018)

[1] http://eslo.huma-num.fr/index.php/pagelarecherche/projets-de-l-equipe-et-sous-corpus/eslo-md
[2] (2016). Modal – Modèles de l'annotation de la Modalité à l'Oral [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr, https://hdl.handle.net/11403/modal.
[3] http://tln.li.univ-tours.fr/Tln_Ancor.html

La question du traitement des corpus oraux comme ESLO est fondamentale et actuelle. La conférence *50 ans de linguistique sur corpus oraux : Apports à l'étude de la variation*[4] qui s'est déroulée du 15 au 17 novembre 2018 à Orléans illustre l'intérêt scientifique de ce genre de questions et démontre la nécessité de disposer d'outils toujours plus efficaces informatiquement et fins linguistiquement parlant pour répondre aux enjeux de la linguistique sur corpus oraux.

Constitution du corpus

Les ESLOs sont actuellement composés des corpus ESLO1 et ESLO2 et ont pour objectif de fournir des données représentatives du français entendu à Orléans.

Le corpus ESLO1 comprend un total de 470 transcriptions représentant près de 320 heures d'enregistrements réparties en 8 modules[5]. Chaque module correspond à un type de situation d'enregistrement différent : des interviews, des communications téléphoniques, des conférences-débats, etc. Le module le plus important est celui des interviews sur questionnaires, aussi appelé Entretiens (157 transcriptions pour 182,5 heures). Chacun des enregistrements de ce module suit le même protocole. Un chercheur identifie un locuteur témoin et convient d'un rendez-vous au domicile de ce dernier. Le chercheur mène la discussion en suivant une trame préétablie tout en essayant d'avoir une conversation la plus naturelle possible. L'objectif de ce module est d'obtenir une collection d'enregistrements à contenu constant avec une gamme de locuteurs sociologiquement représentative. Les bandes sonores collectées sont ensuite numérisées, transcrites et enrichies en métadonnées (sexe, âge, catégorie socioprofessionnelle, etc. du ou des locuteur(s) enregistré(s)).

[4] https://anniveslo-50ans.sciencesconf.org/
[5] cf. Figure 1

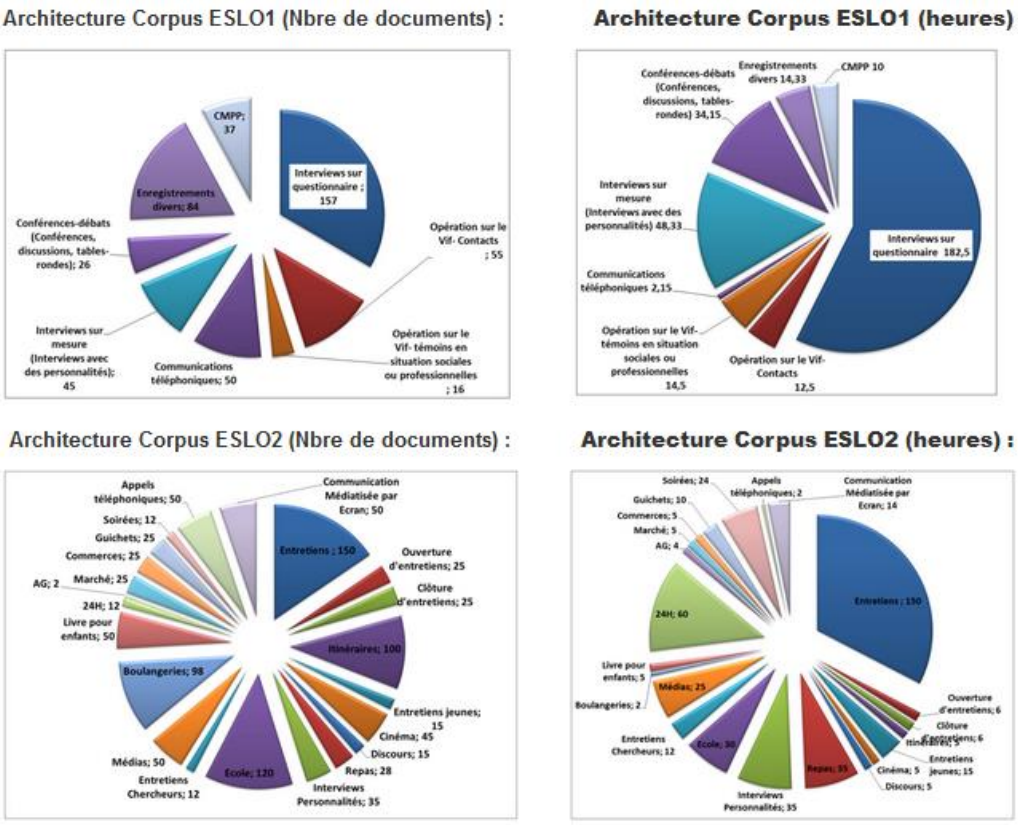

Figure 1 : Composition du corpus ESLO[6]

La constitution du corpus ESLO2 est toujours en cours et comprend à l'heure actuelle 969 enregistrements représentant 446 heures d'enregistrements réparties en 18 modules[7]. L'objectif de produire un corpus comparable à la première campagne de collecte a été atteint. Ainsi, le module Entretiens est lui aussi le plus important du corpus ESLO2 pour une taille similaire au module Entretiens du corpus ESLO1 (150 enregistrements pour 150 heures). Au-delà de la question de la comparabilité, ESLO2 a été enrichi avec d'autres situations d'enregistrements comme avec le module Cinéma où des personnes sont interrogées à la sortie d'un cinéma sur le film qu'elles viennent de voir ou le module Repas dans lequel les interactions entre plusieurs personnes sont enregistrées le temps d'un repas.

Si le protocole de collecte des enregistrements est propre à chaque module, la transcription des enregistrements suit un cahier des charges très précis, détaillé et commun aux ESLO. Cette transcription est considérée comme une première annotation destinée à faciliter la

[6] http://eslo.huma-num.fr/index.php/pagecorpus/pagepresentationcorpus
[7] cf. Figure 1

navigation dans le signal de parole. Les conventions de transcriptions édictées dans le *Guide du Transcripteur et du Relecteurs des ESLO*[8] doit permettre de répondre à une forte contrainte d'interopérabilité : être applicable à n'importe quel type d'enregistrement de parole et pouvoir être utilisable par tout chercheur, quel que soit son objet d'étude. Chaque enregistrement est donc transcrit orthographiquement avec une distinction des tours de parole. La convention de transcription préconise de transcrire sans utiliser de signes de ponctuation ni de majuscules au début des énoncés. Seules exceptions, les points d'interrogation qui différencient les questions et les majuscules pour les noms propres.

Structure et problématique

Cette thèse propose l'analyse de la perception qu'ont les Orléanais de leur ville à travers l'exploitation du corpus ESLO. La chaîne de traitement qui conditionne cette analyse se décompose en plusieurs étapes présentées dans la Figure 2.

La première étape du traitement est la détection des mentions de lieux dans les transcriptions du corpus ESLO. La notion de lieu est considérée du point de vue de la géographie, de la linguistique et du TAL mais sa définition continue à faire débat. La discussion de cette notion sera présentée dans le chapitre 1. L'élaboration du système d'annotation automatique des mentions de lieux dans l'oral transcrit est guidée par des travaux précédents et par l'observation manuelle du corpus présentée dans le chapitre 2. Cette annotation est réalisée selon une approche symbolique associant des règles et la manipulation de ressources lexicales.

L'annotation des mentions de lieux dans le corpus sert d'ancrage pour la deuxième étape du traitement. La perception est une notion subjective dont la définition est discutée à partir des notions d'*émotion*, de *sentiment* et d'*opinion* dans le chapitre 3. A partir des mentions identifiées, une fenêtre d'observation est ouverte pour réaliser une analyse d'opinion, effectuée par apprentissage automatique. Plusieurs classifieurs sont comparés et des perspectives d'utilisation de réseaux de neurones sont explicitées. Cette étape est présentée dans le chapitre 4.

[8] http://eslo.huma-num.fr/index.php/pagemethodologie?id=71doivent

La troisième et dernière étape détaillée dans le chapitre 5 consiste en la création d'une carte de la ville d'Orléans sur laquelle sont placés les lieux mentionnés par les locuteurs du corpus ESLO en fonction des opinions émises et des sentiments éprouvés à leurs sujet. La carte ainsi obtenue est la représentation de la perception qu'ont les Orléanais de leur ville et de leur environnement.

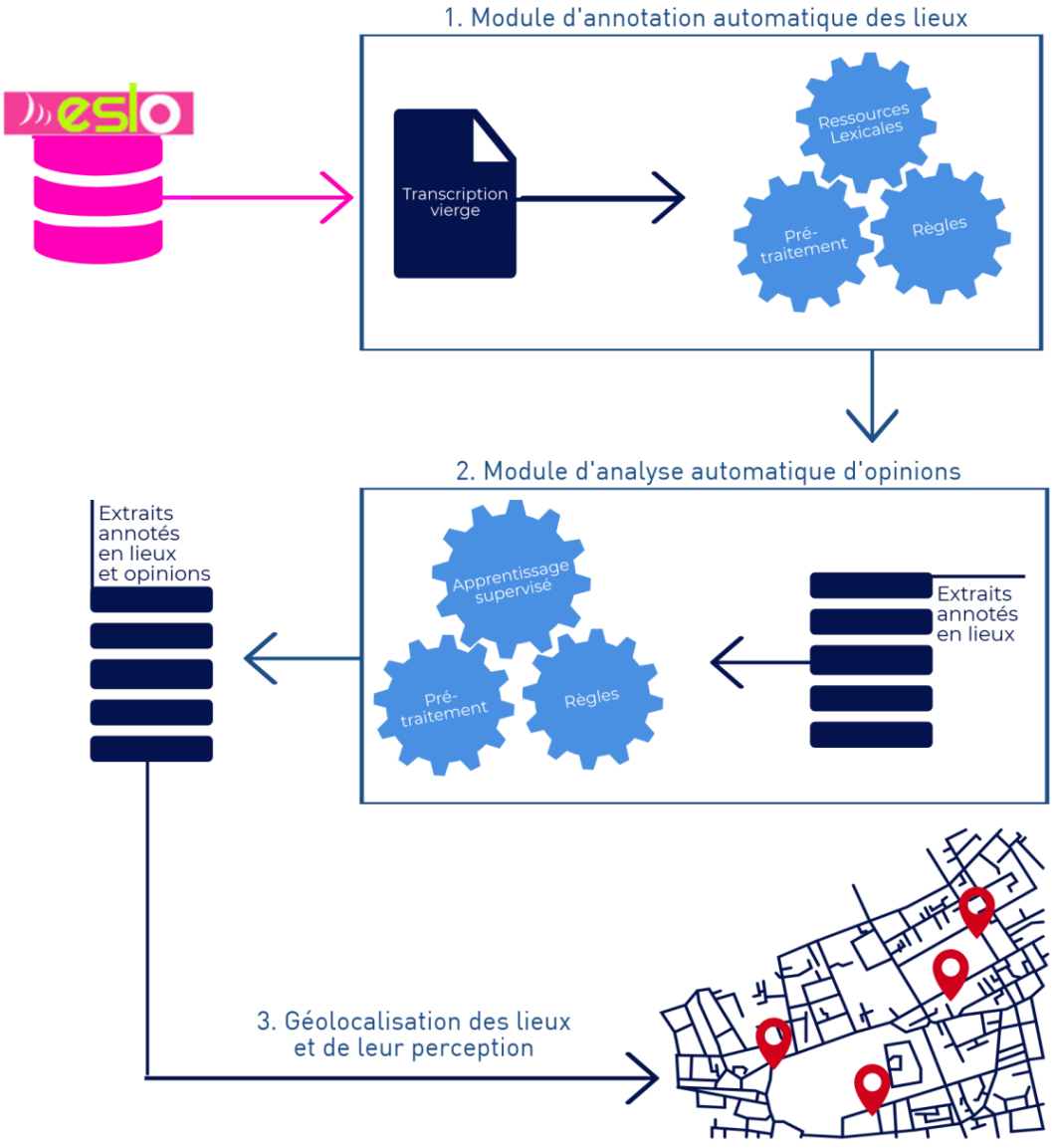

Figure 2 : Schéma général de la chaîne de traitement

Les objectifs de ce travail sont autant applicatifs que théoriques. En premier lieu, il s'agit d'une part de parvenir au traitement informatique de données jusqu'ici peu exploitées. En

effet, la plupart des travaux en TAL s'appuient sur des données écrites (articles de journaux, écrits littéraires ou du Web…). Le corpus traité est composé de transcriptions de conversations orales. Le traitement de l'oral diffère de celui de l'écrit et pose certaines difficultés présentées dans la section 2.4.2.2. 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. Plus largement, l'enjeu applicatif principal 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'ensemble du corpus est disponible en ligne et destiné aux communautés scientifiques intéressées par l'étude du français oral, ainsi qu'au grand public curieux d'en découvrir le contenu. L'autre enjeu 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. La visualisation de cette information 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 et d'autre part sa valeur de témoignage, est révélateur de l'attractivité de la ville. Il permet à des personnes qui voudraient emménager à Orléans d'identifier les quartiers correspondant le mieux à leur mode de vie. Le système élaboré peut aussi servir à des applications dans les domaines du tourisme et de l'urbanisme. Enfin, l'analyse de la carte peut répondre à des problématiques d'aménagement du territoire en mettant par exemple en évidence certaines améliorations demandées par les administrés ou, au contraire, évaluer l'impact des travaux de réfection ou de mise en valeur.

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. A la difficulté de définir cet objet, s'ajoute la problématique de l'analyse de la perception que peut avoir un locuteur de son environnement et comment celle-ci se reflète à l'oral.

L'analyse de la perception d'un lieu commence par l'étude des opinions émises à leur sujet. Cette étude suppose donc une réflexion sur la façon dont quelqu'un donne son opinion, en particulier à propos de son environnement. Cependant, l'analyse de la perception d'un lieu doit aller plus loin que la caractérisation de l'opinion relevée pour proposer une analyse complète. Plus largement, la perception se retrouve dans les habitudes, les actions racontées

par les locuteurs. Percevoir un lieu c'est aussi en avoir une image mentale. La représentation de cette image peut être matérialisée sous la forme de cartes géographiques. Ainsi, la visualisation cartographique des déclarations subjectives des locuteurs est très importante pour rendre compte des différents aspects de leur perception urbaine.

PARTIE 1

IDENTIFICATION DES LIEUX DANS L'ORAL TRANSCRIT

Chapitre 1 : LIEU : UN OBJET A DELIMITER

1.1 Du point de vue de la linguistique

La notion de lieu est l'élément central de notre réflexion et nécessite d'être explicité. Le Trésor de la Langue Française Informatisé (TLFi)[9] considère le lieu comme une « *portion délimitée de l'espace* » en précisant que « *l'espace est déterminé par sa situation dans un ensemble par la chose qui s'y trouve ou l'événement qui s'y produit* ». Le dictionnaire de français Larousse en ligne[10] le définit comme la « *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* ».

Ces définitions s'accordent pour associer le lieu et la notion d'espace. Le TLFi[11] présente l'espace comme une « *distance déterminée* », ou comme la « *distance comprise entre un point et un autre, entre un lieu, un objet et un autre* ». Le Larousse en ligne[12] donne plusieurs définitions au mot *espace* et nous retiendrons les propositions suivantes :

- Propriété particulière d'un objet qui fait que celui-ci occupe une certaine étendue, un certain volume au sein d'une étendue, d'un volume nécessairement plus grand que lui et qui peuvent être mesurés.
- Étendue, surface ou volume dont on a besoin autour de soi.
- Portion de l'étendue occupée par quelque chose ou distance entre deux choses, deux points.
- Étendue, surface, région.

[9] http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?23;s=518124555;r=2;nat=;sol=1;
[10] http://www.larousse.fr/dictionnaires/francais/lieu_lieux/47076
[11] http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?11;s=3175278615;r=1;nat=;sol=0;
[12] http://www.larousse.fr/dictionnaires/francais/espace/31013

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.

Les linguistes se sont aussi intéressés à la notion du lieu majoritairement à travers l'étude des expressions spatiales (Boons, 1987 ; Borillo, 1998 ; Laur, 1991 ; Vandeloise, 1986 ; Le Pesant, 2011, 2012 ; etc.). On peut citer le travail de 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 toponymie s'intéresse à l'histoire, l'évolution, la signification des noms de lieux par rapport à une langue ou une communauté donnée[13]. Les noms propres désignant un lieu sont alors appelés des toponymes. Bouvier (1999) s'interroge sur la structure linguistique des toponymes et de leurs variations dans la pratique quotidienne des usagers. Il définit ainsi deux classes de toponymes :

- les toponymes d'usage, dominants, sinon exclusifs, en dehors des espaces agglomérés. Ils résultent d'un accord implicite, progressivement acquis, des membres d'une collectivité sur la désignation d'un référent qui leur est commun : La Combe, Riouclar, rue de l'Eglise, place du Marché...
- les toponymes que l'on pourrait appeler de création, très rares, sinon inexistants en dehors des agglomérations : créés par la décision d'un pouvoir, qui aujourd'hui est généralement municipal, mais qui a pu être seigneurial dans le passé ou encore d'une autre origine, ils s'imposent en principe à l'usage des habitants et de tous les utilisateurs du lieu : rue Victor Hugo, rue Richelieu, place de la Concorde...

[13] Toponymie. (2019, mai 14). *Wikipédia, l'encyclopédie libre*. Page consultée le 13:38, mai 14, 2019 à partir de http://fr.wikipedia.org/w/index.php?title=Toponymie&oldid=159254232.

Aux toponymes de création, définis presque arbitrairement, peuvent se substituer les toponymes d'usage. Les pratiques du lieu, les habitudes qui y sont liées ou une caractéristique propre à l'endroit ou même l'action du temps, peuvent inciter les usagers du lieu à employer un nom différent du toponyme d'origine pour s'y référer. 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.

1.2 Point de vue du TAL

De nombreux travaux en TAL s'intéressent aux lieux et à leurs spécificités dans l'optique d'un traitement automatique. Les lieux sont le plus souvent identifiés dans le cadre de recherche autour du concept d'entités nommées. Les entités nommées sont tous les éléments du langage qui font référence à une entité unique et concrète comme les noms de personne, d'organisation ou de lieu. Les lieux sont aussi étudiés sous le nom d'entités spatiales. Cette section propose de faire état des travaux existants autour de ces notions en TAL.

1.2.1 Entités nommées

Les chercheurs du domaine du TAL abordent principalement les lieux du point de vue de leur détection automatique dans le cadre de la tâche de reconnaissance des entités nommées (REN). Les entités nommées sont « *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, 2013). Selon Ehrmann (2008) les entités nommées représentent « *toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus* ». Ces entités représentent des objets textuels porteurs de sens généralement classés selon

plusieurs catégories : lieux, personnes, organisations, dates, unités monétaires et pourcentages (Maurel *et al.*, 2011 ; Nadeau & Sekine, 2009 ; Chinchor, 1998).

- […] tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (ie. humain, économique, géographique, etc.)
- […] noms propres au sens classique, noms propres dans un sens élargi mais aussi expressions de temps et de quantité

MUC-A, Chinchor, 1998

La REN consiste en l'annotation et l'extraction d'entités nommées. Selon Eshkol-Taravella (2015 : 10) le processus d'annotation « *consiste dans l'apport d'informations de nature différente* » à des fins variées pour expliquer, commenter, décrire un texte, document, etc.. Annoter aide à décrire, caractériser, expliciter les informations contenues dans un corpus donné. L'annotation des entités nommées consiste en la description des entités porteuses de sens dans un document.

Depuis les années 1990 et les conférences américaines MUC (Message Understanding Conferences), la question de la reconnaissance des entités nommées est incontournable dans le domaine du TAL. De nombreux outils de détection automatique sont dédiés à la REN en anglais comme l'un des modules du Stanford NER (Finkel *et al.*, 2005) ou le POLYGLOT NER présenté dans Al-Rafou *et al.* (2015). Pour le français, nous pouvons citer le module CasEN[14] développé pour le logiciel Unitex[15] (Paumier, 2009).

Nombreux sont les travaux visant la détection de ce type d'entités. Afin de comparer les performances des différents outils, des campagnes d'évaluation sont régulièrement lancées. Le Pevedic & Maurel (2016) décrivent le déroulement des campagnes d'évaluation pour la REN de la manière suivante :

[14] http://tln.li.univ-tours.fr/Tln_CasEN.html
[15] https://unitexgramlab.org/fr

> Une campagne d'évaluation propose donc des annotations à insérer dans un corpus sur lequel tous les systèmes participants seront évalués en termes de rappel et précision (et/ou F-mesure). Un guide d'annotation précise le résultat attendu. Il est accompagné d'un corpus d'entraînement et de test, notamment pour les systèmes à base d'apprentissage. Cette annotation manuelle est elle-même contrôlée par des mesures d'accord inter-annotateurs (Artstein & Poesio, 2008) (Fort *et al.*, 2012).

Diverses campagnes ont eu lieu lors des conférences MUC ou dans le cadre du programme ACE (*Automatic Content Extraction*). En France, les campagnes ESTER[16] et ETAPE[17] sont les plus connues.

La campagne ESTER se décompose en deux phases (ESTER1, 2005 et ESTER2, 2009) et avait pour objectif global de mesurer les performances de système de transcriptions automatiques d'émissions radiophoniques (Gravier et al., 2004). ETAPE présente des objectifs similaires de mesure de performances de systèmes de transcriptions d'émissions radiophoniques. La REN était l'une des tâches évaluées dans les deux campagnes.

Dupont (2017) rappelle que si la REN est une problématique à part entière en TAL et en Recherche d'Information (RI), elle est aussi centrale dans d'autres tâches comme l'extraction de relations, la construction de bases de connaissances, le résumé automatique ou encore la résolution de chaînes de coréférence.

Les définitions proposées restent larges et laissent place à interprétation ce qui pose certaines difficultés pour le traitement des entités nommées. L'une des principales difficultés concerne le choix des catégories à utiliser pour caractériser chaque entité. Fort *et al.* (2009) relèvent que :

> au-delà de la triade "universelle" définie par MUC (PERSONNE, LIEU et ORGANISATION), l'inventaire des catégories à annoter est difficile à stabiliser et à définir. Prenons l'exemple de la catégorie PERSONNE : s'il est évident qu'un nom d'individu tel que Lionel Jospin est à annoter à l'aide

[16] http://www.afcp-parole.org/camp_eval_systemes_transcription/
[17] http://www.afcp-parole.org/etape.html

de cette catégorie, que faut-il faire des Kennedy, de Zorro, des Démocrates ou de St Nicolas ?

En effet, les typologies existantes pour la description des entités nommées diffèrent d'un projet à l'autre et sont loin de faire consensus. Reprenons les exemples présentés par Fort *et al.* (2009) et comparons leurs annotations selon les recommandations des conventions d'annotation des campagnes d'évaluation ESTER2 et Quaero (Rosset *et al.*, 2011 ; Mondary *et al*., 2012) de la campagne ETAPE :

Exemples	ESTER2	ETAPE
Lionel Jospin	<pers.hum>Lionel Jospin</pers>	<pers.ind>Lionel Jospin</pers>
Kennedy	<pers.hum>Kennedy</pers>	<pers.coll>Kennedy</pers>
Zorro	<pers.hum>Zorro</pers>	<pers.ind>Zorro</pers>
Les Démocrates	<pers.hum>Les Démocrates</pers>	<func.coll>Les Démocrates</func>
St Nicolas	<fonc.rel>St</fonc> <pers.hum>Nicolas</pers>	<func.ind>St</func> <pers.ind>Nicolas</pers>

Tableau 1 : Comparaison d'annotation d'entités nommées

Dans la convention d'ESTER2, la classe Personne regroupe les entités désignant des personnes ou des animaux, qu'ils soient réels ou fictifs. Ainsi, *Lionnel Jospin*, *Kennedy*, *Zorro*, et les *Démocrates* sont annotés comme des personnes humaines (<pers.hum>). La convention Quaero propose une approche différente en considérant les Personnes en tant qu'individu ou groupe d'individus. Les entités *Lionel Jospin* et *Zorro* sont donc considérés comme des individus (<pers.ind>) et les *Kennedy* comme un groupe d'individus (<pers.coll>). Dans la convention ESTER2, on fait donc la différence entre les êtres vivants doués de conscience mais aucune distinction ne sera faite lorsque l'on observe un seul individu ou un groupe. Dans la convention Quaero par contre, le type d'être vivant n'est pas considéré, seule l'information sur le nombre d'individus est prise en compte.

L'entité *St Nicolas* est annotée de la même manière dans les deux conventions en tant qu'une personne. La fonction St est considérée comme une fonction religieuse (<fonc.rel>)

par la convention ESTER2 et une fonction individuelle (<fonc.ind>) dans la convention Quaero.

L'une des approches ne prévaut pas sur l'autre mais ces différences démontrent bien les difficultés existantes pour la catégorisation des entités nommées. Les conventions d'annotation sont faites pour pallier cette difficulté mais Fort *et al.* (2009) estiment que « *ces difficultés entraînent un coût supplémentaire et une baisse de qualité de l'annotation* ». Le manque de précision dans la définition même des entités nommées peut laisser place à de l'hésitation chez les annotateurs, faisant apparaître un risque d'incohérence dans une tâche d'annotation plus étendue. L'objectif d'un projet peut tout de même commander quelques adaptations dans la façon de typer des entités nommées. Malgré ces possibles différences de typologies, Fort *et al.* (2009) proposent quelques critères définitoires des entités nommées comme « *l'unicité* » et « *l'autonomie référentielle* ». L'entité doit avoir un référent unique et autonome, dans le sens où l'on doit être capable de l'identifier par sa simple mention, même si cette référence est contextuelle. Par exemple, les noms de ville font référence à un espace géographique déterminé. L'utilisation du nom d'une ville permet de faire le lien avec la réalité qu'elle représente. Plusieurs villes peuvent porter le même nom, mais c'est le contexte qui évacue les possibles ambiguïtés.

1.2.2 Entités spatiales

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. Par contre, les expressions : *l'école Louis Guilloux, la grande rue de Bourgogne* ou *Orléans*, sont bien des lieux et leur nom seul permet de les placer sur une carte. Ils sont donc à la fois des lieux et des entités spatiales.

Lesbeguerries (2007) distingue deux types d'entités spatiales. Les entités spatiales absolues (ESA) qui représentent les informations spatiales les plus « primaires » et les plus proches de la définition des entités nommées de type lieux (ex *: la ville d'Orléans, le campus de la Source*). Les ESA associées à des indications géographiques (ex : *au sud de la ville d'Orléans, près du campus de la Source*) sont qualifiées d'entités spatiales relatives (ESR).

Les entités spatiales continuent à être au cœur de travaux en Recherche d'Information comme en témoigne la variété des contributions à l'atelier Gestion et Analyse des données Spatiales et Temporelles (GAST) [18] organisé lors de la Conférence Internationale sur l'Extraction et la Gestion des Connaissances (ECG)[19], dont la dernière édition date de 2018. Une nouvelle action prospective, intitulée Humanités Numériques Spatialisées[20] a été créée en 2019 sous l'égide du Groupement de Recherche CNRS MAGIS (Méthodes et Applications pour la Géomatique et l'Information Spatiale)[21]. Cette action se consacre tout particulièrement à « l'extraction d'information géographique et analyse spatiale à partir de textes en SHS ». La majorité des travaux dédiés à l'extraction automatique des entités spatiales 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 dans cette thèse s'appuie sur des transcriptions de conversation orales. Ce travail exploite donc des données de nature différente présentant de nouveaux challenges en ce qui concerne l'extraction d'informations géographique.

1.3 Du point de vue géographique

Pour comprendre la notion de lieu, il est nécessaire d'interroger un autre domaine dans lequel cette notion est centrale : la géographie. Le concept de lieu est fondamental en géographie, au point qu'elle a pu être qualifiée de « *science des lieux* » par Paul Vidal de la Blache (1913) au début du XX^{e} siècle. Les questions relatives à la définition du mot *lieu* et même à son utilisation font toujours débat dans le domaine de la géographie. Dans

[18] https://gt-gast.irisa.fr/gast-2018/
[19] https://egc18.sciencesconf.org/
[20] https://projet.liris.cnrs.fr/aphns-magis/
[21] http://gdr-magis.imag.fr/

l'encyclopédie Hypergéo[22], Clerc propose une synthèse. Selon lui, le sens commun aux emplois de *lieu* serait d'être « *des portions déterminées et singulières de l'espace auxquelles sont associées des toponymes* ». Mais depuis les années 70, il souligne que deux acceptions du terme en concurrences précisent sa définition. La première relève de l'analyse spatiale. Dans leur *Introduction à la Géographie Humaine*, Bailly & Béguin (2005) évoquent la difficulté de poser une définition précise du terme et propose de le considérer comme « *une unité spatiale élémentaire dont la position est à la fois repérable dans un système de coordonnées et dépendante des relations avec d'autres lieux dans le cadre d'interactions* ». De ce point de vue, le lieu correspond à un point dans l'espace que l'on peut situer sur une carte indépendamment des relations qu'il peut entretenir avec d'autres lieux. C'est un endroit où l'on peut situer des phénomènes géographiques. Berque, dans le *Dictionnaire de Géographie et de l'espace des sociétés* (Lévy & Lussault, 2013), continue en ce sens en affirmant que le lieu, c'est « *là où quelque chose se trouve ou/et se passe* ». Dans la lignée de Clerc, il considère d'abord que le lieu est indépendant de toutes choses et définissable en lui-même : le lieu est un point abstrait et objectif dans l'espace, relevant de la géométrie. Puis, il rejoint la définition de Béguin en attribuant au lieu une dimension relationnelle : « *le lieu dépend des choses, les choses en dépendent* ». De ces relations, 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.

Diverses nuances existent mais, au delà de l'analyse spatiale pure, ces discussions introduisent toute l'idée de l'existence d'une relation entre un ou des individus avec une portion de l'espace, ou dans une portion de l'espace. Cette idée introduite par Tuan (1977) constitue la deuxième acception du terme lieu et inspire le courant de la *Géographie Humaniste*. A partir de ce postulat, Clerc note que :

> Le lieu et l'homme se fondent mutuellement ; le lieu participe de l'identité de celui qui en est - chacun se définit, et définit son environnement, notamment par son appartenance spatiale - et les individus donnent une identité, et même plus fondamentalement une existence, au lieu. Cette

[22] Pascal CLERC, « LIEU », *Hypergéo* [en ligne], consulté le 13 mars 2018. URL : http://www.hypergeo.eu/spip.php?article214

> relation étroite permet la métaphore de l'enracinement et suppose une dimension temporelle. Le lieu s'inscrit dans la durée ; il est mémoire et temps cristallisé.

Ce sont les interactions de l'homme avec son environnement qui définissent le lieu. Cette définition fait écho aux travaux de Frémont (1980) à propos de l'espace vécu. L'espace vécu est l'endroit où l'homme et son espace « *s'harmonisent* ». L'homme s'y investit et s'approprie son environnement par sa perception et sa pratique. Cette idée que l'homme est intrinsèquement lié au lieu a provoqué l'apparition de nouvelles méthodes de cartographie. Elise Olmedo (2015) définit la « *cartographie sensible* » et la présente comme *« la seule possibilité pour représenter un espace traversé d'affects »*. La géographie s'est ainsi emparée de la relation Homme-Lieu et a vu émerger de nouveaux concepts, de nouvelles méthodes afin d'en analyser tous les aspects.

1.4 Notion de lieu dans ce travail

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. Les ESA considèrent principalement les noms propres faisant référence à des lieux et on parlera d'ESR lorsque les ESA sont associées à des indications géographiques.

L'objectif principal de cette thèse est l'analyse de la perception de la ville Orléans dans le corpus oral ESLO2. Cette perception est notamment représentée sous la forme d'une carte associant les lieux mentionnés par les locuteurs avec les déclarations subjectives qu'ils font à leur sujet. Dans cette perspective, le fait de pouvoir localiser les lieux mentionnés est donc primordial pour atteindre les objectifs définis. Le terme ESA répond directement à cette problématique de géolocalisation mais comme le terme *toponyme*, il recouvre presque

exclusivement les noms propres désignant un espace géographique. Pour la suite de l'analyse, *lieu* sera préféré pour se conformer à la terminologie des recherches en REN mais avec un sens plus large. La définition du lieu combine les différentes approches de la linguistique, du TAL et de la géographie. Le lieu est donc 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.

Chapitre 2 : MODELISATION, ANALYSE ET DETECTION DE LIEUX DANS LES TRANSCRIPTIONS

Les lieux constituent le premier type d'information à identifier dans les transcriptions composant le corpus. L'objet de ce chapitre est la description de cette étape. Il s'agit ici de présenter le développement d'un nouveau système de reconnaissance des lieux. La démarche proposée suit une approche symbolique fondée sur des lexiques et des règles, et prend en compte des caractéristiques propres de l'oral afin de les intégrer dans la chaîne de traitement.

2.1 Méthodologie

Le module de reconnaissance des lieux est composé de plusieurs étapes [23]:

- la modélisation des lieux pour l'élaboration d'une convention d'annotation (A),
- l'annotation manuelle pour la constitution d'un corpus de référence (B),
- le prétraitement du corpus avant son analyse (C),
- la gestion de ressources lexicales pour la génération de variantes de noms de lieux (D),
- l'utilisation des mentions de lieux déjà identifiées pour la résolution de cas simples de coréférences (E),
- l'évaluation du module d'annotation des lieux (G).

[23] cf. Figure 3

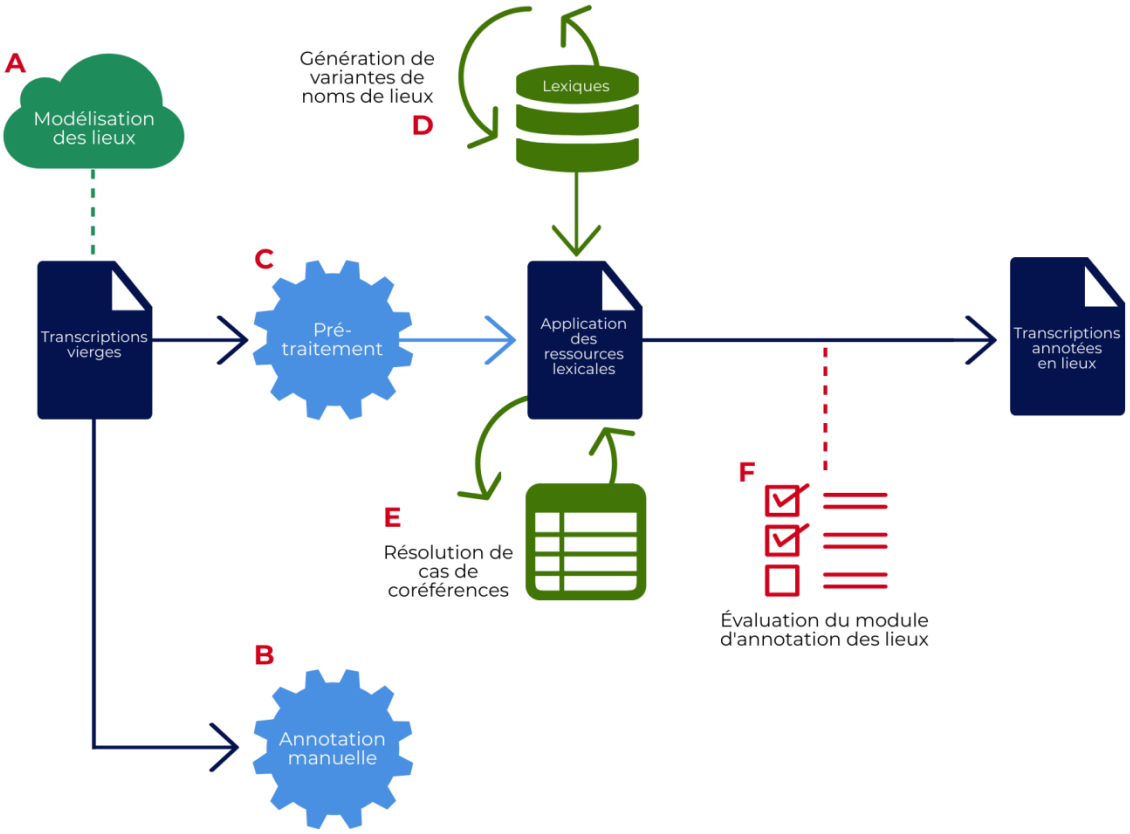

Figure 3 : Méthodologie pour l'annotation des lieux

Avant de procéder au traitement des données, une étape d'observation manuelle du corpus a permis la modélisation des lieux et d'établir des conventions d'annotation (A). Ces conventions guident le travail d'identification des lieux qui est réalisé à partir de transcriptions vierges. Un corpus de référence a été construit manuellement en accord avec des conventions d'annotation préétablies. Le corpus de référence est essentiel pour l'évaluation finale du système (B). Comme évoqué précédemment, la manipulation de conversations orales pose certaines difficultés dans le traitement informatique. Aussi, une étape de pré-traitement est nécessaire pour parer ces difficultés (C). L'annotation des lieux en elle-même s'inscrit dans le cadre des approches symboliques. Le système s'appuie sur l'exploitation de ressources lexicales (D) auxquelles sont associées des règles et des patrons décrivant les contextes d'apparition des lieux. Des opérations sont ensuite réalisées pour essayer de résoudre certains cas de coréférence entre les mentions et d'effectuer l'analyse des variations du nommage d'un même lieu (E). Finalement, le module est évalué à partir du corpus de référence pour valider l'efficacité du système (F).

2.2 Modélisation : conventions d'annotation

Avant de procéder au traitement du corpus, il est nécessaire de s'intéresser à la modélisation de l'information spatiale à identifier. Cette modélisation s'appuie sur les conventions existantes et l'observation manuelle d'un échantillon du corpus général.

Le sous corpus permettant de tester cette modélisation est constitué d'une vingtaine de transcriptions. Son exploration aboutit à l'identification de ressources lexicales et l'élaboration de règles et de patrons nécessaires à la détection des désignations de lieux.

Etablir une convention d'annotation est primordial dans toute tâche d'annotation. Cette convention est un guide présentant les objectifs à atteindre par des annotateurs humains ou des machines dans une tâche d'annotation donnée tout en décrivant la façon de procéder. Son élaboration est fondée sur une observation générale des données afin de déterminer quelles sont les informations pertinentes à relever.

La convention d'annotation des lieux proposée répond à l'objectif final du projet : la cartographie de la perception de leur ville par les Orléanais. Ainsi, la convention requiert le renseignement de trois informations pour caractériser les lieux : le type, la zone géographique et le nom officiel. Elles sont nécessaires pour la réalisation de la carte finale.

2.2.1 Typologie des lieux

2.2.1.1 Typologies existantes

Comme décrit précédemment[24], la tâche de REN est devenue une tâche indépendante en TAL. Celle-ci est régulièrement au centre de différentes campagnes d'évaluations d'outils dédiés à l'extraction d'informations. Plusieurs campagnes comme ESTER ou ETAPE évaluent par exemple l'annotation des entités nommées dans des corpus d'émissions radiophoniques ou télévisuelles. Ces projets présentent chacun des conventions d'annotation des entités nommées qui ont inspiré notre propre convention d'annotation des lieux.

[24] cf. 2.2.1

La campagne d'évaluation ESTER s'est déroulée en deux phases (ESTER1 et ESTER2) ayant pour objectif commun de mesurer les performances de systèmes de transcription d'émissions radiophoniques. L'une des tâches proposées concerne la détection des entités nommées dans le flux de parole. Afin de guider ce travail, des conventions d'annotation sont mises à disposition des participants.

Elles répartissent les entités nommées en sept grandes catégories : Personne, Fonction, Organisation, Lieu, Production Humaine, Date et Heure et Montant (Figure 4).

Convention 1.2	Pour les différentes catégories et sous-catégories identifiées, les valeurs xxx et yyy prennent les valeurs suivantes :
- personne ▪ humain réel ou fictif ▪ animal réel ou fictif	- pers ▪ pers.hum ▪ pers.anim
- fonction ▪ politique ▪ militaire ▪ administrative ▪ religieuse ▪ aristocratique	- fonc ▪ fonc.pol ▪ fonc.mil ▪ fonc.admi ▪ fonc.rel ▪ fonc.ari
- organisation ▪ politique ▪ éducative ▪ commerciale ▪ non commerciale ▪ média & divertissement ▪ géo-socio-administrative	- org ▪ org.pol ▪ org.edu ▪ org.com ▪ org.non-profit ▪ org.div ▪ org.gsp
- lieu ▪ géographique naturel ▪ région administrative ▪ axe de circulation ▪ adresse o adresse postale o téléphone et fax o adresse électronique ▪ construction humaine	- loc ▪ loc.geo ▪ loc.admi ▪ loc.line ▪ loc.addr o loc.addr.post o loc.addr.tel o loc.addr.elec ▪ loc.fac
- production humaine ▪ moyen de transport ▪ récompense ▪ œuvre artistique ▪ production documentaire	- prod ▪ prod.vehicule ▪ prod.award ▪ prod.art ▪ prod.doc
- date et heure ▪ date o date absolue o date relative ▪ heure	- time ▪ time.date o time.date.abs o time.date.rel ▪ time.hour
- montant ▪ âge ▪ durée ▪ température ▪ longueur ▪ surface et aire ▪ volume ▪ poids ▪ vitesse ▪ autre ▪ valeur monétaire	- amount ▪ amount.phy.age ▪ amount.phy.dur ▪ amount. phy.temp ▪ amount. phy.len ▪ amount.phy.area ▪ amount.phy.vol ▪ amount. phy.wei ▪ amount. phy.spd ▪ amount.phy.other ▪ amount.cur

Figure 4 **: Liste des étiquettes de la convention d'annotation ESTER2**[25]

La typologie des lieux dans la convention ESTER comprend cinq catégories[26]. On peut noter qu'une différence est faite entre les axes de circulation et le voies correspondant à une adresse postale. La classe des adresses est elle-même détaillée selon trois sous-catégories qui permettent de faire la distinction entre les adresses postales, téléphoniques

[25] http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf
[26] cf. Tableau 2

et électroniques. Cette distinction assez fine tranche avec l'étiquette assez générale des régions administratives. En effet, aucune différence n'est faite, que l'on annote le nom d'une ville, d'une région ou d'un pays. Pourtant les entités décrites correspondent à des échelles très différentes.

Géographique naturel (<loc.geo>)		Loire, Forêt d'Orléans, Mer du Nord …
Région administrative (<loc.admi>)		Orléans, Ardèche, Maghreb, Corée du Sud …
Axe de circulation (<loc.line>)		RN20, route de Paris…
Adresse	Adresse postale (<loc.addr.post>)	avenue Jean Zay, place Adolphe Cochery …
	Téléphone et fax (<loc.addr.tel>)	02 38 49 47 04 …
	Adresse électronique (<loc.addr.elec>)	*http://www.lll.cnrs.fr/* …
Construction humaine (<loc.fac>)		université d'Orléans, Cathédrale Sainte Croix d'Orléans …

Tableau 2 : Typologie des lieux dans la convention d'annotation ESTER

Pour les besoins de notre analyse, il n'est pas forcément utile de conserver le type *Adresse*, en particulier les adresses téléphoniques et électroniques. Les noms de rue correspondent éventuellement à une adresse postale et les axes de circulation peuvent être considérés sur le même plan. Il est important de faire la distinction entre les *Régions administratives* comme les villes, départements, région etc. Aussi, la nouvelle convention doit utiliser une typologie plus fine que celle proposée par ESTER pour ce type d'entité.

Le programme collaboratif d'innovation et de recherche industrielle sur l'analyse automatique et l'enrichissement de contenus numériques, multimédias et multilingues Quaero (Rosset *et al.*, 2011 ; Mondary et al., 2012), propose en 2011 les conventions d'annotation des entités nommées qui ont été utilisées entre autre par le projet ETAPE. Ce projet s'inscrit dans la continuité des campagnes d'évaluation ESTER visant la mesure des performances des systèmes de transcription d'émissions radiophoniques mais élargit les enjeux scientifiques en y ajoutant la parole spontanée, la parole superposée et la diversité des contenus.

Les conventions décrites sur le site de Quaero concernent l'annotation d'entités nommées plus large que la catégorie des noms propres car elles incluent des « *expressions construites autour de noms communs* » comme *les joueurs de l'équipe de France* ou *les mairies françaises*. Ces conventions tiennent compte également des disfluences de l'oral pour les corpus transcrits où elles peuvent apparaître :

(1) `il avait un restaurant <loc.addr.post>rue de euh Bourgogne</loc.addr.post> (ESLO2_ENT_1023)`

La disfluence *euh* entrecoupe l'entité nommée *rue de Bourgogne*. Malgré la présence de la disfluence qui interrompt la continuité du nom la rue, l'entité doit être annotée dans son ensemble en incluant la disfluence.

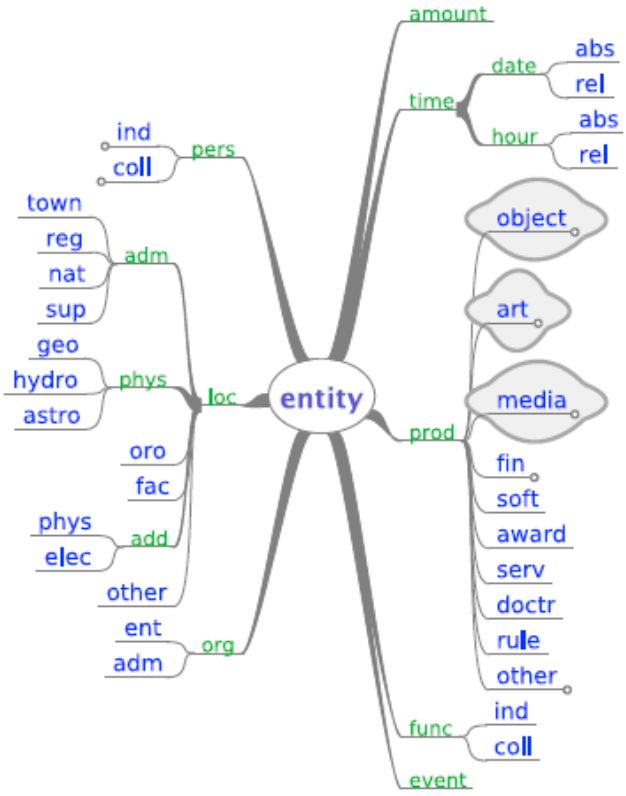

Figure 5 : Hiérarchie des entités dans la convention Quaero

En s'inspirant des thèses de Maud Ehrmann (2008) sur les entités nommées et de Mickaël Tran (2006) sur les noms propres, le jeu d'étiquettes de la convention Quaero a la structure illustrée dans la Figure 4. On distingue huit grandes classes : noms de personnes, de lieux, d'organisation, d'événements, de fonctions, de productions ainsi que des expressions temporelles et quantitatives.

Quelques différences existent entre les conventions proposées par ESTER et celles proposées par Quaero. La différence principale est l'ajout de la catégorie *Event* dans la convention Quaero. On peut aussi noter des différences dans la manière de typer les entités de la classe Lieu[27]. Comme dans la convention d'ESTER on retrouve 5 sous-catégories principales mais celles-ci ne catégorisent pas les lieux de la même façon. Par exemple, on trouve une nouvelle sous-catégorie qui est celle des *Bâtiments* pour toutes les constructions humaines. Les axes de circulation et les voies correspondant à des adresses postales sont regroupés en une seule catégorie. Contrairement à ESTER, les régions administratives sont différenciées selon qu'on a affaire à une ville, une région, un pays ou une entité supranationale.

Localisations administratives	Ville (<loc.adm.town>)	*Orléans, Paris, Lyon, Bordeaux, Vernon…*
	Région (<loc.adm.reg>)	*région Centre Val-de-Loire, Ardèche, Loiret…*
	National (<loc.adm.nat>)	*France, Corée du Sud, Tunisie …*
	Supranational (<loc.adm.sup>)	*Europe, Maghreb, Amérique du Nord …*
Lieux physiques	Terrestre (<loc.phys.geo>)	*Forêt d'Orléans, Etna …*
	Aquatique (<loc.phys.hydro>)	*Loire, Mer du Nord …*
	Astronomique (<loc.phys.astro>)	*Terre, Lune, Saturne …*
Voies (<loc.oro>)		*RN20, avenue Jean Zay, place Adolphe Cochery …*
Bâtiments (<loc.fac>)		*Université d'Orléans, Cathédrale Sainte Croix d'Orléans …*
Adresses	Physique (<loc.add.phys>)	*10 rue de Tours …*
	Electrique (<loc.add.elec>)	*02 38 49 47 04, http://www.lll.cnrs.fr/ …*

Tableau 3 : Typologie des lieux dans la convention Quaero

[27] cf. Tableau 3

Pour la réussite de notre projet, nous avons besoin d'une typologie fine décrivant des lieux de nature très variée. Cette nouvelle typologie est une synthèse étendue des conventions d'annotation d'ESTER2 et de Quaero. Cette nouvelle typologie est détaillée dans le point 3.2.1.2.

2.2.1.2 Types retenus

En vue de l'analyse de l'image d'une ville, il est nécessaire de conserver une typologie fine des lieux urbains. Une finesse qui n'est pas aussi utile pour la description des lieux naturels ou extra-urbains. La nouvelle convention reprend principalement les codes de la convention Quaero en ce qui concerne les lieux que l'on peut découper administrativement (quartiers, villes, pays…). De la même manière, les axes de circulation et les voies correspondant à des adresses postales sont regroupés sous l'étiquette *voie*. Les lieux physiques naturels sont considérés dans leur ensemble comme les lieux géographiques naturels dans la typologie de la convention ESTER2[28].

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…	

Tableau 4 : Nouvelle typologie des lieux (1/2)

[28] cf. Tableau 4

Reste la question de la typologie des Bâtiments ou Construction humaines. Selon la convention d'annotation ESTER2, ces entités sont « des lieux confinés, où l'on peut circuler, comme les maisons, les usines, les stades, les entreprises, les prisons, les musées… ». Elles se distinguent des organisations, soit des institutions ayant des fonctions commerciales ou administratives. Les constructions humaines et les organisations sont des entités souvent difficiles à distinguer. Le contexte d'emploi influence la catégorie dans laquelle est classée l'entité observée. Ainsi dans les exemples 2 et 3 :

(2) alors euh je suis étudiante à **l'université euh d'Orléans** La Source (ESLO2_ENTJEUN_03)

(3) je vais à la à **l'université** (ESLO2_ENTJEUN_03)

le même locuteur mentionne à deux reprises *l'université d'Orléans*. Dans l'exemple 2, les conventions ESTER et Quaero désigneront *l'université d'Orléans* comme une organisation alors que dans l'exemple 3, *l'université* sera considérée comme une construction humaine. Le contexte d'énonciation peut aider à différencier ces deux types d'entités mais ce n'est pas toujours le cas. Dans l'exemple 4 :

(4) bah euh le le le le le gros tort d'Orléans c'est d'avoir mis **l'université** à la Source enfin moi c'est mon avis hein que je vois à Tours euh bah c'est sympa euh euh le enfin ç- ça ça génère une activité euh sympathique ça génère des mouvements de enfin que là d'avoir mis **l'université** euh à à la Source ça ça coupe tout mais euh ça coupe tout quoi et euh ça fait moins d- moins d'émulsion enfin voilà (ESLO2_ENTJEUN_08)

l'université est mentionnée à deux reprises dans un contexte similaire et ambigu. En effet, *l'université* peut être considérée comme une organisation ou une construction. Les deux interprétations sont possibles et le contexte n'est pas suffisant pour en choisir l'une plus que l'autre. On peut tout de même trouver un point commun entre ces deux interprétations : le locuteur fait référence à une entité ayant une existence physique. Au final, que le locuteur fasse référence à la fonction propre de l'entité ou à son état de bâtiment abritant certaines activités, il évoque implicitement la situation géographique de cette entité.

A l'oral, les locuteurs font souvent référence à leur environnement, à la ville en mentionnant les commerces, les institutions, les organisations qui les entourent. Lorsqu'ils donnent leur avis à propos de la ville, ils évoquent par exemple son aménagement, l'offre commerciale, la présence ou l'absence de certains services, etc. La frontière entre construction humaine et organisation peut alors être subtile et la distinction n'être pas toujours pertinente du point de vue de l'analyse de la perception d'une ville par ses habitants. Les opinions émises à propos d'organisations participent aussi à la construction du portrait de la ville et ne peuvent pas être exclues.

La nouvelle convention d'annotation considère les lieux dans un sens plus large que celui prévu par les conventions d'ESTER2 et Quaero. Les entités classiquement réparties entre les classes *Lieu* et *Organisation* de ces conventions seront regroupées dans une classe hybride qui considère plutôt les entités spatiales ayant une fonction particulière. Cette perspective commande d'observer les typologies établies dans les conventions d'ESTER2 et Quaero pour la catégorisation des entités nommées de type *Organisation*.

La convention Quaero propose une simple distinction entre les organisations ayant une fonction commerciale (commerces, écoles, théâtres, etc.) et les organisations ayant une activité principalement administrative (mairies, préfectures, etc.). La convention d'ESTER2 au contraire a fait le choix d'une typologie plus fine constituée de 5 sous-catégories[29].

Politique	org.pol
Educative	org.edu
Commerciale	org.com
Non commerciale	org.non-profit
Média & divertissement	org.div
Géo-socio-administrative	org.gsp

Tableau 5 : Typologie des organisations selon ESTER2

Afin de caractériser les lieux ayant une fonction, le choix a été fait de s'inspirer de la convention d'ESTER2. La nouvelle typologie conserve ainsi les catégories principales

[29] cf. Tableau 5

proposées par ESTER2 en typant les constructions humaines de façon similaire aux organisations[30]. Toutefois, toutes les sous-catégories de la convention d'ESTER2 ne sont pas conservées. Le type *politique* représente les organisations à caractère politique telles que les organisations qui s'occupent des affaires gouvernementales (*partis politiques*, *mairies*, *ministères*, etc.) ou les organisations militaires reliées au gouvernement (ex : *CIA*, *Marine Nationale*…), etc. Pour ce type d'entité on préfèrera employer la catégorie organisation administrative présente dans la convention de Quaero qui semble plus pertinente par rapport aux données du corpus.

Lieux à fonction historique, touristique	type ="monument"
Cathédrale Sainte Croix, Hôtel Groslot…	
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 6 : Nouvelle typologie des lieux (2/2)

En conclusion, la nouvelle convention propose d'annoter les différentes mentions de *l'université* dans les exemples 2, 3 et 4 en tant que lieu à fonction éducative. 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.

2.2.2 Zone géographique

[30] cf. Tableau 6

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[31] 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.

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 7 : Zone géographique

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 5, Paris ne sera pas géolocalisé sur la carte finale contrairement à Orléans.

(5) c'est pas ça pose pas de problème donc euh ce qui manque à <loc type="ville" zone="2" label="Orléans">**Orléans**</loc> je dirais tu peux l'avoir à <loc type="ville" zone="0" zone="Paris">**Paris**</loc> donc c'est vrai que euh
(ESLO2_ENT_1008)

2.2.3 Label officiel

[31] L'agglomération orléanaise compte 22 communes : Boigny-sur-Bionne, Bou, Chanteau, Chécy, Combleux, Fleury-les-Aubrais, Ingré, La Chapelle-Saint-Mesmin, Mardié, Marigny-les-Usages, Olivet, Ormes, Saint-Cyr-en-Val, Saint-Denis-en-Val, Saint-Hilaire-Saint-Mesmin, Saint-Jean-de Braye, Saint-Jean-de-la-Ruelle, Saint-Jean-le-Blanc, Saint-Pryvé-Saint-Mesmin, Saran et Semoy.

Selon Dominguès et Eshkol-Taravella (2013), « *l'écriture des noms de lieux fait appel à des règles complexes qui s'appuient sur des connaissances linguistiques et extralinguistiques* ». Comme évoqué par Bouvier (1999) avec sa définition des toponymes d'usage, les noms de lieux sont le plus souvent créés par la décision d'une autorité et s'imposent aux usagers de ces espaces. Aujourd'hui des organismes sont chargés de surveiller la création et l'évolution des noms de lieux. L'ONU a par exemple créé un Groupe d'Experts des Nations Unies pour les Noms Géographiques (GENUNG) chargé du traitement « des problèmes de normalisation des noms géographiques, et de soumettre des suggestions et recommandations pour une standardisation (principalement linguistique) »[32] au sein d'un État donné. En France, c'est la Commission Nationale de Toponymie (CNT) qui s'occupe de maintenir une cohérence nationale tant sur le plan de la conservation que du développement des toponymes sur le territoire.

Si, comme le rappelle la Commission Toponymique du Québec, « *tout lieu ou entité géographique ne se voit attribuer qu'un seul nom officiel* »[33], il n'empêche que l'on peut s'y référer de différentes façons. Un locuteur peut évoquer un lieu en utilisant un nom non-normalisé. Ainsi, il peut substituer un nom du lieu (*Place Charles de Gaulle*) par un surnom (*Place de l'Arc de Triomphe*) ou une ancienne dénomination (*place de l'Étoile*). Il peut aussi abréger le nom officiel du lieu. D'une manière générale, il n'est pas rare que les noms de villes composés de plusieurs mots soient abrégés. C'est le cas par exemple d'*Aix*, pour *Aix-en-Provence*, *Saint-Germain* pour *Saint-Germain-en-Laye*. Ce type de variation est présent dans le corpus ESLO avec notamment *La Ferté* pour *La Ferté-Saint-Aubin* dans l'exemple 6.

(6) 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)

(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

[32] Groupe d'experts des Nations unies pour les noms géographiques. (2019, mai 24). Wikipédia, l'encyclopédie libre. Page consultée le 10:58, mai 24, 2019 à partir de : http://fr.wikipedia.org/w/index.php?title=Groupe_d%27experts_des_Nations_unies_pour_les_noms_g%C3%A9ographiques&oldid=159528923.

[33] http://www.toponymie.gouv.qc.ca/ct/normes-procedures/criteres-choix/normes-generales.aspx

```
vois où elle est ? la <loc type="voie" label=" rue de la
République">rue de la République</loc> ?
    (ESLO2_iti_06_11)
```

La troncation est un autre phénomène qui peut survenir comme dans l'exemple 7 où la *rue de République* devient *rue de la Rép-*. Différentes opérations concourent à faire varier le nom du lieu de la norme.

La finalité générale du projet est de modéliser la perception de la ville d'Orléans sous la forme d'une carte. Pour placer sur cette carte les lieux mentionnés dans les conversations analysées, il faut disposer de leurs coordonnées géographiques. Ce type d'information peut être récupéré dans des bases de données dédiées à l'information géographiques diffusées librement sur le Web. Cependant, les toponymes stockés dans ces bases de données correspondent aux noms officiels des lieux : les noms d'usage ne sont pas listés.

Même mentionné au moyen d'un nom non-normalisé, chaque lieu identifié doit être annoté et géolocalisé sur la carte finale. Pour que ces variantes soient associées à leurs coordonnées géographiques, c'est-à-dire pour que ces lieux soient associés aux bases de données spécialisées dans l'information géographique, il faut retrouver leur nom officiel. Cette information doit donc figurer dans les annotations effectuées. Ainsi, la valeur de l'attribut label est le nom officiel du lieu, sans aucune modification. Au moment de la création de la carte finale, il permettra de parcourir les entrées des bases de données géographiques pour la géolocalisation.

2.3 Constitution du corpus de référence

La constitution d'un corpus de référence est une étape incontournable dans le développement d'un outil en TAL. Le corpus de référence est composé de quinze transcriptions, distinctes de celles utilisées lors de l'observation manuelle du corpus et parfaitement annotées à la main en accord avec les conventions d'annotation établies. Il prend le rôle de référence dans le sens où il doit servir à évaluer les performances d'un système d'annotation automatique. Au terme de son développement, l'outil annote les transcriptions vierges composant le corpus de référence. L'enjeu est que le système d'annotation automatique parvienne à produire la même annotation. L'annotation réalisée

automatiquement est comparée avec la version de référence réalisée manuellement. Les mesures de Rappel, Précision et F-Mesure sont le plus souvent utilisées pour déterminer les performances de l'outil observé.

Ici le corpus de référence comprend l'annotation de tous les lieux présents dans les transcriptions le constituant. Chacune des informations requises dans la convention d'annotation doit être apposée dans le corpus aux endroits requis. Pour qu'il puisse remplir son rôle de référence, les transcriptions composant le corpus de référence doivent être sélectionnées avant la phase d'observation manuelle du corpus et doivent être annotées à la fin du développement de l'outil dont il doit permettre l'évaluation. Ainsi, son contenu reste inconnu et ne peut pas influencer le développement.

2.3.1 Sélection de données à annoter

Pour constituer le corpus de référence, certains modules du corpus ESLO ont été privilégiés. En effet, en considérant le contexte d'énonciation et les trames guidant les enregistrements, les deux modules Entretiens et Itinéraires ont été choisis pour créer l'échantillon à annoter manuellement.

Les Entretiens consistent en une discussion en face à face entre un chercheur et un locuteur témoin. Le chercheur mène la discussion selon une trame préétablie qui reste assez flexible pour laisser place à la spontanéité du discours du locuteur. D'une manière générale, la trame invite ce dernier à retracer son histoire personnelle, à partager ses habitudes de vie, etc. Chaque témoin est un habitant d'Orléans ou de son agglomération.

Le module Itinéraires regroupe des enregistrements réalisés en pleine rue. Des étudiants ou chercheurs vont à la rencontre de piétons pour leur demander leur chemin comme dans l'exemple suivant :

(8)

FD720 : bonjour excusez-moi de vous déranger je cherche la mairie d'Orléans
MH315 : c'est vers la cathédrale à pied ?
FD720 : oui ou en tram ou en ce que vous voulez [rire] du moment que j'y arrive [rire] (ESLO_iti_06_11)

La question est dans un premier temps posée avec un micro caché. Une fois que le locuteur a répondu, on lui révèle le micro et on lui demande de reformuler sa réponse. Suivent quelques questions sur les habitudes du locuteur dans la ville et son avis sur celles-ci. La collecte a été effectuée dans divers endroits de la ville afin d'interroger des locuteurs représentatifs de la diversité sociologique. Par leur mode de constitution, ces courts enregistrements forment un matériel riche en mentions de lieux relatives à la ville d'Orléans. Au total, dix transcriptions ont été sélectionnées aléatoirement dans les deux modules. Le corpus de référence est ainsi constitué de cinq Entretiens et cinq Itinéraires.

2.3.2 Processus d'annotation manuelle

Le corpus de référence doit être annoté en lieux à partir de la convention d'annotation établie pour les besoins du projet.

Différents outils comme Glozz[34], Analec[35] ou Gate[36] ont été conçus pour l'exploration de corpus et leur annotation. Ces outils sont souvent développés dans le cadre de tâches d'annotation bien particulières et leur optimisation s'en ressent tant dans leur utilisation que dans les formats d'annotation utilisés.

Pour favoriser l'interopérabilité de l'annotation réalisée manuellement, le choix a été fait de ne pas utiliser d'outils en particulier. Les annotations se sont donc faites directement dans l'éditeur de texte Notepad++[37]. Si cette méthode peut sembler fastidieuse, elle permet surtout d'économiser le temps que les annotateurs mettraient à maîtriser l'outil et les fichiers obtenus sont directement exploitables par le module d'annotation automatique des lieux, sans problème de compatibilité.

Des balises XML sont utilisées de la même façon que dans le projet Quaero pour annoter l'ensemble des mentions de lieux. Le format XML permet de normaliser l'annotation et de cette manière de vérifier son homogénéité. La balise générale <loc> compte trois attributs que sont le type de lieu, la zone géographique et le label officiel[38].

[34] http://www.glozz.org/
[35] http://explorationdecorpus.corpusecrits.huma-num.fr/analec-2/
[36] https://gate.ac.uk/
[37] https://notepad-plus-plus.org/fr/
[38] cf. section 2.2

Deux annotatrices sont intervenues (A1 et A2) pour constituer le corpus de référence. Fort (2017) caractérise l'expertise des annotateurs dans une tâche donnée selon trois niveaux : « *l'expertise dans le domaine du corpus* », « *l'expertise dans le domaine de l'annotation* » et « *l'expertise de la tâche* ». Si l'on transpose ces éléments à la tâche d'annotation des lieux dans des transcriptions, ces niveaux sont : le niveau de pratique du corpus ESLO, la connaissance du domaine des entités nommées et de leur annotation et la connaissance de la ville d'Orléans et de ses alentours.

L'annotatrice A1 peut être considérée comme experte dans les trois domaines [39]. L'annotatrice A2 a été recrutée dans le cadre d'un cours. Elle était étudiante en Master de Sciences du Langage à Orléans et se formait plus particulièrement au Traitement Automatique des Langues. Son cursus universitaire lui a donné une bonne connaissance des problématiques liées à l'analyse des corpus oraux. En formation au TAL au moment de la phase d'annotation, elle ne peut ni être considérée comme experte dans le domaine des entités nommées ni comme ayant suffisamment d'expérience sur des tâches d'annotation manuelle.

2.3.3 Accord Inter-Annotateur

Avant de réaliser l'annotation manuelle du corpus de référence, il faut s'assurer de l'efficacité de la convention d'annotation dans son rôle de guide et de la capacité des annotatrices à réaliser la même tâche de façon cohérente. Un test d'annotation a été réalisé sur trois transcriptions extraites du module Itinéraire et un accord inter-annotateur a été calculé.

Le calcul de l'accord inter-annotateur (AIA) consiste en la comparaison de l'annotation d'un même segment de données par deux personnes. Le score obtenu doit permettre de déterminer si les deux personnes observées sont capables de réaliser une même tâche, sans que cela ne soit dû au hasard. En l'occurrence, l'AIA doit montrer dans quelle mesure les annotatrices A1 et A2 sont en accord ou en désaccord dans la tâche d'annotation des lieux en fonction de conventions d'annotation préétablies.

[39] L'annotatrice A1 est l'auteure de cette thèse.

Plusieurs métriques existent pour évaluer cet accord. La plus simple est celle de l'accord observé (P_o) qui consiste à dénombrer le nombre d'items pour lesquels les annotateurs sont en accord. C'est donc la part d'annotation *a* pour laquelle les deux annotatrices sont d'accord par rapport au total d'annotation *N* :

$$P_o = a/N$$

Après annotation des trois transcriptions composées de 679 tours de paroles, A1 a réalisé 89 annotations alors que A2 en a réalisé 100. Parmi ces annotations, l'accord observé est donc :

$$P_o = a/N = 89/100 = 0{,}89 \text{ soit } 89\%$$

Les désaccords observés concernent généralement des différences d'empan au niveau de l'annotation comme dans l'exemple 9.

(9)

A1 : [rire] celle celle d'<loc type="ville" zone="2" label="Orléans">**Orléans**</loc> centre

A2 : [rire] celle celle d'<loc type="ville" zone="2" label="Orléans centre-ville">**Orléans centre**</loc> (ESLO_iti_02_03)

A2 a inclus l'adjectif centre dans son annotation. L'expression *Orléans centre* peut être considérée comme une ESR composée de l'ESA *Orléans* et de l'indication géographique *centre*. La tâche ne requiert que l'annotation des lieux, sans inclure les indications géographiques. On peut aussi noter que les annotatrices sont d'accord à 90% dans l'attribution d'un label officiel aux entités identifiées. Les différences observées s'expliquent le plus souvent par ces différences d'empan.

La différence dans le nombre d'annotations se retrouve dans des cas comme celui de l'exemple 10. L'annotatrice A1 n'a pas annoté l'entité *statue de Jeanne d'Arc* contrairement à l'annotateur A2. La *statue de Jeanne d'Arc* est un objet ayant un emplacement géographique bien précis dans la ville. Elle est la caractéristique principale

de la place du Martroi, l'une des places principales d'Orléans et est l'un des points de rendez-vous privilégiés des Orléanais.

(10)

A1 : la la statue de Jeanne d'Arc hein c'est quand même le truc d'<loc type="ville" zone="2" label="Orléans">**Orléans**</loc> hein euh

A2 : la la <loc type="monument" zone="2" label="statue de Jeanne d'Arc">**statue de Jeanne d'Arc**</loc> hein c'est quand même le truc d'<loc type="ville" zone="2" label="Orléans">**Orléans**</loc> hein euh (ESLO_iti_02_03)

Néanmoins, cette statue ne peut pas être considérée comme un lieu. En effet, malgré la dimension géographique qui lui est attachée, elle reste un élément de décor urbain qui se situe sur la place du Martroi. On choisira donc de ne pas annoter ce type d'entité.

		A1							
		Ville	Voie	Monument	Admin	Educatif	Commerce	Ncommerce	Total
A2	Ville	55	-	-	-	-	-	-	55
	Voie	-	5	-	-	-	-	-	5
	Monument	-	-	3	-	-	-	-	3
	Admin	-	-	1	0	-	-	-	1
	Educatif	-	-	-	-	4	-	-	4
	Commerce	-	-	-	-	-	19	-	19
	Ncommerce	-	-	-	-	-	-	2	2
	Total	55	5	4	0	4	19	2	89

Tableau 8 : Matrice de confusion des types de lieux annotés

L'évaluation de l'accord porte aussi sur les attributs du type de lieu et la zone géographique et a été réalisée grâce à une autre métrique : le Kappa de Cohen (Cohen, 1960).

Le Kappa $_{k}$ évalue « *la proportion de désaccords prévisibles qui ne se produisent pas, ou bien la proportion d'accord après que l'accord aléatoire a été retiré de la considération* »[40]. Cette mesure normalise l'accord observé et relativise cet accord en calculant un accord aléatoire, c'est-à-dire un accord dû au hasard. Elle calcule donc un rapport entre la probabilité d'accord observé (P_o) entre deux annotateurs et la probabilité d'un accord aléatoire (P_e) :

$$k = P_o - P_e / 1 - P_e$$

La probabilité (P_e) est la somme des annotations d'une même étiquette par un premier annotateur multiplié par les annotations d'un deuxième annotateur, divisé par le nombre total d'annotations *N*. Dans le cas de l'évaluation de l'accord à propos du type de lieu dans l'annotation des lieux, la formule est :

$$P_e = (A1_{ville} \times A2_{ville / N}) + (A1_{voie} \times A2_{voie / N}) + \ldots + (A1_{ncommerce} \times A2_{ncommerce / N})$$

A partir de la matrice de confusion des types de lieux annotés présentée dans le tableau 8, le k a pu être calculé :

$$P_o = i/N = (55 + 5 + 3 + 0 + 4 + 19 + 2) / 89 = 88/89 = 88$$

$$P_e = (55 \times 55 / 89) + (5 \times 5 / 89) + \ldots (2 \times 2 / 89) = 38,68$$

$$k = 88 - 38,68 / 1 - 38,68 = 0,98$$

Le Kappa de Cohen obtenu est de 0,98 pour le type des lieux dans l'annotation. Landis et Koch (1977) propose une grille de lecture faisant état de l'art pour l'interprétation de la mesure obtenue[41]. Le calcul du Kappa de Cohen pour la comparaison des annotations de A1 et 12 est égal à 0.81. Si l'on se rapporte à la grille de lecture, l'accord est considéré comme excellent.

[40] « *The coefficient is simply the proportion of chance expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration* » (Cohen, 1960)
[41] cf. Tableau 9

Accord	Kappa
Excellent	≥ 0,81
Bon	0,80 – 0,61
Modéré	0,60 – 0,41
Médiocre	0,40 – 0,21
Mauvais	0,20 – 0,0
Très mauvais	< 0,0

Tableau 9 : Grille d'interprétation de Landis & Koch du Kappa de Cohen[42]

Un seul désaccord est observé dans le tableau 8. Ce désaccord concerne les étiquettes *admin* et *monument* :

(11)

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

(ESLO_iti_02_03)

Comme représenté dans l'exemple 11, les deux annotatrices A1 et A2, résidants à Orléans, ne sont pas d'accord sur la façon de catégoriser le lieu *mairie d'Orléans*. Une partie de l'actuelle mairie d'Orléans est située à l'Hôtel Groslot, un ancien hôtel particulier du XVI^{e} siècle. Celui-ci sert principalement à accueillir des réceptions ou pour les célébrations de civils. Il est aussi ouvert à la visite tout au long de l'année. Le reste de la mairie se trouve de l'autre côté de la rue dans un bâtiment datant du XVII^{e} siècle. Les deux bâtiments présentent un très fort attrait touristique. Cet attrait a semble-t-il induit en erreur l'annotatrice A2. La mairie reste tout de même une entité à fonction principalement

[42] Source : J. Richard Landis et Gary G. Koch : The measurement of observer agreement for categorical data. *Biometrics, 33(1):159–174, 1977. ISSN 0006341X. URL http:*//www.jstor.org/stable/2529310

administrative, aussi, la mention mairie d'Orléans doit être annotée comme l'a fait l'annotatrice A1 : en tant qu'administration.

		A1			
		0	1	2	Total
A2	0	14	1	-	15
	1	-	-	18	18
	2	-	1	55	56
	Total	14	2	73	89

Tableau 10 : Matrice de comparaison des types de lieux annotés

L'accord autour de l'annotation de la zone géographique correspondant aux lieux identifiés a aussi été évalué au moyen du Kappa de Cohen à partir de la matrice présentée dans le tableau 10 :

$$P_o = i/N = (14 + 0 + 55) / 89 = 69/89 = 88$$

$$P_e = (A1_0 \times A2_{0/N}) + (A1_1 \times A2_{1/N}) + (A1_2 \times A2_{2/N})$$

$$P_e = (14 \times 14 / 89) + (5 \times 5 / 89) + (2 \times 2 / 89) = 48,7$$

$$k = 88 - 38,68 / 1 - 38,68 = 0,50$$

Le kappa k vaut 0,50. Selon la grille d'évaluation de Landis et Koch, c'est un accord modéré. Ce score s'explique surtout par les 18 désaccords pour lesquels A1 a indiqué la zone 2 (Orléans) alors que A2 a indiqué la zone 1 (l'agglomération). Les 18 désaccords correspondent à la même erreur. Ils concernent tous des lieux situés dans le quartier de La Source à Orléans. Ce vaste quartier se situe à l'extrémité Sud de la ville et a la réputation d'être très excentré. Ce sont sûrement ces attributs qui ont induit en erreur A2 au point de considérer ce quartier comme une ville distincte d'Orléans. Il fallait donc préférer la zone 2.

Après la comparaison des annotations et l'analyse des scores d'AIA obtenus, une concertation a eu lieu entre les deux annotatrices afin de leur permettre d'annoter le corpus de référence.

2.3.4 Analyse quantitative du corpus de référence

Quinze transcriptions ont donc été sélectionnées et annotées manuellement par deux annotatrices afin de constituer un corpus de référence. Dix d'entre-elles sont extraites du module Entretiens d'ESLO2 et et les cinq autres proviennent du module Itinéraires. Parmi ces transcriptions, 2 292 mentions de lieux ont été annotées[43].

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.

Transcription	Total
ESLO2_ENT_1016	311
ESLO2_ENT_1018	210
ESLO2_ENT_1019	162
ESLO2_ENT_1026	361
ESLO2_ENT_1028	193
ESLO2_ENT_1039	225
ESLO2_ENT_1042	182
ESLO2_ENT_1005	206
ESLO2_ENT_1050	175
ESLO2_ENT_1051	158
ESLO2_iti_10_05	13
ESLO2_iti_11_01	28
ESLO2_iti_11_06	23
ESLO2_iti_12_04	22
ESLO2_iti_13_02	23
Total	2292

[43] cf. Tableau 11

Tableau 11 : Nombre de mentions de lieux dans les transcriptions du corpus de référence

La répartition des lieux mentionnés en fonction de la typologie définie est affichée dans la figure 6. 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 24 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 10 lieux à fonction non commerciale et un seul lieu supranational sont présents dans le corpus de référence.

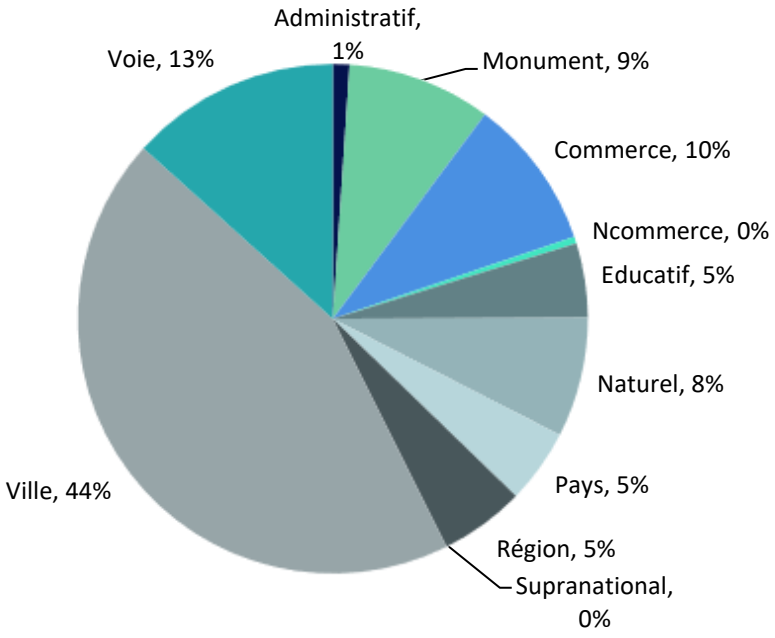

Figure 6 : Répartition des types de lieux dans le corpus de référence

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.

Label	Nombre	Proportion
Identiques	1456	64%
Variantes	834	36%

Tableau 12 : Part des variantes et des noms officiels de lieux dans le corpus de référence

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 subies des modifications de la part des locuteurs représentent 36 % des lieux annotés[44].

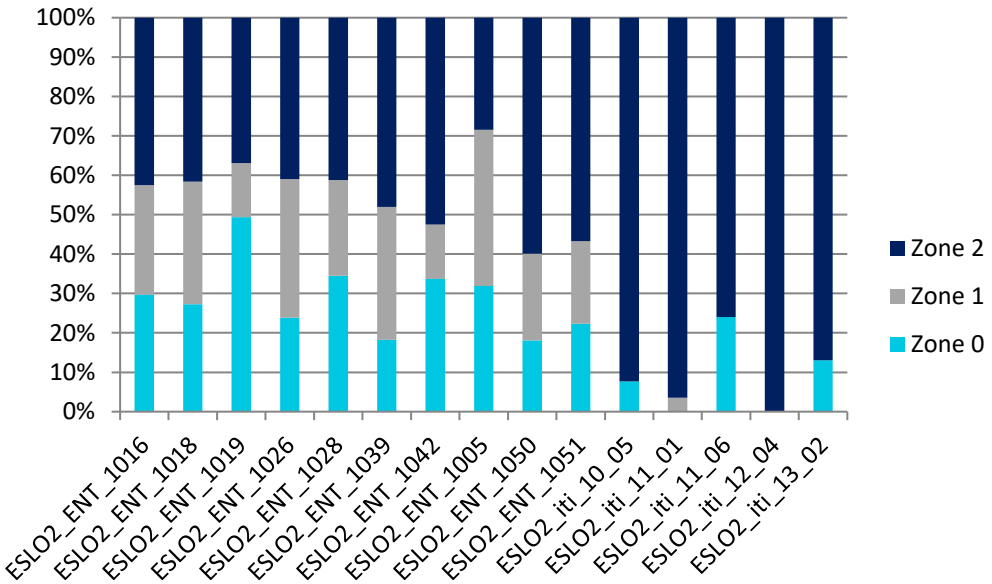

Figure 7 : Répartition des lieux selon la zone géographique dans le corpus de référence

[44] cf. Tableau 12

D’un point de vue géographique, la répartition des étiquettes stipulant la zone géographique montre que les Orléanais se concentrent sur les lieux en zone 2, relatifs à la ville d’Orléans (46 % des lieux mentionnés). Dans l’agglomération orléanaise, zone 1, et hors agglomération, zone 2, les indications de lieux représentent 54 % des mentions à parts égales. Dans la figure 7, on peut constater que les lieux de la zone 2 représentent au minimum 30 % des lieux mentionnés dans une transcription. Les lieux mentionnés dans les transcriptions extraites du module Itinéraire correspondent presque tous à des lieux de cette même zone en sorte qu’ils constituent des mentions pertinentes pour une étude sur la perception de la ville.

2.4 Module d’annotation automatique des lieux

Pour modéliser la perception de la ville d’Orléans au travers des enregistrements du corpus ESLO, la première grande étape est le repérage de toutes les mentions de lieux. On rappelle qu’un lieu est un espace déterminé que l’on peut mesurer et dont l’existence est intrinsèquement lie l’action de l’homme qui les nomme et les caractérise. Si des normes existent pour le nommage des lieux, chacun est à même de faire varier cette norme et de se référer aux espaces de son environnement par des moyens détournés. Pour ce travail, on utilise une définition extensive des lieux qui sélectionne les entités qui correspondaient aux classes Lieu et Organisation dans des travaux en REN. Il est aussi primordial de pouvoir placer sur la carte finale les lieux repérés.

En vue du développement d’un nouveau module d’annotation, des conventions d’annotation spécifiques ont été établies. A partir de ces conventions, deux annotatrices ont annotés manuellement un corpus de référence afin, d’une part, de valider l’efficacité de la convention d’annotation et, d’autre part, de permettre l’évaluation du module d’annotation automatique. Cette section présente le développement du module d’annotation automatique des lieux dans l’oral transcrit.

2.4.1 Etat de l’art

A l'heure du Big Data et du numérique, les méthodes et les outils disponibles évoluent continuellement pour traiter toujours plus de données, dont la nature varie continuellement. Les méthodes permettant la détection de mentions de lieux s'inscrivent la plupart du temps dans les travaux en REN. Nouvel *et al.* (2015 : 82) font l'état de ces différentes méthodes qui peuvent être classées selon trois catégories : des méthodes statistiques, symboliques et hybrides.

2.4.1.1 Méthodes statistiques

Les méthodes statistiques sont des approches souvent fondées sur l'apprentissage automatique. Le système d'apprentissage automatique apprend à partir d'un exemple qu'on lui a fourni comment à son tour réaliser la tâche qu'on lui aura demandé de réaliser. En l'occurrence, à partir d'un échantillon de données similaire au corpus de référence constitué, le système doit apprendre à détecter des mentions de lieux dans le reste du corpus. Pour obtenir de meilleurs résultats, cette approche nécessite une grande quantité de données pour constituer le modèle d'apprentissage. Ce corpus peut être constitué manuellement ou en utilisant une approche symbolique comme étape préalable à la correction manuelle.

Nouvel *et al.* (2015) décrivent différents systèmes comme les modèles par classes majoritaires qui consistent à « *déterminer la classe de chaque mot en considérant la classe qui lui est majoritairement associée dans le corpus d'apprentissage* », des modèles à décisions contextuelles s'appuyant sur des modèles génératifs comme les modèles de Markov à états cachés ou encore des modèles fondés sur des champs markoviens conditionnels (ou CRF pour Conditionnal Random Fields). Cette dernière méthode est à même, à la fois, de « *tenir compte du contexte pour prendre des décisions […] et de multiples indices* » définis lors d'un pré-traitement afin de « *déterminer quelle séquence d'étiquettes est la plus vraisemblable pour un texte donné* ». Le module pour la REN Stanford NER[45] (Finkel *et al.*, 2005) est un exemple de système fondé sur l'utilisation de CRF. D'autres systèmes peuvent s'appuyer sur des arbres de décisions (Borthwick *et al.*, 1998) ou des Machines à Vecteurs de Support (SVM) (Isozaki & Kazawa, 2002).

[45] https://nlp.stanford.edu/software/CRF-NER.html

L'inconvénient principal de ce type de méthode est la nécessité de disposer d'une grande quantité de ressources annotées pour créer les modèles d'apprentissages, c'est la raison pour laquelle les approches symboliques sont aussi utilisées dans des tâches de REN.

2.4.1.2 Méthodes symboliques

Les approches symboliques s'appuient sur des règles ou des patrons décrivant les contextes d'emploi ou la structure des entités recherchées à partir de ressources lexicales et de dictionnaires préalablement construits.

2.4.1.2.1 Travaux existants

Parmi les approches symboliques, plusieurs approches à bases de règles peuvent être citées. Stern & Sagot (1996) par exemple, proposent un système de REN traitant d'une part la détection et le typage des entités nommées, et la désambiguïsation et la résolution des entités nommées d'autre part. L'approche symbolique est couteuse en temps puisqu'élaborée manuellement par des experts de la tâche à réaliser. Cependant, elle présente l'avantage d'être adaptable à tout type de situation. Elle permet de gérer par exemple le cas des noms de lieux non-normalisés. En mêlant règles et ressources lexicales, Dominguès & Eshkol-Taravella (2013 ; 2015 ; 2017) traitent de la question des écarts par rapport à la norme pour le nommage des lieux dans les écrits du Web. Elles introduisent ainsi la notion de *lieux subjectifs* : des mentions de lieux subjectivement réappropriés par ceux qui les emploient. Une base de données dédiée à l'information géographique, BDNyme[46], est associée à des patrons et des règles pour l'identification de toponymes standards et non-standards. Ces règles décrivent les différents procédés lexicaux, syntaxiques et pragmatiques qu'une personne peut mettre en œuvre pour exprimer sa perception d'un lieu.

[46] https://documentation.ensg.eu/index.php?lvl=categ_see&id=18116

Au nombre des outils disponibles qui utilisent les méthodes symboliques pour la REN, on compte la ressource CasEN. Comme la plupart des outils dédiés à la REN, CasEN traite des données écrites. Pour vérifier l'hypothèse que ces outils ne sont pas directement applicables sur des données orales, l'efficacité de CasEN pour la détection des entités nommées de type lieu a été testée sur le corpus ESLO.

2.4.1.2.2 Test de CasEN

La ressource open-source CasEN (Maurel *et al.*, 2011) est dédiée à la REN dans des textes écrits en français. En accord avec les conventions Quaero et grâce au logiciel Unitex, la cascade CasEN utilise des ressources lexicales et des grammaires locales. Une cascade est une série de règles appliquées dans un certain ordre sur le corpus. Chaque règle s'appuie sur les conclusions de la précédente.

CasEN n'est pas développé initialement pour traiter des données orales, mal adapté à l'annotation d'entités nommées de type lieu dans un corpus comme ESLO. Pour le vérifier, nous avons évalué l'annotation faite par CasEN sur le corpus de référence établi spécifiquement à cette fin[47]. L'évaluation a été réalisée en termes de Précision (P), Rappel (R) et F-mesure (F). Ces mesures s'appuient sur la comparaison de corpus de référence présenté dans la section 2.3, et l'annotation automatique réalisée par le système à évaluer sur les données vierges composant le corpus de référence.

La mesure du Rappel représente la part des détections pertinentes par rapport à la totalité des détections que le système est censé effectuer. La Précision montre la part de détections pertinentes par rapport à l'ensemble des détections réalisées par le système. Ces deux mesures prennent en compte le nombre d'éléments correctement (Vrais positifs - VP) et incorrectement (Faux positifs - FP) identifiés ainsi que le nombre d'éléments qui n'ont pas été identifiés (Faux négatifs – FN). La F-mesure combine le Rappel et la Précision pour informer sur les performances générales du système :

[47] cf. section 2.3

$$P = VP / (VP + FP)$$

$$R = VP / (VP + FN)$$

$$F = 2 \times (P \times R) / (R + P)$$

En l'occurrence, pour la tâche d'annotation des lieux dans l'oral, CasEN obtient un Rappel de 0,27, une Précision de 0,96 et une F-mesure de 0,47. Le Rappel indique le taux de non-détection d'un élément par un outil. Plus ce taux est proche de 0, moins les résultats sont satisfaisants et à l'inverse 1 est un score parfait. Les résultats de l'évaluation montrent que le système détecte peu de lieux présents dans le corpus. À l'inverse des résultats obtenus au niveau du Rappel, les résultats de la Précision, qui montrent le taux des lieux détectés correctement, sont excellents. On peut en déduire que le système reconnaît peu de lieux présents dans le corpus. La grande majorité des détections manquantes dans l'annotation de CasEN concerne les mentions de lieux différentes de la norme, soit les noms abrégés (comme *La Ferté* au lieu de *La Ferté-Saint-Aubin*) ou tronqués (comme *rue de la Rép-* pour *rue de la République*). Les noms de lieux non-normalisés requièrent des traitements spécifiques. L'utilisation de CasEN aurait nécessité de très nombreuses adaptations et pour cette raison, le système n'a pas été retenu pour répondre aux objectifs de l'identification des lieux dans le corpus oral.

2.4.1.3 Méthodes hybrides

Les méthodes hybrides combinent les méthodes symboliques à base de règles et les méthodes statistiques à base d'apprentissage. Dans ce cadre, Lesbeguerries (2007) propose un système de reconnaissance des ESA dans un corpus écrit, fondé sur des méthodes symboliques s'inspirant de travaux en REN. Dans un premier temps, il recycle des Systèmes d'Information Géographiques (SIG), qui sont des systèmes d'informations « *conçus pour recueillir, stocker, traiter, analyser, gérer et présenter tous les types de*

données spatiales et géographique »[48] et les utiliser comme des lexiques afin d'identifier les ESA et les associer aux coordonnées géographiques correspondantes.

- La dénomination des entités n'est a priori pas le meilleur moyen de les identifier de manière unique mais c'est le seul disponible. En effet aucune règle d'unicité n'existe sur la dénomination des lieux.

Lesbeguerries, 2007 : 55

L'application mécanique d'un lexique sur le corpus ne suffit évidemment pas pour identifier l'ensemble des ESA qui s'y trouvait, mais c'est forcément la première étape à suivre dans l'élaboration d'un tel système. Pour compléter ces lexiques, les méthodes par apprentissages peuvent être utilisées. En l'occurrence Lesbeguerries propose de définir les ESA avec des caractéristiques statistiques et de les appliquer dans un processus de classification SVM, une méthode d'apprentissage basée sur des statistiques.

Les travaux plus récents de Zenasni (2016) s'intéressent à l'identification d'entités spatiales dans un corpus de messages courts (SMS et tweets). Un dictionnaire d'entités spatiales est appliqué sur le corpus afin d'identifier les lieux. Mais les caractéristiques propres aux écrits peu standards composant le corpus nécessitent en plus l'emploi de règles fondées sur des mesures de similarités et les caractéristiques lexicales des lieux afin d'identifier les mentions qui diffèrent de la norme proposée. Les mesures de similarité, comme la distance de Levenshtein (1965) servent à la comparaison de chaînes de caractères afin de mesurer jusqu'à quel degré ces deux chaînes diffèrent ou non entre-elles. Dans un système automatisé de REN, une telle mesure peut permettre de faire des rapprochements entre deux termes si l'on considère que ce qui les différencie n'est pas suffisamment significatif. Par exemple, une telle mesure peut être utilisée pour comparer *Montpellier* et *Montepllier* où la seule différence est l'interversion entre les lettres t et p : le score de la distance de Levenshtein est donc 1. Pour la comparaison de *Montpellier* avec la variante *Montpel*, la distance de Levenshtein vaut 3 puisqu'il y a une différence de trois lettres entre les deux

[48] Système d'information géographique. (2018, octobre 22). *Wikipédia, l'encyclopédie libre*. Page consultée le 15:44, octobre 22, 2018 à partir de http://fr.wikipedia.org/w/index.php?title=Syst%C3%A8me_d%27information_g%C3%A9ographique&oldid=153282415.

séquences. Ces scores ne mesurent que le nombre de différences constatées entre deux chaînes de caractères. Lors de leur intégration, l'enjeu est de définir les seuils qui permettent d'interpréter les scores obtenus.

2.4.1.4 Approche retenue

Les approches utilisées pour la REN ou l'identification des entités spatiales sont très variées. Les systèmes dédiés au traitement de l'écrit ne sont pas développés pour prendre en compte les spécificités de l'oral comme le montre l'expérience menée avec CasEN. Le travail présenté s'appuie sur l'exploitation d'un corpus d'enregistrement transcrit pour en extraire la perception qu'ont les Orléanais de leur ville. La nature des données et la finalité du projet commandent l'élaboration d'un nouveau système de détection des noms de lieux à même de renseigner les informations requises dans la convention d'annotation et indépendant d'autres outils afin de faciliter le passage à l'analyse d'opinion.

Le nouveau système s'inspire des approches décrites précédemment. Il s'efforce de prendre en compte les difficultés liées à la détection de lieux et celles inhérentes au traitement de l'oral. Ainsi s'appuie-t-il sur l'utilisation de ressources lexicales à la manière de Lesbeguerries (2007) et des cascades de CasEN. Des bases de données dédiées à l'information géographique sont utilisées pour identifier les noms officiels présents dans le corpus. Pour identifier ensuite les noms non-normalisés, des règles sont définies pour gérer les variations possibles qui posent déjà des difficultés à l'écrit.

2.4.2 Difficultés de la détection automatique des lieux

2.4.2.1 Problèmes généraux

D'une manière générale, la tâche de REN présente certaines difficultés. L'un des problèmes principaux réside dans la polysémie de certaines entités, dont les lieux font partie. Un exemple bien connu est celui du terme *orange* : désigne-t-il la couleur, le fruit, la ville du Sud de la France ou la compagnie de téléphonie française ? La présence ou l'absence d'une majuscule, ou le contexte d'emploi (*aller à Orange, souscrire chez*

Orange, *manger une orange*, *la chemise orange*) peuvent aider à faire la différence entre ces sens mais l'ambiguïté persiste souvent (*J'adore Orange ! Les oranges oranges*).

Parfois, les noms de lieux sont employés en tant que personne. On observe cet emploi dans les articles journalistiques : *La Maison Blanche a déclaré que…*, *Paris lance un appel à Washington*, etc. Dans ces cas-là, le contexte apporte de l'ambiguïté dans la tâche de typage des entités nommées. Dans *Paris a déclaré que…* : est-il question de la ville de Paris en tant que lieu, en tant qu'organisation ou bien en tant que personne prénommée Paris ? La convention Quaero recommande aux annotateurs de se référer au contexte de l'entité pour désambiguïser son type. Cependant, si le doute ne peut être levé, l'annotateur peut lister les types entre lesquels l'entité est ambiguë dans un attribut class-set attribué à une balise générale <entity>. Si l'annotateur ne peut pas déterminer si Paris est une ville, une organisation ou une personne, il notera :

```
<entity class-set="loc.adm.town org.adm pers.ind">
                                                    Paris
                                                </entity> a déclaré que…
```

Selon la convention d'annotation Quaero, cette option n'a jamais été utilisée par les annotateurs.

Une variation dans le nommage des lieux est une difficulté supplémentaire. Que ce soit à l'oral ou à l'écrit, la nature morphologique des mots désignant un lieu peut varier. Les noms de lieux peuvent être composés d'un ou de plusieurs mots (*place du Général de Gaulle*, *université d'Orléans*, etc.), être liés ou non par un trait d'union (*le quartier Orléans-La Source* ou *Orléans La Source*, *Saint-Jean-de-la-Ruelle*, etc.), etc. Le nommage des lieux répond à certaines conventions définies arbitrairement, ce qui n'empêche pas pour autant les locuteurs de faire varier cette norme. Ainsi, on peut indifféremment employer les noms *place de l'Arc de Triomphe* ou la *place de l'Étoile* pour se référer à la *place Charles De Gaulle* à Paris. Il n'existe qu'une seule place de ce genre donc cela reste possible pour un système automatisé de déterminer qu'il s'agit du même endroit et d'attribuer les coordonnées géographiques correspondantes. Cependant, à quelle rue fait-on référence si l'on évoque la *rue de l'église* ou la *rue de la gare* lorsque l'on se trouve dans une ville disposant de plusieurs églises ou gares ? Dans ces cas-là, si le contexte n'est pas

suffisamment explicite, il devient très complexe, voire impossible dans certains cas de relier ces expressions à un emplacement exact. Matérialiser le portrait de la ville d'Orléans en géolocalisant les lieux identifiés dans le corpus ESLO suppose de résoudre cette problématique de variation des noms de lieux. Que la mention de lieu soit conventionnelle ou non, il faut dans tous les cas pouvoir associer ces mentions avec les coordonnées géographiques correspondantes.

Le repérage automatique est complexe, plus encore dans le cas des corpus oraux. Comme démontré par Brando et. al, (2016), ce type de corpus peut être riche en entités nommées mais surtout en mentions de lieux génériques (ex : *un endroit très beau, le long de la grande rue*). Pour reconnaître ces types de lieux qui se rapprochent des ESR, les systèmes de REN existants ne sont pas adaptés à l'oral.

2.4.2.2 Les difficultés inhérentes au traitement de l'oral

Eshkol-Taravella (2015) rappelle que l'écrit « *se présente au destinataire comme un produit final alors que l'oral est un produit en cours d'élaboration* ». L'une des caractéristiques principales de l'oral est la présence de disfluences (Blanche-Benveniste *et al.* 1990 ; Riegel *et al.* 1994). Les disfluences sont les « *marques typiques des énoncés en cours d'élaboration* » (Dister, 2007). Ces éléments rompent le flux de parole de façon aléatoire et révèlent les traces de construction du discours en temps réel (Blanche-Benveniste *et al.* 1990).

Parmi les disfluences, on peut retrouver des marqueurs discursifs (*quoi*, *enfin*, …), des onomatopées ou interjections (*ah, euh, hum*, …), des répétitions[49], des reformulations[50], ou encore des amorces de mots[51].

(12) vous êtes obligée déjà **de de** traverser la Loire (ESLO2_ENT_1009)

(13) et y a pas de questions pièges enfin je veux dire c'est une discussion normale euh ne t'inquiète pas (ESLO2_ENTJEUN_04)

[49] cf. exemple 12
[50] cf. exemple 13
[51] cf. exemple 14

(14) 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)

Dans l'exemple 14, l'expression *cons-* est l'amorce du lieu conservatoire. A cet instant, le locuteur hésite comme en témoigne l'onomatopée *euh*, puis, il se reprend pour donner la forme complète *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 14, 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. A l'oral, il est extrêmement rare que quelqu'un fasse allusion à la cathédrale en utilisant son nom complet. Son nom officiel est : Cathédrale Sainte-Croix. Dans l'exemple 14, le locuteur utilise les éléments du nom complet mais les donne dans le désordre. Il y fait allusion d'abord par le nom propre, et précise ensuite le type du monument.

La relation entre une personne et un lieu rend possible l'appropriation, la personnalisation d'un lieu par un individu (Dominguès & Eshkol-Taravella, 2015) comme dans l'exemple 15.

(15) on regardait **notre cathédrale** tous les deux (ESLO2_iti_07_01)

Le locuteur emploi le déterminant possessif *notre* pour se référer à la *cathédrale*. Cette énonciation marque un attachement particulier à l'objet, en l'occurrence la cathédrale, qu'il observe. Avec l'utilisation de *notre*, c'est comme si la cathédrale leur appartenait. Le possessif est un indice de la perception cet habitant par rapport la ville d'Orléans.

Les noms de lieux peuvent aussi être tronqués ou abrégés. Un mot tronqué est un mot dans lequel on a supprimé une ou plusieurs syllabes. Dans l'exemple 7, dans le nom de la *rue de la République*, le mot *République* a été tronqué en *Rép-* suite à la suppression des syllabes *ublique*. Dans l'exemple 6 par contre, *La Ferté* est la version abrégée de *La Ferté-Saint-*

Aubin. L'abréviation consiste ici en la suppression d'un ou plusieurs mots dans une expression de lieux composée.

Des surnoms peuvent être substitués au nom conventionnel. Ils sont porteurs d'indices sur la perception du locuteur et source d'erreurs au moment de la détection automatique des différentes mentions. Par exemple, dans :

(16) 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 hyperonyme pour se référer à son environnement. Dans l'exemple suivant :

(17) 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)

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 donné au moment de son inauguration en 1763. Nombre d'Orléanais continuent à se référer à ce pont en utilisant son ancien nom.

(18) *en gros euh sous* ***les*** **Arcades** (ESLO2_ENTJEUN_04)

Enfin, *les Arcades* correspond à la *rue Royale,* une artère d'Orléans[52], bordée sur toute sa longueur par des galeries à arcades. Cette spécificité architecturale a conduit les Orléanais à se référer à cette rue en substituant son nom officiel par une appellation plus imagée. On observe une véritable réappropriation nominale du lieu.

[52] cf. exemple 18

Un enjeu pour le TAL est de reconnaitre un lieu mentionné en discours de différentes manières : par le nom officiel, le nom ancien ou un surnom. La difficulté principale dans le repérage automatique est de tenir compte de toutes les variations possibles de la dénomination et être capable de détecter la coréférence entre plusieurs mentions du même lieu pour la détection automatique de lieux dans les transcriptions, les noms de lieux tronqués, généralisés, personnalisés, et les disfluences doivent être inclus dans la chaîne de traitement proposée.

2.4.3 Rappel méthodologique pour l'annotation des lieux dans l'oral transcrit

L'annotation automatique des lieux dans l'oral transcrit se fait en plusieurs étapes[53] décrites dans la section 2.1. Chacune prend en compte les problèmes soulevés par l'extraction des lieux à l'oral. Le système répond aux exigences de la convention d'annotation établie pour la construction de la carte finale de la perception de la ville d'Orléans qui comporte trois informations : le type de lieu, la zone géographique et le label officiel. La finalité du projet guide le développement de l'outil.

[53] cf. Figure 8

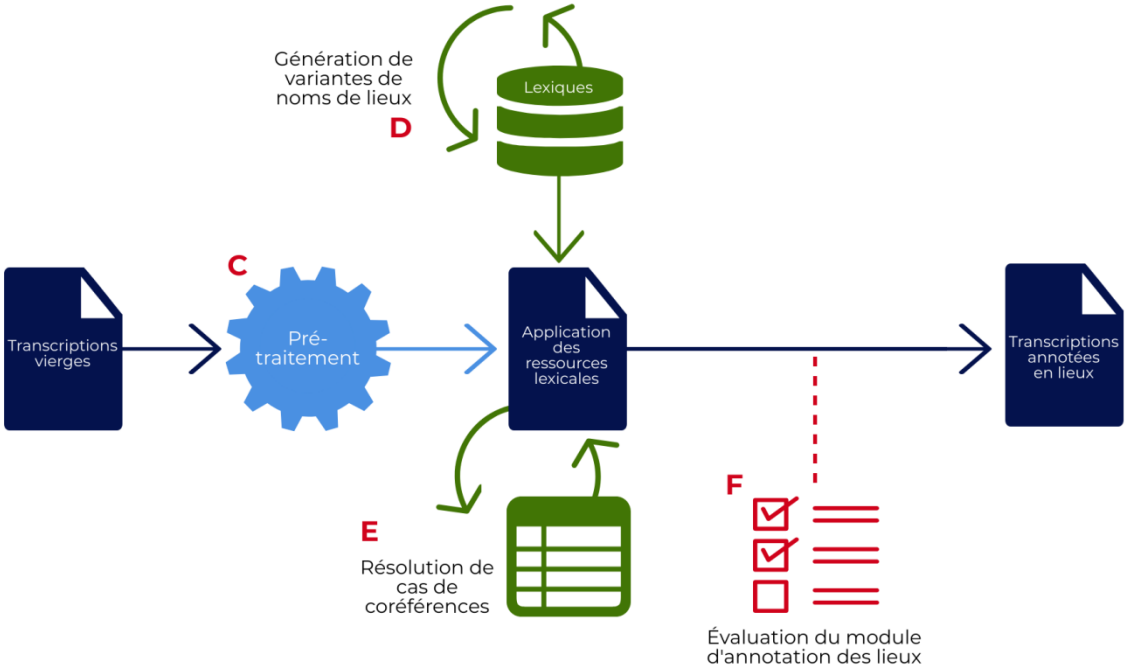

Figure 8 : Extrait de la méthodologie générale pour l'annotation des lieux

Une première étape de prétraitement (C) met en forme les données pour l'analyse et s'efforce de résoudre, au moins en partie, la question des disfluences. Pour les noms de lieux abrégés, tronqués ou remplacés par un surnom, des ressources lexicales sont collectées à partir de base de données géographiques et enrichies en nouvelles mentions de lieux (D). Dans une approche symbolique, des règles sont appliquées sur le corpus pour contribuer à l'identification de noms de lieux non normalisés. Lorsqu'un lieu est identifié, mais que sa mention est inconnue des bases de données parce que le nom n'est pas normalisé, celui-ci est comparé avec les détections précédemment effectuées. Grâce à des règles et des mesures de similarité, le système peut faire des liens de coréférence et retrouver le label officiel du lieu à identifier (E). Enfin, le module d'annotation est évalué par comparaison avec le corpus de référence. Chacune des étapes nécessaires à l'identification des noms de lieux, normalisés ou non, est décrite dans la section suivante.

2.4.4 Etapes du traitement

2.4.4.1 Prétraitement

Afin de permettre la détection des mentions de lieux contenues dans les transcriptions du corpus ESLO, les fichiers font l'objet de deux types de prétraitements : au niveau de la structure de la transcription en elle-même et au niveau des disfluences.

Les transcriptions du corpus ESLO2 se présentent sous la forme de fichier .trs structuré au format XML. La figure 9 présente un extrait d'une transcription. Dans celui-ci, la balise <Turn> délimite les tours de parole. Lorsque deux locuteurs parlent en même temps, les tours de parole se superposent. Les chevauchements apparaissent dans les balises <Who>. L'alignement d'un enregistrement sonore et de la transcription correspondante est garanti par les balises temporelles <Sync>. La plupart des métadonnées relatives à la transcription sont stockées dans l'arborescence du fichier (nom du transcripteur, date de l'enregistrement, etc.). Toutes ces métadonnées ne sont pas utiles à cette étape de la détection des lieux.

```
<Turn speaker="spk1" startTime="11.37" endTime="15.072">
<Sync time="11.37"/>
depuis combien de temps est-ce que vous êtes originaire d'Orléans
ou de sa région ?
</Turn>
<Turn speaker="spk1 spk2" startTime="15.072" endTime="15.871">
<Sync time="15.072"/>
<Who nb="1"/>
ou
<Who nb="2"/>
hm hm
</Turn>
```

Figure 9 : Extrait d'une transcription au format .trs

Afin de simplifier les documents analysés, nous avons décidé de modifier le format du fichier pour ne conserver que les informations pertinentes à ce stade. Le contenu des tours de parole, le code du locuteur et le code temporel de chaque tour. Toutes les autres sont déportées et stockées pour une utilisation ultérieure.

```
[ch_CD2]  **11.37 :** depuis combien de temps est-ce que vous
êtes originaire d'Orléans ou de sa région ?
[ch_CD2 et RL2]  **15.072 : ** ou  hm hm
```

Figure 10 : Format simplifié de la transcription

La figure 10 présente l'extrait de la figure 9 dans une version simplifiée : chaque ligne équivaut à un seul tour de parole. Dans cette forme simplifiée, chaque ligne commence par l'identifiant du locuteur (*[ch_CD2]*), suivi par le code temporel (***11.37 :***) et par le contenu du tour de parole (*depuis combien de temps est-ce que vous êtes originaire d'Orléans ou de sa région ?*).

En cas de chevauchement, les tours de parole ne sont pas distingués. Ils apparaissent sur la même ligne. Les codes des deux locuteurs apparaissent en début de ligne (*[ch_CD2 et RL2]*). Ce traitement est appliqué seulement au moment de l'annotation des lieux et à l'étape suivante, l'analyse d'opinion. Les chevauchements de parole apparaîtront dans la carte finale représentant la perception de la ville d'Orléans afin de rester au plus près de l'enregistrement où est extraite l'opinion identifiée. Comme pour les tours de parole avec un locuteur unique, le code des deux locuteurs est suivi du code temporel (***15.072 :***) et de la concaténation des deux tours de paroles qui se chevauchent (*ou hm hm*)

Comme nous l'avons vu, les disfluences sont des caractéristiques spécifiques de l'oral. Les onomatopées ou les interjections[54] peuvent apparaître à l'intérieur du nom d'un lieu. Par exemple dans :

(19) b- c'est la grande région euh c'est la grande région **euh** Centre (ESLO2_ENT_1034)

où *euh* entrecoupe *grande région Centre*. Ce type de disfluences est supprimé puis replacé à la fin de l'étape d'annotation des lieux.

2.4.4.2 Ressources lexicales utilisées

Dans la lignée des travaux exploitant des approches symboliques en REN (Lesbeguerries 2007 ; Maurel et al, 2011 ; Dominguès et Eshkol-Taravella, 2013, 2015) des ressources

[54] Liste des onomatopées : ah,oh,aha,atchoum,aïe,baf,bah,bam,bap,ben,beuh,bim,bla,bloum,boah,bof,boh,bouh,boum,bé,cha,chpoum,chtac,clac,clou,ding,euh,gla,gna,gnin,hein,hop,hou,hu,hum,hé,la,leu,lin,mah,maille,miam,mouais,moui,of,ouah,ouais,ouf,ouille,oïe,pam,pff,pim,piouc,plaf,ploum,pof,poh,pop,pouf,pouh,poum,snif,teu,vouf,vouh,voum,vroum,wahou,ya,yeah,yoh,zut

lexicales ont été constituées afin de guider l'annotation des lieux. Trois bases de données ont été utilisées : Geonames[55], GEOFLA[56] et Data.gouv.fr[57].

- **[GEONAMES] :** base de données collaboratives géographiques distribuée gratuitement sous licence Creative Commons. Ce sont les utilisateurs de la base qui ajoutent des données, les améliorent ou les corrigent.
- **[GEOFLA] :** base de données mise à disposition librement par l'Institut National de l'Information Géographique et Forestière (IGN). Elle est dédiée aux applications de géomarketing et de cartographie statistiques et thématiques. Elle permet, à des échelles nationales et régionales, de situer toute information thématique, d'analyser des données statistiques et de gérer des déplacements routiers. Cette base contient la liste de toutes les zones correspondent à un découpage administratif en France.
- **[DATA.GOUV.FR] :** portail de données ouvertes gouvernementales françaises lancé en décembre 2011. L'ouverture des données d'intérêt public vise à encourager la réutilisation des données au-delà de leur utilisation première par l'administration. La plateforme data.gouv.fr recense les jeux de données accessibles ainsi que les réutilisations qui en sont faites.

Chacune de ces bases de données référencie des noms de lieux normalisés, associés à des coordonnées géographiques permettant une géolocalisation sur une carte. La base Géonames est internationale. Elle permet de récupérer les noms de villes, de terroirs en France et à l'étranger. Cependant, les informations propres à la France contenue dans la base ne sont pas suffisamment précises. La base GEOFLA pallie ce manque en fournissant la liste de toutes les régions, villes, cantons et départements du territoire français. Enfin la plateforme Data.gouv.fr permet d'obtenir les informations géographiques spécifiques à Orléans. Grâce à cette plateforme, nous avons accès à la liste des rues d'Orléans, des zones piétonnes, des associations et des entreprises ayant un numéro SIREN dans la métropole orléanaise. Tout comme GEOFLA et Géonames, chacune des entrées de la base est associée

[55] https://www.geonames.org/
[56] http://professionnels.ign.fr/geofla
[57] https://www.data.gouv.fr/fr/

à des coordonnées géographiques. Celles-ci sont nécessaires pour la constitution de la carte finale.

Ces bases de données géographiques sont utilisées comme des lexiques pour participer à l'annotation des lieux mentionnés dans le corpus. 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. Les ressources lexicales sont donc enrichies en variantes de noms de lieux qui permettent une annotation plus exhaustive et analogue aux habitudes de nommage des lieux par les locuteurs.

2.4.4.3 Traitement des abréviations

2.4.4.3.1 Génération d'abréviations

Les noms de lieux divergents de la norme sont nombreux à l'oral. L'abréviation d'un nom de lieu est un phénomène récurent à l'oral. L'abréviation consiste en le raccourcissement d'un groupe de mots par la suppression d'un ou plusieurs mots le composant. Les lieux composés de plusieurs mots, peuvent être aussi abrégés dans le discours oral [58]. L'observation de mentions de lieux dans le corpus a permis de constater quelques cas d'abréviations récurrents au niveau des noms de villes et de voies.

Une récurrence apparaît dans la manière d'abréger les noms de voie. Dans tous les cas observés, 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.

(20) on voit l'état de la la **place De Gaulle** aussi euh (ESLO2_ENT_1031)

[58] cf. exemple 6 : *La Ferté* pour *La Ferté Saint Aubin*

Dans l'exemple 20, 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 par le locuteur. Cette information a été jugée facultative par le locuteur. Il sait qu'il peut supprimer ces éléments du nom du lieu et tout de même se faire comprendre par son interlocuteur.

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*).

A partir de ces constatations, des variantes de noms de voies ont été générées. Tous les noms de voies référencés dans les ressources lexicales ont été abrégés pour créer la liste de toutes les formes possibles de voies après la suppression de termes entre le mot type et le dernier mot du nom de la voie. De cette manière, un mot est supprimé dans le nom officiel de la voie pour créer une nouvelle variante. La suppression est effectuée jusqu'à ce qu'il ne reste que le dernier mot ou groupe prépositionnel et le terme caractérisant le nom de la voie. Le tableau 13 présente des exemples de variantes de noms de voies. Ainsi, la *place du Cheval Rouge* pourrait être abrégée comme *Place Cheval Rouge* ou même *Place Rouge*. La *place du Marché de la Madeleine* pourrait devenir par exemple la *place de la Madeleine* ou la *place Madeleine*.

Type	Nom Officiel	Variante 1	Variante 2	Variante 3	Variante 4
place	Place du Commerce	Place Commerce			
place	Place du Cheval Rouge	Place Cheval Rouge	Place Rouge		
place	Place du Jardin des Plantes	Place Jardin des Plantes	Place des Plantes	Place Plantes	
place	Place du Marché de la Madeleine	Place Marché de la Madeleine	Place de la Madeleine	Place la Madeleine	Place Madeleine

Tableau 13 : Génération d'abréviations de noms de voies

Ce procédé permet l'identification du nom de lieu abrégé *place De Gaulle* et surtout, de le lier à son label officiel : *Place du Général de Gaulle*. Au-delà de l'enjeu de la détection des

formes non-normalisées des noms de lieu, cette méthode permet aussi de relier ces formes divergentes avec leurs formes conventionnelles.

Il existe une exception : celle des rues comportant le mot faubourg. Un faubourg est une ancienne dénomination pour désigner des quartiers entourant une ville. Aujourd'hui, cette expression seule est désuète et les seules traces qu'il en reste sont 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. Lorsqu'un nom de voie contient le terme faubourg, les mots le précédent sont supprimés. Ainsi, pour la *rue du Faubourg Saint-Jean* et la *rue du Faubourg Madeleine*, les variantes *faubourg Saint-Jean* et *faubourg Madeleine* sont respectivement ajoutées comme nouvelles entrées dans les ressources lexicales.

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. Si pour les voies, c'est le ou les derniers termes qui sont conservés et les autres potentiellement supprimés, le processus est inverse pour les noms de ville. Dans le cas des noms de ville, ce sont le ou les premiers termes qui sont conservés, et ce sont les derniers qui peuvent être supprimés. On peut l'observer dans l'exemple 6 avec *La Ferté-Saint-Aubin* abrégée en *La Ferté*. On aurait pu trouver *Rio* pour *Rio de Janeiro*, *Aix*, pour *Aix-en-Provence*, *Sully*, pour *Sully-sur-Loire*, *Saint-Cyr* pour *Saint-Cyr-en-Val*. La règle de génération de variante de noms de ville consiste donc en la suppression des derniers termes ou groupes prépositionnels composants le nom.

2.4.4.3.2 Application des ressources lexicales enrichies en abréviations

Les ressources lexicales extraites des bases de données contenant les noms de lieux ont été enrichies en nouvelles mentions. La projection de l'ensemble du lexique constitué sur les

tours de parole du corpus est une opération très lourde pour le système d'annotation. Aussi, quelques règles ont été définies pour déclencher cette application des ressources lexicales.

Le déclencheur principal de l'application des ressources lexicales est la majuscule. Tous les noms propres sont transcrits avec une majuscule selon les conventions d'ESLO. Une grande majorité de noms de lieux sont des noms propres. Ces mots constituent donc de bons candidats pour la détection des lieux. Lorsqu'une majuscule est détectée, on n'applique pas toutes les ressources lexicales mais seulement celles qui commencent par une majuscule. De cette façon, le système peut identifier les villes, départements, cantons, etc., sous leur nom normalisé ou abrégé grâce à l'enrichissement des ressources lexicales en abréviations.

En ce qui concerne les noms de voies, le système s'appuie sur la liste des déclencheurs :

> allée, ancienne, avenue, belle, boulevard, chemin, cité, cloitre, clos, cour, esplanade, faubourg, galerie, grande, guichet, halles, impasse, jardin, levee, parc, passage, place, pont, quai, quartier, résidence, route, rue, ruelle, sentier, square, venelle, voies.

Cette liste a été constituée à partir des données extraites de Data.gouv.fr et plus particulièrement de la liste des noms de voies d'Orléans. Les termes caractérisant le type de voies sont toujours placés en tête du nom de la voie. La liste des noms de type de voies est ainsi constituée des premiers mots identifiés dans les noms de rue. Le mot *faubourg* a été ajouté pour répondre à l'ambiguïté soulevée dans le point 2.4.4.3.1. et permettre l'identification par exemple de *faubourg Saint-Jean* ou *faubourg Madeleine*. Si l'un de ces mots est présent dans le tour de parole, on considère qu'il pourrait introduire le nom d'une voie et on procède à l'application des ressources lexicales correspondantes.

(21) oh [pf] mon lieu préféré dans la ville alors dans les questions euh c'est la **rue de la République** je pense (ESLO2_iti_12_01)

Dans l'exemple 21, le système doit permettre la détection de la *rue de la République*. Le déclencheur *rue* est identifié et enclenche l'application de la liste des rues d'Orléans. La rue de la République étant listée dans les ressources lexicales, la mention repérée est annotée par le système de la manière suivante :

(21) oh [pf] mon lieu préféré dans la ville alors dans les questions euh c'est la <loc type="voie" zone="2" label="rue de la République">**rue de la République**<loc> je pense (ESLO2_iti_12_01)

La *rue de la République* est une *voie*, située à Orléans, donc dans la zone *2* et son nom normalisé est aussi *rue de la République*.

L'application de ce nouveau lexique enrichi en abréviations ne suffit pas pour autant à identifier l'ensemble des lieux présents dans le corpus. De plus, si aucun lieu n'est détecté à partir des ressources lexicales alors qu'un déclencheur avait été identifié, cela ne signifie pas encore que le tour de parole observé ne contient pas de nom de lieu. En effet, le tour de parole peut contenir un nom de lieu non-normalisé. Pour atteindre une plus grande exhaustivité dans l'annotation et identifier les noms de lieux divergents de la norme, des patrons sont établis.

2.4.4.4 Traitement des 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 de découvrir de nouvelles mentions dans le corpus, l'enjeu principal autour de la détection des noms de lieux tronqués est de pouvoir relier la forme tronquée avec 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 lieux tronqués des patrons sont établis. Un patron sert à décrire le contexte morphosyntaxique de l'objet à identifier. En l'occurrence, les patrons définis ici décrivent le contexte d'apparition ou la forme des mentions de lieux. Le module d'annotation utilise donc ces patrons pour analyser les tours de parole les uns après les autres.

Les patrons permettant la détection d'un lieu tronqué 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 autre déclencheur éventuellement identifié dans le tour de parole jusqu'au tiret. Dans le cas où il n'y a pas de mot déclencheur comme pour les voies, le point de départ de la fenêtre d'observation peut être la majuscule précédant le tiret. S'il n'y a ni mot déclencheur, ni majuscule avant, la fenêtre d'observation s'étend depuis le premier mot du tour de parole jusqu'au tiret. Dans l'exemple 6, la mention *rue de la Rép-* est la version tronquée de la *rue de la République*.

(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 **rue de la Rép-** tu vois où elle est ? la **rue de la République** ? (ESLO2_iti_06_11)

L'étape d'application des ressources lexicales ne permet l'identification que de la version normalisée de la *rue de la République*. Dans cet exemple se trouve deux autres occurrences du terme rue. Celles-ci peuvent être le point de départ du nom d'un autre lieu qui serait absent des ressources lexicales. De plus, on constate qu'un tiret se trouve après la deuxième occurrence : on peut donc construire une fenêtre d'observation pour tenter d'identifier le nom tronqué d'une voie. La fenêtre d'observation s'étend donc de la deuxième occurrence du mot *rue* jusqu'au mot tronqué *Rép-*, soit : *rue de la Rép-*.

A cause de la présence d'un tiret, l'hypothèse est que la fenêtre d'observation contient un nom de lieu tronqué, c'est-à-dire le nom normalisé d'un lieu dont on a supprimé les dernières syllabes. Pour retrouver le nom normalisé d'origine, les ressources lexicales sont artificiellement tronquées. On va essayer de reproduire la troncation observée dans la fenêtre d'observation sur le lexique et voir si l'on retrouve deux séquences identiques.

Pour ce faire, on commence par calculer la taille de la fenêtre d'observation en nombre de caractère sans inclure le tiret. La séquence *rue de la Rép-* a une longueur de treize

caractères. Cette information implique que le nom normalisé de la rue doit dépasser les treize caractères. Ainsi, la troncation artificielle est réalisée sur tous les noms de rue et leurs variantes dépassant les treize caractères. Toutes les séquences correspondantes sont tronquées à partir du treizième caractère[59]. Les noms de voies ainsi tronqués peuvent être appliqués sur la fenêtre d'observation.

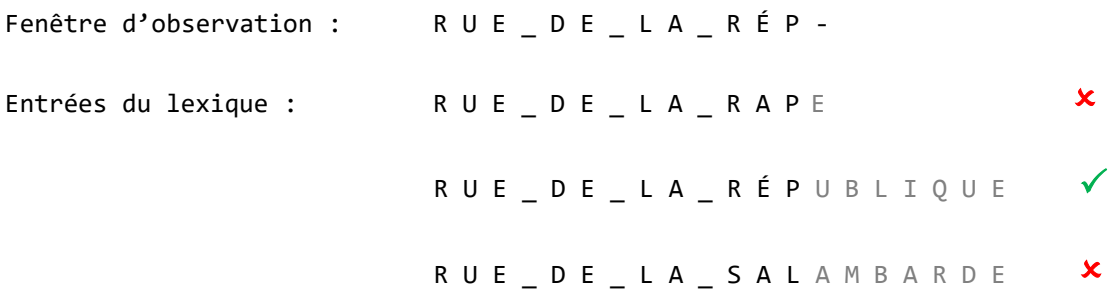

Figure 11 : Réduction de la fenêtre d'observation pour la détection d'une voie

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-*.

Le principe de troncation artificielle est utilisé de la même façon pour annoter tous les autres types de lieu. Dans l'exemple 22, il est question du restaurant Mc Donald's[60], mentionné sous la forme tronquée *Mc Do-*.

(22) je sais pas je [pf] non bah si on va plus souvent au shawarma ou au **Mc Do-** que euh dans un vrai resto (ESLO2_ENT_1001_C)

Le système détecte dans cet exemple un tiret, ce qui déclenche la création d'une fenêtre d'observation permettant de déterminer si le mot tronqué est un nom de lieu ou non. Le tour de parole ne comporte pas de mot déclencheur et le terme tronqué commence par une

[59] cf. Figure 11
[60] Comme il a été indiqué plus haut (cf. 2.2.1), ce travail ne distingue pas les lieux des organisations.

majuscule : la fenêtre d'observation s'étend donc de cette majuscule jusqu'au tiret, soit : *Mc Do-*. La longueur en caractères de la fenêtre d'observation est calculée. A partir de celle-ci, une troncation artificielle est réalisée sur toutes les entrées des ressources lexicales commençant par une majuscule et d'une longueur supérieure à celle de la longueur de la fenêtre d'observation. En l'occurrence, une seule correspondance existe dans l'ensemble des ressources lexicales interrogées : *Mc Donald's*, dans la liste des commerces de la ville. La démarche est la même pour l'exemple 23 :

(23) à côté y a Gifi euh y a Casto- euh ce genre de choses donc ouais c'est très varié (ESLO2_ENTJEUN_09_C)

La troncation artificielle des ressources lexicales permet de relier la forme tronquée *Casto-* à la forme entière *Castorama*. De la même façon que dans le traitement de l'exemple 24, on ne trouve qu'une seule correspondance entre la forme tronquée contenue dans la fenêtre d'observation et les ressources lexicales. Mais que se passe t-il si plusieurs entrées dans les ressources lexicales correspondent à la séquence présentée dans la fenêtre d'observation ? Comment déterminer quelle entrée des ressources lexicales est la forme originale de la version tronquée ?

Dans l'exemple 24, le locuteur s'interrompt avant de prononcer le nom complet de la ville Olivet. Le système détecte dans ce tour de parole la présence d'un tiret. A partir de ce tiret, il crée la fenêtre d'observation devant déterminer si le mot tronqué est un nom de lieu ou non. En l'occurrence, la fenêtre d'observation est : *Oliv-*.

(24) ouais Olivet les bords d'Oliv- euh les les moulins d'Olivet (ESLO2_iti_05_04)

A partir de la fenêtre d'observation établie, les ressources lexicales sont tronquées artificiellement[61]. Cette fois, deux entrées, *Olivese* et *Olivet*, correspondent à la séquence observée une fois tronquées.

[61] cf. Figure 12

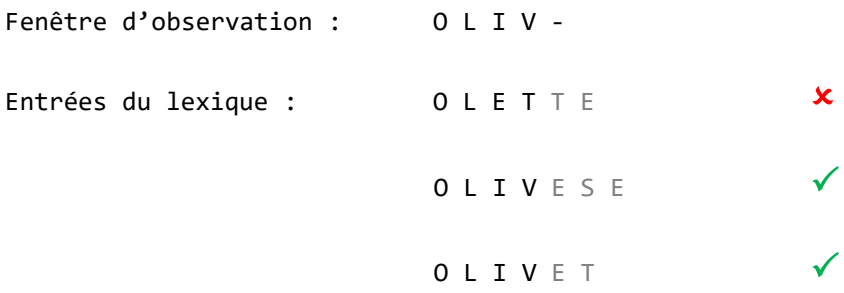

Figure 12 : Réduction de la fenêtre d'observation pour la détection d'une voie

En l'état, le système ne peut pas déterminer si le locuteur se réfère à l'une ou l'autre ville. Ce problème existe aussi pour les noms de lieux abrégés. Que ce soit pour les noms tronqués ou abrégés, la question est de savoir comment retrouver le nom d'origine d'un lieu lorsque plusieurs entrées dans les ressources lexicales peuvent correspondre. Pour résoudre cette question, il est nécessaire de créer des patrons supplémentaires utilisés dans le module de traitement des coréférences présentées dans le chapitre suivant.

2.4.4.5 Traitement partiel 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 (Grobol et. al, 2018 ; Muzerelle *et al.*, 2014 ; Desoyer *et al.*, 2015) ou DEMOCRAT[62] (Landragin, 2016 ; Landragin et. al, 2018 ; Grobol et. al, 2019) étudient les chaînes de coréférences et travaillent à l'élaboration de systèmes automatisés dédiés à leur identification[63].

Le plus souvent, les coréférents sont des pronoms comme dans l'exemple :

[62] http://www.lattice.cnrs.fr/Projet-ANR-DEMOCRAT
[63] Mais aucun de ces outils n'est disponible pour l'instant pour pouvoir le tester sur nos données.

(25) 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 (ESLO2_ENT_1010)

Dans cet exemple il est question de la *rue de Bourgogne*. Le locuteur mentionne ce lieu évoqué précédemment avec un nom générique *cette rue-là*. Le pronom *elle* fait lui aussi référence à cette rue. La question de la résolution de coréférences est une problématique à part entière dans le domaine du TAL. Cette question s'éloigne de l'objet de notre étude, aussi, nous avons fait le choix de ne pas approfondir le traitement des coréférences. Ainsi, les cas concernant des pronoms, souvent sujets à ambiguïtés, ne sont pas traités par notre module.

Cependant, 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*.

Il existe tout de même des cas ambigus dans lesquels il est difficile de retrouver le nom officiel d'un lieu.

(26) 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 (ESLO2_ENT_1003)

Dans l'exemple 26, trois villes sont mentionnées : *Saint-Jean-de-la-Ruelle*, *Ingré* et *Saint-Jean*. La ville *Saint-Jean-de-la-Ruelle* est d'abord mentionnée via son nom conventionnel puis sous la forme abrégée *Saint-Jean*. Comme illustré dans la figure 13, les villes *Saint-Jean-de-la-Ruelle* et *Ingré* sont d'abord reconnues par le système grâce à l'application des ressources lexicales. Chacune de ces deux villes sont donc annotées dans le tour de parole. Elles sont par la suite extraites pour être stockées avec leurs attributs dans une liste répertoriant les différents lieux annotés dans l'ensemble de la conversation, dans leur ordre d'apparition.

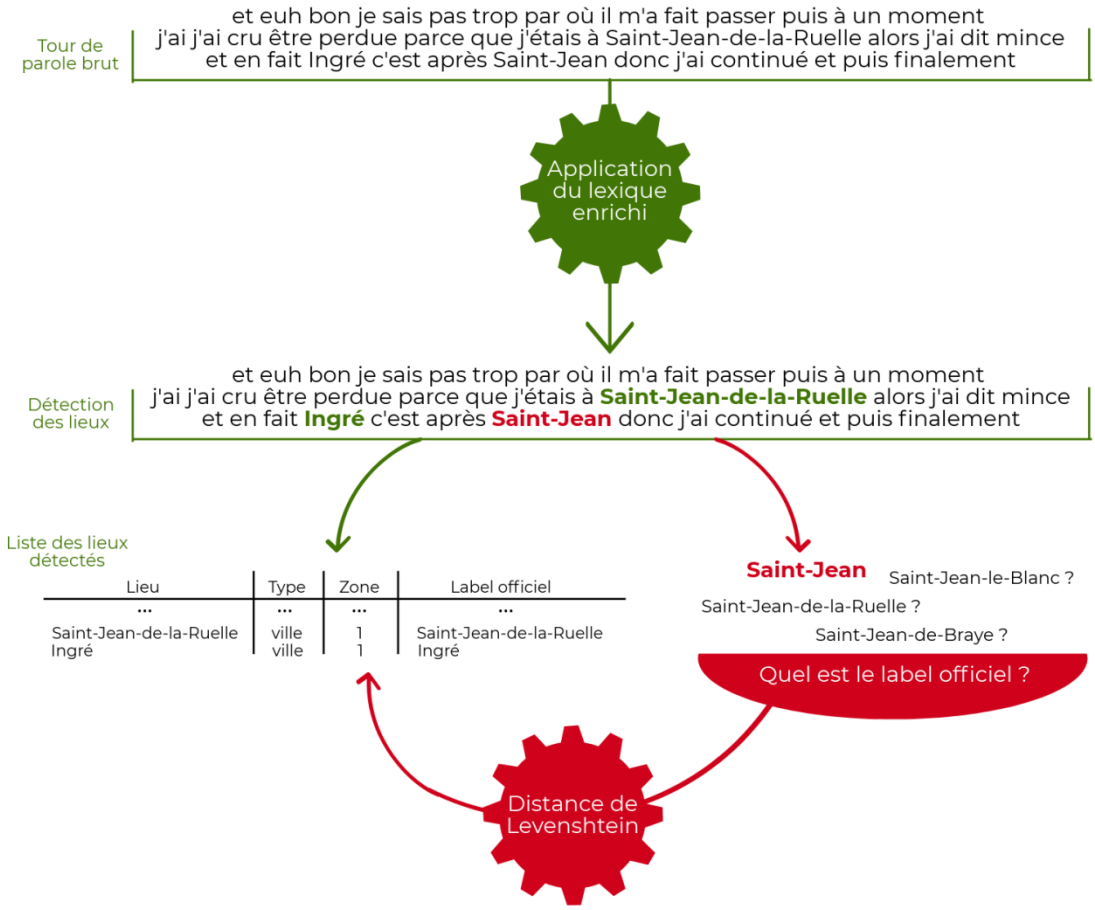

Figure 13 : Résolution de coréférence dans un tour de parole

De la même façon, la forme abrégée *Saint-Jean* est présente dans les ressources lexicales et est aussi reconnue par le système. Le problème est que dans ces ressources, il existe 172 noms de ville pouvant être abrégés en *Saint-Jean*. Etant donné le contexte d'énonciation des enregistrements qui poussent les locuteurs à parler d'Orléans, on pourrait supposer que les locuteurs parlent de lieux proches de la région orléanaise. Une telle supposition permettrait de réduire la recherche aux villes situées dans le même département : le Loiret (45). Néanmoins, cette hypothèse est réductrice puisque même s'ils y sont incités, les locuteurs ne sont pas forcés de ne parler que d'Orléans. De plus, dans le cas de l'abréviation *Saint-Jean*, quel serait le label officiel à lui attribuer en sachant que les villes *Saint-Jean-de-la-Ruelle*, *Saint-Jean-de-Braye* et *Saint-Jean-le-Blanc* font partie de l'agglomération orléanaise ?

Le raccourcissement du nom de la ville crée de l'ambiguïté et seul le contexte de la conversation en cours peut résoudre ce problème. L'hypothèse que nous émettons dans ce

travail est que le locuteur mentionne le nom du lieu d'une manière complète la première fois qu'il s'y réfère. Dans l'exemple 26, on voit bien que le locuteur commence par parler de la ville *Saint-Jean-de-la-Ruelle* avant d'utiliser la forme abrégée *Saint-Jean*. La forme conventionnelle du lieu est employée dans le même tour de parole que sa version abrégée, mais elle aurait très bien pu apparaître plus en amont de la conversation.

Pour résoudre ces cas de coréférences, le système va compter sur la volonté du locuteur d'être compris par son interlocuteur. Le locuteur va faire en sorte de désambiguïser son discours en donnant la forme conventionnelle du lieu avant d'utiliser une abréviation. Dès lors, si 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 le fichier pour retrouver sa première mention complète et établir un lien de coréférence. La liste des lieux déjà identifiés est utilisée comme une nouvelle ressource lexicale avec laquelle sont comparées les différentes correspondances.

Mais comment déterminer quel lieu précédemment mentionné est le label officiel recherché ? Il s'agit ici d'être en mesure d'évaluer la similarité entre deux chaînes de caractères. Pour cela, on utilise la distance de Levenshtein (1965). Cette distance mathématique permet de mesurer la similarité entre deux chaînes de caractères en comptabilisant le nombre d'ajout, de suppression ou d'interversion de caractères. La distance de Levenshtein vaut la somme des différences constatées entre les deux chaînes de caractères comparées. Dans sa forme classique, cette distance fait des comparaisons caractère par caractère.

Ainsi, dans l'exemple 26, pour retrouver le label officiel de *Saint-Jean*, une première comparaison est effectuée avec le dernier lieu identifié, soit *Ingré*. La distance de Levenshtein est utilisée pour comparer ces deux séquences[64].

[64] cf. Figure 14

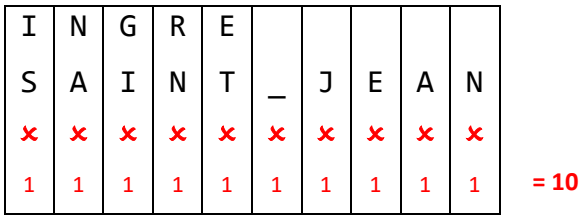

Figure 14 : Application de la distance de Levenshtein – Ingré & Saint-Jeazn

La séquence *Ingré* n'est composée que d'un seul mot alors que la séquence *Saint-Jean* en comporte deux. Dans le cas où la nouvelle mention dont on essaye de retrouver la référence (*Saint-Jean*) est composée de plus d'un mot, on procède à une application de la distance de Levenshtein en comparant mot par mot. Ainsi, on détermine quels mots sont différents ou identiques aux autres dans les deux mentions de lieux. Plus il y a de mots, plus le seuil sera grand car il y a plus de chance d'y avoir des différences. Au contraire, moins les expressions contiennent de mots, moins il y a de chance qu'il y ait une différence entre les deux expressions donc le seuil sera bas. Pour ce travail, le seuil correspond au nombre de mots composant le nom du lieu précédemment identifié (*Ingré* ou *Saint-Jean-de-la-Ruelle*) et qui est comparé avec la nouvelle mention dont on essaye de retrouver la référence (*Saint-Jean*).

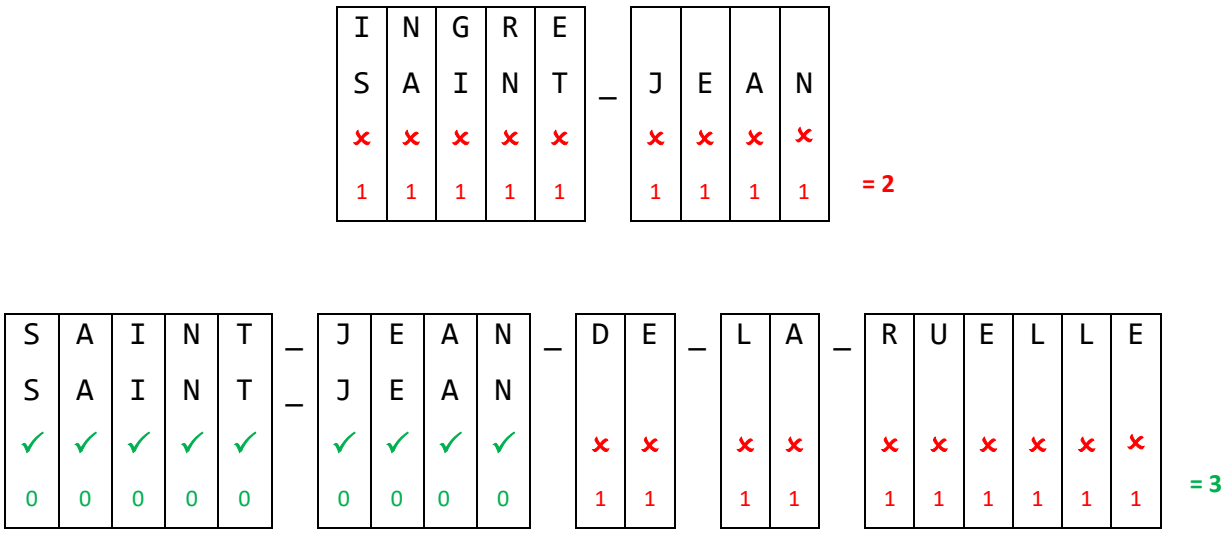

Figure 15 : Application de la distance de Levenshtein par mots

La comparaison de *Ingré* et *Saint-Jean* donne une distance de 2 : un mot manque et un autre est différent de celui avec qui il est comparait[65].

Dans cet exemple, le seuil à ne pas dépasser étant 1, *Ingré* est exclue comme forme officielle de *Saint-Jean*. Désormais, la comparaison entre Saint-Jean-de-la-Ruelle et Saint-Jean ne vaut plus que 3 : les deux premiers mots sont identiques, et les 3 mots suivants sont manquants. Ce score est inférieur à 5, le seuil établi à partir de la longueur en mots de la séquence Saint-Jean-de-la-Ruelle. Saint-Jean-de-la-Ruelle est donc validée comme nom officiel de Saint-Jean qui peut être annoté de la manière suivante :

(27) 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 <loc type="ville" zone="1" label="Saint-Jean-de-la-Ruelle ">**Saint-Jean-de-la-Ruelle**</loc> alors j'ai dit mince et en fait <loc type="ville" zone="1" label="Ingré">**Ingré**</loc> c'est après <loc type="ville" zone="1" label="Saint-Jean-de-la-Ruelle ">**Saint-Jean**</loc> donc j'ai continué et puis finalement (ESLO2_ENT_1003)

Dès que le système identifie un lieu non-normalisé pour lequel il existe plusieurs correspondances, le système a recours à cette méthode. Si l'on reprend l'exemple 24 avec *Oliv-* qui n'est composé que d'un seul terme : le système contrôle les lieux précédemment mentionné et applique la forme classique de la distance de Levenshtein pour remonter à la forme complète *Olivet*, mentionné deux tours de parole en amont. Ici la comparaison est seulement calculée au niveau des caractères et non au niveau des mots.

La méthode décrite fonctionne pour le traitement des noms de lieux abrégés et tronqués mais elle trouve ces limites lorsqu'il s'agit du traitement de nouvelles mentions de lieux. En effet, les locuteurs du corpus ESLO peuvent employer des surnoms ou même créer de nouveaux noms pour se référer à leur environnement. Ces surnoms ou nouveaux noms ne sont généralement pas construits à partir des mots composant le nom officiel du lieu auquel ils font référence. A partir du moment où les termes employés dans les variantes des noms de lieu n'ont plus de points communs avec ceux composant la référence, on ne peut plus faire le lien entre les deux. Cette difficulté impacte surtout la création de la carte finale et

[65] cf. Figure 15

pose la question de comment seront représentés ce type de noms de lieux sur cette dernière. Néanmoins, ce type de mentions est tout particulièrement intéressant pour réaliser l'analyse de la perception d'Orléans. Aussi, des traitements spécifiques sont réalisés pour identifier ces surnoms ou nouvelles mentions.

2.4.4.6 Patron de détection de nouvelles mentions

A l'oral, un locuteur peut transformer les noms des lieux auxquels il se réfère. Il peut créer de nouvelles mentions en tronquant ou abrégeant les noms conventionnels de ces lieux. Des surnoms peuvent aussi être utilisés à la place des noms conventionnels des lieux. Selon le TLFi[66], un surnom est une « *appellation familière ou pittoresque que l'on substitue au véritable nom d'une personne* ». C'est une dénomination qui se substitue généralement à un nom propre. Généralement, les surnoms ne sont pas construits à partir du nom officiel du lieu correspondant. Les mécanismes utilisés pour créer ces surnoms sont divers. Les locuteurs peuvent s'appuyer sur des caractéristiques esthétiques du lieu comme pour la *ville rose* pour Toulouse ou l'*île de Beauté* pour la Corse. L'histoire du lieu peut aussi amener à la création de surnom comme la *cité phocéenne* pour Marseille et son passé commercial avec les grecs. Enfin, les activités se déroulant dans l'endroit peuvent aussi servir d'inspiration comme pour la *rue des agences immobilières* à Orléans pour la rue Bannier ou *la capitale du cerf-volant* pour Dieppe. On ne retrouve pas la trace du nom officiel dans le surnom comme on la trouve dans une abréviation ou une troncation. Il faut donc s'appuyer sur d'autres indices pour identifier ces surnoms de lieux.

Ce type d'appellation est très souvent révélateur de l'image que peut avoir un lieu et il est donc primordial de les prendre en compte dans ce travail qui vise à analyser la perception qu'ont les orléanais de leur ville. L'observation du corpus a permis d'établir quelques règles pour l'identification de ces nouvelles mentions.

De la même façon que pour la détection des noms de voies, l'analyse se fonde sur la recherche de déclencheur dans le discours. Ainsi, le système s'appuie sur un lexique de noms communs caractérisant des lieux pour identifier les nouvelles mentions. Ce lexique a été constitué manuellement. Il est constitué de 1 649 entrées et regroupe la liste des noms de voies, des noms communs désignant des organisations (*boulangerie*, *école*, *maison*,

[66] http://stella.atilf.fr/Dendien/scripts/tlfiv5/advanced.exe?8;s=2389796505;

cathédrale, etc.), et des termes désignant les lieux répondant à un découpage administratif (*ville*, *capitale*, *région*, etc.). Chacune des entrées du lexique est typée en fonction de la convention d'annotation établie. C'est-à-dire que les termes *ville*, *cité*, *capitale*, etc., seront étiquetés comme des villes. Les termes *boulangerie*, *magasin*, *boutique*, etc., comme des commerces, *rue*, *place*, *venelle*, etc., comme des voies, etc.

Le lexique est appliqué sur chaque tour de parole. Lorsque l'un des termes du lexique est identifié, on applique des règles pour observer le contexte du déclencheur et étendre éventuellement l'annotation. Dans l'exemple 27, le système est capable d'identifier les lieux *Place d'Arc*, *rue de la République* et *rue Royale* grâce aux méthodes précédemment décrites. Après leur identification, le système applique le lexique destiné à identifier des surnoms de lieux ou des nouvelles mentions. Dans cet exemple, le système détecte le terme *avenue*.

(28) ben moi j'aime beaucoup faire euh **Place d'Arc rue de la République rue Royale** euh l'**avenue de la cathédrale** aussi euh voilà (ESLO2_iti_03_01)

A partir du déclencheur *avenue*, le système contrôle les mots suivant pour déterminer si le locuteur mentionne seulement le terme *avenue* ou bien une séquence plus longue. Pour cela, il observe les termes suivant le déclencheur, considéré ici comme la tête d'un potentiel groupe nominal[67]. Si le mot suivant commence par une majuscule, ou s'il est une préposition (PREP) ou un déterminant (DET), on suppose que ces termes font partie du nom du lieu. Tant que les mots suivants le déclencheur appartiennent à ces trois catégories, on inclut ces termes dans le nom du lieu. Si le dernier terme identifié est une préposition ou un déterminant, on inclut le terme suivant dans l'annotation et on arrête le traitement.

[67] cf. Figure 16

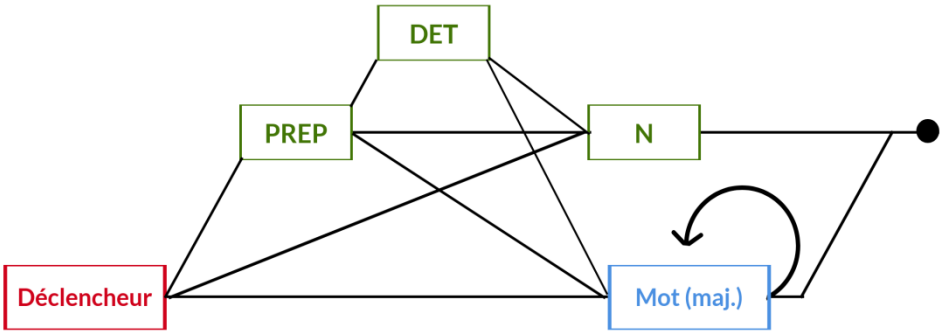

Figure 16 : Patron de détection de nouvelles mentions

Ainsi dans l'exemple 27, le déclencheur *avenue* est suivi par la préposition *de*, elle-même suivie par le déterminant *la*. Le mot suivant *cathédrale* ne répond à aucune des trois conditions mais le dernier terme identifié étant le déterminant *la*, il est inclus dans l'annotation. Le système a donc identifié le lieu *avenue de la cathédrale*[68].

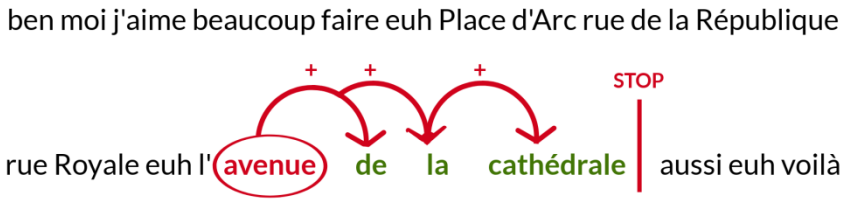

Figure 17 : Annotation d'une nouvelle mention de lieu

A ce stade du traitement, le système est capable de donner le type de lieu, le déclencheur *avenue* étant annoté comme une voie dans le lexique. Cependant, il ne peut pas donner avec certitude la zone ou le label officiel du lieu. Par défaut, on attribuera à la nouvelle mention la même zone que le dernier lieu identifié. On suppose que le locuteur parle de lieux répondant à la même zone géographique pour garder une cohérence dans son discours. Dans cet exemple 27, l'*avenue de la cathédrale* sera placée dans la zone 2 comme le dernier lieu

[68] cf. Figure 17

identifié, *rue Royale*. Enfin, la nouvelle mention identifiée est considérée comme le nom officiel du lieu. L'*avenue de la cathédrale* est donc annotée de la manière suivante :

```
<loc type="voie" zone="2" label="avenue de la cathédrale">
                                                    avenue de la cathédrale
                                                                    </loc>
```

Cette étape du traitement bénéficie à la qualité de l'analyse d'opinion réalisée par la suite sur les lieux identifiés. Par contre, elle n'est pas satisfaisante du point de vue de la création de la carte finale. Si le système ne peut remonter jusqu'au nom officiel du lieu, il ne peut pas récupérer les coordonnées géographiques de ce lieu afin de le placer sur cette carte. Aussi, cette difficulté suppose un traitement différent de ce type de lieu lors de la dernière étape de la modélisation de la perception de la ville d'Orléans.

2.4.5 Evaluation du module de détection des lieux

2.4.5.1 Méthodologie de l'évaluation

La détection automatique des désignations des lieux est réalisée grâce à diverses ressources et afin d'attribuer à chaque lieu une zone géographique, de le typer, et de retrouver son label officiel. Afin de rendre compte de la validité du système développé, il est nécessaire de procéder à son évaluation en suivant les différentes étapes présentées dans la Figure 18.

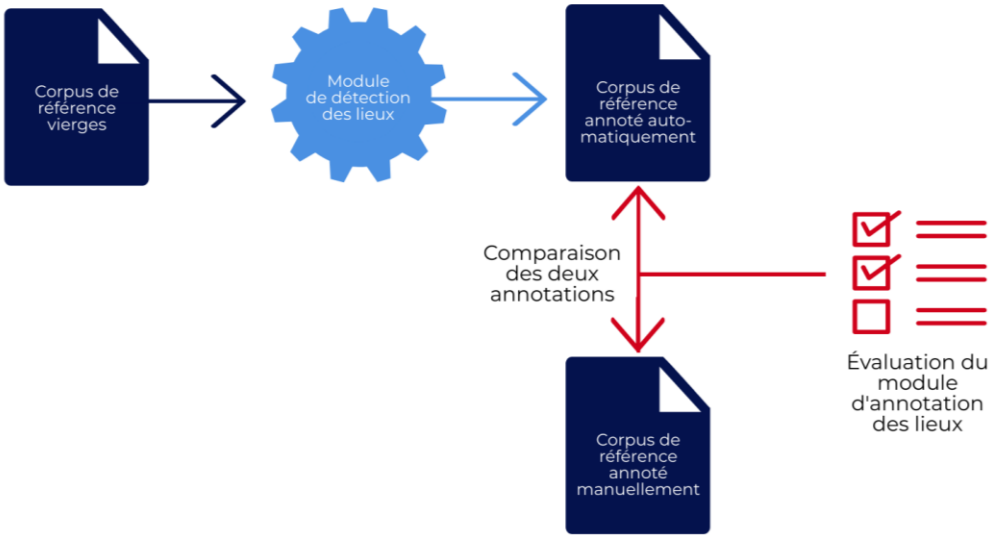

Figure 18 : Méthodologie de l'évaluation du module de détection des lieux

L'évaluation s'appuie sur le corpus de référence présenté dans la section _. La première étape est de faire annoter une version vierge d'annotation transcriptions composant le corpus de référence par le module de détection automatique des lieux. La version annotée automatiquement par le module est ensuite comparée avec la version du corpus de référence annotée manuellement. Plus la version annotée automatiquement est similaire à celle annotée manuellement, plus le module est considéré comme efficace. Afin de mesurer les différences entre ces deux versions, nous utilisons les mesures de Rappel, Précision et F-Mesure, déjà utilisée pour l'évaluation des performances du module CasEN sur nos données[69]. Grâce à ces mesures, nous rendrons compte des performances du module dans la tâche générale de détection de lieux et dans le renseignement des trois informations requises lors de l'annotation.

2.4.5.2 Evaluation générale de la détection des désignations de lieux

La première évaluation porte sur la tâche générale de détection des lieux dans des transcriptions, sans prendre en considération à ce stade de la typologie choisie par le système pour chacun des lieux. L'objectif est d'observer les capacités du module à détecter

[69] cf. section 2.4.1.2.2

des localisations tout en respectant les limites de chaque entité. C'est-à-dire que le système doit être capable de marquer le bornage du nom et de distinguer les différents noms de lieux lorsqu'un tour de parole en contient plusieurs[70]. Pour cette tâche, le module obtient un Rappel de 0,90, une Précision de 0,93 et une F-Mesure de 0,91[71].

Rappel	Précision	F-Mesure
0,90	0,93	0,91

Tableau 14 : Evaluation générale de l'évaluation de la détection des lieux

Le Rappel quantifie la part de détections pertinentes parmi l'ensemble des détections réalisées par le système. En l'occurrence, le Rappel (0,90) de notre module est considéré comme satisfaisant. Ce score démontre que la plupart des détections attendues ont été opérées, comme dans l'exemple 29 dans lequel trois lieux sont cités (*Orléans*, la *rue de la République* et la *rue Royale*).

(29) 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)

(30) 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)

Néanmoins, certaines détections manquent comme le bateau restaurant *l'Inexplosible* mentionné dans l'exemple 30. Les absences de détection s'expliquent principalement par l'absence d'exhaustivité des ressources lexicales employées. Si des lieux ne sont pas listés

[70] cf. exemple 27
[71] cf. Tableau 14

comme le bar *l'Inexplosible* mentionné dans l'exemple 28, le module ne peut pas les détecter.

Dans l'exemple 31, il est question du *Campo Santo,* un cloître accueillant régulièrement des événements organisés par la mairie d'Orléans. Ce nom n'est pas répertorié et il ne peut donc pas être relevé.

(31) 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)

La Précision, qui représente le degré de pertinence de l'annotation, est-elle 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 31, si le système n'a pas identifié le *Campo Samto,* il a annoté *Campo* comme une ville en renvoyant à la commune corse Campo qui est référencée dans la base de données GEOFLA[72].

(32) je l'ai repassé parce que <loc type="monument" zone="1" label="mine de rien">mine de rien</loc> c'est quand même pas si facile que ça (ESLO2_ENT_1026)

Dans des cas marginaux, les règles établies pour l'identification des lieux ont pu être source de bruits lors de l'annotation. C'est le cas dans l'exemple 32 où le module a annoté l'expression *mine de rien*. En effet, *mine* a été reconnu comme un terme déclencheur pour l'application des patrons permettant l'identification de nouvelles mentions de lieu. La préposition *de* et le pronom *rien* ont été associés au déclencheur *mine*.

[72] Cf. section 2.4.4.2

D’une manière générale, le module présente de bonnes performances dans 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.

L’objectif de ce projet est restreint à l’analyse des lieux situés à Orléans. Or, le système réalise plus d’annotations que nécessaire en considérant d’autres lieux extérieurs à la ville. L’ajout de ces lieux permet d’obtenir une plus grande quantité de données utiles pour l’étape suivante de l’analyse d’opinion. Chaque évaluation peut être effectuée en fonction de l’attribut de la zone géographique[73].

Evaluation	Rappel	Précision	F-Mesure
Zone 0	0,84	0,96	0,90
Zone 1	0,91	0,93	0,92
Zone 2	0,93	0,97	0,95

Tableau 15 : Evaluation du module de détection des lieux en fonction de la zone géographique

L’évaluation du module établie en fonction de la zone géographique suit la même tendance que l’évaluation générale avec une Précision plus forte que le Rappel. Si les scores de Rappel pour les lieux des zones 1 et 2 (0,91 et 0,93) sont proches du Rappel de l’évaluation générale (0,93), le score concernant les lieux de la zone 0 est moins satisfaisant (0,84). Ce score signifie que le silence est plus important dans l’annotation des lieux situés dans la zone 0, soit en dehors d’Orléans. Là encore, l’explication du nombre plus faible de détections s’explique par la constitution des ressources lexicales. Celles-ci ont été élaborées en privilégiant l’identification de lieux relatifs à Orléans, aussi sont-elles moins précises en ce qui concerne la zone 0. L’ajout de bases de données géographiques supplémentaires pourrait contribuer à améliorer ce score.

Malgré cette dégradation du Rappel pour les lieux de la zone 0, la Précision est quant à elle excellente (0,96). L’annotation de ces lieux est donc de très bonne qualité. Celle-ci est même meilleure lorsqu’il s’agit des lieux de la zone 2 (0,97), soit les lieux situés à Orléans,

[73] cf. Tableau 14

ce qui est de bon augure pour la suite du travail. La F-mesure de 0,95 pour la détection des lieux de la zone 2 confirme que le module parvient à détecter la plupart des lieux attendus.

2.4.5.3 Evaluation des attributs caractérisant les lieux identifiés

Après l'évaluation des capacités générales du système à détecter les lieux, nous proposons d'évaluer sa capacité à les caractériser au travers de l'observation des trois informations renseignées lors de l'annotation : la zone géographique, le type et le label officiel[74].

	Type			Zone			Label		
Evaluation	Rappel	Précision	F-Mesure	Rappel	Précision	F-Mesure	Rappel	Précision	F-Mesure
Générale	0,92	0,89	**0,90**	0,91	0,93	**0,92**	0,90	0,85	**0,88**
Zone 0	0,85	0,95	**0,90**	0,83	0,92	**0,88**	0,83	0,85	**0,84**
Zone 1	0,95	0,96	**0,96**	0,92	0,95	**0,93**	0,90	0,87	**0,89**
Zone 2	0,94	0,86	**0,89**	0,94	0,96	**0,95**	0,93	0,90	**0,91**

Tableau 16 : Evaluation des attributs en fonction des zones géographiques

D'une manière générale, les scores de F-Mesure pour le renseignement des attributs type (0,90), zone (0,92) et label (0,88) sont satisfaisants. On peut même observer une amélioration des scores à proportion de la proximité de la zone géographique d'Orléans, ce qui est cohérent avec l'évaluation générale et les constatations faites à propos des ressources lexicales.

Une difficulté de premier ordre concerne la résolution par le système de la caractérisation des lieux dont la mention diffère du nom officiel. En ce qui concerne l'attribut du label, la F-Mesure de 0,88 montre la capacité du système à parvenir à relier la variante d'un nom de lieu à son nom officiel comme dans l'exemple 31 dans lequel le *faubourg Bannier* est correctement relié à la *rue du faubourg Bannier*.

[74] cf. Tableau 15

(33) y a des boulangeries <loc type="voie" zone="0" label="rue du Faubourg Bannier">faubourg Bannier</loc> (ESLO2_ENT_1042)

Cependant, on peut remarquer que les difficultés du système dans la tâche d'attribution du label officiel se concentrent sur les lieux de la zone 0, soit en dehors d'Orléans et son agglomération. La F-Mesure de 0,84 n'est pas aussi satisfaisante que les autres mesures. Dans l'exemple 34, *Paris* est correctement annoté et caractérisé comme une *ville*, de zone *0* dont le nom officiel est *Paris*. Par contre, si l'entité *fac de langues* est bien identifiée comme un lieu à fonction éducative de la zone 0, le label n'est pas correct. Ce lieu a été reconnu grâce au patron dédié au repérage des nouvelles mentions mais le système n'est pas parvenu à le relier à son nom officiel. Dans ce cas, le système considère que le label correspond au nom identifié dans la transcription. Ensuite, il attribue au lieu la même zone que celle de la dernière mention détectée en l'occurrence la zone 0, comme la ville de Paris. Malgré l'incapacité du système à retrouver le label officiel, les autres attributs sont correctement renseignés.

(34) peut-être sur <loc type="ville" zone="0" label="Paris">**Paris**</loc> ou une <loc type="educatif" zone="0" label="fac de langues">**fac de langues**</loc> elle sait pas très bien(ESLO2_ENT_1052)

PARTIE II

LIEU ET PERCEPTION

Chapitre 3 : PERCEPTION, OPINION, SENTIMENT, EMOTION : DES NOTIONS SUBJECTIVES

3.1 Définitions

On définit communément la subjectivité en opposition à l'objectivité. La subjectivité est définie dans le TLFi[75] comme la « qualité (inconsciente ou intérieure) de ce qui appartient seulement au sujet pensant » tandis que l'objectivité est la « qualité de ce qui existe en soi, indépendamment du sujet pensant ». De Landsheere (1979), repris dans le *Grand Dictionnaire* de l'Office Québécois de la Langue Française[76], définit l'objectivité comme le « caractère de ce qui donne une image non déformée des organismes et des choses, ou de ce qui les décrit et les juge, sans parti pris ». Cette définition met en lumière les caractéristiques de la subjectivité. Cette dernière correspondrait à tout ce qui n'est pas objectif, soit tout ce qui ne résulte pas d'une simple observation du réel mais qui donne une image des choses influencée par une appréhension intime. Ce qui est subjectif est propre à un individu. Le TLFi confirme cette assertion en présentant la subjectivité comme « la qualité de ce qui ne donne pas une représentation fidèle de la chose observée ». Une description subjective ne rend pas compte d'une réalité observée mais d'une réalité ressentie.

Kerbat-Orecchionni (2009 : 125-130) définit plusieurs types de subjectivité :

- La **subjectivité affective** qui considère les expressions qui « indiquent que le sujet d'énonciation se trouve émotionnellement impliqué dans le contenu de son énoncé ». Elle donne pour exemple les expressions « Cette pénible affaire, cette triste réalité », dans lesquelles on retrouve la trace des émotions de l'énonciateur.

[75] http://stella.atilf.fr/Dendien/scripts/tlfiv5/advanced.exe?8;s=1771473780;
[76] http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8462502

- La **subjectivité interprétative** s'intéresse au processus de dénomination d'un objet qui consiste à « choisir au sein d'un paradigme dénominatif ; […] orienter dans une certaine direction analytique, l'objet référentiel ». Le fait même de pouvoir choisir entre plusieurs étiquettes pour identifier son environnement est subjectif : « aucun item lexical ne saurait être utilisé en toute objectivité ». Selon le contexte, une même expression ne sera pas interprétée de la même façon.
- La **subjectivité modélisatrice** qui renvoie aux « expressions qui spécifient le mode d'assertion (constatif, hypothétique, obligatif, etc.) des propositions d'énoncés, et le degré d'adhésion (forte, réticente, nuancée) ».
- La **subjectivité axiologique** qui se rapporte aux idéologies et de ce que l'on croit en savoir pour dégager « la source évaluative de l'objet qui supporte l'évaluation positive ou négative, et du degré d'intensité avec lequel elle se formule ».

Cette typologie révèle des liens entre subjectivité et d'autres notions. La subjectivité affective renvoie explicitement aux concepts d'*émotion* ou de *sentiment*, tandis que la subjectivité modélisatrice et la subjectivité axiologique se rapportent au concept de *modalité*. En exposant l'idée que chacun peut trouver une interprétation différente aux termes employés dans certains contextes, la subjectivité interprétative peut se rapprocher de l'*opinion*.

La notion centrale de cette étude est la perception. Partager sa perception d'un objet est un processus subjectif dans lequel chacune des notions citées joue un rôle plus ou moins important. Il est donc nécessaire de définir ces différentes notions afin de rendre compte de la notion de perception.

3.1.1 Emotions

Dans le Larousse en ligne[77], une émotion est définie comme une « réaction affective transitoire d'assez grande intensité, habituellement provoquée par une stimulation venue de l'environnement ». Ces stimulations provenant de l'environnement correspondent aux captations d'énergie réalisées par les récepteurs sensoriels qui entrent en jeu dans l'expérience de perception. Le TLFi[78] reprend cette idée en ajoutant celle de « conduite réactive, réflexe, involontaire vécue simultanément au niveau du corps ». L'individu ne choisit pas d'éprouver des émotions, celles-ci s'imposent à lui dans une expérience physique.

De nos jours, les émotions restent un concept difficile à définir. Les premiers travaux majeurs sur la question des émotions sont ceux de Darwin (1872). Il pose le postulat que les émotions sont *innées* et *universelles*. Selon Luminet (2008), Darwin considère que les émotions ne sont pas le produit d'un individu, humain comme animal, et ne nécessitent pas d'apprentissage. Les émotions « émergent indépendamment de leur utilité pour la survie de l'espèce » et seraient plutôt « les traces d'habitudes ou pratiques anciennes » (Luminet, 2008 : 22). Ces théories ont inspiré William James qui selon Nugier véhicule l'idée que « faire l'expérience d'une émotion c'est d'abord faire l'expérience des changements corporels ou physiologiques qui l'accompagnent » (2009 : 9). Selon James, les émotions ont une existence physique qui se trouve dans les changements d'états du corps et qui appelle une réaction. Il donne l'exemple de la peur que l'on peut ressentir face un ours qui nous surprendrait lors d'une promenade en forêt. Le fait de percevoir le changement de notre rythme cardiaque, de notre respiration, des tremblements, etc. fait que nous avons peur et que nous aurons une certaine attitude en réaction à cette émotion. Pour James (cité dans Nugier, 2009 : 9) :

> Nous ferions l'expérience d'émotions parce que notre corps a évolué pour répondre automatiquement et de façon adaptative aux aspects de l'environnement qui auraient une signification pour nous en termes de survie.

[77] https://www.larousse.fr/dictionnaires/francais/%c3%a9motion/28829?q=%c3%a9motion#28701
[78] http://stella.atilf.fr/Dendien/scripts/tlfiv5/advanced.exe?8;s=701919120;

L'émotion revêt un caractère *adaptatif* grâce auquel l'organisme peut s'adapter aux exigences et contraintes de son environnement.

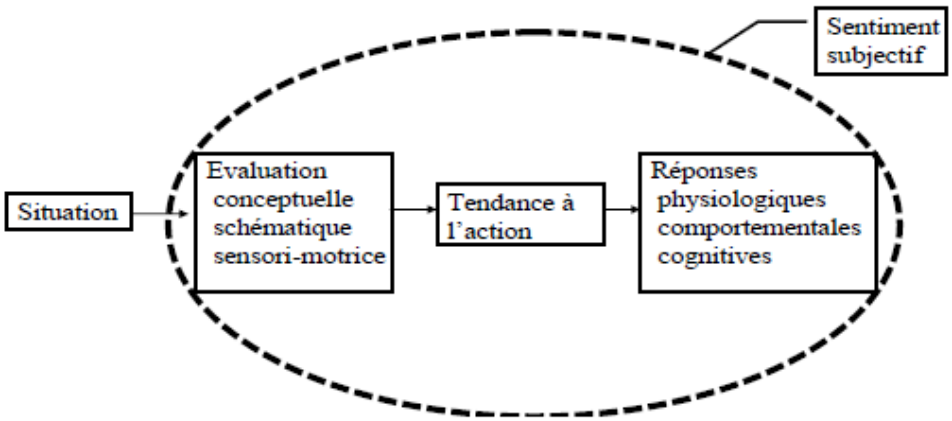

Figure 19: Les composantes du processus émotionnel - Extrait de P. Philippot. (2007)

Dans *Emotions et psychothérapie* (2013), Philippot schématise les composantes émotionnelles[79]. Peu importe la situation, tout individu attribue une valeur émotionnelle à cette situation. Cette activité est le produit d'une double activité cognitive : il y a émotion lorsqu'un sujet attribue à une situation sociale une valeur émotionnelle, et lorsqu'il sélectionne ensuite une stratégie pour réagir de façon adaptée dans ce contexte précis. Cette « évaluation émotionnelle » s'appuie sur les récepteurs sensoriels pour créer une représentation conceptuelle de la situation, dont l'interprétation ne peut être dissociée des connaissances ou des valeurs de l'individu (Charaudeau, 2000). Elles sont liées à « un savoir de croyances polarisé autour de valeurs socialement constituées » (*op. cit.* : 131) et font donc l'objet d'un jugement moral. Dans une perspective cognitiviste, Scherer (cité dans Philippot, 2007) définit cinq dimensions pour décrire ce processus :

- Evaluation de la *nouveauté* pour déterminer si la situation en présence présente des éléments nouveaux ou non pour l'individu. « Ce qui se passe est-il familier, soudain, prévisible ? » (Philippot, 2009 : 5).

[79] cf. Figure 19

- Evaluation de la *valence* pour juger si les stimuli reçus ont une valeur plutôt positive ou négative. « Ce qui se passe est-il positif versus négatif, attirant versus aversif ? » (*op. cit.* : 6).
- Evaluation du *rapport au but* pour estimer la pertinence de la situation par rapport aux buts que poursuit l'individu à cet instant. « Ce qui a été détecté est pertinent pour certains buts ou besoins de l'individu » (*op. cit.* : 6).
- Evaluation du *potentiel de maîtrise* pour indiquer dans quelle mesure l'individu contrôle la situation et peut savoir comment celle-ci pourrait évoluer pour finalement déterminer s'il pourra en tirer parti. Il faut « établir les ressources dont on dispose pour tirer au mieux parti de cet événement, qu'il facilite ou qu'il vienne faire obstacle au but » (*op. cit.* : 9).
- Evaluation de *l'accord avec les normes* pour remettre en perspective la situation et les conclusions des précédentes évaluations avec les valeurs personnelles de l'individu et celle de la société dans laquelle il évolue. « Cette dimension comprend deux facettes : l'évaluation par rapport aux normes externes ou sociales et l'évaluation par rapport aux normes internes ou personnelles » (*op. cit.* 11)

Chacune de ces évaluations dépend de la précédente et l'ensemble des conclusions obtenues donnent une signification émotionnelle à la situation observée. A partir de cette signification, l'organisme aura une « tendance à l'action », c'est-à-dire qu'il se prépare à interagir avec son environnement. Cette tendance « amorce les différents systèmes de l'organisme en vue de soutenir un certain type d'action » (Frijda 1986, Kuipers & Ter Schure, 1989, cité dans Philippot, 2007). Frijda (cité dans Philippot 2007) identifie huit tendances à l'action : l'*approche positive*, l'*agression*, la *panique*, le jeu, l'*inhibition*, le rejet, la *soumission* et la *dominance*. Ces huit tendances composent le « bagage émotionnel inné dont disposeraient les humains pour organiser leurs réponses émotionnelles » (Philippot, 2007 : 20).

Les réponses émotionnelles sont physiologiques, comportementales et cognitives. Elles sont la réalisation de l'action amorcée lors de la phase de tendance à l'action. On parle aussi d'*expressions émotionnelles* et celles-ci peuvent être verbales ou non-verbales. En effet, les émotions peuvent s'exprimer à travers les expressions faciales, la posture générale du corps et le langage. Dans le cas des expressions faciales, on étudiera quels muscles du visage

entrent en jeu dans l'expression de différentes émotions, quelles règles régissent ces mouvements, quel peut être l'impact de la culture etc. (Ekman, 1973 ; Izard, 1969, 1990 ; Levenson *et al.*, 1999 ; Dantzer, 2002). Au-delà du visage, on peut suivre les mêmes questionnement pour les appliquer à l'étude de la posture du corps (Bloch *et al.*, 1987 ; Duclos *et al.*, 1989 ; Coulson, 2004). Du point de vue du langage, la prosodie est elle aussi révélatrice des émotions d'un individu. Selon que la voix monte, descend, est plus ou moins forte, l'émotion exprimée n'est pas la même. On peut aussi aborder les émotions en étudiant les situations d'énonciation (Furukawa, 1996), ou en analysant les structures de verbes d'émotions (Novakova *et al.* 2014) ou des *émotimots* et *émotiphrases*[80] (Bottineau, 2013).

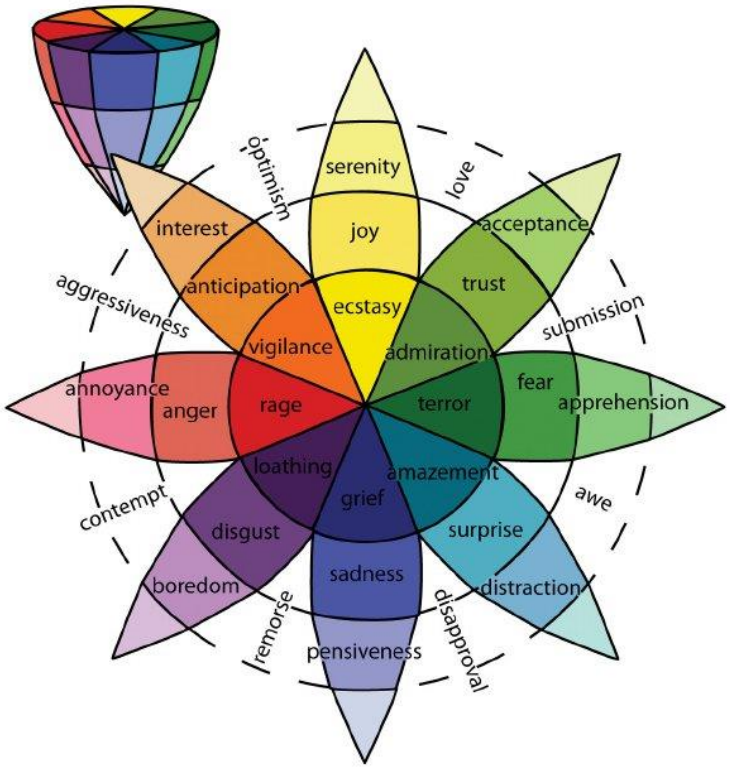

Figure 20 : Roue des émotions de Plutchik – Extraite de Plutchik & Kellerman (1980)

Les protocoles pour étudier les émotions en fonction de la façon dont elles sont exprimées sont multiples, mais une question reste commune : celle de la typologie des émotions. Plusieurs typologies existent dans la littérature. L'une des premières typologies existantes

[80] Interjections, expressions et phrases figées à caractère interjectif.

est celle de Descartes (1728) qui définit six émotions : l'*admiration*, l'*amour*, la *haine*, le *désir*, la *joie* et la *tristesse*. L'une des typologies les plus citées aujourd'hui est surement celle de Ekman (1973) qui présente six émotions de *base* : la *tristesse*, la *joie*, la *colère*, la *peur*, le *dégoût* et la *surprise*. Il distingue ces émotions auxquelles correspondent des expressions faciales comprises universellement au sens de Darwin, des émotions dites *secondaires*, qui résultent de la combinaison des émotions de base. Plutchik & Kellerman (1980) considère plutôt huit émotions de base – la *joie*, la *peur*, le *dégoût*, la *colère*, la *tristesse*, la *surprise*, la *confiance* et *l'anticipation* – dont ils proposent une représentation sous la forme d'une roue des émotions[81]. Les huit émotions de base ou *primaires* composent le cœur de la roue. Chacun des rayons de la roue représente les déclinaisons de chaque émotion primaire en émotions *secondaires*. D'une manière générale, les émotions sont organisées des plus intenses (moyeu) au moins intense (extérieur de la roue).

Ces dernières années, les études concernant les émotions se sont multipliées. Ekman (2016) s'interroge sur le degré de consensus pouvant exister autour de la manière de concevoir ou catégoriser les émotions. Pour cela, il présente les résultats d'un sondage réalisé auprès de 248 chercheurs sélectionnés en fonction de leurs problématiques de recherche. L'une des premières questions aborde la typologie des émotions. Parmi une liste de 18 émotions, les participants devaient sélectionner celles qui selon eux ont été empiriquement établies. Il en ressort un fort accord à propos de 5 émotions : la colère (*anger*, 91%), la peur (*fear*, 90%), le dégoût (*disgust*, 86%), la tristesse (*sadness*, 80%), la joie (*happiness*, 76%). Ekman déduit de ce fort accord que la preuve de l'importance de ces cinq émotions est robuste. L'objectif d'Ekman est de minimiser l'importance des désaccords persistant dans la communauté scientifique pour plutôt mettre l'accent sur les consensus observés.

3.1.2 Sentiments

[81] cf. Figure 20

Les définitions des dictionnaires Larousse en ligne[82] et du TLFi[83] sont révélatrices des connexions et des ambiguïtés qui existent entre les notions d'émotions, de sentiment et d'opinions. Si le Larousse définit le sentiment comme la « connaissance plus ou moins claire, donnée d'une manière immédiate », il le définit aussi comme une « opinion, avis que l'on a sur quelque chose » et comme un « état affectif complexe et durable lié à certaines émotions ou représentations ». Le terme sentiment est ici considéré comme le synonyme d'opinion. Pourtant ces deux notions ne recouvrent pas les mêmes concepts.

Dans le TLFi, le sentiment serait la « conscience que l'on a de soi et du monde extérieur », ou le fait d'être « dans un état d'inconscience » qui relèverait du « domaine des sens (excepté la vue et l'ouïe) » et du « domaine de l'intellect, de l'intuition ». On retrouve dans ces définitions le lien entre émotions et sentiment. Si les émotions sont des réactions spontanées qui se manifestent physiquement et brièvement, les sentiments s'appuient sur les émotions pour s'élaborer cognitivement dans la durée.

Reboul [84] propose de définir le terme *sentiment* dans l'*Encyclopédie Universalis* en commençant par faire état des conceptions de la notion par différents auteurs. Il cite Malebranche pour qui « le sentiment est la perception confuse des choses et de soi-même », Janet qui réduit le sentiment « aux émotions fondamentales », ou encore Alain qui considère le sentiment comme « l'acte par lequel la volonté assume les passions et les transfigure ». Si Reboul ne remet pas fondamentalement en cause leurs propos, il considère néanmoins que ce que ces auteurs décrivent ne relève pas du sentiment « faute d'avoir accepté au départ l'ambiguïté foncière du mot " sentiment " ». Face à cette ambiguïté, Reboul propose à son tour une définition du sentiment :

> Le sentiment est avant tout l'acte et le résultat du *sentir*, lequel désigne la prise de conscience immédiate, sans intermédiaire, sans distance, des choses et de nous-mêmes ; l'objet du sentiment est toujours ce qui nous « touche ». À partir de là, sentiment signifie *conscience*.

[82] https://www.larousse.fr/dictionnaires/francais/sentiment/72138?q=sentiment#71335
[83] http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?60;s=701919120;r=3;nat=;sol=1;
[84] Olivier REBOUL, « **SENTIMENT** », *Encyclopædia Universalis* [en ligne], consulté le 27 juin 2019. URL : http://www.universalis.fr/encyclopedie/sentiment/

A cause de cette difficulté à distinguer l'émotion du sentiment, les typologies des sentiments proposées dans la littérature recoupent souvent celle des émotions. Goossens (2005) évacue cette difficulté en utilisant le terme *SENTIMENT* en majuscule pour englober les deux concepts. Ainsi, il considère à la fois la *colère*, l'*amour*, la *joie*, la *tristesse* et la *honte* comme des sentiments et des émotions. Les travaux d'Ekman et Plutchik précédemment cités identifient pourtant la *colère*, la *joie* et la *tristesse* comme des émotions.

Hotyat (1973, cité dans le *Grand Dictionnaire* de l'Office Québécois de la Langue Française) définit les sentiments comme « état affectif plus stable que l'émotion et centré consciemment vers des êtres ou des systèmes de valeurs ». Il donne aussi des exemples de type de sentiments comme : « la sympathie, l'antipathie et leurs modalités ; les sentiments sociaux, comme l'esprit de corps, le civisme, le patriotisme ; les sentiments impersonnels ou idéaux comme la soif de justice ou de vérité ; l'amour du beau, le sentiment religieux ». Cette typologie n'est pas exhaustive mais montre bien qu'il est possible de distinguer les émotions des sentiments, en particulier au niveau typologique même si cet exercice reste difficile.

Les définitions de la notion de *sentiment* sont souvent mises en perspective avec l'ambiguïté existant avec celles d'*émotion*. Cosnier (1974, cité dans Poncet, 2007) considère que l'émotion « se rapporte à un état psycho-organique ayant comme caractéristique son intensité, sa brièveté et la scission brutale des fonctions mentales et physiologiques qu'elle induit » et que le sentiment serait au contraire un « état affectif mentalisé ». Damasio (2006) différencie aussi les émotions des sentiments en s'appuyant sur le caractère *public* du premier et le caractère *privé* du deuxième. C'est-à-dire que le terme *sentiment* désignerait « l'expérience mentale et privée d'une émotion », alors que le terme *émotion* désignerait « l'ensemble de réponses qui, pour bon nombre d'entre elles, sont publiquement observables ».

Le sentiment serait le résultat de l'analyse de sensations expérimentées qui permettrait d'avoir une connaissance direct de soi-même. Dans cette acception, les émotions et les sentiments pourraient donc représenter les deux dimensions de la perception présentées par Barbaras (2009). Les émotions, qui découlent de l'expérience des sens, pourraient servir de mode d'accès à la réalité, et les sentiments, qui correspondent à l'état dans lequel se retrouve un individu face à ses émotions, serait l'épreuve que se fait l'individu de la réalité décrites par ses émotions.

Les émotions et les sentiments relèvent tous les deux d'expériences internes dont l'expression peut transparaitre de manière spontanée dans des indices verbaux comme non verbaux. Enfin Orth[85] (cité dans Beaunis, 1904 : 656) considère aussi le sentiment comme un « un phénomène psychique indépendant et qui doit être considéré comme un élément constitutif de la conscience de même que la sensation ».

3.1.3 Opinions

Selon le *Larousse* en ligne, une opinion est un « jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense » ou l'ensemble « des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc. ». L'opinion n'est pas conçue instantanément, elle est le fruit d'une réflexion que l'on exprime, partage. Le *TLFi* définit aussi l'opinion comme :

- Jugement personnel que l'on porte sur une question, qui n'implique pas que ce jugement soit obligatoirement juste.
- Jugement, manière de penser dénotant une orientation particulière.
- Point de vue, position précise que l'on a dans un domaine particulier.

Une opinion est propre à un individu et non une réponse immédiate à un stimulus comme pourrait l'être une émotion. Formuler une opinion implique d'avoir une certaine connaissance de l'objet appréhendé afin de porter un jugement à son propos. A partir de l'analyse de plusieurs éléments, un individu prend position par rapport à l'objet et en a une certaine opinion. Cette opinion est personnelle et n'est pas donc pas obligatoirement « juste ». L'opinion est une conviction personnelle qui n'a pas vocation à être vraie ou fausse.

La question de la définition de l'opinion se pose depuis l'époque des philosophes de la Grèce Antique. Dans ses dialogues, Platon oppose l'opinion ou *doxa*, à la science. Dans *La théorie platonicienne de la doxa*, Lafrance & Brisson (2015) reprennent la définition de Platon du terme *doxa*. Ce dernier présente deux sens : « celui d'apparence − ce qui

[85] Beaunis H. Orth Gefühl und Bewusstseinslage. In: L'année psychologique. 1904 vol. 11. pp. 654-656.

m'apparaît objectivement –, et celui d'opinion − ce qui subjectivement me semble être le cas ». Platon considère même l'opinion comme un type de connaissance inférieur à la science et oppose ainsi l'opinion au concept de vérité.

Martin & White (2005) pose la théorie de l'*appraisal* afin d'expliciter comment un locuteur construit et communique des jugements de valeurs et modalise son message. Cette théorie est le résultat de la combinaison de trois systèmes :

- L'attitude qui examine le sentiment du locuteur envers l'entité ou la proposition comme bonheur, malheur, contentement, mécontentement.
- L'engagement qui renvoie aux signifés par lesquels les locuteurs reconnaissent ou ignorent la diversité des points de vue que leurs messages mettent en jeu.
- La graduation permet de graduer le degré d'intensité de l'évaluation dans le discours.

Selon Zhang (2015), la difficulté de la notion d'opinion tient au fait qu'elle ouvre deux champs d'investigation différents et complémentaires puisqu'elle possède simultanément une dimension dite *épistémique* et une dimension dite *axiologique*. La dimension épistémique renvoie à « la qualification d'un énoncé par rapport à un critère de connaissance » (*op. cit.* : 39). Elle représente la distinction possible entre des énoncés « porteurs de vérité », dans le sens objectif, et les énoncés « de l'ordre de la croyance », c'est-à-dire subjectif. La dimension axiologique renvoie quant à elle à « l'évaluation à proprement dite, à l'expression d'un jugement de valeur, c'est-à-dire à l'indication de ce qui vaut, ce qui est bon ou mauvais ». Le jugement se fait dans un certain référentiel, selon une certaine norme. Il ne s'agit pas de décrire des faits mais plutôt de distinguer quelles actions peuvent être considérées comme souhaitable ou non.

Enfin, Ferrand (2016 : 103) rappelle les deux postulats les plus répandus sur lesquels se fondent de nombreuses recherches traitant des opinions :

- un acteur aurait *une* certaine représentation d'un objet, *un* certain savoir profane sur une question, *une* certaine opinion face à une incertitude ;

- c'est l'esprit de l'acteur qui serait le siège de sa pensée, de sa capacité à juger et opiner, et donc de son savoir et de son opinion sur un objet ou sur une question donnés.

Il critique ces deux assertions et préfère considérer que « en situation d'incertitude, un acteur réfléchit et parvient à prendre position grâce aux discussions que permet une relation avec un autre acteur ». Dans ces conditions, l'origine de l'opinion ne se trouve pas seulement dans l'esprit de l'individu mais serait « le processus dialogique constitutif de la relation ». L'opinion ne se construit pas seulement dans l'individualité. De plus, cette relation de discussion peut connecter l'individu à des groupes socialement constitués, porteurs de savoirs et de valeurs variés à propos d'un même objet ou d'une même question. Ainsi, l'opinion privée est indissociable de l'opinion publique.

Selon Habermas *et al.* (1988, cité par Farge,2013), l'opinion publique « renvoie à un public constitué par des personnes privées faisant usage de la raison ». La convergence d'opinions privées forme l'opinion publique. Mesurer l'opinion publique reste difficile. Blondiaux (2016) s'est donné pour objectif de décrire comment s'est constituée l'opinion publique historiquement et tente surtout d'apporter une réponse à la question : « Comment en est-on venu à accepter l'équivalence entre opinion publique et résultats de sondages ? ». Il discute l'importance de l'opinion publique dans l'idéologie politique démocratique et l'amalgame fait entre résultats de sondage et l'opinion de la société. Blondiaux estime que la naissance des sondages en 1935 aux Etats-Unis a contribué à transformer la notion d'opinion publique de « concept ambigu » en « construit mesurable ». Depuis, l'opinion publique est devenue ce que mesurent les sondages. Bourdieu (1979) remet en cause l'existence même de l'opinion publique représentée par les résultats de sondages en critiquant les méthodes utilisées pour la collecter. L'opinion publique serait donc un « artefact, construit à partir des résultats publiés de réponses aux questions de sondage agrégées ».

Privée ou publique, l'opinion présente des enjeux majeurs pour les régimes démocratiques. Dans l'introduction correspondante à l'entrée « opinion publique » de l'*Encyclopedia Universalis*[86], Champagne considère l'existence de la notion comme indiscutable puisque devenue « banale » dans les démocraties actuelles et lui attribue « indiscutablement une

[86] Patrick CHAMPAGNE, « OPINION PUBLIQUE », *Encyclopædia Universalis* [en ligne], consulté le 13 juillet 2019. URL : http://www.universalis.fr/encyclopedie/opinion-publique/

réalité sociale ». L'existence de l'opinion et surtout sa diffusion a une importance telle dans notre société que la liberté d'opinion est reconnue par la Déclaration Universelle des Droits de l'Homme de 1948 :

> Tout individu a droit à la liberté d'opinion et d'expression, ce qui implique le droit à ne pas être inquiété pour ses opinions et celui de chercher, de recevoir et de répandre, sans considération de frontières, les informations et les idées par quelque moyen d'expression que ce soit.
>
> Article 19

Face aux enjeux que présentent les opinions dans la société actuelle, les chercheurs en TAL se sont aussi penchés sur le traitement automatique de ce type d'expressions. Ainsi, la reconnaissance automatique de la subjectivité, soit des émotions, des sentiments et des opinions, occupe une place majeure dans la recherche en TAL.

3.1.4 TAL et subjectivité

Du point de vue du TAL, la subjectivité est le plus souvent abordée sous l'angle de l'analyse de sentiment (*sentiment analysis*) et de la fouille d'opinion (*opinion mining*). Comme le précise Marchand (2015), les noms de ces deux domaines peuvent être utilisés de manière interchangeable. Les différences de définitions mises en lumières entre les notions d'émotions, de sentiments et d'opinions[87] sont souvent lissées dans les travaux en TAL traitant de subjectivité. Ainsi, la détection d'opinion désigne généralement le classement des donnés selon « une axiologie positif/négatif » tandis que l'analyse d'opinions concerne « l'étude des émotions telles que la peur, la colère ou la joie » (Marchand, 2015 : 9). Liu (2012) propose de modéliser l'opinion en un quintuplet[88] d'informations qui considère :

- L'entité à propos de laquelle l'opinion est émise.
- La caractéristique de l'entité qui est ciblée par l'opinion.

[87] cf. section 3.1.1 ; section 3.1.2 ; section 3.1.2

[88] Dérivation du terme *n-uplet* désignant un assemblage de *n* éléments dans lesquels on peut dénombre le premier, le deuxième… jusqu'au *n*-ième élément.

- La qualification (positive ou négative) de la caractéristique.
- L'émetteur de l'opinion.
- L'instant auquel l'opinion est exprimée.

Pour Liu, le processus idéal d'extraction d'opinion consisterait à transformer un texte libre en une série de quintuplets pour ensuite pouvoir quantifier et comparer les opinions. Dans Pang et Lee (2008 : 6), on peut trouver la liste des qualités essentielles que doit posséder un outil destiné à de la fouille d'opinions :

> Selon Dave *et al.*, l'outil idéal d'extraction d'opinion « traiterait un ensemble de résultats de recherche pour un article donné, générant une liste d'attributs du produit (qualité, caractéristiques, etc.) et en agrégeant les opinions sur chacun d'entre eux (mauvais, mixte, et bon) ». [89]

Les méthodes employées pour l'analyse de la subjectivité dans ces différents corpus sont multiples. Par exemple, Zhang (2015) propose une chaîne de traitement à base de règles pour mener une veille d'entreprise dans un corpus de presse économique et financière avec un intérêt particulier pour l'analyse des opinions sous l'angle de leur polarité et de leur intensité. La formalisation de connaissances linguistiques sous forme de règles, de lexiques et de réseaux lexicaux et « est directement exploitable dans une approche symbolique alors qu'il faudrait un travail supplémentaire parfois important pour identifier les indices exploitables dans une approches probabilistes ». Les approches symboliques en fouilles d'opinions s'appuient le plus souvent sur des dictionnaires de mots polarisés. Dans ces dictionnaires, une polarité (positive ou négative) ou même des émotions sont associées à chacune des entrées. Lorsque ces dictionnaires sont inclus dans une chaîne de traitement, un score de subjectivité peut être attribué au document analysé en comptabilisant les mots porteurs d'opinion ou de polarité. De nombreuses ressources existent pour décrire l'anglais comme le General Inquirer [90], le lexique de l'Opinion Finder System [91] ou le

[89] *According to Dave* et al., *the ideal opinion-mining tool would "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)".*
[90] http://www.wjh.harvard.edu/~inquirer/homecat.htm
[91] https://mpqa.cs.pitt.edu/opinionfinder/

SentiWordNet[92] qui attribue une polarité à chacun des *synsets* composant la base de données WordNet[93]. Peu de ressources similaires existent pour le français. Le NRC Emotion Lexicon (Mohammad & Turney, 2010, 2013) est une liste de 14182 mots anglais annotés en polarité et selon sept émotions : *colère*, *peur*, *anticipation*, *confiance*, *surprise*, *tristesse*, *joie*, et *dégoût*. Ce lexique a été traduit en français afin de constituer le French Expanded Emotion Lexicon (FEEL)[94] par Abdaoui *et al.* (2016). La traduction a été validée manuellement par un traducteur professionnel et le nombre d'émotions caractérisant chaque entrée du lexique a été réduit à cinq (*colère*, *peur*, *surprise*, *tristesse*, *joie*, *dégoût).*

Aujourd'hui, les méthodes d'apprentissage supervisé sont omniprésentes dans les recherches en fouille d'opinion et analyse de sentiments. La combinaison d'outils linguistiques et d'outils de classification utilisés en fouille de texte est largement exploitée pour détecter l'opinion d'un texte (Harb *et al.* 2008). Ces méthodes sont souvent utilisées dans des challenges nationaux. La conférence d'évaluation DEFT (Grouin & Forest, 2012) a régulièrement mis à l'honneur la question de la détection de la subjectivité dans diverses sources de données.[95] En 2017 et en 2018 il était par exemple question d'extraire des opinions et des sentiments dans des corpus de tweets.

Avec l'essor du Web 2.0, l'intérêt pour les opinions et les états affectifs des internautes qui s'y expriment en temps réel est grandissant. Internet regorge de déclarations d'utilisateurs à propos d'une infinité de sujets exprimées sous toutes les formes possibles. Les enjeux relatifs à l'exploitation de ce type de données sont aussi nombreux que leur variété est grande et présentent donc d'importants challenges pour le TAL (Mohammad, 2017). Dans ce contexte, Boullier & Lohard (2012) évoquent par exemple l'existence de travaux analysant des critiques de cinéma (Pang *et al.*, 2002 ; Pang & Lee, 2008) ou des critiques télévisuelles et vidéoludiques (Gillot, 2010). Aujourd'hui, Twitter et plus largement les sites d'avis de consommateurs sont largement étudiés (Pak & Paroubek, 2010 ; Blake *et al.*, 2010 ; Eichstaedt *et al.*, 2015 ; Kumar & Bala, 2016 ; Sindhu & Vadivu, 2019). Les travaux actuels en TAL se fondent principalement sur les données issues du Web et s'intéressent peu aux avis exprimés à l'oral. Les données permettant ce genre d'analyse sont certes moins facilement disponibles mais présentent tout autant d'enjeux pour la

[92] https://github.com/aesuli/sentiwordnet
[93] https://wordnet.princeton.edu/
[94] http://www.lirmm.fr/~abdaoui/FEEL
[95] Editions : 2007, 2009, 2015, 2017 et 2018 - https://deft.limsi.fr/

problématique de la fouille d'opinion et de l'analyse de sentiments. De ce point de vue, ESLO constitue un corpus très utile et intéressant pour faire émerger de nouvelles approches dans le domaine.

3.2 Notion de *perception*

Il est communément admis que l'Homme dispose de cinq sens – la vue, l'ouïe, le toucher, le goût et l'odorat – grâce auxquels il est capable d'appréhender le monde qui l'entoure. Des récepteurs sensoriels, portés par des organes sensoriels dotés de mobilité (la main pour les récepteurs tactiles, les oreilles pour les récepteurs auditifs, etc.), réagissent aux stimulations de l'environnement en fonction de ces cinq sens (Gibson, 1966). Les récepteurs sensoriels agissent alors comme des « capteurs d'énergies véhiculant de l'information sur le monde » et « sont le point de départ du processus perceptif » (Luyat, 2014 : 16). L'analyse des informations collectées par les récepteurs de sens permet ainsi d'accéder à la connaissance de l'objet perçu. Le dictionnaire *Larousse* en ligne[96] ainsi que le *TLFi*[97], définissent la perception comme l'action de « percevoir par les organes des sens » ou comme une « idée, compréhension plus ou moins nette de quelque chose ». Ces acceptions sont précisées dans le *TLFi* avec l'idée que la perception est une « opération psychologique complexe par laquelle l'esprit, en organisant les données sensorielles, se forme une représentation des objets extérieurs et prend connaissance du réel » ou est une action « de prendre connaissance par l'intuition, par l'intelligence ou l'entendement ». Ces définitions démontrent que la perception est le résultat de la mise en relation d'un sujet avec un objet : depuis le processus de collecte d'information par le biais des sens jusqu'à la compréhension de l'objet qui en résulte.

La perception est donc caractérisée par une double dimension. En ce sens, Barbaras (2009) présente d'un côté la perception comme un « mode d'accès à la réalité telle qu'elle est en elle-même » et de l'autre côté, il donne à la perception un caractère *sensible*, dans le sens où elle est « l'épreuve que *je* [un individu] fais de la réalité » (2009 : 8). Un individu perçoit la réalité telle qu'elle existait avant qu'il ne la regarde et c'est par l'intermédiaire des récepteurs de sens qu'il peut *éprouver* son environnement. La perception est la conciliation

[96] https://www.larousse.fr/dictionnaires/francais/perception/59399?q=perception#59036
[97] http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?13;s=701919120;r=1;nat=;sol=2;

entre observation de la réalité et interprétation des observations réalisées. On retrouve l'idée de sensible chez Pradines (1946), pour qui « la perception est la représentation des choses situées dans l'espace, à travers de simples impressions sensibles ». La notion de *perception* est souvent associée à celle de *sensible* ou de *sensation*. En effet, lorsqu'une information sensorielle est détectée par un récepteur sensoriel, une sensation se produit. La perception recouvre un processus plus large par lequel le cerveau sélectionne, organise et interprète ces sensations. Selon Lechevallier (1995), les sensations sont la part sensorielle de la perception et sont associées à une interprétation « empreinte de subjectivité » des stimuli reçus.

> Dans la philosophie classique, la sensation est constituée par les qualités spécifiques d'un stimulus dans un système sensoriel donné, par exemple, dans le système visuel : la direction, la forme, les couleurs. La sensation concourt à distinguer un objet de son environnement, à maintenir sa permanence sensorielle dans son intégralité.
>
> Bernard Lechevallier, 1995 : 9

La question du lien entre *sensation* et *perception* anime les différents courants ayant abordé la notion de *perception*. Barbaras (2009) reprend par exemple la conception empiriste de Locke selon laquelle « la connaissance vient toute de l'expérience qui est faite de sensations ». Ce sont les sensations qui sont à l'origine de la connaissance : d'abord chronologiquement puisque nous commençons par avoir des sensations, puis logiquement car ce sont les sensations qui nous permettent d'élaborer nos idées. Locke déclare que « connaître ne [lui] semble rien d'autre que percevoir la connexion et la convenance ou le désaccord et la disconvenance entre n'importe lesquelles de nos idées » (*Essai,* IV, I, 2). Pour Locke, les idées que l'on peut avoir découlent des sensations que l'on éprouve. Aussi, la perception est l'action de l'esprit afin d'établir des liens entre les informations collectées par les sens afin de donner naissance aux idées. Il se place ainsi dans la lignée de Hume & Lévy-Bruhl (1740) selon qui toute idée dérive d'une impression. On ne peut avoir d'idées sans l'intermédiaire des organes de sens.

Les théories empiristes s'opposent aux théories intellectualistes qui considèrent que les idées préexistent dans notre esprit. Descartes (1724) par exemple présente la théorie que

les idées sont innées. Si tel est le cas, alors les organes des sens ne font que réveiller en nous les idées correspondantes et utiles à la construction de la pensée. Il illustre ses propos avec l'exemple de l'observation d'un morceau de cire avant et après chauffage. Une fois fondue, la cire n'a plus la même forme ou texture que lorsqu'elle était encore solide. Les sens seuls ne décrivent plus la même réalité, pourtant l'esprit sait tout de même qu'il s'agit du même objet. La perception est donc un processus intellectuel, une interprétation active et logique d'informations sensorielles. On ne perçoit pas par les sens, c'est la conscience qui donne une cohérence aux sens.

De nos jours, l'approche dominante est celle de la psychologie cognitive pour laquelle l'humain est envisagé comme un organisme dont la tâche principale est de traiter les informations qui proviennent de son environnement et de les intégrer afin de construire des représentations et des connaissances. Les mécanismes de la perception mais aussi de la mémoire, de l'intelligence ou encore de la conscience font partie des études menées dans ce cadre. Dans *Cognitive Psychology* (1967), Neisser définit le terme *cognition* :

> Le terme "cognition" renvoie à tous les processus par lesquels le stimulus sensoriel est transformé, élaboré, mémorisé, retrouvé et réutilisé. Le terme englobe ces processus, même lorsqu'ils opèrent en l'absence d'une stimulation pertinente, comme dans les images et les hallucinations. Des termes tels que sensation, perception, imagerie, rétention, mémorisation, résolution de problèmes et réflexion, entre autres, font référence à des étapes ou aspects hypothétiques de la cognition.[98]

Lorsque nous percevons un objet, ses propriétés ne se révèlent pas à nous immédiatement : l'information est traitée, mémorisée et forme une représentation qui pourra être retrouvée et réutilisée. Les sens permettent d'observer la forme, la couleur, l'odeur ou encore le goût que peut avoir un objet. Pour donner une cohérence à cette liste

[98] « *The term "cognition" refers to all processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used. It is concerned with these processes even when they operate in the absence of relevant stimulation, as in images and hallucinations. Such terms as sensation, perception. imagery. retention. recall. problem-solving. and thinking. among others. refer to hypothetical stages or aspects of cognition.* »

de caractéristiques, des processus cognitifs interviennent nécessairement dans la structuration et l'interprétation des informations.

Pour Lemaire & Didierjean, le système cognitif est comme un « système de traitement de l'information actif et non passif » (2018 : 26). Les informations ne sont pas enregistrées passivement mais plutôt manipulées, transformées par un « système symbolique actif ». Le domaine du traitement de l'information est issu de la psychologie cognitive. Selon Fortin & Rousseau (2015), l'approche du traitement de l'information a « comme principale caractéristique de considérer les processus mentaux comme une succession d'étapes » qui aboutiront à la mémorisation de l'information, offrant ainsi la possibilité de transférer cette information et de la récupérer ultérieurement. Dans une vision quantitative de l'information, Shannon (1948) considère le traitement de l'information comme la transformation d'une information *latente*, en une information *manifeste*. Du point de vue de la perception, il s'agit dans un premier temps d'identifier de manière formelle les informations nécessaires pour caractériser la perception, pour ensuite implémenter les algorithmes qui les rendront explicites et manipulables dans un système d'information.

3.3 Perception dans ce travail

La perception est une notion vaste que la linguistique, la psychologie et même les sciences de l'information ont tenté de circonscrire. L'étude des émotions, des sentiments et des opinions participe à la compréhension de la perception en décrivant ses différentes facettes.

Nous considérons la perception comme le processus de traitement d'informations recueillies à l'aide de récepteurs sensoriels dans le but de créer des représentations et des connaissances à propos de l'objet perçu. Percevoir un objet passe d'abord par son expérimentation physique et sensorielle. L'analyse immédiate de ces impressions, qui se rapprochent des émotions, construit les sentiments et prépare la construction d'une opinion à propos de l'objet. L'ensemble de ces traitements compose la perception et révèle la façon dont un individu appréhende l'objet perçu. La perception est une expérience qui varie d'un individu à l'autre et qui peut être captée lorsque l'individu décide de la partager. C'est cet instant de partage de la perception qu'il s'agit de détecter pour dresser le portrait de la ville d'Orléans.

Les travaux existant en TAL proposent des outils qui ont fait leurs preuves pour identifier la subjectivité et la polarité dans des ressources textuelles. Du point de vue de l'analyse de la perception à propos d'un lieu, les études sont rares. Le projet principal en TAL est Senterritoire[99], avec le développement d'OPILAND (*OPinion mIning from LAND-use planning documents*), une méthode pour « l'identification de la perception des territoires par la fouille de textes » (Kergosien *et al.* 2013, 2014, 2015). Ce projet consiste en l'extraction semi-automatique des entités nommées de type lieu et organisation, puis dans l'identification des opinions relatives à ces entités et de leur polarité.

Si les tâches classiques en TAL de la détection des opinions et de la polarité participent à l'analyse de la perception d'un lieu, elles ne suffisent pas à rendre compte de l'image qu'ont les habitants de leur ville. La méthodologie présentée propose de partir des méthodes classiques de détection de la subjectivité et de la polarité dans des données de nature différente, afin de proposer une nouvelle typologie de la perception dans la perspective d'un traitement automatisé plus fin.

Chapitre 4 : TRAITEMENT DE LA PERCEPTION RELATIVE A UN LIEU

4.1 Méthodologie générale

L'analyse de la perception relative à Orléans s'appuie sur les techniques d'apprentissage automatique supervisé et suit plusieurs étapes[100] :

- la modélisation de la perception pour l'élaboration d'une convention d'annotation (A),
- la division des transcriptions annotées en lieux en extraits susceptibles d'être porteurs d'éléments subjectifs (B),

[99] http://www.msh-m.fr/programmes/programmes-2013/senterritoire/
[100] cf. Figure 21

- l’annotation manuelle d’un échantillon des extraits définis pour la constitution d’un corpus d’entraînement, de test et d’évaluation (C),
- détection automatique de la perception par apprentissage supervisé (D)
- l’évaluation du module de détection de la perception (E),

De la même façon que pour le traitement des lieux, une étape de modélisation de la perception est nécessaire afin d’établir des conventions d’annotations (A). L’étape de détection des lieux mentionnés dans les conversations d’ESLO2 sert d’ancrage pour l’analyse de la perception. A partir de cette détection, des règles permettent de sélectionner des segments de transcriptions susceptibles d’être porteurs de la perception des locuteurs à propos des lieux mentionnés (B). La détection de la perception consiste d’abord en la distinction des segments subjectifs des segments objectifs. Il s’agit ensuite de déterminer le caractère positif ou négatif des segments jugés subjectifs. Les traitements réalisés pour la reconnaissance de la perception s’inscrivent dans le cadre de l’apprentissage automatique supervisé et suppose l’exploitation de données de référence annotées manuellement. Les segments de transcriptions nouvellement définis sont annotés manuellement en fonction des conventions d’annotation établies (C). Le sous-corpus obtenu est divisé en trois sous-corpus d’entraînement, de test et d’évaluation. Différents prétraitements sont appliqués sur le corpus d’entrainement et des classifieurs sont entrainés afin de détecter la subjectivité et la polarité dans les segments analysés (D). Enfin, les différents modèles entrainés sont évalués grâce au corpus d’évaluation afin de sélectionner le plus performant d’entre eux.

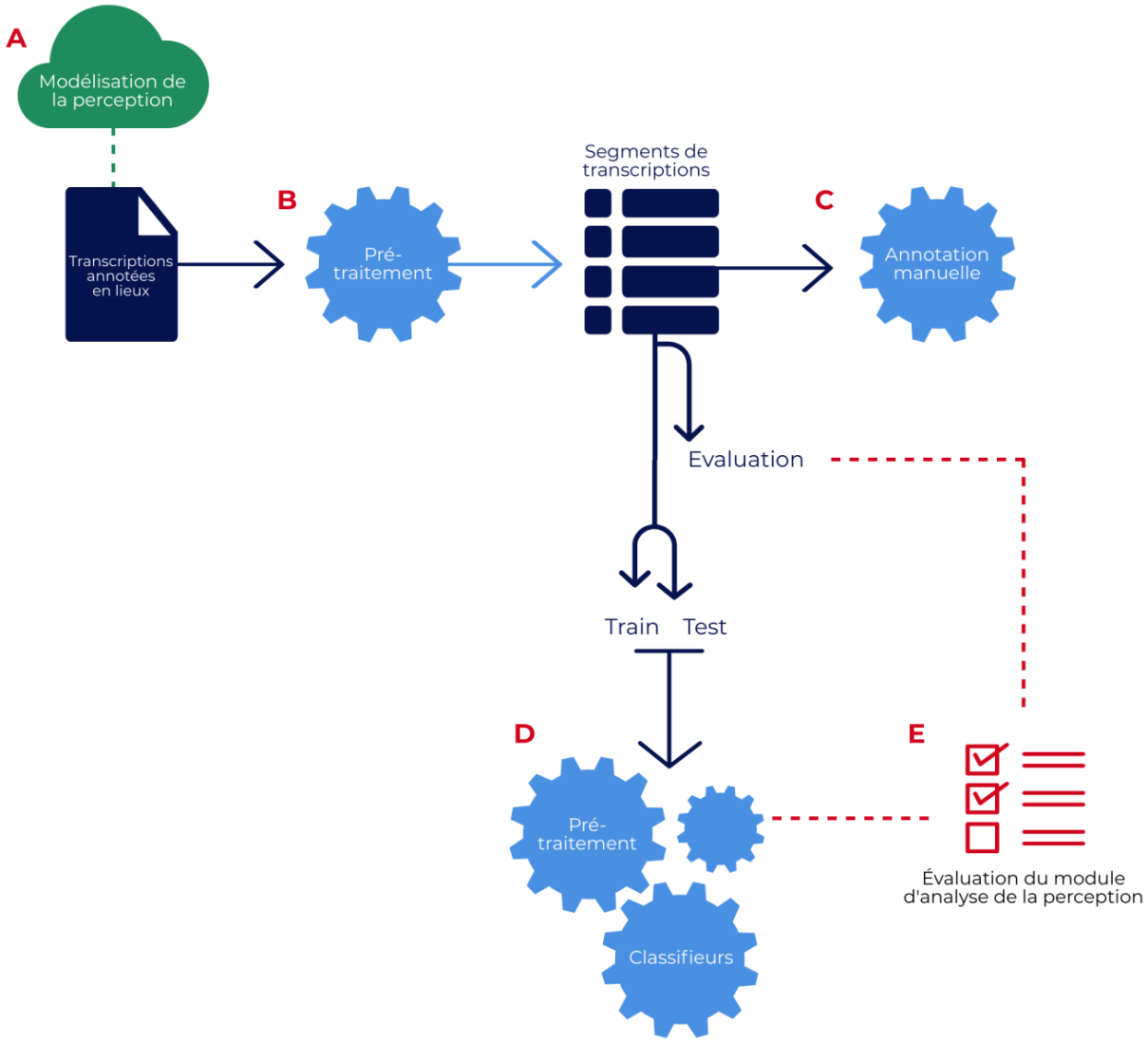

Figure 21 : Méthodologie pour l'analyse de la perception

4.2 Modélisation de la perception : conventions d'annotation

La méthodologie employée pour modéliser la subjectivité exprimée à propos d'un lieu est la même que celle que nous avions suivie pour la modélisation de l'information spatiale. A partir de l'exploration d'un échantillon du corpus annoté en lieux, des conventions d'annotation ont été établies. L'expression de la perception est un processus cognitif complexe qui peut transparaitre dans les émotions, sentiments ou opinions des locuteurs.

Les typologies possibles pour rendre compte de la perception sont multiples. Comme évoqué précédemment[101], la plupart des travaux en TAL portant sur l'analyse de sentiments et d'opinions s'intéressent d'abord à la classification des données en fonction de leurs polarité. Notre étude s'appuie sur des données rarement, voire jamais exploitées dans des travaux en TAL pour ce type d'analyse. Nous proposons donc d'explorer cette problématique et de nous intéresser à la détection de la subjectivité dans le français parlé et à la classification des énoncés en fonction de leur polarité. Cette étude de la subjectivité et de la polarité des énoncés est la première étape vers l'analyse de la perception.

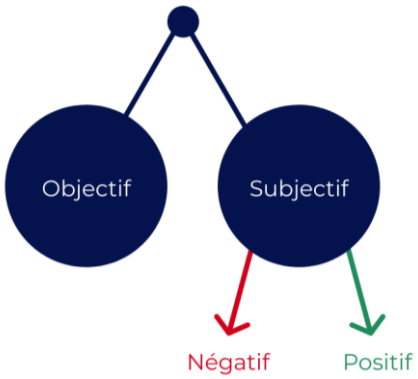

Figure 22 : Typologie de la subjectivité

La typologie utilisée pour caractériser la subjectivité et la polarité dans le corpus est présentée dans la Figure 22. Dans un premier temps, les énoncés objectifs doivent être distingués des énoncés subjectifs. Ensuite, la nature de la subjectivité est précisée au regard de la polarité. Chaque énoncé jugé subjectif est annoté en fonction de son orientation positive ou négative.

4.2.1 Objectivité et subjectivé

Comme défini dans la section 3.1, la subjectivité est le caractère de ce qui se rapporte à une conscience individuelle, par opposition à l'objectivité qui est la concordance entre plusieurs observations, une description du réel sans a priori. En fonction de ces définitions, il s'agira

[101] cf. section 3.1.4

dans l'annotation de commencer par faire la distinction entre les énoncés objectifs et les énoncés subjectifs.

(35) pendant ces ces cinq ans-là après j'ai repris des études à Paris pour passer mon DESS parce que je n'avais pas mon DESS et normalement là où j'avais été embauchée à Saint Jean-de-la-Ruelle (ESLO2_ENT_1039)

Beaucoup d'énoncés ne sont pas porteurs d'éléments subjectifs, ce qui les rend non pertinents pour l'analyse de la perception. Dans l'exemple 35, le locuteur décrit son parcours personnel et à cette occasion, mentionne *Paris* et *Saint-Jean-de-la-Ruelle* pour situer son récit. Dans ces extraits, il n'est pas à proprement parler question de ces deux villes. Le locuteur fait simplement état de ce qui lui est arrivé de façon très factuelle. On ne retrouve aucune trace d'émotion ou de sentiment, aucun marqueur d'intensité et aucune opinion n'est exprimée. Dans ces conditions, le tour est considéré comme *objectif*. De la même façon, l'exemple 36 ne présente aucun élément subjectif, et encore moins à propos du lieu *quartier* qui a été identifié. Ce tour de parole doit lui aussi être annoté comme *objectif*.

(36) et pratiques plurilingues comme on appelle ça est-ce que vous entendez parler d'autres langues que le français dans dans le quartier ou (ESLO_ENT_1018)

Les lieux *médiathèque*, *bibliothèque* et *Orléans* ont été identifiés dans l'exemple 37. Cette fois, on peut retrouver différents éléments subjectifs. L'expression méliorative *très bien* et les adjectifs *grande* et *centrale* qualifient les lieux *médiathèque* et *bibliothèque*. Le locuteur commente les services et l'offre de la bibliothèque.

(37) oui c'est-à-dire que c'est la l- la plus grande enfin c'est la seconde **médiathèque** après la **bibliothèque** centrale oui donc celle d'**Orléans** oui non non très bien et puis ils ont un choix de livres euh vraiment parce que s'ils l'ont pas là ils peuvent

bon ils font des échanges entre eux si on l'a pas on peut le commander (ESLO2_ENT_1016)

En décrivant la médiathèque d'Orléans, le locuteur de cet extrait ne reste pas factuel. Il choisit de mettre l'accent sur certaines caractéristiques du lieu, il le positionne par rapport à d'autre et le qualifie même de *très bien*. Cet énoncé doit donc être annoté comme *subjectif*.

4.2.2 Polarité

Lorsqu'un énoncé est jugé subjectif, l'annotateur qualifie cette subjectivité en lui attribuant une polarité. Nous proposons de garder une typologie générale qui ne conserve que les degrés positif et négatif, sans distinction d'intensité.

Les étiquettes positive et négative sont en totale opposition. On considère comme positive toute expression en faveur de quelque chose, qui peut avoir une connotation approbatrice, méliorative ou bien qui véhicule des sentiments et des émotions comme la joie, le bonheur, ou le plaisir. Les énoncés positifs peuvent mentionner des éléments jugés bénéfiques, souhaitables, constructifs, etc. Dans la notion de positif, il y a l'idée de bien, voire de mieux par rapport à une référence. L'inverse exact de ces affirmations décrit des expressions jugées négatives. Les énoncés négatifs renvoient aux notions de désapprobation, de désavantage, de mécontentement, de regret etc.

Dans l'exemple 38, le locuteur parle des endroits dans lesquels il a l'habitude de faire ses courses. En l'occurrence, il explique qu'il aime faire ses courses dans des grandes surfaces comme *Leclerc* et *Auchan*, qu'il juge très pratiques. Cet énoncé porte l'avis positif du locuteur à propos de ces magasins : ce sont des lieux dans lesquels il choisit de se rendre, il le fait avec plaisir et il choisit de partager cette information avec son interlocuteur. Cet énoncé doit donc être annoté comme positif.

(38) Le- **Leclerc Leclerc Auchan** oui les grandes surfaces je suis très **grandes surfaces** oui c'est c'est très pratique (ESLO2_ENT_1016)

A l'inverse, le locuteur de l'exemple 39 déplore le manque de vie étudiante à *Orléans*. Cette affirmation implique que l'énoncé doit être annoté comme négatif, tout comme l'exemple 40. Le locuteur ne fait pas qu'expliquer le fonctionnement du stationnement dans la ville, il ajoute à sa description une connotation négative sur la *rue Bannier* dans laquelle on peut se faire *aligner*. Le locuteur accentue son mécontentement en répétant *ils passent tous les tous les tous les jours*.

(39) je on était quand même assez étonné que bon cette faculté s- c'est vrai qui était sur **la Source** et vous aviez **aucune vie étudiante sur Orléans** je sais pas je sais pas si c'est toujours pareil mais
(ESLO2_ENT_1016)

(40) ouais c'est payant en fait c'est gratuit à partir de dix-neuf heures donc les horaires de débauche quoi donc ça ça va bien par contre le week-end si tu la laisses **tu te fais aligner quoi rue Bannier ils passent tous les tous les tous les jours**
(ESLO2_ENT_1019)

4.3 Préparation du corpus avant l'analyse

Afin d'analyser la perception à propos des lieux, il n'est pas nécessaire d'analyser l'intégralité des transcriptions. En effet, l'ensemble des conversations analysées ne portent pas seulement sur la ville d'Orléans et donc, les déclarations subjectives émises par les locuteurs au fil de la conversation ne concernent pas forcément des lieux. Pour optimiser le traitement et isoler les déclarations subjectives au sujet d'un lieu de celles qui portent sur un autre sujet, les transcriptions sont segmentées en fonction des lieux identifiés.

L'hypothèse est faite que lorsqu'un locuteur exprime sa perception de son environnement, il va mentionner le nom de lieux pour situer son discours. Sur cette hypothèse, des règles sont utilisées pour définir des fenêtres d'observation autour des lieux précédemment détectés[102]. Ainsi, pour chaque lieu détecté, une segmentation est réalisée afin de répondre à deux exigences. Elle doit être suffisamment restrictive pour s'assurer que le contenu de

[102] cf. Figure 23

la conversation porte bien sur le lieu observé. Elle doit aussi être complète dans le sens où la fenêtre d'observation établie doit être suffisamment riche en contenu pour pouvoir contenir l'expression de la perception.

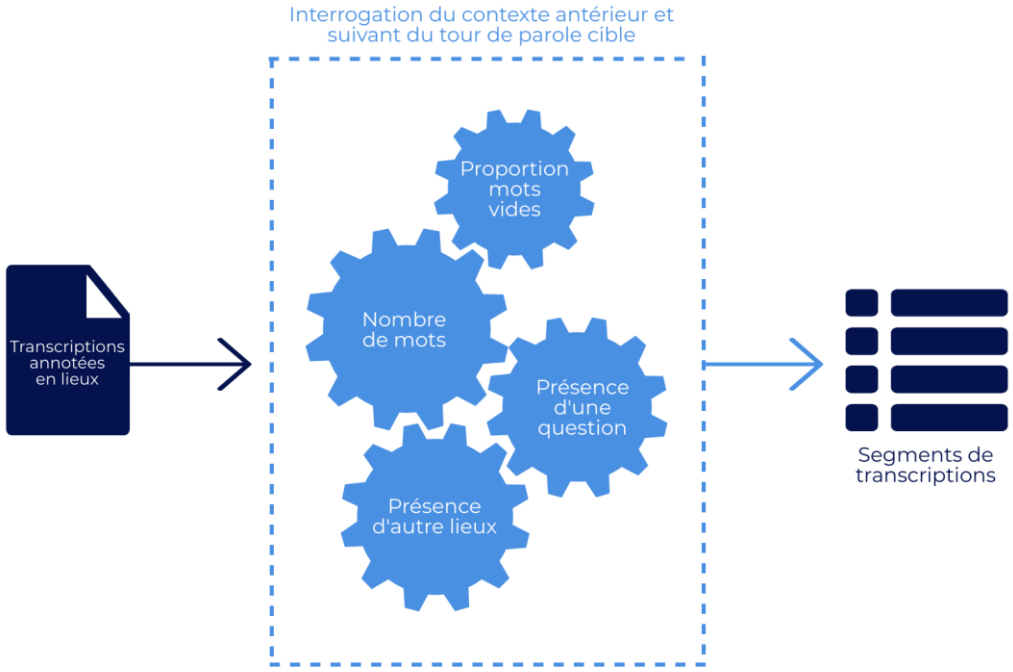

Figure 23 : Segmentation des transcriptions annotées en lieux

Différents points sont considérés pour guider la segmentation : le nombre de mots, la présence d'autres mentions de lieux et de question dans le tour de parole analysé et son contexte. Chacun de ces points sont détaillés dans les sections suivantes.

4.3.1 Nombre de mots :

Dans un premier temps, on considère la longueur en nombre de mots des tours de paroles. Leur taille peut être très variable, tout comme la nature de leur contenu. Un grand nombre de mots n'est pas forcément synonyme de richesse dans le contenu. Un tour de parole peut contenir beaucoup de mots vides (articles, prépositions, etc.) et de nombreuses disfluences,

ce qui à dilue la pertinence de son contenu. Ainsi, pour considérer un tour de parole comme suffisamment riche, un ratio est calculé entre le nombre de mots vides et ceux qui ne le sont pas. Nous considérons que le tour est suffisamment riche s'il contient plus de trente mots qui ne sont ni des disfluences, ni des mots vides. Dans le cas contraire, il est nécessaire de considérer les tours de paroles antérieurs et suivants.

4.3.2 Présence d'autres lieux :

Pour répondre à l'exigence de restriction et obtenir la fenêtre d'observation la plus pertinente possible, il est nécessaire de s'intéresser à la présence d'autres lieux. La mention de plusieurs lieux dans un même contexte peut créer de l'ambiguïté au moment de l'analyse de la perception.

(41) bah il doit y avoir des trucs jolis je pense hein y a il me semble en plus qu'il y a une forêt pas loin la **forêt d'Orléans** (ESLO2_ENT_1047)

Dans l'exemple 41, le lieu *forêt d'Orléans* a été identifié. Le tour de parole est composé de trente mots parmi lesquels se trouvent des disfluences (*bah*, *hein*) et des mots vides (*il*, *en*, *qu'*, etc.) ce qui signifie que le tour est considéré comme trop court. Il s'agit donc d'interroger le contexte du tour dans la limite de deux tours précédents et deux tours suivants.

(41) LD47: plutôt ouais il paraît qu'y a des euh c'est un petit coin sympa je sais pas hm
ch_MP10: des jolies promenades à faire ouais ouais ouais

LD47: bah il doit y avoir des trucs jolis je pense hein y a il me semble en plus qu'il y a une forêt pas loin la **forêt d'Orléans**

LD47: donc euh oui enfin sinon j'aime bien me promener mais bon comme je suis tout seul sur Orléans euh que mes collègues de boulot euh la plupart je m'entends bien avec mais ils ont déjà une f- une vie de famille des enfants euh une femme euh
(ESLO2_ENT_1047)

On peut remarquer que les deux tours de parole précédents ne mentionnent pas d'autres lieux. Ils sont donc jugés pertinents pour la description de la *forêt d'Orléans* et sont ajoutés à la fenêtre d'observation. Si l'on observe le tour suivant, on remarque que celui-ci mentionne le lieu *Orléans*, soit un lieu différent du lieu cible *forêt d'Orléans*. Dans ce cas, nous choisissons de l'écarter. De cette façon, nous sommes en mesure d'éliminer les tours de parole dans lesquels il n'est plus question du lieu cible.

De nombreux tours de parole contiennent plusieurs noms de lieux. Dans ces cas là, on crée une fenêtre d'observation commune à tous les lieux mentionnés dans le tour.

Ainsi, dans l'exemple 42, cinq lieux sont mentionnés : *université*, *la Source*, *Orléans*, *fac* et *ville*. Les deux tours de parole suivants et précédents pourront être ajoutés à la fenêtre d'observation s'ils ne contiennent pas d'autres noms de lieux que les cinq présents dans le tour cible. En l'occurrence, les quatre tours considérés mentionnent *Orléans* et le nom commun *ville* : ils sont donc ajoutés à la fenêtre d'observation.

(42) JR18: oui oui oui hm moi qui connais quand même oui depuis très longtemps maintenant euh oui je trouve que quand même Orléans euh bouge un peu quand même et euh donc c'est bien quoi on arrête d'être euh même si c'est encore ça reste quand même un peu une ville morte quand même je trouve
ch_NS3: hm hm

> JR18: euh on a toujours dit que le problème de l'**université** à **la Source** et laisser **Orléans** euh ça a toujours été ouais je fais partie des gens ouais j'aurais préféré quand même que l'**université** la **fac** tout ça se trouvent un petit peu plus sur **Orléans** qu'à **la Source** euh ça aurait sûrement amené un plus à la **ville**

ch_NS3: hm hm
JR18: [bb] euh mais bon je trouve que les Orléanais changent un peu quand même m- mais que c'est quand même pas une ville euh top au niveau euh ça bouge sans plus on va dire quoi ça bouge plus que ça bougeait mais voilà quoi
(ESLO2_ENT_1018)

4.3.3 Présence de questions

Les tours de parole qui sont des questions ont un traitement particulier. Selon les conventions de transcription du corpus ESLO, les questions sont marquées avec un point d'interrogation. Dans le cas d'un tour de parole simple jugé trop court, nous choisissons de ne pas ajouter à la fenêtre d'observation les tours précédents et suivants qui seraient des questions.

Cependant, bon nombre de lieux sont mentionnés dans des questions qui peuvent appeler une réponse subjective à propos du lieu. Aussi, lorsqu'un lieu est mentionné dans une question, nous avons décidé d'y adjoindre le tour de parole suivant, voire le deuxième tour de parole suivant mais d'écarter les tours précédents pour établir la fenêtre d'observation.

(43) ch_LA11: est-ce qu'il y a une façon de parler propre à Orléans ? (ESLO2_ENT_1041)

Dans l'exemple 43, le locuteur ch_LA11 pose une question à propos d'*Orléans*. Par rapport au nombre de mots, cette question est considérée comme trop courte et il est nécessaire de regarder le contexte. Puisque le lieu est mentionné dans une question, on choisit d'écarter les tours de parole précédents qui seront moins susceptibles d'être pertinents par rapport aux tours suivants qui contiendront la réponse. En l'occurrence, les deux tours de parole suivants sont :

(43)
ch_LA11: est-ce qu'il y a une façon de parler propre à Orléans ?
BC41: ouais [rire] oui y a des mots #1 y a des mots y a m- ah de ma génération à moi
ch_LA11: oui y a des mots ? (ESLO2_ENT_1041)

Le tour suivant est considéré comme pertinent pour faire partie de la fenêtre d'observation finale. Par contre, le deuxième tour suivant est une question. Celui-ci ne sera donc pas ajouté à la fenêtre d'observation correspondant au lieu *Orléans*.

A partir de ces traitements, les segments définis sont associés aux informations recueillies lors de la détection des lieux et sont stockés dans un fichier au format CSV.

4.4 Constitution du corpus de référence

4.4.1 Sélection de données à annoter

La détection de la perception est réalisée par apprentissage automatique supervisé. L'objectif de l'apprentissage automatique est de comprendre la structure des données et de les intégrer dans des modèles destinés à répondre à la problématique posée. Comme évoqué dans la section 2.4.1.1, les algorithmes d'apprentissages automatiques supervisés s'appuient sur des données préalablement annotées pour apprendre à réaliser la tâche qu'on

leur a attribuée. Pour que ces méthodes soient efficaces, il faut disposer d'un grand nombre de données annotées manuellement pour former un corpus de référence. Lors de la phase d'apprentissage automatique, le corpus de référence est divisé en trois ensembles utiles à différentes étapes de la création du modèle :

- le **corpus d'entrainement** qui se compose au minimum d'une paire composée d'un vecteur d'entrée et du vecteur de sortie correspondant, communément appelé cible. Le modèle s'entraine à associer chaque vecteur d'entrée avec l'étiquette attendue.
- le **corpus de validation** qui est utilisé pour tester l'efficacité du modèle entrainé sur un nouveau jeu de données. La répartition des cibles le composant est souvent proportionnelle à celle du corpus d'entrainement. Il fournit une évaluation impartiale de l'ajustement d'un modèle en comparant les décisions prises par le modèle avec la référence annotée manuellement.
- le **corpus de test** ou *holdout dataset* est un ensemble de données sans a priori sur la répartition des cibles, qui n'a jamais été utilisé dans le corpus d'entrainement ou de validation. Lorsque l'évaluation réalisée sur le corpus de validation indique que le modèle est suffisamment stable, on réalise une nouvelle évaluation sur le corpus de test pour confirmer les scores obtenus.

Dans cette perceptive, trente transcriptions ont été sélectionnées pour être annotées manuellement en fonction des conventions d'annotation établies et constituer un nouveau corpus de référence. Les quinze premières transcriptions retenues pour former les trois ensembles de données sont celles qui constituaient déjà le corpus de référence utilisé pour évaluer la détection des lieux[103]. Dans cette série de transcriptions, 2290 lieux avaient été annotés. A cette série s'ajoutent quinze autres transcriptions, elles-aussi issues des modules Entretiens et Itinéraire d'ESLO2. Parmi ces nouvelles transcriptions, 2224 mentions de lieux ont été annotées grâce au module développé. Au total, 4516 lieux ont été identifiés dans ce nouveau sous-corpus.

[103] cf. section 2.3

A partir de ces détections, les trente transcriptions sont segmentées en fonction des règles décrites dans la section précédente. Le traitement permet de créer 3178 fenêtres d'observation qui seront utilisées pour l'analyse de la perception.

4.4.2 Processus d'annotation manuelle et son évaluation

Le corpus de référence doit être annoté manuellement en fonction de la convention établie pour les besoins de l'analyse de la perception. Cette annotation est réalisée sur les fenêtres d'observations définies lors de l'étape de segmentation du corpus d'entraînement et d'évaluation. Elle doit renseigner sur le caractère subjectif ou non-subjectif de chacune des fenêtres d'observation et, le cas échéant, sur sa polarité. Les annotations ont été directement réalisées dans le fichier CSV produit après la phase de segmentation. La structure de ce fichier est illustrée dans le Tableau 17.

Chaque ligne du tableau représente une fenêtre d'observation, soit un segment de transcription dans lequel a été annoté un lieu. L'information du *code segment* et du *locuteur* servent à garder une trace de la place du segment dans le reste de la transcription. Les colonnes *label*, *type* et *zone* correspondent aux métadonnées recueillies à propos du lieu cible lors de la phase de détection des lieux. L'annotation de la subjectivité et de la polarité est faite dans la colonne *Polarité.* Si le segment est jugé objectif, il est annoté comme *none*. S'il est jugé subjectif, on annotera directement la polarité correspondante (*positif* ou *négatif*).

Pour réaliser l'annotation du corpus, deux annotatrices (A1 et A2) ont été mobilisées. L'annotatrice A1 peut être considérée comme experte dans la tâche d'annotation de la perception[104]. L'annotatrice A2 est étudiante en troisième année de Licence de Psychologie à l'Université de Poitiers. Son cursus universitaire l'a initiée aux concepts de subjectivité ainsi qu'aux méthodologies d'annotation dans le cadre d'un protocole de recherche. En formation au moment de la phase de l'annotation, elle reste considérée comme une débutante avertie pour réaliser le travail demandé.

Code segment	Locuteur	Segment	lieu	label	type	zone	Polarité

[104] L'annotatrice A1 est l'auteure de cette thèse.

ESLO2_iti_10_03_TDP82	VJ757	mon lieu préféré dans la ville ça serait quand même du quartier cathédrale euh/voilà justement oui c'est très joli	Cathédrale	cathédrale Sainte-Croix d'Orléans	monument	2	Positif
ESLO2_ENT_1047_TDP63	ch_MP10 et LD47	arrivé à Orléans vous êtes arrivé voilà ouais tout de suite ici oui tout de suite ici un gros coup de chance parce que bon/je trouve qu'il est pas dégueulasse comme appart-	Orléans	Orléans	ville	0	Positif
ESLO2_ENT_1047_TDP71	LD47 et ch_MP10	rue du Martroi ouais le seul petit défaut ouais/c'est les places de parking c'est vrai que là euh	rue du Martroi	Rue du Martroi	voie	0	negatif
ESLO2_ENT_1047_TDP706	LD47	c'est dans les Alpes du nord/d'accord	nord	Nord	region	2	none

Tableau 17 : Structure du fichier CSV annoté en subjectivité et polarité

Avant de procéder à l'annotation complète du corpus de référence, un accord inter-annotateur a été mesuré selon la procédure décrite dans la section 3.3.3. Cinquante segments ont été aléatoirement sélectionnés et annoté par les deux annotatrices. Un Kappa de Cohen de 0,85 a été calculé à partir de la comparaison des deux annotations. Selon la grille d'interprétation proposée par Landis et Koch (1997), ce score est considéré comme excellent. Après une concertation entre les deux annotatrices, le reste du corpus de référence a pu être annoté.

4.4.3 Analyse quantitative du corpus de référence

Trente transcriptions ont donc été sélectionnées et annotées manuellement par annotatrice afin de constituer le corpus de référence qui sera utilisé pour entraîner et évaluer le modèle développé pour la détection de la perception.

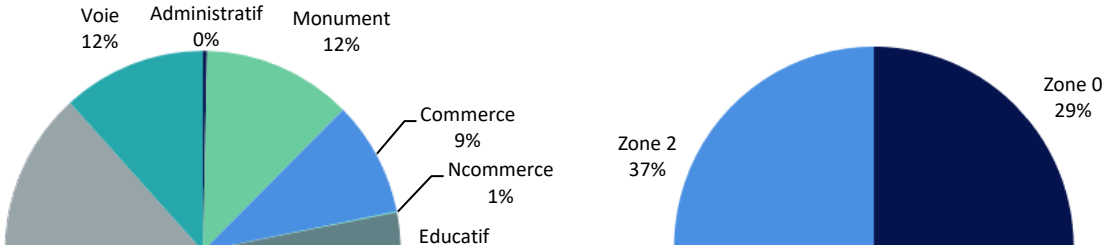

Figure 24 : Répartition des types de lieux et des zones dans le corpus de référence

La figure 24 présente la répartition des lieux parmi les 3178 entrées du corpus de référence. On peut constater une répartition plutôt équilibrée entre les trois zones géographiques définies. La majorité des segments mentionnent tout de même une majorité de lieux situés à Orléans (37%, soit 1175 segments). Les types de lieux suivent une répartition similaire à celle décrite dans le corpus de référence utilisé pour l'évaluation du module de détection des lieux. On compte une majorité de *villes* 44%, de *voies* 12%, de *monuments* 12% et de *commerces* 9%. Les lieux *administratifs* et *supranationaux* restent rares.

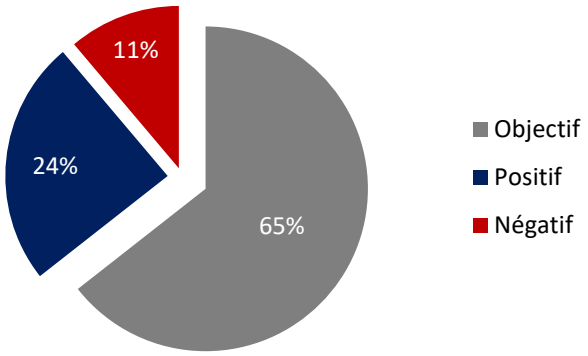

Figure 25 : Répartition de la subjectivité et de la polarité dans le corpus de référence

En ce qui concerne la subjectivité, 1112 expressions sont subjectives et représentent 35% des segments du corpus de référence[105]. Sur le total des segments, 24% ont une connotation positive et 11% ont une connotation négative. Sur les 1175 expressions concernant des lieux situés à Orléans, 479 sont subjectives.

4.5 Détection de la perception par apprentissage supervisé

4.5.1 Schéma général des étapes de traitement

La détection de la perception par apprentissage supervisé est réalisée sur les segments de transcription définis autour des noms de lieux détectés. Un échantillon de segments a été annoté manuellement en fonction des conventions d'annotation de la perception décrites dans la section 4.2. Le sous-corpus ainsi obtenu est divisé en trois sous-corpus d'entraînement, de validation et de test. L'entraînement du modèle est réalisé à partir du sous-corpus d'entrainement et de validation, avant d'être évalué avec le sous-corpus de test.

La détection de la perception est réalisée en deux temps. D'abord, les segments sont annotés selon leur caractère subjectif ou objectif. Ensuite, les segments jugés subjectifs sont annotés selon l'opposition positif/négatif. Que ce soit pour l'une ou l'autre tâche de détection, différentes étapes sont nécessaires pour que l'algorithme puisse s'entraîner à réaliser les tâches demandées. La première étape consiste en l'application de prétraitements sur les données afin de collecter différentes informations linguistiques. La lemmatisation et l'étiquetage morpho-syntaxique (POS Tagging) et le calcul de scores de polarité d'émotions sont présentés dans la section 4.5.2. Les segments analysés sont des données textuelles. Pour que l'algorithme puisse les interpréter, celles-ci doivent être transformées dans une représentation vectorielle. Plusieurs méthodes existent pour obtenir la forme algébrique du texte et seront discutées dans la section 4.5.3. Enfin, des classifieurs sont aussi testés pour réaliser la détection de la subjectivité et de la polarité dans la section 4.5.4.

[105] cf. Figure 25

La façon de combiner les prétraitements et les *features* (traits) lors de la classification permet d'atteindre les deux objectifs en suivant la même méthodologie.

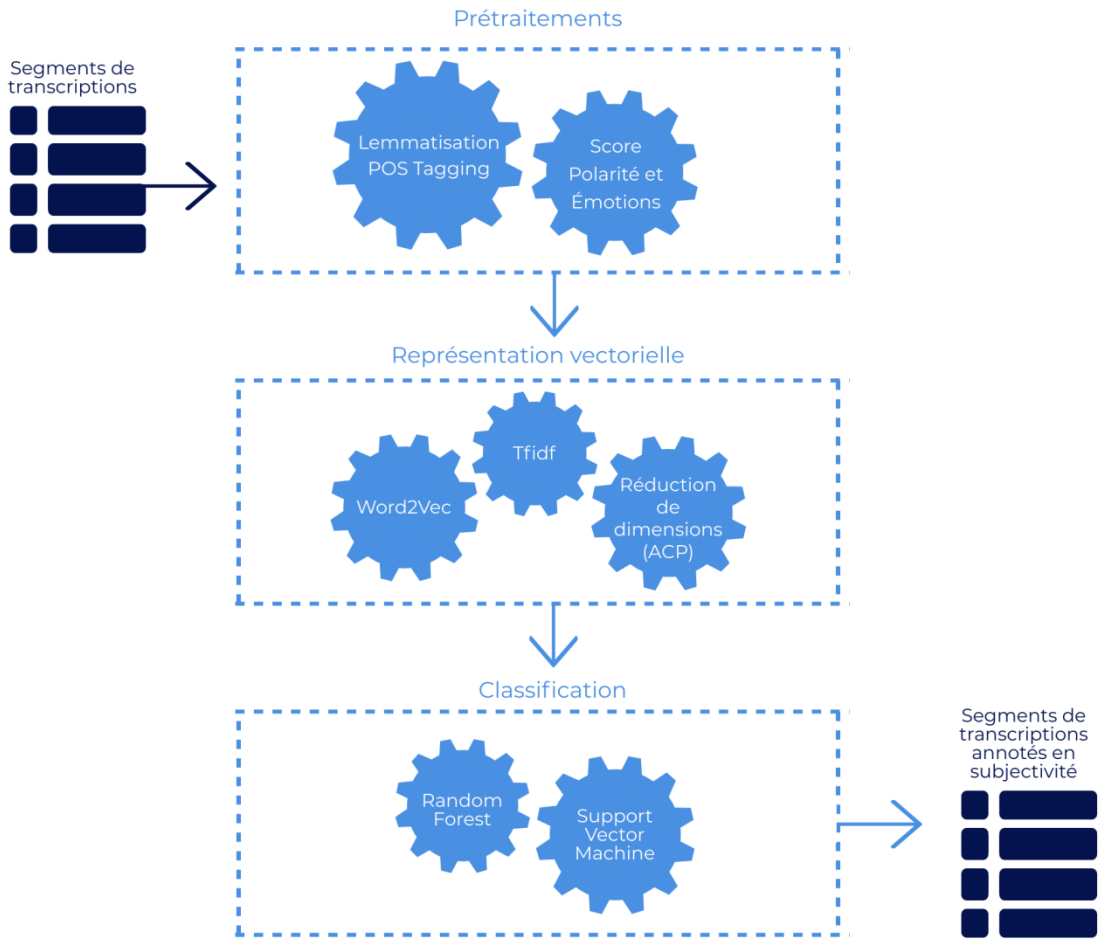

Figure 26 : Processus de détection de la subjectivité et de la polarité

4.5.2 Prétraitements et traits linguistiques

Le nettoyage et le prétraitement des données sont des tâches essentielles dans un processus d'apprentissage automatique pour obtenir des modèles prédictifs performants. Selon les tâches réalisées, les données exploitées proviennent de sources différentes et peuvent contenir des anomalies ou des valeurs incorrectes compromettant la qualité du jeu de données. Pour éviter de traiter des données erronées, il faut impérativement analyser les données, détecter les éventuelles anomalies et déterminer quelles sont les étapes de prétraitement et de nettoyage les plus appropriées.

Dans le cas d'exploitation de données textuelles issues du Web, ces questions sont particulièrement importantes. Lors de la récupération des données, certaines entrées peuvent être incomplètes ou mal formées. Par exemple, selon l'objectif du modèle entraîné, la présence d'émojis ou la répétition de signes de ponctuation peuvent bruiter l'analyse et donc nécessiter des traitements particuliers. Il peut être aussi nécessaire de normaliser l'accentuation ou la casse des textes analysés.

Notre étude exploite des transcriptions de conversations orales. Celles-ci sont transcrites orthographiquement, ne contiennent pas de signes de ponctuation[106], et ne sont donc pas concernées par les problématiques classiques de prétraitement des données. Il n'est donc pas utile pour la tâche de détection de la perception de contrôler la casse, l'accentuation ou la présence de caractères particuliers comme les émojis.

Les prétraitements que nous réalisons consistent plutôt en la sélection de traits linguistiques (*features*). Ces traits linguistiques seront donnés en entrée lors de l'entraînement de l'algorithme.

4.5.2.1 Lemmatisation et étiquetage morpho-syntaxique

Le premier prétraitement réalisé est la lemmatisation des segments analysés. La lemmatisation consiste à retrouver la forme canonique de tous les mots composant une phrase. Ainsi, les marques d'accord, de genre ou temps disparaissent du texte. La plupart des modèles d'apprentissage automatique utilisent comme *feature* le nombre de mots des textes analysés. L'intérêt de la lemmatisation est de pouvoir compter les formes fléchies de chaque mot plutôt que de compter distinctement un même verbe, conjugué de plusieurs façons. Dans l'exemple 44, on trouve les formes *l'* et *les* du déterminant *le*. Après lemmatisation, ces formes ne seront pas considérées comme deux déterminants différents mais bien comme deux occurrences du déterminant *le*.

(44) j'en fais avec qu'une une amie j'en fais y a ma petite voisine qui a huit ans qui vient qui s'est qui est accro à

[106] Hormis les points d'interrogation pour marquer les questions.

```
l’origami donc je lui montre aussi qu’est-c- je vais tous les
quinze jours    (ESLO2_ENT_1039)
```

Plusieurs bibliothèques Python permettent de lemmatiser des textes en français :

- **NLTK** (Natural Language Toolkit)[107] : bibliothèque open source dont le modèle français a été entrainé sur un corpus du journal *Le Monde* (Loper & Bird, 2002).
- **Spacy**[108] : bibliothèque open source développée par Honnibal (2015). Les modèles pour le français ont été entraînés sur le corpus UD French Sequoia[109] et le WikiNER[110].
- **Treetaggerwrapper**[111] : implémentation de l’outil Treetagger[112] (Schmidt, 1994) dans une bibliothèque Python par Pointal en 2004.

D’une manière générale, les librairies permettant la lemmatisation sont très performantes pour traiter l’anglais et moins pour ce qui est du français. La lemmatisation de quelques exemples par les différents lemmatiseurs cités permet de mettre en lumière leurs différences de performance.

Forme attendue	Je en faire avec que un ami je en fais y avoir mon petit voisin qui avoir huit an qui venir qui se être qui être accro à le origami donc je lui montrer aussi que être - c- je aller tout le quinze jour
Lemmatisation par Spacy	je en fai avec que un un amie je en faire y avoir mon petit voisin qui avoir huit an qui venir qui se être qui être accro à le origami donc je luire montre aussi que être - c- je aller tout le quinz jour

[107] https://www.nltk.org/index.html
[108] https://spacy.io/
[109] Corpus français, provenant d'*Europarl*, du corpus l'*Est Republicain*, de la *Wikipedia Fr*, et de l'agence européenne du médicament (documents extraits du corpus EMEA) manuellement annotée pour les catégories morpho-syntaxiques et la structure syntagmatique, en suivant les guides d'annotation du French Treebank. https://universaldependencies.org/treebanks/fr_sequoia/index.html
[110] Annotation en entités nommées de l’intégralité de Wikipédia. http://schwa.org/projects/resources/wiki/Wikiner
[111] https://treetaggerwrapper.readthedocs.io/en/latest/
[112] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Lemmatisation par Treetagger	Je en faire avec que un ami je en fais y avoir mon petit voisin qui avoir huit an qui venir qui se être qui être accro à le origami donc je lui montrer aussi que être - c- je aller tout le quinze jour

Tableau 18 : Comparaison de la lemmatisation réalisée par Spacy et Treetgagerwrapper

Dans le tableau 18, on peut voir les lemmatisations réalisées par Treetgagerwrapper et Spacy sur l'exemple 44. Dans la lemmatisation faite par Spacy, les terminaisons des mots *fais* et *quinze* ont été supprimées (*fai, quinz*) mais les formes obtenues ne sont pas des lemmes possibles. Le terme *amie* n'a pas été ramené à sa forme au masculin (*ami*) et le mot *lui* n'a pas été reconnu comme un pronom et a été lemmatisé en tant que le verbe *luire*. La lemmatisation opérée par Treetagger correspond à la forme attendue et ne présente pas d'anomalie.

Segment lemmatisé	Segment étiqueté en POS
Je en faire avec que un ami je en fais y avoir mon petit voisin qui avoir huit an qui venir qui se être qui être accro à le origami donc je lui montrer aussi que être - c- je aller tout le quinze jour	PRO:PER PRO:PER VER:pres PRP KON DET:ART NUM DET:ART NOM PRO:PER PRO:PERVER PRO:PER VER:pres DET:POS ADJ NOM PRO:REL VER:pres NUM NOM PRO:REL VER:pres PRO:REL PRO:PER VER:pres PRO:REL VER:pres NOM PRP DET:ART NOM ADV PRO:PER PRO:PER VER:pres ADV KON ADV ADJ PRO:PER VER:pres PRO:IND DET:ART NUM NOM

Tableau 19 : Lemmatisation et étiquetage morpho-syntaxique par Treetgagerwrapper

La lemmatisation opérée par Treetgagerwrapper semble plus fiable et est donc préférée pour notre développement. L'intérêt supplémentaire dans ce choix est que Treetgagerwrapper permet aussi de réaliser l'étiquetage morpho-syntaxique (*Part-of-Speech Tagging* – POS) des textes. Chaque mot du texte est associé aux informations grammaticales le caractérisant (genre, nombre, partie du discours). Ainsi, chaque segment analysé composant le corpus de référence est d'une part lemmatisé puis annoté en POS comme illustré dans le tableau 19.

4.5.2.2 Score de subjectivité

Comme évoqué dans la section 3.1.4, des dictionnaires de termes annotés en polarité et en émotions ont été constitués dans le cadre d'approches symboliques en fouille d'opinion et réutilisés pour des tâches d'apprentissage automatique supervisé. Le lexique FEEL compile un dictionnaire de 14127 termes français annotés en polarité (positif et négatif) et en émotions (joie, peur, tristesse, colère, surprise et dégoût).

Grâce à ce lexique, il est possible de calculer un score de subjectivité pour une phrase donnée. Ce score est égal au nombre total de mots qui contiennent de la subjectivité dans le document. La bibliothèque python PyFeel[113] propose de calculer ce score en exploitant le lexique FEEL. Nous avons repris cette librairie pour calculer un score de subjectivité pour chacun de nos segments. Si l'on observe les exemples suivants :

(45) je préfèrerais faire les bords de la Loire c'est vrai que c'est agréable euh c'est agréable (ESLO2_ENT_1018)

(46) mais ici ouais trop plat oui oui ils sont moins intéressés oui oui/ça va comme hier qu'on a fait une heure en forêt ça allait mais faire trois heures de marche à pied là même au bord de la Loire euh c'est bon quoi/oui/ça leur convient (ESLO2_ENT_1016)

PyFeel donne les scores suivants :

Ex.	Positif	Négatif	Colère	Dégoût	Peur	Joie	Tristesse	Surprise
12	0.7	0.0	0.0	0.0	0.0	0.7	0.0	0.7
13	0.0	0.6	0.0	0.0	0.1	0.0	0.2	0.1

Ces exemples sont des déclarations dans lesquelles la polarité est assez claire. L'exemple 45 est positif et les scores obtenus par PyFeel vont dans ce sens, de la même façon que pour l'exemple 46 qui a une connotation plutôt négative. Si ces scores sont représentatifs de la polarité exprimée dans ces exemples, on ne peut pas être certain que ce soit toujours le cas.

[113] https://github.com/AdilZouitine/pyFeel

En effet, le calcul du score de subjectivité s'appuie uniquement sur le lexique sans considération du contexte.

(47) si y avait vraiment la crue de la Loire ça serait une catastrophe (ESLO2_ENT_1016)

Dans l'exemple 47, le locuteur parle de Loire et du risque de crue qui serait une *catastrophe* pour la ville. Le terme *catastrophe* a une connotation fortement négative que l'on pourrait éventuellement retrouver dans le terme crue. Pourtant, PyFeel donne la description suivante de cet exemple :

Ex.	Positif	Négatif	Colère	Dégoût	Peur	Joie	Tristesse	Surprise
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Aucune émotion n'a été identifiée dans cet extrait pour la raison que le terme *catastrophe* n'est pas référencé dans le lexique. Le score de subjectivité n'est donc pas représentatif du segment. Les seules mesures obtenues par PyFeel ne sont pas suffisantes pour analyser la polarité et la subjectivité dans les segments du corpus de référence. Mais leur intégration dans le processus d'apprentissage automatique peut à minima donner des indices intéressants à l'algorithme à propos du caractère objectif ou subjectif de chaque segment. Ainsi, nous appliquons PyFeel sur le corpus de référence pour obtenir le score général de subjectivité de chaque segment. Dans cette série de mesures nous distinguons deux informations :

- Le score de polarité, qui représente les scores *Positif* et *Négatif*.
- Le score d'émotion, qui représente les scores *Colère*, *Dégoût*, *Peur*, *Joie*, *Tristesse et Surprise*.

Ces deux scores sont utilisés en tant que *features* lors de l'entraînement du modèle pour la détection de la subjectivité et de la polarité.

4.5.3 Entraînement du modèle : vectorisation et classifieurs

4.5.3.1 Représentation vectorielle des données

Les algorithmes d'apprentissage automatique traitent des données numériques. Afin de pouvoir traiter des données textuelles, il est essentiel de procéder à une *vectorisation* du texte. La vectorisation consiste en la transformation du texte en liste de nombres (ou vecteurs) afin de faciliter la manipulation des données par les algorithmes de classification. Les vecteurs ont généralement la taille du vocabulaire du corpus analysé, soit le nombre de mots uniques présents dans le corpus total. Il existe différentes manières de représenter vectoriellement des documents.

L'approche la plus simple est celle du *One-Hot Encoding* qui consiste à représenter le texte en fonction de la présence ou non de chaque terme dans ce texte par rapport à la liste de tous les mots présents dans le corpus. Cette approche, dite en sac de mots (*Bags of Words* – BOW), ne tient pas compte de l'ordre des mots dans la représentation.

- Orléans est une ville du Loiret.
- Tours est une ville de l'Indre-et-Loire

Si l'on considère le corpus composé des phrases a) et b), le vocabulaire de ce corpus est constitué de dix mots uniques : *Orléans*, *est*, *la*, *préfecture*, *du*, *Loiret*, *Tours*, *de*, *l'*, *Indre-et-Loire*. Peu importe le nombre de mots composant les documents analysés, la taille du vecteur, ou *dimension*, correspond à la taille du vocabulaire du corpus total. En l'occurrence, chaque vecteur a une dimension de dix et a la forme suivante :

	Orléans	est	une	ville	du	Loiret	Tours	de	l'	Indre-et-Loire
a)	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
	1	1	1	1	1	1	0	0	0	0

b)	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓
	0	1	1	1	0	0	1	1	1	1

Tableau 20 : Représentation vectorielle des phrases a) et b)

Le *One-Hot Encoding* est une représentation simple des documents qui ne tient compte que de la présence ou non de chaque terme composant le vocabulaire. Dans ce court exemple, la dimension du vecteur est limitée, ce qui nous permet d'obtenir des matrices denses, c'est-à-dire que les vecteurs sont majoritairement composés de 1. Mais dans un ensemble de données de la taille de notre corpus de référence réunissant plus de 40 000 termes différents, on obtiendra plutôt des matrices creuses, contenant beaucoup de zéros et donc moins d'informations. Pour redonner du sens aux vecteurs, le *One-hot Encoding* est souvent associé à des mesures comme le TF-IDF qui s'intéresse au nombre d'occurrences de chaque terme dans le corpus. D'autres approches proposent de réduire le nombre de dimensions des vecteurs ou encore de prendre en compte le contexte d'apparition des termes comme avec Word2Vec. Ces éléments ont donc été pris en compte pour entraîner notre modèle et sont détaillés dans les sections suivantes.

4.5.3.1.1 TF-IDF

Le TF-IDF (Salton & Buckley, 1998), est une méthode de pondération permettant de déterminer l'importance d'un mot en fonction du document dans lequel il se trouve mais aussi en fonction de tous les autres documents composant le corpus. Le TF-IDF est employé pour tenter d'accorder une pertinence lexicale à un terme dans un document donné que l'on n'avait pas dans une représentation *One-Hot*. On cherche donc à appliquer une relation entre un document, et un ensemble de documents partageant des similarités en matière de mots clés. Il s'agit d'une certaine manière de trouver un rapport entre quantité et qualité lexicale à travers un ensemble de documents.

Plus précisément, le TF-IDF est le produit de deux fréquences :

- Term Frequency (TF) qui représente le nombre d'occurrences d'un terme au sein d'un même document.
- Inverse Document Frequency (IDF) qui fait le rapport entre le nombre total de document composant le corpus et le nombre de documents du corpus qui contiennent le terme analysé.

Plus un terme est fréquent dans un corpus, moins il sera considéré comme distinctif. Au contraire, si un terme est rare dans le corpus, alors sa valeur discriminante sera plus grande. Ainsi, pour une requête avec un terme X, un document a plus de chances d'être pertinent comme réponse à la requête, si ce document présente une occurrence de ce terme, et que ce terme est rare dans d'autres documents reliés au premier.

Le modèle classique du TF-IDF considère les mots un par un. La notion de n-gramme aide à relativiser cette représentation en proposant au modèle de considérer les mots seuls mais aussi en tant qu'ensembles constitués de *n* mots. Les séquences composées de deux mots, ou bi-gramme, sont fréquemment utilisés en TAL dans des tâches de fouille de textes.

La vectorisation de notre corpus en fonction du TF-IDf est réalisée grâce à la bibliothèque Python Scikit-learn (Pedregosa *et al.*, 2011) [114] . Comme l'approche One-Hot, la représentation vectorielle du TF-IDF est une approche BOW qui ne prend pas en compte le contexte d'apparition des termes observés. Les vecteurs obtenus restent de grande dimension et sont donc des matrices vides composées d'une majorité de zéros. Aussi, il existe des libraires comme TruncatedSVD[115] qui permettent de réduire les dimensions des vecteurs et de rendre la matrice plus dense.

4.5.3.1.2 Word2vec

Word2Vec est un outil développé par Mikolov *et al.* (2013) pour réaliser des plongements de mots ou *word embedding*. Cette méthode d'apprentissage automatique est une technique de représentation vectorielle qui vise à représenter les mots d'un document selon leur contexte d'apparition.

[114] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer
[115] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

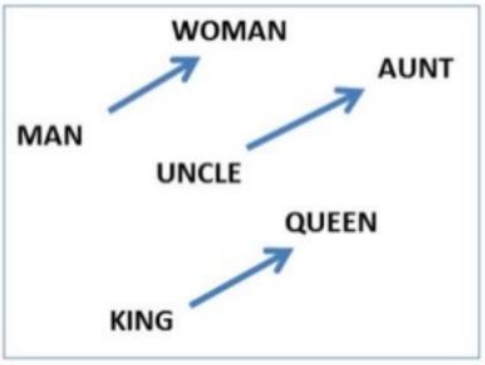

Figure 27 : Exemples de vecteurs dans Word2Vec - Extrait de Mikolov *et al.* (2013)

Dans la figure 27, on observe les représentations de plusieurs termes dans Word2Vec. Grâce à l'analyse de leurs contextes d'apparition, Word2Vec peut établir un lien sémantique entre les termes *man* et *woman*, *uncle* et *aunt*, et *king* et *queen*.

Les vecteurs de Word2Vec sont considérés comme denses, car ils disposent d'une dimension prédéfinie dans l'algorithme. Généralement cette dimension est comprise entre 100 et 300. Avec cette technique, il est possible de représenter une séquence de mots, comme une phrase, en s'appuyant sur leur représentation vectorielle, en faisant la somme de chaque composante des vecteurs et en le divisant par le nombre de termes additionnés.

4.5.3.2 Classifieurs

Différents outils et modèles statistiques sont utilisés pour réaliser des tâches de classification par apprentissage supervisé. Les plus utilisés en TAL dans des tâches de fouille d'opinion sont les machines à vecteurs supports (ou SVM pour *Support Vector Machines*) ou les forêts d'arbres décisionnels (*Random Forest Classifier*). La bibliothèque libre Python Scikit-learn[116], développée par Pedregosa *et al.* (2011) permet de manipuler ces classifieurs et de les intégrer dans un processus d'apprentissage automatique. Cette bibliothèque propose notamment des outils pour évaluer les performances des classifieurs utilisés mais aussi pour en optimiser le paramétrage.

[116] https://scikit-learn.org/stable/

L'implémentation d'un classifieur nécessite le réglage d'hyperparamètres i.e. de paramètres dont les valeurs sont définies avant le début du processus d'apprentissage. Le choix des valeurs est déterminant pour la qualité de l'apprentissage. Grâce au module GridsearchCV[117] disponible dans Scikit-Learn, nous pouvons déterminer quels sont les hyperparamètres les plus efficaces pour réaliser la classification attendue.

4.5.3.2.1 Random Forest Classifier

Le classifieur Random Forest, proposé par Breiman (2001), est un algorithme qui effectue un apprentissage à partir de multiples arbres de décision entraînés sur des sous-ensembles de données aléatoirement constitués. Chaque arbre de décision cartographie les observations des sous-ensembles de données et tire des conclusions sur la valeur cible des données. Dans le modèle prédictif, les attributs des données qui sont déterminés par l'observation sont représentés par les branches, tandis que les conclusions sur la valeur cible des données sont représentées dans les feuilles. Lors de l'apprentissage d'un arbre, les données source sont divisées en sous-ensembles en fonction d'un test de valeur d'attribut, qui est répété récursivement sur chacun des sous-ensembles dérivés. Une fois que le sous-ensemble d'un nœud a la valeur équivalente à sa valeur cible, le processus de récursion est terminé.

A partir de la documentation officielle de Scikit-Learn pour l'utilisation de Random Forest[118], nous avons sélectionné plusieurs hyperparamètres à faire varier pour optimiser l'entraînement du modèle :

- *n_estimators* : nombre d'arbres dans la forêt. Valeurs testées : 4, 6, 9.
- *max_features* : Valeurs testées : log2, sqrt, ou la valeur par défaut.
- *max_depth* : profondeur maximale de l'arbre. Valeurs testées : 5, 10, 15.
- *min_samples_split* : nombre minimum d'échantillons requis pour créer un nouveau nœud dans l'arbre. Valeurs testées : 2, 3, 5.
- *min_samples_leaf* : nombre minimum d'échantillons au niveau d'un nœud dans l'arbre tel qu'à « n'importe quelle profondeur <il> ne sera considéré

[117] https://scikit-learn.org/stable/modules/grid_search.html
[118] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

que s'il laisse au moins des échantillons d'apprentissage *min_samples_leaf* dans chacune des branches gauche et droite »[119]. Valeurs testées : 2, 5, 8.

4.5.3.2.2 SVM

Les SVM, pour *Support Vector Machines* (machines à vecteurs supports ou séparateurs à vastes marges) sont des méthodes performantes en TAL qui ont l'avantage d'être efficaces pour classifier des données de grandes dimensions. Issues d'analyses mathématiques précises et avancées du problème de l'apprentissage, elles ont été introduites par Vapnik (1992). Un SVM est un séparateur binaire, c'est-à-dire qu'il vise à séparer les données étiquetées en deux sous-ensembles distincts. Puisque c'est un problème de classification, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle et s'appuie sur l'utilisation de fonctions dites noyau (*kernel*) pour atteindre une séparation optimale des données.

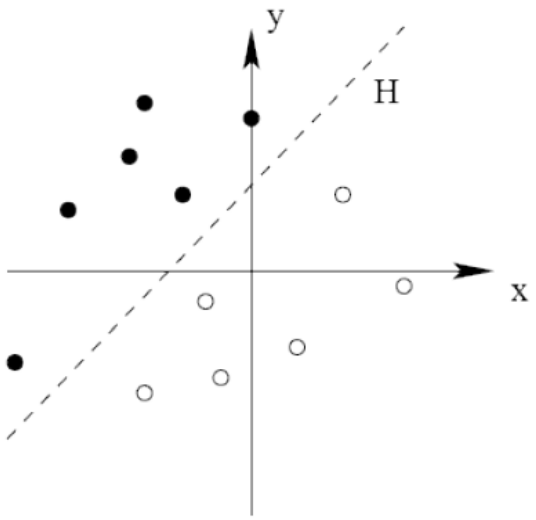

Figure 28: Illustration du focntionnement d'un SVM – Extrait de Hasan & Boris (2006)

Selon Hasan & Boris (2006), le but d'un SVM est de « trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes ». Dans la figure 28, le classificateur linéaire ou *hyperplan*, noté H, sépare les deux ensembles de points noirs et

[119] « *A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches*. »

blancs. Dans le cas où les données ne sont pas linéairement séparables, l'algorithme utilise le *kernel trick* (« astuce du noyau »), qui consiste à représenter les données dans un espace de dimension plus grand. Un hyperplan est ensuite dessiné afin de discriminer les classes.

Pour définir les hyperparamètres nécessaires à l'entrainement du modèle, nous nous sommes référée à la documentation officielle relative à l'utilisation de SVM dans Scikit-Learn[120]. Ainsi, nous avons fait varier les paramètres suivants :

- *C* : paramètre de pénalité. Valeurs testées : 1, 10, 100, 1000.
- *gamma* : coefficient influençant le noyau du SVM. Valeurs testées : 1e-3, 1e-4, ou la valeur par défaut.
- *kernel* : linéaire, sigmoïde, ou la valeur par défaut.

4.5.4 Expériences réalisées

Afin d'entraîner un modèle destiné à la détection de la subjectivité puis de la polarité contenues dans des segments de transcription, différentes expériences ont été menées. Un corpus de référence a d'abord été annoté manuellement selon des conventions d'annotation préétablies. Des traits linguistiques ont été définis à partir de ce corpus pour l'entraînement du modèle. Ces traits, qui sont la lemmatisation et l'étiquetage morpho-syntaxique des segments et le calcul de scores de polarité et d'émotion, préparent l'étape de vectorisation des données textuelles. Le TF-IDF et Word2Vec sont des méthodes possibles pour créer une représentation vectorielle des segments à analyser. Enfin deux classifieurs, RandomForest et SVM, ont été retenus pour réaliser la classification attendue.

L'entraînement du modèle consiste à identifier quelle combinaison de *features*, de méthodes de vectorisation du texte et de classifieurs permet d'atteindre les objectifs de détection de la subjectivité et de la polarité dans le corpus. Les différentes expériences réalisées sont décrites dans les sections suivantes et sont évaluées grâce aux mesures de Rappel, Précision, F-mesure et la macro average. La macro average fait la moyenne entre les mesures de Rappel et Précision en donnant un poids égal à chaque classe qu'il faut

[120] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

détecter. Elle représente l'efficacité globale du modèle à classer les segments en fonction de la subjectivité.

4.5.4.1 Subjectivité

Pour procéder à la classification des segments définis autour de la détection des lieux, en fonction de leur caractère objectif ou subjectif, nous avons réalisé différents prétraitements, en sélectionnant des méthodes pour représenter vectoriellement nos données textuelles et deux classifieurs pour les traiter. Le corpus de référence annoté en subjectivité contient 3178 segments, soit 1112 segments subjectifs et 2066 segments objectifs. Le corpus de référence est divisé en trois sous-corpus : 80% de l'ensemble constitue le corpus d'entrainement, 10% constitue le corpus de validation et les 10% restants constituent le corpus de test. Le corpus de test est réservé pour l'évaluation finale du modèle retenu. Le modèle est entrainé sur le corpus d'entrainement puis évalué sur le corpus de validation.

Les expériences ont été menées en deux temps. D'abord, nous avons évalué l'impact de la méthode de vectorisation du texte sur les résultats de la classification des segments selon s'ils sont lemmatisés ou non. Cinq méthodes de vectorisation ont été évaluées : TF-IDF seul, avec ou sans bi-grammes, avec ou sans réduction de dimensions et Word2Vec. Chaque méthode a été employée avec Random Forest et SVM sur les segments lemmatisés et non lemmatisés. Le tableau 21 résume les résultats obtenus.

Les différentes méthodes de vectorisation des textes présentent des résultats plutôt similaires, qu'elles soient utilisées avec Random Forest ou SVM. D'une manière générale, Random Forest présente de meilleurs résultats sur l'analyse des segments bruts tandis que SVM fonctionne mieux avec les segments lemmatisés. Random Forest est un classifieur adapté pour les problèmes multi-classes et qui fonctionne généralement de façon satisfaisante avec les mélanges de caractéristiques numériques et catégorielles. SVM est plutôt optimisé pour traiter les cas de classification binaire en cherchant l'hyperplan le plus pertinent. La dimension des vecteurs des segments lemmatisés est moins grande que celle des segments sans prétraitements. Puisque ces segments lemmatisés contiennent moins d'éléments, ils sont plus simples à classer en fonctions des deux cibles objectif et subjectif par SVM.

Le Rappel de la détection des segments objectifs est satisfaisant dans l'ensemble et est toujours supérieur à celui de la détection des segments subjectifs. Peu importent les prétraitements réalisés ou non sur les segments ou la méthode de vectorisation employée, il semble plus évident pour SVM et Random Forest de décider qu'un segment est objectif. Le Rappel pour la détection des segments subjectifs est en moyenne égal à 0,48 ce qui signifie que le modèle ne détecte pas la moitié des éléments qu'il aurait dû classer. Nous ne retrouvons pas la même tendance dans la comparaison des scores de Précision. Au contraire, que ce soit avec ou Random Forest, les scores de Précision pour l'attribution de l'étiquette subjectivité ou objectivité sont équivalents et plutôt satisfaisants. En moyenne, les deux classifieurs catégorisent correctement 70% des segments.

			Segments brutes				Segments lemmatisés			
Représentation vectorielle	Classifieur	Cible	Macro Average	Précision	Rappel	F-mesure	Macro Average	Précision	Rappel	F-mesure
TF-IDF	Random Forest	subjectif	0,72	0,79	0,42	0,55	0,72	0,71	0,51	0,59
		objectif		0,70	0,93	0,80		0,72	0,86	0,78
	SVM	subjectif	0,70	0,76	0,39	0,51	**0,75**	**0,76**	**0,53**	**0,62**
		objectif		0,69	0,92	0,79		**0,73**	**0,89**	**0,8**
TF-IDF + Bi-gramme (1, 2)	Random Forest	subjectif	0,73	0,75	0,52	0,62	0,71	0,71	0,50	0,59
		objectif		0,72	0,88	0,8		0,72	0,86	0,78
	SVM	subjectif	0,71	0,71	0,48	0,57	0,73	0,79	0,46	0,58
		objectif		0,71	0,87	0,78		0,71	0,92	0,8
TF-IDF + Réduction de dimensions	Random Forest	subjectif	0,71	0,72	0,47	0,57	0,70	0,71	0,44	0,55
		objectif		0,71	0,88	0,79		0,70	0,88	0,78
	SVM	subjectif	0,72	0,66	0,62	0,64	0,71	0,65	0,62	0,63
		objectif		0,75	0,78	0,77		0,75	0,77	0,76
TF-IDF + Bi-gramme (1, 2) + Réduction de dimensions	Random Forest	subjectif	0,66	0,78	0,22	0,35	0,68	0,79	0,26	0,4
		objectif		0,65	0,96	0,77		0,66	0,95	0,78
	SVM	subjectif	0,71	0,71	0,51	0,59	0,7	0,68	0,51	0,58
		objectif		0,72	0,86	0,78		0,72	0,84	0,77
Word2vec	Random Forest	subjectif	0,71	0,67	0,54	0,60	0,69	0,65	0,51	0,57
		objectif		0,73	0,82	0,77		0,71	0,81	0,76
	SVM	subjectif	0,71	0,7	0,48	0,57	0,69	0,67	0,47	0,55
		objectif		0,71	0,86	0,78		0,70	0,84	0,76

Tableau 21 : Résultats de la détection de la subjectivité par SVM et Random Forest en fonction des représentations vectorielles des segments bruts ou lemmatisés.

Les macro average sont comprises entre 0,66 et 0,75. On peut constater que l'utilisation du TF-IDF avec les bi-grammes et la réduction de dimension est la méthode la moins efficace lorsqu'elle est utilisée avec le classifieur Random Forest (0,66 de macro average) mais aussi avec SVM (0,68). La combinaison de traits la plus efficace pour Random Forest est celle du TF-IDF dans laquelle on ajoute les bi-grammes (0,73 de macro average).

Nous considérons que la combinaison la plus efficace pour identifier la subjectivité est celle qui utilise le TF-IDF seul pour vectoriser les segments lemmatisés avec le classifieur SVM (0,75 de macro average). On peut expliquer ce résultat par le fait que cette méthode permet de créer les vecteurs les plus réduits et donc qu'elle est mieux adaptée avec SVM.

Représentation vectorielle	Classifieur	Cible	Macro Average	Précision	Rappel	F-mesure
TF-IDF	SVM	subjectif	0,75	0,76	0,53	0,62
		objectif		0,73	0,89	0,8

Tableau 22 : Baseline pour la détection de la subjectivité

On aurait pu penser que la réduction de dimension aurait été meilleure avec SVM. En l'occurrence, on constate que la macro average est moins bonne qu'avec le TF-IDF seul. On peut tout de même considérer les scores obtenus comme corrects du fait que les scores obtenus sont plus lisses. On peut supposer que la réduction de dimension facilite le classement des segments par SVM et que cela confère de la stabilité au modèle. Cependant, ce gain de stabilité se fait au détriment de la Précision : plus de segments sont détectés mais ils sont moins bien catégorisés. Nous préférons conserver le modèle présentant les meilleures performances en général et surtout la meilleur Précision pour privilégier la qualité de la détection. Le tableau 22 présente la *baseline*[121] du modèle sélectionné.

SVM + TF-IDF		Segments lemmatisés			
Features	Classe	Macro Average	Précision	Rappel	F-mesure
Score polarité	subjectif	0,70	0,63	0,62	0,63
	objectif		0,75	0,76	0,75
Score polarité + POS	subjectif	0,75	0,69	0,54	0,61
	objectif		0,76	0,87	0,82

[121] Valeur de référence pour le suivi des performances d'un modèle.

Score émotions	subjectif	0,72	0,66	0,63	0,65
	objectif		0,76	0,79	0,77
Score émotions + POS	subjectif	0,75	0,68	0,58	0,62
	objectif		0,78	0,85	0,81
Score polarité + Score émotions	subjectif	0,74	0,68	0,51	0 ,58
	objectif		0,76	0,87	0,81
Score polarité + Score émotions + POS	subjectif	**0,76**	**0,70**	**0,55**	**0,62**
	objectif		**0,77**	**0,87**	**0,82**

Tableau 23 : Résultats de la détection de la subjectivité en fonction des *features* utilisés

La deuxième phase de l'expérimentation consiste en l'ajout des *features* précédemment définis[122]. Nous pouvons supposer que l'utilisation des scores de polarité et d'émotions calculés pour chaque segment ainsi que l'étiquetage morphosyntaxique (POS) permettront d'améliorer l'efficacité du modèle et de dépasser la *baseline*. Le tableau 23 présente les résultats de ces nouvelles expérimentations.

4.5.4.2 Polarité

La détection de la polarité suit la même procédure que celle de la détection de la subjectivité. Le modèle est entrainé à partir des segments du corpus de référence catégorisés comme subjectifs et ne considère donc que les segments marqués comme positif ou négatifs. Le nouveau corpus de référence est constitué de 762 segments positifs contre 350 segments négatifs. De la même façon que pour l'identification de la subjectivité, cette part du corpus de référence est divisée dans les mêmes proportions en un corpus d'entrainement (80%), un corpus de validation (10%) et un corpus de test (10%).

La première série d'expériences réalisée sert à identifier la méthode la plus efficace pour vectoriser les segments en fonction de leur réduction en lemmes ou non. Le tableau 24 présente les différentes méthodes de vectorisation utilisées avec les classifieurs SVM et Random Forest. Les macro average sont comprises entre 0,67 et 0, 74. Si ces moyennes sont similaires à celles obtenues pour la détection de la subjectivité, dans le détail, ces scores sont moins satisfaisants.

[122] Cf. section 5.5.2

En effet, les mesures de Rappel pour la détection des segments négatifs sont très faibles ce qui signifie que très peu de détections sont réalisées. Au maximum, seuls 36% des segments négatifs sont identifiés. La qualité de ces identifications est meilleure mais reste moyenne. Néanmoins, la détection des segments positifs est plus satisfaisante : le Rappel oscille entre 0,84 et 1 ce qui est excellent. La grande majorité des segments positifs est détectée et, en moyenne, 70% des détections faites sont pertinentes. Ces résultats s'expliquent très probablement par la composition du corpus de référence. La classe positive étant majoritaire, le classifieur dispose de plus d'exemples pour apprendre à la détecter au détriment de la classe négative. On peut faire l'hypothèse que l'équilibrage de la composition du corpus de référence, et donc du corpus d'entrainement par l'ajout de segments négatifs supplémentaires, permettrait l'amélioration du Rappel.

			Segments brutes				Segments lemmatisés			
Représentation vectorielle	Classifieur	Cible	Macro Average	Précision	Rappel	F-mesure	Macro Average	Précision	Rappel	F-mesure
TF-IDF	Random Forest	positif	0,69	0,69	0,99	0,81	0,69	0,70	0,96	0,81
		négatif		0,50	0,02	0,01		0,54	0,11	0,19
	SVM	positif	0,68	0,74	0,84	0,78	**0,74**	**0,75**	**0,92**	**0,83**
		négatif		0,49	0,34	0,40		**0,65**	**0,36**	**0,45**
TF-IDF + N-gramme (1, 2)	Random Forest	positif	0,72	0,72	0,99	0,83	0,71	0,71	0,98	0,82
		négatif		0,82	0,15	0,25		0,70	0,11	0,20
	SVM	positif	0,72	0,75	0,89	0,81	0,73	0,75	0,90	0,82
		négatif		0,58	0,34	0,43		0,61	0,36	0,45
TF-IDF + Réduction de dimensions	Random Forest	positif	0,70	0,70	0,99	0,82	0,71	0,71	0,99	0,83
		négatif		0,67	0,07	0,12		0,86	0,10	0,18
	SVM	positif	0,69	0,72	0,90	0,80	0,70	0,73	0,90	0,81
		négatif		0,50	0,23	0,31		0,55	0,26	0,36
TF-IDF + N-gramme (1, 2) + Réduction de dimensions	Random Forest	positif	0,70	0,70	0,99	0,82	0,70	0,70	1	0,82
		négatif		1	0,05	0,09		1	0,05	0,09
	SVM	positif	0,72	0,75	0,87	0,81	0,72	0,74	0,92	0,82
		négatif		0,57	0,38	0,46		0,61	0,28	0,38
Word2vec	Random Forest	positif	0,67	0,70	0,89	0,79	0,68	0,69	0,96	0,80
		négatif		0,42	0,18	0,25		0,40	0,07	0,11
	SVM	positif	0,69	0,69	0,99	0,81	0,70	0,71	0,98	0,82
		négatif		0,50	0,02	0,03		0,70	0,11	0,20

Tableau 24 : Résultats de la détection de la polarité par SVM et Random Forest en fonction des représentations vectorielles des segments brutes ou lemmatisés

Les tendances observées pour la détection de la polarité sont valables pour les deux classifieurs. L'écart entre les performances de SVM et Random Forest est très faible. On peut tout de même noter que le SVM est plus efficace pour détecter les segments négatifs lorsque ceux-ci sont lemmatisés.

Représentation vectorielle	Classifieur	Cible	Macro Average	Précision	Rappel	F-mesure
TF-IDF	SVM	positif	0,74	0,75	0,92	0,83
		négatif		0,65	0,36	0,45

Tableau 25 : Baseline pour la détection de la polarité

Ainsi, nous considérons encore une fois que la vectorisation des segments lemmatisés par le TF-IDF associés à un SVM est la meilleure méthode pour traiter la polarité (0,74 de macro average). Ce modèle ne présente pas le meilleur Rappel pour la détection des segments positifs mais il est considéré comme le plus stable et constitue la *baseline* de l'apprentissage[123]. Partant de cette conclusion, nous procédons de nouveau à la sélection des *features* les plus pertinents pour l'optimisation du modèle. Le tableau 26 présente les résultats des différentes expériences menées autour de la sélection des *features*.

SVM + TF-IDF		Segments lemmatisés			
Features	Classe	Macro Average	Précision	Rappel	F-mesure
Score polarité	positif	0,73	0.76	0.90	0.82
	négatif		0.62	0.38	0.47
Score polarité + POS	positif	0,75	0.77	0.90	0.83
	négatif		0.65	0.43	0.51
Score émotions	positif	0,73	0.76	0.90	0.82
	négatif		0.62	0.38	0.47
Score émotions + POS	positif	0,75	0.77	0.90	0.83
	négatif		0.65	0.43	0.51
	positif	0,73	0.76	0.90	0.28

[123] cf. Tableau 25

Score polarité + Score émotions	négatif		0.62	0.38	0.47
Score polarité + Score émotions + POS	positif	**0,76**	**0.78**	**0.90**	**0.83**
	négatif		**0.66**	**0.46**	**0.53**

Tableau 26 : Résultats de la détection de la polarité en fonction des *features* utilisés

Les dynamiques observées lors de l'entrainement du modèle de détection de la subjectivité se retrouvent dans celui de la détection de la polarité. Les expériences dans lesquelles seuls les scores de polarité et d'émotions sont utilisés montrent des performances inférieures à la *baseline* définie. L'ajout de l'information des POS améliore légèrement la macro average par rapport à la *baseline* en améliorant significativement le Rappel et la Précision pour la détection des segments négatifs. Le modèle utilisant les scores de polarité, d'émotions et les POS donne une macro average de 0,76 dans laquelle la F-mesure de la détection des segments positifs reste stable par rapport à la *baseline* et la F-mesure de la détection des segments négatifs est améliorée à 0,53 (anciennement 0,45).

4.5.5 Modèle retenu

Les différentes expériences précédemment décrites ont permis d'établir un modèle d'apprentissage pour la détection de la subjectivité et la détection de la polarité. L'utilisation du classifieur SVM, la représentation vectorielle des segments lemmatisés avec le TF-IDF et l'utilisation de tous les features disponibles constitue le modèle le plus efficace pour réaliser les deux tâches de détection. Pour confirmer les performances de ce modèle, celui-ci est évalué sur les corpus de test définis. Le tableau 27 présente les résultats de cette nouvelle évaluation.

Représentation vectorielle	Features	Classifieur	Cible	Macro average	Précision	Rappel	F-mesure
TF-IDF	Score polarité + Score émotions + POS	SVM	subjectif	0,77	0,69	0,56	0,63
			objectif		0,77	0,89	0,83
			positif	**0,76**	**0,78**	**0,91**	**0,82**
			négatif		**0,67**	**0,46**	**0,54**

Tableau 27 : Evaluation du modèle pour la détection de la subjectivité et de la polarité

Les nouvelles mesures réalisées confirment les observations faites lors de l'entrainement du modèle. Le modèle est plus efficace pour la détection de la subjectivité (0,77 de macro average), ce qui peut s'expliquer par le fait que cette classe dispose du plus grand nombre de données. Il présente des performances similaires pour le traitement de la polarité (0,76 de macro average), avec notamment un excellent Rappel de 0,91 pour la détection des segments positifs.

D'une manière générale nous remarquons que, moins il y a de traitements appliqués aux segments lors de leur vectorisation, meilleurs sont les résultats. En effet, les méthodes de vectorisation limitant le nombre de dimensions des vecteurs profitent à la qualité de la classification réalisée par SVM. La piste de la réduction de dimension des vecteurs doit être approfondie pour améliorer les performances du classifieur. Dans le même temps, l'association des scores de polarité et d'émotions et des POS participe à l'optimisation du modèle. Il semble aussi nécessaire d'équilibrer la composition du corpus de référence, en ajoutant en particulier des segments polarisés négativement. Une nouvelle phase d'annotation manuelle doit donc être réalisée. Enfin d'autres *features*, comme le nombre de mots composant les segments, la position des termes porteurs dans le segment, devront être identifiés pour améliorer l'apprentissage du modèle.

4.6 Analyse de la perception

La perception est considérée comme le processus de traitement d'informations recueillies à l'aide de récepteurs sensoriels dans le but de créer des représentations et des connaissances à propos de l'objet perçu. C'est une expérience subjective qui varie d'un individu à l'autre au gré de ses émotions, sentiments et opinions. Afin d'analyser la perception, il s'agit d'abord de détecter la subjectivité et caractériser la polarité dans les segments de conversations mentionnant des lieux. Ces étapes permettent de détecter les segments du discours dans lesquels la perception peut être exprimée que ce soit positivement ou négativement. La détection de la subjectivité et de la polarité oriente l'analyse de la perception mais ne suffit pas à en rendre compte. Pour aller plus loin dans la réflexion, il faut approfondir la typologie de la perception pour en comprendre la nature et préparer son traitement automatique.

4.6.1 Propositions typologiques de la perception

4.6.1.1 Cible de la perception

Du point de vue de la fouille d'opinion, l'un des axes d'étude possibles est celui de la cible de la perception émise, c'est-à-dire de la thématique abordée. En effet, il serait intéressant de pouvoir détecter quels sont les éléments mis en avant par les locuteurs pour partager leurs perception de l'environnement.

Les déclarations analysées sont faites sur de l'oral, ce qui implique que le discours se construit au fur et à mesure qu'il est énoncé. Les données sont les pensées brutes des locuteurs qu'ils partagent avec un tiers. Même si la trame de certains enregistrements guide la conversation et incite les locuteurs à parler de leurs lieux de vie et en général d'Orléans, les arguments pour le faire ne sont pas imposés. A l'instant où un locuteur est interrogé sur le sujet, il fait le choix, instantanément, d'évoquer certains éléments plutôt que d'autres. Le fait de mettre l'accent sur certaines thématiques pour décrire la ville est révélateur de la représentation que se fait la personne de son environnement. Ces choix faits spontanément, sans réflexion préalable, ne sont pas anodins ; ils illustrent leur perception.

L'exploration manuelle du corpus révèle que l'expression de la perception à propos d'une ville se fait en fonction de thématiques précises. Celles-ci peuvent être classées en quatre catégories : urbanisme, sociale, économique et historico-culturelle.

4.6.1.1.1 Urbanistique

Les locuteurs décrivent Orléans d'un point de vue spatial, urbanistique. Selon le TLFi, l'urbanisme désigne l'ensemble « des sciences, des techniques et des arts relatifs à

l'organisation et à l'aménagement des espaces urbains »[124]. L'une des principales problématiques concerne l'accessibilité. Au sens physique du déplacement dans l'espace, comme capacité à accéder à un lieu, en fonction de l'organisation des transports, de la disponibilité d'infrastructures, d'équipements adaptés, etc. Dans cette thématique, la ville est perçue sous l'angle de son accessibilité générale et de l'organisation des espaces qui la composent.

(48) le le Pathé qui est sur les bords de Loire c'est pas du tout pratique d'accès c'est vraiment donc il reste vraiment maintenant plus que les Carmes donc c'est bien pour les Carmes ils vont travailler (ESLO2_ENT_1020)

Ainsi, un individu peut décrire un commerce sans s'intéresser aux services qu'il propose mais en critiquant plutôt son emplacement dans la ville comme dans l'exemple 48. Dans l'exemple 49, la *place De Gaulle* est caractérisée par les travaux en cours à l'instant où le locuteur s'y réfère.

(49) donc vous prenez le tram à Saint-Marceau ou là sur le côté plus loin là à combien deux cents mètres vous arrêtez à Place De Gaulle c'est **une grande place qui est en train d'être aménagée puisqu'on installe le tram**
(ESLO2_iti_08_01)

(50) voilà le campus il est **loin** (ESLO2_ENTJEUN_09)

La thématique de l'urbanisme regroupe les déclarations subjectives qui ont trait avant tout à la situation au sens géographique des espaces les uns par rapport aux autres. Dans l'exemple 50, *le campus* est considéré comme éloigné du centre-ville. Ce n'est pas simplement une déclaration factuelle : c'est non seulement une appréciation personnelle mais aussi la caractéristique du lieu que le locuteur choisit de mettre en avant pour le décrire.

[124] stella.atilf.fr/Dendien/scripts/tlfiv5/advanced.exe?8;s=2189235945

4.6.1.1.2 Economique

La description de la ville ne se retrouve pas seulement dans ses espaces physiques mais aussi dans ses institutions, ses commerces et les services qui s'y trouvent. Certaines décisions politiques ou certaines implantations d'activités influent sur l'image de la ville. Dans l'exemple 51, le locuteur évoque les activités qu'il mène à Orléans plutôt qu'à Olivet en justifiant ce choix par l'importance de la *politique sportive* d'Orléans.

(51) nos activités sont plus sur Orléans sont plus centrées sur Orléans que que sur Olivet parce qu'à Orléans **y a beaucoup de de politique sportive au niveau basket** au niveau
(ESLO2_ENT_1015)

Les lieux sont perçus par le biais de leurs fonctions, et donc par l'usage qu'ont les personnes de ces lieux, des habitudes qu'ils y ont prises ou non. A quel endroit vont-ils pratiquer leurs activités mais aussi à quel endroit ne vont-ils pas ? Quels sont les services qui caractérisent la ville, en particulier d'un point de vue économique ? Ainsi dans l'exemple 52, il est question de l'offre commerciale d'Orléans et des préférences du locuteur lorsqu'il s'agit de faire ses courses.

(52) ouais ça c'est vrai que c'est vrai que **ça manque peut-être un peu de commerces** parce que c'est vrai qu'il y a y a quand même du monde ouais et c'est vrai qu'au point de vue enfin après faut voir parce que y a y a **avant y avait Marché plus maintenant c'est Carrefour city oui c'est vrai que c'est plus sympa et c'est moins cher**
(ESLO2_ENT_1050)

La perception de l'activité citadine se retrouve dans le dynamisme économique, industriel ou associatif de la ville et chez les agents qui sont les acteurs de ces dynamiques.

4.6.1.1.3 Sociale

Une ville est avant tout un espace de vie dans lequel évoluent des individus organisés en communauté(s). L'évaluation qui en est faite peut directement concerner ses habitants, leur réputation ou leur mode de vie. Au-delà des services, des événements qui se produisent dans la ville, celle-ci peut être décrite sous l'angle des personnes qui y vivent, des relations qui s'y nouent, de l'animation sociale.

(53) donc y avait ça puis les fêtes de Loire je trouve ça sympa parce que c'est bien quand même d'avoir **une ville qui bouge là elle bouge un peu plus quand même qu'avant** (ESLO2_ENT_1050)

(54) donc le centre ville rue de la Lionne à coté de la place du Martroi rue Bannier tout ce qui est rue piétonne rue commerciale donc **c'est un quartier quand même assez calme sans vraiment de souci** (ESLO2_ENT_1026)

Dans les exemples 53 et 54, il est question de l'ambiance de la ville et de l'un de ses quartiers, de son animation. Un lieu animé est un lieu dans lesquels les gens se rendent pour des raisons diverses et variées. Il ne s'agit pas de s'intéresser aux activités qui les poussent à se réunir dans le lieu mais plutôt de garder un regard général sur l'importance de l'occupation humaine de l'endroit. Le locuteur considère dans l'exemple 66 que la *ville bouge* parce qu'il y a de l'animation au sens d'activité humaine lors des *fêtes de Loire*. Au contraire, le locuteur de l'exemple 54 considère le *quartier* qu'il décrit comme calme malgré sa proximité avec la *rue commerciale* que l'on suppose dynamique économiquement. Cet exemple illustre bien la différence entre les deux thématiques économique et sociale. Il n'y a pas forcément de corrélation entre les activités, les services proposés et les personnes qui en bénéficient.

4.6.1.1.4 Historico-culturelle

Une dimension essentielle dans la description d'une ville concerne son histoire et sa culture. Ces éléments font partie intégrante de la perception du locuteur. Celui-ci se projette dans son environnement en considérant les activités qui s'y sont déroulées, se déroulent ou se dérouleront.

(55)

ch_AC7: et pourtant euh les Orléanais euh ont mauvaise réputation puisque tu es pas d'Orléans tu peux le dire quand on parle des Orléanais tu connais les termes euh que l'on emploie euh
EW15: non
ch_AC7: les **chiens d'Orléans** les non tu as jamais entendu parler ?
EW15: ah oui oui si si j'ai entendu parler j- euh
ch_AC7: parce que tous les gens qui arrivent à Orléans disent que les Orléanais sont pas toujours trop aimables trop accueillants
EW15: oui c'est vrai oui ils sont que ils sont assez froids que
(ESLO2_ENT_1015)

Le patrimoine culturel et historique de la ville correspond à l'ensemble des biens, matériels et immatériels, ayant une importance artistique ou historique avérée. Les locuteurs se réfèrent à ce patrimoine en racontant les légendes et histoires le concernant. Dans l'exemple 55, il est question des *chiens d'Orléans*, un surnom donné aux Orléanais. Même si l'origine de l'expression est encore discutée, sa signification reste stable. Elle est utilisée pour décrire le caractère des Orléanais souvent jugé froid. Bien que dans cet exemple il est question de décrire les Orléanais, le locuteur le fait par le prisme d'une anecdote historique.

(56) c'était par rapport à dans la rue Bourgogne à des lieux où y aurait eu des femmes enfin c'était la rumeur c'était qu'une rumeur hélas où des des des des personnes étaient étaient kidnappées quoi en quelque sorte quoi disparaissaient et y a plusieurs interprétations dont toujours la même hélas et je crois qu'Orléans est bien payée pour ça c'est toujours un espèce de fond antisémitisme qui qui revient à la surface de temps en temps je pense que le plus gros drame avant a été l'épisode de la guerre

et de et de la de la après le Front Populaire et de la comment de la du meurtre de Jean Zay parce qu'il est quand même d'ici c'est quand même un un grand monsieur qu'on n'honore pas assez de de mon point de vue moi j'ai eu la chance de de fréquenter et toujours d'ailleurs Hélène Mouchard-Zay et quand même quelqu'un qui porte fort cette affaire mais que Orléans honore pas suffisamment cet cet homme et cette famille parce que je pense que heureusement qu'y a encore des gens comme et toute son association et tout ça néanmoins on a quand même attendu beaucoup d'années pour donner des noms à des des structures importantes mais bon c'est comme ça on n'est jamais prophète dans son pays

(ESLO2_ENT_1004)

Dans l'exemple 56, le locuteur raconte la *Rumeur d'Orléans*. Une rumeur, apparue sur fond d'antisémitisme en 1969, selon laquelle des jeunes femmes disparaissaient après être entrées dans les cabines d'essayage de boutiques tenues par des juifs. Ces femmes étaient prétendument enlevées pour être livrées à un réseau de prostitution. Cette histoire a marqué la ville et reste dans les esprits comme le montre le témoignage du locuteur de l'exemple 56. Un témoignage qui est aussi marqué par l'évocation de personnalités importantes de la ville comme Jean Zay et sa fille Hélène Mouchard-Zay. Les différents événements historiques qui se sont déroulés dans la ville ont participé à la dessiner telle qu'elle est aujourd'hui et les monuments de la ville en sont les premiers témoins jusque dans leur dénomination[125].

La perception de la ville d'Orléans a donc pour objet des éléments relatifs à l'organisation, l'accessibilité, le dynamisme ou encore l'histoire de la ville. Cependant, la catégorisation des thématiques abordées ne suffit pas non plus pour décrire les subtilités de la perception et elle doit être dépassée. Notre objectif est de caractériser la manière dont les locuteurs font part de leur perception, selon quelles modalités et dans quelles conditions. Pour répondre à cette problématique, une typologie de la nature de la perception est présentée dans la section suivante.

[125] cf. exemple 17, section 2.4.2.2

4.6.1.2 Nature de la perception

Pour décrire la ville, les locuteurs abordent différentes thématiques mais de quelle manière le font-ils ? Au-delà de l'objet décrit, la perception peut passer à travers les habitudes des individus, des actions qu'ils réalisent, de leurs goûts ou de ce qu'ils expérimentent par les sens. Ces éléments décrivent la nature de la perception et recoupent les différentes thématiques pouvant être abordées.

Pour décrire la nature de la perception, nous considérons trois catégories : la perception sensorielle, l'expérience personnelle et l'expérience collective.

4.6.1.2.1 Perception sensorielle

Les sens font partie intégrante du processus perceptif comme le montre la définition de la notion de *perception*[126]. Cette catégorie considère le positionnement d'un individu par rapport à son environnement via le prisme de ses sens. En faisant principalement appel aux sens de la vue, de l'ouïe et de l'odorat, l'individu peut transmettre les expériences internes qu'il fait des lieux dont il parle. Dans l'exemple 57, le bruit est la première caractéristique présentée par le locuteur pour décrire son environnement.

(57) le quartier est plutôt **bruyant** alors par rapport aux aux enfants (ESLO2_ENTJEUN_05)

A partir du sensoriel apparaissent les émotions et sentiments, par quoi le locuteur peut faire part de ses goûts, de ses préférences ou de ses aversions.

(58) une histoire il vit ce ce fleuve oui oui oui tout à fait oui oui ouais et j'avais appris énormément de choses avec lui oui oui **c'est un beau fleuve** (ESLO2_ENT_1016)

[126] cf. section 3.2

(59) mon lieu préféré dans la ville ça serait quand même du quartier cathédrale euh voilà justement oui **c'est très joli** (ESLO2_iti_10_03)

Ainsi, la Loire est un *beau fleuve*, le quartier de la cathédrale est *très joli*. Pour arriver à ce constat, le locuteur a analysé ce qu'il a expérimenté sensoriellement par la vue.

4.6.1.2.2 Expérience individuelle

La perception peut aussi passer par les habitudes prises dans le lieu décrit ou les activités qui y sont menées. Le fait d'énoncer quelles activités une personne peut ou ne peut pas faire dans un certain endroit contribue à en dresser le portrait.

(60) donc je vais je vais voir ailleurs quoi puis bon pour les pour les pubs qui sont rue de Bourgogne tout ça je je bouge aussi quoi on reste dans le oui dans le voilà centre-ville en fait c'est vraiment le centre-ville c'est ça oui finalement c'est là qu'y a tout c'est là qu'y a tout hein (ESLO2_ENT_1047)

Le locuteur partage dans l'exemple 73 ses habitudes en termes de sorties dans *la rue de Bourgogne* avant d'ajouter que le centre-ville *c'est là qu'y a tout*. C'est cette rue en particulier que le locuteur a choisi de mentionner pour parler de ses habitudes : il associe une certaine pratique à une certaine rue. Dans l'exemple 48, le cinéma Pathé n'étant pas d'accès facile, le locuteur préfère le cinéma des Carmes. De même que dans les exemples 38 et 52 où il est question des endroits où faire ses courses. Le locuteur met en avant ses propres actions. Dans l'exemple 53, on retrouve l'idée d'une activité qui caractérise un lieu. Seulement, l'angle de vue n'est plus personnel. Au contraire, le locuteur prend de la distance et généralise : Orléans est une ville qui bouge grâce notamment aux Fêtes de Loire. Quand bien même il est question d'un groupe de personnes, cela reste le discours de celui qui parle. Ce n'est pas la perception du groupe qui est présentée, mais l'idée que s'en fait le locuteur. Ainsi, le locuteur de l'exemple 61 rapporte sa perception du centre-ville par le prisme des sorties tardives des personnes qui s'y rendent.

(61) y a qu'Orléans qui fait ça parce que d'une part ils sont dans le centre-ville donc les gens peuvent venir à pied pas besoin de prendre leur voiture donc ils peuvent repartir avec un coup dans le nez sans risque enfin plus ou moins sans risque (ESLO2_ENT_1026)

L'expérience personnelle peut à la fois être exprimée directement avec l'emploi de la première personne du singulier mais elle peut aussi être exprimée indirectement comme dans cet exemple.

4.6.1.2.3 Expérience collective

La perception d'un objet peut aussi être partagée du point de vue des représentations ou des connaissances que l'individu a à son propos. Percevoir un objet c'est en avoir une image mentale. Cette image se construit à partir des émotions et sentiments ressentis mais aussi à partir des connaissances ou de la mémoire collective relative à l'objet. Dans :

(62) Michel met souvent un petit trait de euh de vinaigre euh dessus sur l'omelette oui c'est il parait que c'est orléanais je sais pas (ESLO2_ENT_1063)

Ce n'est pas une expérience dans un lieu qui est décrite. C'est plutôt le récit d'un comportement qui pourrait être attribué aux Orléanais. Le locuteur emploie l'expression *il parait*, ce qui indique qu'il n'a pas de certitude et qu'il partage plutôt d'une image collective d'un comportement particulier. Il rapporte des expériences qu'il n'a pas lui-même vécues mais qui sont révélatrices d'une représentation de la ville dans les yeux de la communauté. On retrouve dans cette catégorie les anecdotes à propos de la *Rumeur d'Orléans* ou des *chiens d'Orléans* comme dans les exemples 55 et 56 ou encore les comparaisons que les locuteurs font entre Orléans et d'autres villes comme Paris ou Tours.

4.6.2 Perspectives de traitement de la perception

Les recherches en TAL à propos de la subjectivité s'intéressent majoritairement à la détection de la polarité dans des données écrites. Dans ce contexte, nous avons d'abord procédé au classement des segments établis autour des lieux identifiés en fonction de leur caractère objectif ou subjectif. Puis, grâce au même modèle, nous avons identifié la polarité présente dans les segments considérés comme subjectifs. Les résultats du modèle pour ces deux tâches sont encourageants et même plutôt satisfaisants.

Néanmoins, la détection de ces informations ne suffit pas à décrire la perception qu'ont les Orléanais de leur ville. Pour rendre compte de l'image qu'ils en ont, il est nécessaire de considérer d'autres informations que la polarité. Une réflexion typologique sur la perception en tant que telle a été menée afin de décrire sur quoi elle porte, à travers quoi ou de quelle façon elle s'exprime. A partir d'une exploration manuelle du corpus, nous proposons deux axes complémentaires selon lesquels on peut décrire la perception.

Dans un premier temps, nous nous sommes intéressée à la cible même de la perception en décrivant les types d'objets mentionnés pour décrire la ville. Nous avons établi que les thématiques abordées dans les déclarations subjectives à propos de la ville se concentrent autour de l'urbanisme, des activités économiques, des relations sociales et du patrimoine historico-culturel. Dans un second temps, nous nous intéressons aux moyens par lesquels les locuteurs partagent leur perception et nous nous proposons de caractériser la nature de la perception. Au-delà de la catégorisation de ces thématiques, les locuteurs utilisent des procédés particuliers pour partager leur perception. Ils peuvent faire appel à leurs sens, se placer du point de vue des habitudes et des pratiques en fonction des lieux décrits ou bien des représentations et connaissances qu'ils ont les concernant.

Cette réflexion prospective donne un premier cadre pour une analyse approfondie de la perception. Sur le plan technique, les méthodes envisagées pour la détection de la subjectivité et de la polarité ne permettent pas en l'état la détection des cibles ou de la nature de la perception. Des travaux en TAL sur la détection de la cible d'opinion (Levy *et al.*, 2014 ; Lark, 2015 ; Ouertatani *et al.*, 2018 ; Villaneau *et al.*, 2019) doivent être examinés pour encadrer la construction d'un modèle pour la détection des thématiques de la perception. La mise en place d'un protocole d'apprentissage supervisé pour la détection de la cible et de la nature de la perception suppose une phase d'annotation manuelle. L'identification des indices représentatifs de ces deux typologies grâce à l'exploration manuelle des données permettrait de guider le choix d'un classifieur, d'une méthode de

vectorisation et des *features* les plus pertinents pour entraîner un nouveau modèle et atteindre les nouveaux objectifs établis.

PARTIE III

MODELISATION
DE LA PERCEPTION DE LA VILLE D'ORLEANS

Chapitre 5 : VISUALISATION DE LA PERCEPTION

5.1 Enjeux et problématique de la visualisation de données

Selon Karoui et al. (2014 : 1), le *Big Data* résulte de « nouveaux usages liés à la prolifération d'Internet » et aux « progrès de la technique » qui contribuent à « une avalanche de données produites à grande échelle et souvent de manière non structurée ». L'essor du Web 2.0 et des nouvelles technologies a contribué à rendre disponibles des « masses » de données dans des formes et des modalités différentes. L'abondance des données, ainsi que le développement d'outils d'analyse et de modélisation dédiés, ouvrent d'immenses perspectives dans les domaines de la recherche en général.

Comment procéder pour extraire les informations pertinentes de ce « déluge de données » (Sebillot, 2015 : 1) ? De nombreuses recherches en TAL tentent de répondre à cette question dans l'optique d'aider l'utilisateur à se retrouver dans l'ensemble d'informations à sa disposition. Ces techniques sont par exemple employées pour résoudre des problématiques en traduction automatique ou en résumé automatique. Des informations sont extraites de grands corpus pour répondre à des besoins spécifiques. Si les outils d'aide à l'analyse de corpus atteignent aujourd'hui une certaine maturité, ceux-ci trouvent leurs limites pour mettre « en évidence des faits importants, des relations entre les faits et surtout leur interprétation » (Poibeau, 2014).

L'une des réponses possibles à ces limites peut se trouver dans les problématiques de représentation de l'information. La visualisation des informations extraites participe à l'émergence de nouvelles connaissances à propos des objets étudiés. La visualisation d'informations est un domaine pluridisciplinaire qui trouve ses origines dans l'informatique. Son objet est l'étude de la représentation visuelle des données, principalement abstraites, sur une interface graphique. Pour Hascoët & Beaudouin-Lafon (2001 : 3), le but de la visualisation d'information est « d'exploiter les caractéristiques du système visuel humain pour faciliter la manipulation et l'interprétation de données

informatiques variées ». Selon eux, la visualisation d'information peut remplir plusieurs missions :

- « exploration rapide d'ensembles d'informations inconnues » (*op. cit.* : 3).
- « mise en évidence de relations et de structures dans les informations » (*op. cit.* : 3).
- « mise en évidence de chemins d'accès à des informations pertinentes » (*op. cit.* : 3).
- « classification interactive des informations » (*op. cit.* : 3).

Dans ce domaine, il s'agit de passer de la donnée à la connaissance au moyen de l'image. Pour cela, on peut utiliser des tableaux ou des graphiques mais aussi des cartes mentales, des arborescences, des nuages de mots, etc. La représentation visuelle permet d'avoir une vision synthétique de l'information.

La question de la visualisation des informations est essentielle du point de vue de la géographie. L'élaboration de cartes ou la mise en place de Systèmes d'Information Géographique est un enjeu majeur pour la représentation de l'information spatiale. Ces problématiques s'intègrent dans une logique transdisciplinaire en cherchant à mettre en évidence les liens qu'entretient l'Homme avec l'espace. Ainsi, des Groupes d'Intérêt Spécial (GIS) s'organisent autour des *Geo-humanities* que la DARIAH[127] présente ainsi :

> Le groupe de travail Géo-Humanités établira un registre qui utilisera la technologie du Web sémantique pour interroger les ensembles de données et les services géospatiaux en fonction de leur pertinence et de leur applicabilité à une région géographique ou à un domaine de recherche en sciences humaines précis. [128]

[127] https://www.dariah.eu/2018/03/22/geo-humanities-a-new-dariah-working-group-dealing-with-geo-spatial-data/

[128] « *The working group Geo-Humanities will establish a registry that will use semantic web technology to allow querying geospatial datasets and services in terms of their relevance and applicability for a specific geographic region and/or humanities research domain* »

L'objectif de la dernière étape de notre travail est de représenter visuellement les résultats des analyses réalisées jusqu'ici. Cette étape commande de réfléchir aux moyens les plus adaptés pour permettre la visualisation de la perception. Il s'agit de mettre en valeur les informations extraites de notre corpus afin de faciliter leur manipulation et les rendre accessibles. Dans une autre mesure, la visualisation des lieux et des annotations réalisées contribue faire émerger les relations qui unissent les différentes informations détectées.

5.2 Représentation de la perception d'Orléans

5.2.1 Nuages de mots

Les nuages de mots (ou nuages de tags) sont un moyen de représentation visuelle d'un document. Ils découlent directement des méthodes d'apprentissage de type *word embedding*. Les espaces vectoriels obtenus grâce aux plongements de mots sont très utiles pour des tâches d'apprentissage supervisées mais leur contenu est assez peu lisible en l'état par l'humain. L'utilisation de nuages de mots facilite la navigation dans les espaces lexicaux en permettant leur visualisation.

Classiquement, les données textuelles sont représentées sous la forme de tableaux et de listes. Selon Boullier & Crépel (2009), les nuages de mots sont un dispositif graphique permettant de « s'orienter, […] de hiérarchiser et en tout cas d'attirer l'attention sur des indices en produisant des différences ». En fonction de la requête formulée, le nuage de mots met en valeur les termes jugés les plus pertinents en faisant varier leur forme et leur couleur. Le système peut par exemple mettre en valeur les mots les plus souvent mentionnés dans un document donné ou bien faire émerger les mots les plus proches d'un autre à l'intérieur d'un espace vectoriel. La forme du nuage de mots peut faire émerger de nouvelles connaissances extraites du document représenté.

Différents outils permettent de créer des nuages de mots. EDF a par exemple développé l'outil Wordsurf (Suignard, 2017) à partir du modèle Word2vec, afin de traiter les questions posées à son agent conversationnel Laura. Le module Wordcloud2.js[129], diffusée sous

[129] https://github.com/amueller/word_cloud

licence libre, permet de générer des nuages de mots. Cette librairie exclue les mots vides et se base sur leade chaque terme pour représenter un document : plus un terme est mentionné, plus il est inscrit en grand.

Dans la figure 29, nous avons utilisé Wordcloud pour créer un nuage de mots représentant les segments ciblant des lieux situés à Orléans dans notre corpus de référence. Puisque les segments analysés ciblent Orléans, on constate logiquement la prédominance du nom de la ville mais aussi du vocabulaire citadin avec des termes comme *quartier*, *ville* et *place*. La saillance du mot Loire illustre son importance. Nous savons que les segments du corpus de référence sont majoritairement objectifs et que les segments subjectifs sont plutôt positifs. On retrouve cette information dans les mots qui ressortent du nuage avec la présence d'adjectifs à connotation positive comme *bien*, *petit*, *vrai* et *grand*.

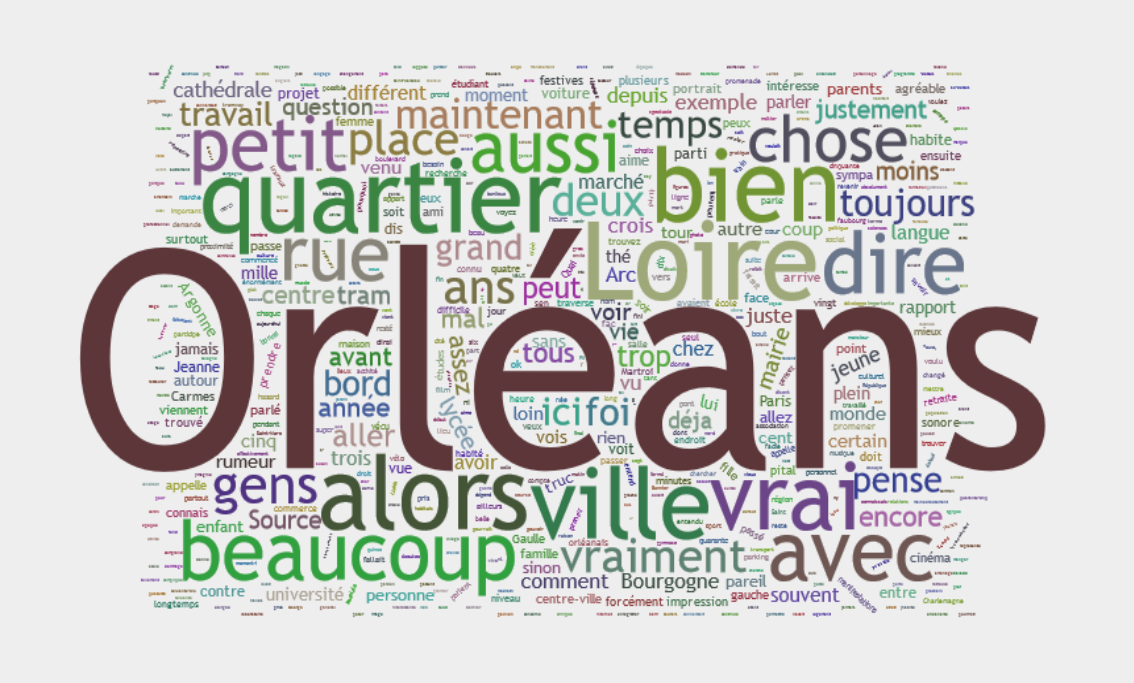

Figure 29 : Nuage de mots des lieux situés à Orléans

Ce type de représentation permet d'identifier d'un seul coup d'œil les termes les plus représentatifs d'un document. Une autre possibilité de visualisation des informations extraites du corpus est la représentation cartographique. Celle-ci implique de disposer des coordonnées géographiques des lieux à représenter. Or, comme décrit dans la section

2.4.4.6, certains noms de lieux sont des noms communs que nous ne pouvons pas rapprocher du nom propre. Dans ces cas-là, la représentation cartographique est impossible. Les nuages de mots peuvent donc être une alternative pour observer le contenu des déclarations faites à propos de ces lieux.

Dans les exemples 54 et 57, il est question d'un *quartier* d'Orléans mais le système n'est pas en mesure de déterminer lequel précisément. Ces déclarations, pourtant jugées subjectives, ne pourront pas être projetées dans la représentation cartographique. Le nuage de mots constitue donc une alternative pour visualiser les informations subjectives à propos de ce type de lieu. La figure 30 présente le nuage correspondant aux segments ciblant un *quartier* dans la zone 2 considérant les lieux d'Orléans. Le nuage révèle par exemple une forte thématique sociale dans les segments analysés avec beaucoup d'adjectifs comme *agréable*, *social*, *difficile*, ou encore *sympa* qui décrivent l'ambiance dans les quartiers évoqués et des adverbes comme *justement*, *beaucoup* et *vraiment* viennent appuyer ces adjectifs.

Figure 30 : Nuage de mots des segments pour le terme *quartier*

Les nuages de mots peuvent être aussi utilisés pour représenter les déclarations en fonction des autres informations disponibles sur les lieux comme leur zone ou leur type. La figure 31 est la représentation des segments ciblant des lieux considérés comme naturels. Le terme le plus représentatif de ce type est la *Loire*, ce qui est cohérent avec le nuage des lieux situés à Orléans[130]. On trouve ensuite les termes *bord* et *quai*, souvent associés à la Loire mais aussi au *Loiret*. Le périmètre des segments ciblant des lieux naturels ne se limitant par à Orléans, on retrouve des termes comme *mer* et *plage*. On peut aussi remarquer que ce type de lieu est souvent décrit par le biais des activités réalisées par les locuteurs avec les termes *promenade*, *balade*, *vacances* et *festival*.

Figure 31 : Nuage de mots pour les lieux de type *naturel*

Cette méthode de visualisation des données permet d'en avoir une vision globale et peut servir de premier moyen d'accès à leur contenu. Pour avoir une vision plus précise des éléments observés, il faut utiliser une autre méthode. Les outils les plus indiqués pour représenter de l'information géographique sont les Systèmes d'Information Géographique.

[130] cf. Figure 29

5.2.2 Système d'Information Géographique

Un SIG est un outil utilisé en géographie, conçu pour recueillir, stocker, analyser, gérer et présenter tous types de données spatiales et géographiques. La visualisation de la perception par les Orléanais de leur ville peut se concrétiser avec la mise en place d'un Système d'Information Géographique (SIG) et la création de cartes. Pour cela, nous utilisons l'outil ArcGIS Online, développé par Esri et disponible en ligne. Dans cette section, nous présentons la préparation des données en vue de leurs projections dans le SIG avant de proposer une analyse des dynamiques géographiques émergeant dans le SIG par rapport à la perception d'Orléans.

5.2.2.1 Préparation des données

L'ensemble des traitements étudiés jusqu'ici ont été réalisés dans la perspective de l'implémentation finale du SIG destiné à représenter la perception qu'ont les Orléanais de leur ville.

L'étape initiale de détection des lieux[131] présentait un double objectif. Le premier était d'identifier tous les lieux présents dans les transcriptions des conversations analysées, de les décrire grâce à différentes informations dans le but de préparer l'analyse de la perception. En parallèle, le deuxième objectif était de préparer les données extraites pour les rendre interprétables dans le SIG final. L'information essentielle à obtenir pour répondre à cet objectif est la détermination des coordonnées géographiques de chaque lieu identifié. Des bases de données dédiées à l'information géographique ont été utilisées comme lexique afin d'identifier les lieux dans le corpus et récupérer leurs coordonnées géographiques. Cependant, l'exploration manuelle du corpus a révélé certaines difficultés relatives au nommage des lieux à l'oral. La complexité principale concerne le fait que les locuteurs peuvent se référer à des lieux sans utiliser leur nom officiel[132]. Des règles ont donc été élaborées pour identifier les lieux et lorsque cela est possible, les associer à leurs coordonnés géographiques malgré les variations dans la désignation. Enfin, chaque lieu a

[131] cf. Partie I
[132] cf. section 2.4.2

été catégorisé en fonction de son type [133] et de la zone géographique à laquelle il appartient[134].

A partir des lieux identifiés, des segments sont définis pour jouer le rôle de fenêtres d'observations lors de l'analyse de la perception. Grâce à un modèle entraîné par apprentissage automatique supervisé, le caractère subjectif et la polarité de chaque segment sont identifiés.

En ajoutant à ces informations les métadonnées relatives à l'enregistrement dont sont extraites les déclarations subjectives faites à propos des lieux d'Orléans nous disposons de toutes les données essentielles à la construction du SIG.

5.2.2.2 ArcGIS Online

Développé par la société Esri (*Environmental Systems Research Institute*, Inc.)[135], ArcGIS est une suite de logiciels dédiés au traitement de SIG. Comme affiché sur son site officiel, ArcGIS est un « système complet qui permet de collecter, organiser, gérer, analyser, communiquer et diffuser des informations géographiques ». Cette plateforme largement utilisée pour des applications industrielles ou scientifiques permet de gérer, éditer et diffuser des informations géographiques au moyen de navigateurs Web, d'applications pour mobile et ordinateur.

ArcGIS Online[136] est une version simplifiée de ArcGIS. Disponible en ligne et gratuit, il permet de créer et partager simplement des cartes. Ce service est utilisé dans ce travail pour créer le système dans lequel est présentée la perception de la ville d'Orléans à partir des données extraites grâce aux traitements mis en place pour notre étude[137].

La figure 32 présente une capture d'écran d'ArcGIS Online et de la carte présentant la perception d'Orléans. Plusieurs couches d'informations sont projetées dans le système et permettent de figurer les éléments suivants :

[133] cf. section 2.2.1
[134] cf. section 2.2.2
[135] https://www.esri.com/fr-fr/home
[136] https://www.arcgis.com/index.html
[137] SIG pour la visualisation de la perception de la ville d'Orléans – http://arcg.is/1v9DaC

- L’ensemble des lieux détectés en fonction de leur type
- La densité des mentions de lieux à Orléans
- Le nombre de mentions de villes
- La polarité associée aux lieux d’Orléans

La signification de chacune des couches est présentée dans la section suivante.

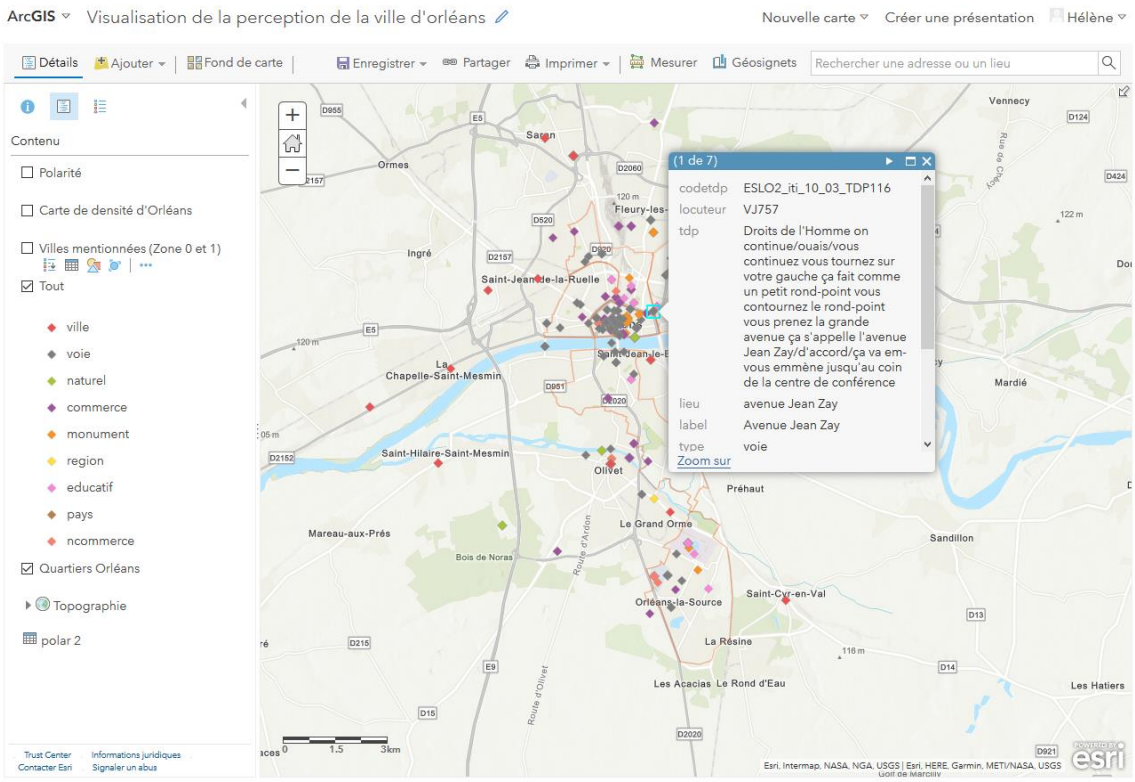

Figure 32 : Interface ArcGIS Online

5.2.3 Représentation cartographique de la perception

Différents types d’informations ont pu être projeté dans ArcGIS Online afin de visualiser les données extraites des différents traitements mis en place. Ces informations organisées en couches recouvrent des périmètres différents.

La première couche nommée *Tout*, affichée dans la figure 32, présente l’ensemble des segments établis à partir de la détection des lieux, peu importe l’emplacement de ces lieux.

Chaque point sur la carte représente un lieu et est colorisé en fonction du type qui lui a été attribué comme indiqué dans la légende visible sur la gauche de l'outil. Les lieux sont associés aux informations qui les décrivent (zone, type, polarité, etc.) et aux déclarations faites à leur sujet. On peut voir dans la figure 32 les informations relatives à l'avenue Jean Zay que l'on voit catégorisée comme une voie. Le système indique en haut de la fenêtre contextuelle que l'on peut accéder à sept déclarations à propos de ce lieu. Dans l'exemple affiché, on voit bien la déclaration, le code du locuteur qui l'a réalisée et le nom de l'enregistrement dont elle est extraite.

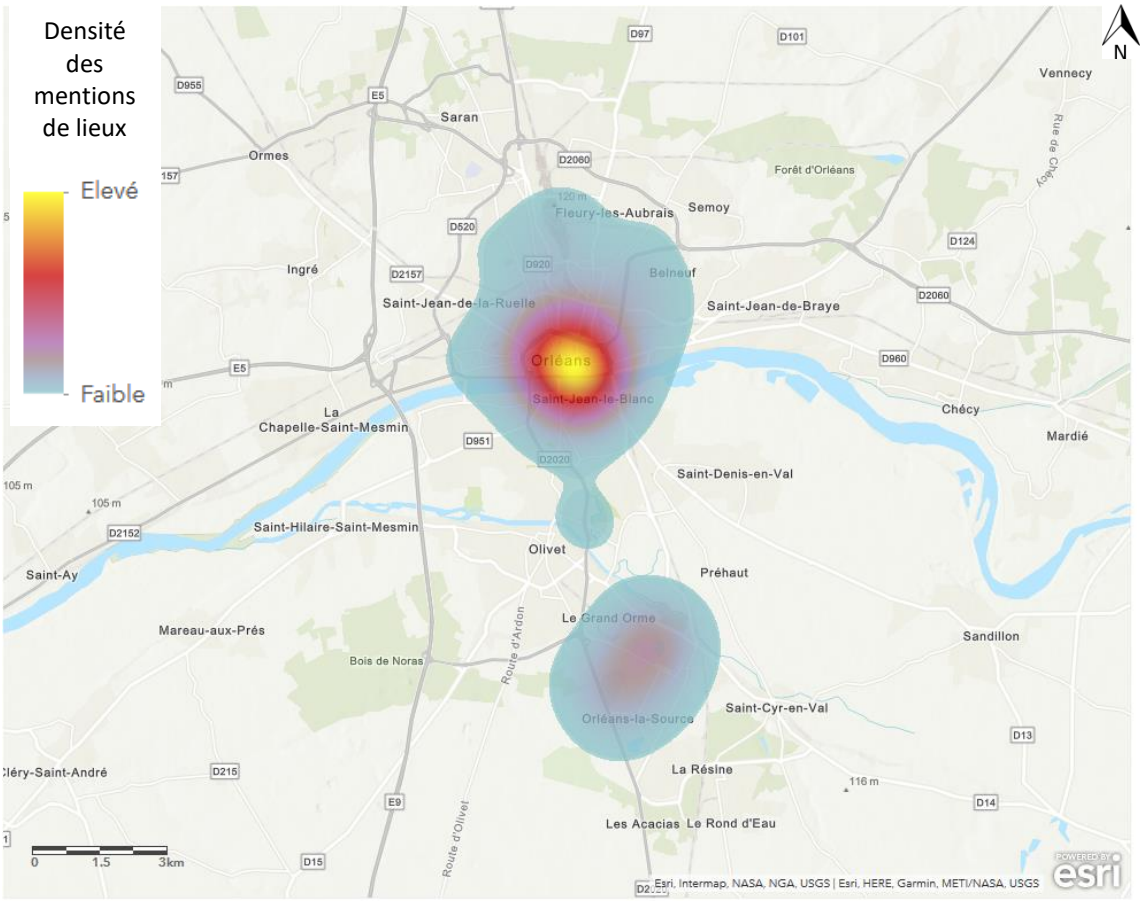

Figure 33 : Carte de densité des lieux d'Orléans

Lorsqu'une couche contient un grand nombre d'entités, l'affichage individuel de chaque entité sur la carte est souvent inutile. En l'état, on ne voit pas la différence entre les lieux qui n'ont été mentionnés qu'une seule fois et ceux qui l'ont été des dizaines de fois. Dans cette représentation, les entités se superposent, ce qui empêche de les distinguer clairement.

Même lorsqu'elles ne se superposent pas, il est généralement difficile, voire impossible, d'extraire visuellement les informations significatives lorsque des centaines ou des milliers de points sont affichés en même temps.

Pour résoudre ce problème, il est possible de générer une carte de densité (ou *heat map*). Une carte de densité représente la densité géographique des entités à représenter sur une carte à l'aide de zones colorées représentant ces points. Les zones les plus grandes sont celles où la concentration de points est la plus importante. La figure 33 illustre la carte de densité des lieux situés à Orléans : plus la couleur s'éloigne du bleu pour aller vers le jaune, plus la concentration de lieux identifiés est importante. On remarque que deux zones se distinguent. La première, et la plus dense, est située au niveau du centre-ville. C'est de ce quartier et des lieux qui s'y trouvent dont les locuteurs parlent le plus. La deuxième zone se situe au niveau du quartier Orléans La Source. Ces quartiers concentrent la plupart des services de la ville et il semble donc logique que ce soit de ces zones-là dont les gens parlent. La carte de densité révèle aussi un espace qui semble délaissé par les locuteurs, c'est-à-dire la partie sud du quartier Saint-Marceau. Aucun lieu n'a été ne serait-ce que mentionné dans cette partie du quartier. Nous avons établi que les locuteurs se réfèrent aux espaces dans lesquels ils expérimentent, dont on leur parle. Ainsi, si cette partie de la ville n'est pas mentionnée, cela signifie que les locuteurs n'y vivent pas d'expériences particulières.

Une autre stratégie peut être employée pour représenter les entités les une par rapport aux autres en fonction de leurs occurrences. Ainsi, les villes représentées dans la couche nommée *Villes mentionnées (Zone 0 et 1)* le sont en fonction du nombre de fois où elles ont été mentionnées : plus le point est grand, plus le nombre est élevé. Cette carte nous permet de confirmer le fait que les locuteurs parlent avant tout d'Orléans et des villes des alentours. Mais on peut aussi constater que Paris et Tours sont souvent mentionnés. En effet, la comparaison entre Tours et Orléans et entre Paris et Orléans est récurrente lorsqu'il s'agit pour les locuteurs de décrire Orléans.

Enfin, une dernière couche présente la polarité exprimée dans les segments identifiés. Deux informations sont simultanément représentées : le nombre d'occurrence des lieux et leur polarité.

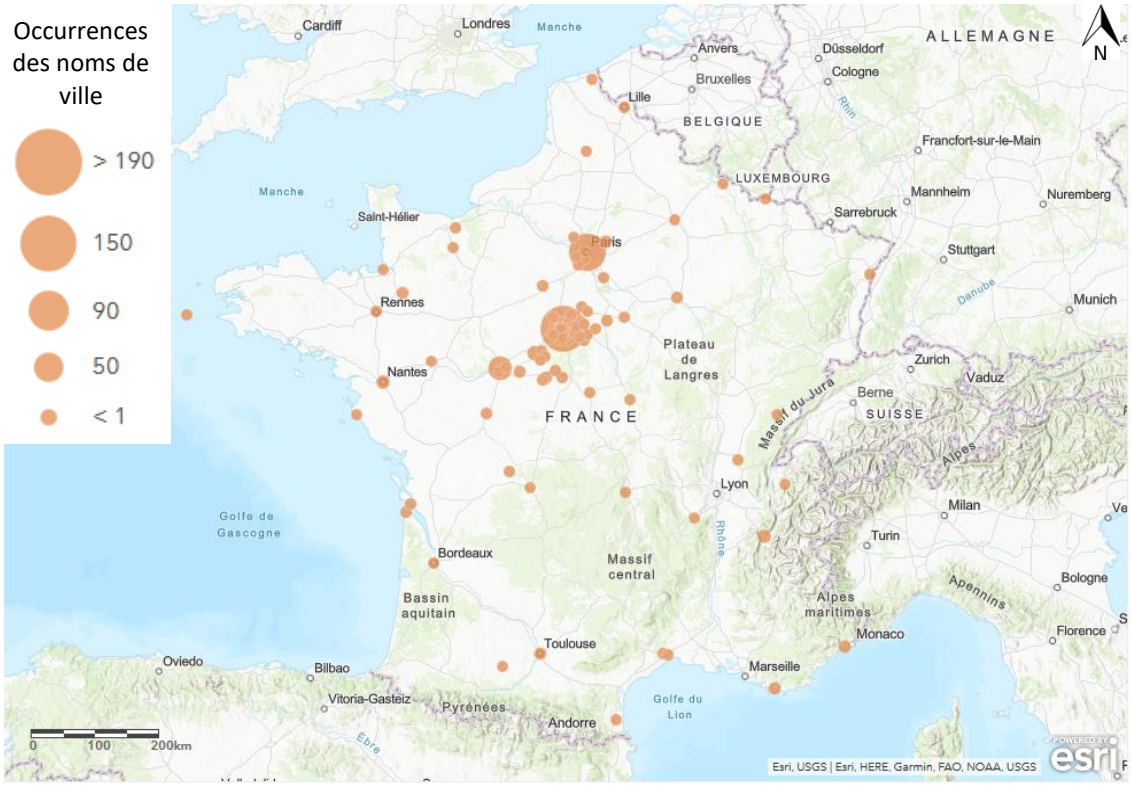

Figure 34 : Répartition des villes sur le territoire en fonction de leurs occurrences

En ce qui concerne la polarité, un même lieu peut être à la fois perçu comme positif ou comme négatif. On ne peut donc pas simplement représenter les segments positifs par une couleur et les segments négatifs par une autre. Nous avons donc choisi de calculer un ratio pour faire la part entre les deux polarités pour une même entité. Pour calculer ce ratio, nous calculons le pourcentage des segments positifs par rapport au total des segments considérés comme subjectif. Ensuite, ce pourcentage est divisé par 20 de manière à ce que la valeur obtenue soit comprise entre 0 et 5. Ainsi, plus le score obtenu est proche de 0, plus la polarité générale du lieu est négative. Au contraire, plus le score s'approche de 5, plus la polarité générale est positive.

Afin de relativiser ce score, la représentation cartographique tient aussi compte du nombre de segments subjectifs relevés à propos du lieu. C'est-à-dire que de la même façon que pour les villes, plus le lieu est mentionné, plus le point est gros. En effet, on ne peut pas considérer de la même façon le ratio de polarité d'un lieu pour lequel on a une seule déclaration subjective et le score d'un lieu en ayant une dizaine.

La figure 35 présente la polarité des lieux situés dans le centre ville d'Orléans. On y observe le point représentant la Loire qui est bien plus volumineux que les autres et auquel sont associés quarante segments positifs pour onze segments négatifs. Le ratio vaut donc 3,92 ce qui est considéré comme plutôt positif. Sur la carte, on trouve des points comme la Cathédrale Sainte-Croix d'Orléans qui comptent bien moins d'occurrences que la Loire et pour lesquels le ratio vaut 5. Le score de 5 reste excellent mais en ajoutant l'information du nombre d'occurrences, on peut considérer comme tout aussi positif le ratio obtenu pour la Loire.

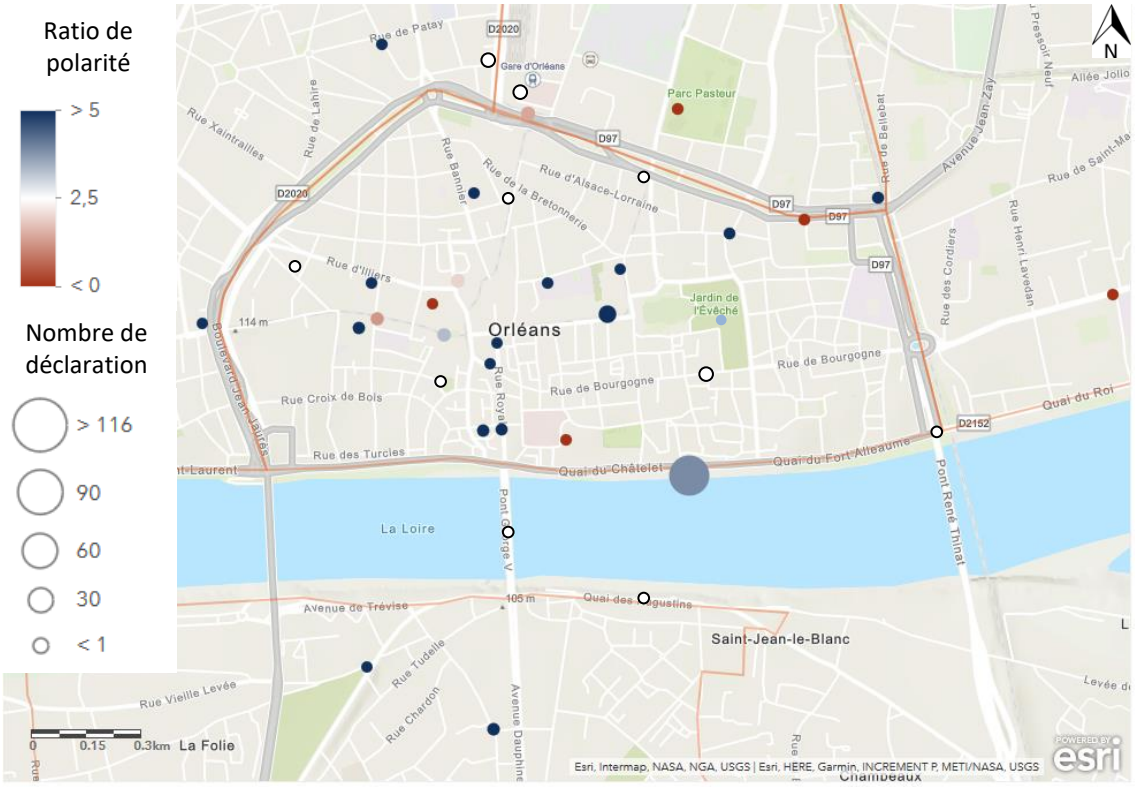

Figure 35 : Répartition de la polarité à propos des lieux du centre-ville d'Orléans

5.2.4 Perspectives pour la visualisation de la perception d'Orléans

Bien que les données projetées sur la carte ne soient qu'une partie infime du corpus ESLO, celles-ci permettent déjà de visualiser certaines dynamiques géographiques ou de confirmer la pertinence des typologies de la perception qui ont été établies.

Grâce à l'utilisation de nuages de mots et à l'implémentation d'un SIG, différentes dimensions de cette perception sont matérialisées. La représentation cartographique replace les lieux dans l'espace. Elle illustre la manière dont s'articulent les lieux les uns par rapport aux autres en montrant le nombre de fois où ils ont été mentionnés et en mettant en évidence la polarité qui leur est associée. Les nuages de mots font ressortir le contenu des segments en faisant apparaître le lexique le plus représentatif du lieu ciblé. Leur utilisation permet en plus de décrire les lieux qui ne sont pas géolocalisables ou de les analyser en fonction des catégories définies par la typologie des lieux et la typologie de la perception.

Les nuages de mots pourraient être intégrés au SIG pour renforcer l'importance du discours relatif aux objets géographiques décrits. Cette combinaison serait très utile pour représenter les segments annotés en fonction des typologies de la cible et de la nature de la perception. La représentation en nuages de mots mettrait l'accent sur le lexique employé en fonction de la thématique abordée ou de la manière dont le locuteur parle du lieu. On pourrait par exemple observer quel type de lieu est le plus souvent perçu par le biais des sens ou plutôt par les actions réalisées ou encore si certaines thématiques sont plus ou moins susceptibles d'être polarisées et dans quelle mesure.

Cette possibilité n'est pas implémentable dans la version en ligne d'ArcGIS. Mais, dans l'éventualité du développement d'une application dédiée, cette option serait intégrable. A chaque espace localisé peut être associé un nuage de mots construit à partir des segments le ciblant. Des liens seraient établis entre les termes communs aux différents lieux. L'utilisateur pourrait se servir de ces liens pour naviguer dans la carte et parcourir les différentes déclarations extraites d'ESLO.

Au-delà de l'articulation des lieux dans l'espace, d'autres informations pourraient être ajoutées dans le SIG pour fournir plus d'éléments de contexte aux déclarations extraites.

Par exemple, on peut envisager des liens entre les segments projetés sur la carte et les transcriptions d'où ils sont extraits. Le corpus ESLO est disponible en ligne et est notamment archivé sur la plateforme COCOON (COllections de COrpus Oraux

Numériques)[138]. COCOON est une plateforme qui « accompagne les producteurs de ressources orales, pour créer, structurer et archiver leurs corpus ». Chacun des enregistrements d'ESLO est décrit par une notice et associé à sa transcription. Etablir un lien entre les segments et leurs transcriptions d'origine permettrait de replacer les déclarations des locuteurs dans le contexte de la conversation.

Des liens pourraient aussi être établis avec des bases de données externes à ESLO contenant des informations objectives à propos du lieu identifié, par exemple, les déclarations sur la Cathédrale Sainte-Croix d'Orléans associées à la page Wikipédia correspondante. L'ajout de données objectives à propos du lieu apporterait une nouvelle perspective en replaçant le lieu dans un référentiel plus large que celui de la conversation. Le fait de prendre en compte les témoignages d'ESLO et d'y associer l'ensemble des informations des bases de données donnerait une dimension anthropologique et sociologique à la carte produite. Un tel traitement est techniquement possible et la suite de ce travail prévoit sa concrétisation.

[138] https://cocoon.huma-num.fr/exist/crdo

CONCLUSION ET PERSPECTIVES

Contribution

Les travaux présentés dans cette thèse s'inscrivent dans le cadre pluridisciplinaire des Humanités Numériques en associant la linguistique, le Traitement Automatique des Langues, et de la géographie. Le travail initié s'interroge sur l'exploitation nouvelle de données linguistiques dans un corpus à dimension sociolinguistique avec l'objectif d'en extraire du contenu subjectif. En effet, à partir de l'exploitation du corpus d'enregistrements sonores ESLO, la finalité de cette étude est de modéliser, détecter et visualiser la perception des locuteurs de la ville d'Orléans. Pour cela, différentes techniques du TAL ont été utilisées.

Dans la première partie, la notion de *lieu* a d'emblée été positionnée par rapport à plusieurs approches disciplinaires qui reflètent sa complexité définitoire. La définition du lieu combine les approches lexicales de la linguistique, les réflexions en géographie sur la relation qui unit l'homme à l'espace et les questionnements typologiques dans le domaine de la reconnaissance d'entités nommées (REN) et la détection des entités spatiales (ES). Le lieu est ici 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.

Les lieux sont le premier type d'informations à identifier dans le corpus de transcriptions pour ancrer l'analyse de la perception. La démarche proposée suit une approche symbolique fondée sur des lexiques et des règles élaborées grâce à une analyse approfondie des lieux et de leur nommage. Si des normes existent pour le nommage des lieux, chacun est à même de faire varier cette norme et de se référer aux espaces de son environnement par des moyens détournés. A l'oral en particulier, les noms de lieux peuvent être abrégés, tronqués ou il peut leur être substitués de nouvelles expressions. Le module de détection automatique mis en place s'appuie sur des lexiques extraits de bases de données dédiées à l'information géographique qu'il manipule grâce à des règles qui prennent en compte les caractéristiques propres de l'oral. Les lieux sont ainsi identifiés tel qu'ils sont nommés, c'est-à-dire sous la

forme utilisée quotidiennement par les gens. De plus, les conventions d'annotation des lieux anticipent des traitements ultérieurs en associant les variantes de noms de lieux à leur forme conventionnelle afin de préparer l'étape finale de géolocalisation des lieux détectés.

Le module développé est évalué et obtient une F-Mesure de 0,91 avec un Rappel de 0,90 et une Précision de 0,93. Des performances satisfaisantes qui montrent l'efficacité du module de détection *lieux* à gérer les difficultés liées à la reconnaissance de ce type d'entité et qui sont complexifiées par les spécificités de l'oral.

La deuxième partie de la thèse apporte un éclairage sur la notion de subjectivité et surtout sur l'articulation des émotions, des sentiments et des opinions avec la notion de *perception*. Percevoir un objet passe par son expérimentation physique et sensorielle. La perception est le processus de traitement d'informations collectées par les sens permettant de créer des représentations et des connaissances à propos de l'objet perçu. Elle est une expérience qui varie d'un individu à l'autre et qui peut être captée lorsque l'individu décide de la partager.

A partir de la détection des lieux, les transcriptions sont divisées en segments susceptibles d'être porteurs d'éléments subjectifs à propos des lieux ciblés. Ces segments sont analysés par apprentissage automatique supervisé afin de détecter leur caractère subjectif ou objectif ainsi que leur polarité positive ou négative. Différentes expériences ont été menées afin d'entraîner un modèle à même de répondre à cet objectif. Des traits linguistiques, qui sont la lemmatisation et l'étiquetage morpho-syntaxique des segments et le calcul de scores de polarité et d'émotion, ont été sélectionnés. Des méthodes de vectorisation des données textuelles et des classifieurs ont été comparées. Ces expériences ont permis de définir un modèle obtenant une macro average de 0,77 pour la tâche de détection de la subjectivité et de 0,76 pour celle de détection de la polarité dans les segments analysés.

La détection de la subjectivité et de la polarité oriente l'analyse de la perception mais ne suffit pas à en rendre compte. Pour aller plus loin, la typologie de la perception est approfondie. La première proposition typologique concerne la cible de la perception. L'exploration manuelle du corpus révèle en effet que l'expression de la perception à propos d'une ville se fait en fonction de thématiques précises. Celles-ci peuvent être classées en quatre catégories : urbanistique, sociale, économique et historico-culturelle. Afin de mieux comprendre les mécanismes de la perception, une deuxième typologie est proposée pour caractériser la manière dont les locuteurs font part de leur perception, selon quelles

modalités et dans quelles conditions. La perception peut être abordée du point de vue des habitudes des individus, des actions qu'ils réalisent ou de ce qu'ils expérimentent par les sens. Ces éléments décrivent la nature de la perception et recoupent les différentes thématiques pouvant être abordées. Pour décrire la nature de la perception, nous considérons trois catégories : la perception sensorielle, l'expérience personnelle et l'expérience collective.

Afin de confronter les segments subjectifs extraits des transcriptions avec les différents éléments détectés à leur sujet, des techniques de visualisation de l'information sont employées. Une représentation visuelle des données permet d'avoir une vision synthétique de l'information mais aussi de créer de la connaissance en passant du texte à l'image. Ainsi, la visualisation des lieux et des annotations réalisées contribue à faire émerger les relations qui unissent les différentes informations détectées. Il s'agit de mettre en valeur les informations extraites de notre corpus afin de faciliter leur manipulation et les rendre accessibles.

La troisième et dernière partie décrit les méthodes employée pour visualiser graphiquement la perception de la ville d'Orléans par les locuteurs du corpus ESLO. Des nuages de mots sont générés pour faire ressortir le contenu des segments en mettant en valeur le lexique le plus représentatif du lieu ciblé. Un Système d'Information Géographique (SIG) replace les lieux dans l'espace et illustre la manière dont s'articulent les lieux les uns par rapport aux autres en montrant notamment le nombre de fois où ils ont été mentionnés et en mettant en évidence la polarité qui leur est associée.

Perspectives

La démarche pluridisciplinaire présentée exploite les techniques variées du TAL (les méthodes symboliques et les méthodes d'apprentissage supervisé), les outils de la géographie (SIG) et les approches de la linguistique de corpus. Cette démarche est généralisable sur d'autres données et peut être appliquée sur des enregistrements portant sur d'autres villes.

Les retombées de ce travail peuvent être nombreuses. En premier lieu, la carte finale obtenue offre une nouvelle manière d'accéder au corpus ESLO qui se présente comme le

portrait sonore de la ville d'Orléans. La matérialisation de ce portrait de la ville d'Orléans ancre d'une part la dimension patrimoniale et anthropologique du corpus, et d'autre part, le témoignage qu'il représente. Ce témoignage est révélateur de l'attractivité de la ville et peut permettre à des personnes qui voudraient emménager à Orléans d'identifier les quartiers correspondant le mieux à leur mode de vie. Le système élaboré peut aussi servir pour des applications dans les domaines du tourisme et de l'urbanisme. Les touristes sélectionnent les lieux mis en avant par les orléanais eux-mêmes pour organiser leur séjour. L'analyse de la carte peut répondre à des problématiques d'aménagement du territoire en mettant par exemple en évidence certaines améliorations demandées par les administrés ou au contraire évaluer l'impact que des travaux ont pu avoir sur ces derniers

Bien que quelques améliorations puissent encore être réalisées, nous avons montré la faisabilité de nos propositions sur des données authentiques. Ainsi, des perspectives s'ouvrent au terme de cette étude :

- L'élaboration des typologies de la cible et de la nature de la perception est la première étape vers leur détection automatique. De nouvelles expériences vont être menées pour identifier des traits linguistiques pertinents pour cette nouvelle analyse afin d'entrainer un nouveau modèle.

- D'un point de vue technique, le SIG mis en place est limité par les fonctionnalités de la plateforme en ligne ArcGIS Online. L'amélioration de la visualisation de la perception d'Orléans passe par le développement d'une application dédiée. Cette nouvelle application permettrait :
    - → D'intégrer les nuages de mots générés pour préciser la représentation des lieux géolocalisés dans le SIG. Ces nuages auront aussi la fonction de faciliter la navigation dans le système ainsi que la manipulation des données représentées.
    - → Etablir des liens entre les segments représentés et les transcriptions dont ils sont extraits, archivées sur la plateforme COCOON. Cette liaison permet de réintroduire le segment dans son contexte conversationnel.

- L’analyse de la perception peut aussi être appréhendée d’un point de vue temporel. Les locuteurs peuvent décrire un même lieu dans des temporalités différentes. Il serait intéressant de pouvoir détecter cette information pour suivre l’évolution de la perception d’un même lieu dans le temps. On pourrait ainsi envisager une analyse micro-diachronique de la perception de la ville en se basant sur le discours des locuteurs ou en comparant le corpus ESLO1, réalisé à la fin des années 70, avec le corpus ESLO2 constitué à partir de 2014.

- Plus largement, la méthodologie décrite dans cette thèse pourrait être utilisée pour analyser la perception au sujet d’autres villes ou même de lieux de nature différente. L’extension de l’étude de la perception pourra permettre la comparaison de différents espaces géographiques évoqués dans des corpus de nature similaire.

BIBLIOGRAPHIE

ABDAOUI, Amine, AZÉ, Jérôme, BRINGAY, Sandra et PONCELET, Pascal, 2017. Feel: a french expanded emotion lexicon. Language Resources and Evaluation. Vol. 51, n° 3, pp. 833–855.

ABOUDA, Lotfi et BAUDE, Olivier, 2006. Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. Le cas des ESLO. In : *Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation*. 2006.

ABOUDA, Lotfi et SKROVEC, Marie, 2017. Alternance futur simple/futur périphrastique: variation et changement en français oral hexagonal. Revue de Sémantique et Pragmatique. Vol. 41, n° 41-42, pp. 155–179.

ABOUDA, Lotfi et SKROVEC, Marie, 2018. Pour une micro-diachronie de l'oral: le corpus ESLO-MD. In : SHS Web of Conferences. EDP Sciences. pp. 11004.

AL-RFOU, Rami, KULKARNI, Vivek, PEROZZI, Bryan et SKIENA, Steven, 2015. Polyglot-NER: Massive Multilingual Named Entity Recognition. Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015. avril 2015.

ARTSTEIN, Ron et POESIO, Massimo, 2008. Inter-coder agreement for computational linguistics. Computational Linguistics. Vol. 34, n° 4, pp. 555–596.

BAILLY, Antoine et BEGUIN, Hubert, 2005. Introduction à la géographie humaine (éd. 8). Paris: Armand Colin.

BARBARAS, Renaud, 2009. La perception: essai sur le sensible. Vrin.

BAUDE, Olivier et DUGUA, Céline, 2011. (Re) faire le corpus d'Orléans quarante ans après: quoi de neuf, linguiste? Corpus. n° 10, pp. 99–118.

BEAUNIS, Henri-Etienne, 1904. Orth Gefühl und Bewusstseinslage. L'Année psychologique. Vol. 11, n° 1, pp. 654–656.

BLAKE, Brian P, AGARWAL, Nitin, WIGAND, Rolf T et WOOD, Jerry D, 2010. Twitter Quo Vadis: Is Twitter bitter or are tweets sweet? In : 2010 Seventh International Conference on Information Technology: New Generations. IEEE. 2010. pp. 1257–1260.

BLANCHE-BENVENISTE, Claire, BILGER, Mireille, ROUGET, Christine, VAN DEN EYNDE, Karel, MERTENS, Piet et WILLEMS, Dominique, 1990. Le français parlé (études grammaticales). Sciences du langage.

BLOCH, Susana, ORTHOUS, Pedro et SANTIBAÑEZ-H, Guy, 1987. Effector patterns of basic emotions: A psychophysiological method for training actors. Journal of Social and Biological Structures. Vol. 10, n° 1, pp. 1–19.

BLONDIAUX, Loïc, 2016. La fabrique de l'opinion. Une histoire sociale des sondages. Le Seuil.

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. Editions Ophrys.

BORTHWICK, Andrew, STERLING, John, AGICHTEIN, Eugene et GRISHMAN, Ralph, 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In : Sixth Workshop on Very Large Corpora.

BOTTINEAU, Didier, 2013. OUPS! Les émotimots, les petits mots des émotions: des acteurs majeurs de la cognition verbale interactive. Langue française. n° 4, pp. 99–112.

BOULLIER, Dominique et CREPEL, Maxime, 2009. La raison du nuage de tags: format graphique pour le régime de l'exploration? Communication langages. n° 2, pp. 111–125.

BOULLIER, Dominique et LOHARD, Audrey, 2012. Opinion mining et Sentiment analysis: Méthodes et outils. OpenEdition Press.

BOURDIEU, Pierre, 1979. Public opinion does not exist. Communication and class struggle. Vol. 1, pp. 124–130.

BOUVIER, Jean-Claude, 1999. Odonymes d'agglomération entre l'écrit et l'oral. Nouvelle revue d'onomastique. Vol. 33, n° 1, pp. 303–310.

BRANDO, Carmen, DOMINGUÈS, Catherine et CAPEYRON, Magali, 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. ACM. 2016. pp. 7.

BREIMAN, Leo, 2001. Breiman and Cutler's Random Forest for Classification and Regression. R package version 4.5–16. In : The R Project Online.

CHARAUDEAU, Patrick, 2000. Une problématisation discursive de l'émotion. Les émotions dans les interactions. pp. 125–155.

CHINCHOR, N. A., 1998. Overview of MUC-7/MET-2. In : *Proceedings of the 7^th Message Understanding Conference (MUC7)*. S.l. : s.n. 1998.

COHEN, Jacob, 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement. Vol. 20, n° 1, pp. 37–46.

COULSON, Mark, 2004. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. Journal of nonverbal behavior. Vol. 28, n° 2, pp. 117–139.

GROUIN, Cyril et FOREST, Dominic, 2012. Expérimentations et évaluations en fouille de textes: Un panorama des campagnes DEFT. Lavoisier.

DAMASIO, Antonio R., 2006. L'erreur de Descartes: la raison des émotions. Odile Jacob.

DANTZER, Robert, 2002. Can farm animal welfare be understood without taking into account the issues of emotion and cognition? Journal of Animal Science. Vol. 80, n° E-suppl_1, pp. E1–E9.

DARWIN, Charles, 1872. The expression of emotions in animals and man. London: Murray.

DE LANDSHEERE, Gilbert, 1979. Comment les maitres enseignent.

DESCARTES, René. *Les Méditations métaphysiques de René Descartes touchant la première philosophie*. Huet, 1724.

DESCARTES, René, 1728. *Les passions de l'âme*.

DÉSOYER, Adèle, LANDRAGIN, Frédéric, TELLIER, Isabelle, LEFEUVRE, Anaïs et ANTOINE, Jean-Yves, 2015. Les coréférences à l'oral: une expérience d'apprentissage automatique sur le corpus ANCOR.

DISTER, Anne, 2007. De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données VALIBEL. PhD Thesis. Thèse de doctorat, Université catholique de Louvain, Louvain-la-Neuve.

DOMINGUÈS, Catherine et ESHKOL-TARAVELLA, Iris, 2013. Repérer des toponymes dans les titres de cartes topographiques.

DOMINGUÈS, Catherine et ESHKOL-TARAVELLA, Iris, 2015. Toponym recognition in custom-made map titles. International Journal of Cartography. Vol. 1, n° 1, pp. 109–120.

DOMINGUÈS, Catherine et ESHKOL-TARAVELLA, Iris, 2017. Écriture des toponymes en français: variations entre normes et usages. Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie. n° 110, pp. 151–170.

DUCLOS, Sandra E., LAIRD, James D., SCHNEIDER, Eric, SEXTER, Melissa, STERN, Lisa et VAN LIGHTEN, Oliver, 1989. Emotion-specific effects of facial expressions and postures on emotional experience. Journal of Personality and Social Psychology. Vol. 57, n° 1, pp. 100.

DUPONT, Yoann, 2017. La structuration dans les entités nommées. PhD Thesis. Université Sorbonne Paris Cité.

EHRMANN, Maud, 2008. Les Entitées Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation. PhD Thesis.

EICHSTAEDT, Johannes C., SCHWARTZ, Hansen Andrew, KERN, Margaret L., PARK, Gregory, LABARTHE, Darwin R., MERCHANT, Raina M., JHA, Sneha, AGRAWAL, Megha, DZIURZYNSKI, Lukasz A., SAP, Maarten et OTHERS, 2015. Psychological language on Twitter predicts county-level heart disease mortality. Psychological science. Vol. 26, n° 2, pp. 159–169.

EKMAN, Paul, 1973. Cross-cultural studies of facial expression. Darwin and facial expression: A century of research in review. Vol. 169222, n° 1.

EKMAN, Paul, 2016. What scientists who study emotion agree about. Perspectives on Psychological Science. Vol. 11, n° 1, pp. 31–34.

ESHKOL-TARAVELLA, Iris, BAUDE, Olivier, MAUREL, Denis, HRIBA, Linda, DUGUA, Céline et TELLIER, Isabelle, 2011. Un grand corpus oral «disponible»: le corpus d'Orléans 1 1968-2012.

ESHKOL-TARAVELLA, Iris, 2015. La définition des annotations linguistiques selon les corpus: de l'écrit journalistique à l'oral. PhD Thesis.

FARGE, Arlette, 2013. Le goût de l'archive. *Le Seuil*.

FERRAND, Alexis, 2016. Appartenances multiples, opinion plurielle. Presses universitaires du Septentrion.

FINKEL, Jenny Rose, GRENAGER, Trond et MANNING, Christopher, 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In : Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics. 2005. pp. 363–370.

FORT, Karën, EHRMANN, Maud et NAZARENKO, Adeline, 2009. Vers une méthodologie d'annotation des entités nommées en corpus?

FORT, Karën, NAZARENKO, Adeline et ROSSET, Sophie, 2012. Modeling the complexity of manual annotation tasks: a grid of analysis.

FORTIN, Claudette et ROUSSEAU, Robert, 2015. Psychologie cognitive: une approche de traitement de l'information. PUQ.

FRÉMONT, Armand, 1980. L'espace vécu et la notion de région. Travaux de l'Institut de Géographie de Reims. Vol. 41, n° 1, pp. 47–58.

FRIJDA, Nico H., KUIPERS, Peter et TER SCHURE, Elisabeth, 1989. Relations among emotion, appraisal, and emotional action readiness. Journal of personality and social psychology. Vol. 57, n° 2, pp. 212.

FRIJDA, Nico H., 1986. The emotions. Cambridge University Press.

FUCHS, Catherine et HABERT, Benoit, 2004. Le traitement automatique des langues: des modèles aux ressources. Le Français Moderne-Revue de linguistique Française. Vol. 72, n° 1.

FURUKAWA, Naoyo, 1996. Grammaire de la prédication seconde: forme, sens et contraintes. Duculot.

GIBSON, James Jerome, 1966. The senses considered as perceptual systems.

GILLOT, Sabine, 2010. La place de la posture dans le diagnostic et les décisions thérapeutiques. PhD Thesis. UHP-Université Henri Poincaré.

GOOSSENS, Vannina, 2005. Les noms de sentiment. Esquisse de typologie sémantique fondée sur les collocations verbales. Lidil. Revue de linguistique et de didactique des langues. n° 32, pp. 103–121.

GOUVERT, Xavier, 2008. Problèmes et méthodes en toponymie française: essais de linguistique historique sur les noms de lieux du Roannais. PhD Thesis. Paris 4.

GRAVIER, Guillaume, BONASTRE, Jean-François, GEOFFROIS, Edouard, GALLIANO, Sylvain, MCTAIT, Kevin et CHOUKRI, Khalid, 2004. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. Proc. Journées d'Etude sur la Parole (JEP).

GROBOL, Loïc, LANDRAGIN, Frédéric et HEIDEN, Serge, 2017. Interoperable annotation of (co)references in the Democrat project. In : Thirteenth Joint ISO-ACL Workshop on Interoperable Semantic Annotation [en ligne]. Montpellier, France : ACL Special Interest Group on Computational Semantics (SIGSEM) and ISO TC 37/SC 4 (Language Resources) WG 2. septembre 2017. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01583527

GROBOL, Loïc, TELLIER, Isabelle, DE LA CLERGERIE, Éric Villemonte, DINARELLI, Marco et LANDRAGIN, Frédéric, 2018. ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations.

HABERMAS, Jürgen. et DE LAUNAY, Marc Buhot., 1988. L'espace public: archéologie de la publicité comme dimension constitutive de la société bourgeoise [en ligne]. Payot.

Critique de la politique. ISBN 978-2-228-88013-8. Disponible à l'adresse : https://books.google.fr/books?id=2NWrQgAACAAJ

HARB, Ali, PLANTIÉ, Michel, DRAY, Gerard, ROCHE, Mathieu, TROUSSET, François et PONCELET, Pascal, 2008. Web Opinion Mining: How to extract opinions from blogs? In : Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology. ACM. pp. 211–217.

HASAN, Mohamadally et BORIS, Fomani, 2006. Svm: Machines à vecteurs de support ou séparateurs à vastes marges. Rapport technique, Versailles St Quentin, France. Cité. pp. 64.

HASCOET, Mountaz et BEAUDOUIN-LAFON, Michel, 2001. Visualisation interactive d'information. Revue I3. Vol. 1, n° 1, pp. 77–108.

HONNIBAL, Matthew, 2015. spaCy: Industrial-strength Natural Language Processing (NLP) with Python and Cython.

HUME, David et LÉVY-BRUHL, Lucien, 1946. Traité de la nature humaine. Aubier Paris.

ISOZAKI, Hideki et KAZAWA, Hideto, 2002. Efficient support vector classifiers for named entity recognition. In : Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics. pp. 1–7.

IZARD, Carroll E, 1969. The emotions and emotion constructs in personality and culture research. Handbook of modern personality theory. pp. 496–510.

IZARD, Carroll E, 1990. Facial expressions and the regulation of emotions. Journal of personality and social psychology. Vol. 58, n° 3, pp. 487.

KAROUI, Myriam, DAVAUCHELLE, Grégoire et DUDEZERT, Aurélie, 2014. Big data. Mise en perspective et enjeux pour les entreprises. Ingénierie des Systèmes d'Information. Vol. 19, n° 3, pp. 73–92.

KERBRAT-ORECCHIONI, Catherine, 2009. L'énonciation: de la subjectivité dans le langage. Armand Colin.

KERGOSIEN, Eric, LAVAL, Bernard, ROCHE, Mathieu et TEISSEIRE, Maguelonne, 2014. Are opinions expressed in land-use planning documents? International Journal of Geographical Information Science. Vol. 28, n° 4, pp. 739–762.

KERGOSIEN, Eric, MAUREL, Pierre, ROCHE, Mathieu et TEISSEIRE, Maguelonne, 2013. OPITER: Fouille de données d'opinion pour les territoires. Spatial Analysis and GEOmatics (Sagéo'13), Brest.

KERGOSIEN, Eric, MAUREL, Pierre, ROCHE, Mathieu et TEISSEIRE, Maguelonne, 2015. Senterritoire pour la détection d'opinions liées à l'aménagement d'un territoire. Revue Internationale de Géomatique.Vol. 25, n° 1, pp. 11–34.

KUMAR, Monu et BALA, Anju, 2016. Analyzing Twitter sentiments through big data. In : 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE. pp. 2628–2631.

LAFRANCE, Yvon et BRISSON, Luc, 2015. La théorie platonicienne de la doxa [en ligne]. Les Belles Lettres. Collection d'études anciennes. ISBN 978-2-251-40333-5. Disponible à l'adresse : https://books.google.fr/books?id=_4sFogEACAAJ

LANDIS, J. Richard et KOCH, Gary G., 1977. The measurement of observer agreement for categorical data. biometrics. pp. 159–174.

LANDRAGIN, Frédéric, DELABORDE, Marine, DUPONT, Yoann et GROBOL, Loïc, 2018. Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours [en ligne]. Université de Rennes. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01797982

LANDRAGIN, Frédéric, 2016. Conception d'un outil de visualisation et d'exploration de chaînes de coréférences. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Jun 2016, Nice, France. pp.109-120. ⟨halshs-01329414⟩

LARK, Joseph, MORIN, Emmanuel et SALDARRIAGA, Sebastián Peña, 2015. CANÉPHORE: un corpus français pour la fouille d'opinion ciblée.

LAUR, Dany, 1991. Sémantique du déplacement et de la localisation en français: une étude des verbes, des prépositions et de leurs relations dans la phrase simple. PhD Thesis. Toulouse 2.

LE PESANT, Denis, 2011. Problèmes de morphologie, de syntaxe et de classification sémantique dans le domaine des prépositions locatives.

LE PESANT, Denis, 2012. Essai de classification des prépositions de localisation. In : SHS Web of Conferences. EDP Sciences. pp. 921–937.

LE PEVEDIC, Solenn et MAUREL, Denis, 2016. Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI. Corela. Cognition, représentation, langage. n° 14-2.

LE SQUÈRE, Roseline, 2006. Analyse des perceptions, usages et fonctions des toponymes actuels des territoires ruraux et urbains de Bretagne. Cahiers de sociolinguistique. n° 1, pp. 81–99.

LECHEVALIER, Bernard, EUSTACHE, Francis et VIADER, Fausto, 1995. Perception et agnosies: Séminaire Jean-Louis Signoret. De Boeck Supérieur.

LEMAIRE, Patrick et DIDIERJEAN, André, 2018. Introduction à la psychologie cognitive. De Boeck Supérieur.

LESBEGUERIES, Julien, 2007. Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé. PhD Thesis.

LEVENSHTEIN, Vladimir, 1965. Leveinshtein distance.

LEVENSON, Robert W., 1999. The intrapersonal functions of emotion. Cognition & Emotion. Vol. 13, n° 5, pp. 481–504.

LÉVY, Jacques et LUSSAULT, Michel, 2013. Dictionnaire de géographie et de l'espace des sociétés.

LEVY, Ran, BILU, Yonatan, HERSHCOVICH, Daniel, AHARONI, Ehud et SLONIM, Noam, 2014. Context Dependent Claim Detection. In : Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland : Dublin City University and Association for Computational Linguistics. août 2014. pp. 1489–1500.

LIU, Bing, 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 2012. Vol. 5, n° 1, pp. 1–167.

LOCKE, John, 2006. Essai sur l'entendement humain: livres III et IV. Vrin.

LOPER, Edward et BIRD, Steven, 2002. NLTK: the natural language toolkit. arXiv preprint cs/0205028.

LOUSTAU, Pierre, GAIO, Mauro et NODENOT, Thierry, 2008. Interprétation automatique d'itinéraires à partir d'un corpus de récits de voyages pilotée par un usage pédagogique.

LUMINET, Olivier, 2008. Psychologie des émotions: confrontation et évitement. De Boeck Supérieur.

LUYAT, Marion, 2014. La perception. Dunod.

MARCHAND, Morgane, 2015. Domaines et fouille d'opinion: une étude des marqueurs multi-polaires au niveau du texte. PhD Thesis.

MAUREL, Denis, FRIBURGER, Nathalie, ANTOINE, Jean-Yves, ESHKOL, Iris et NOUVEL, Damien, 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. Traitement automatique des langues. Vol. 52, n° 1, pp. 69–96.

MICHEL, Lussault, 2007. L'homme spatial. La construction sociale de l'espace humain. Paris, Seuil.

MIKOLOV, Tomas, CHEN, Kai, CORRADO, Greg et DEAN, Jeffrey, 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

MOHAMMAD, Saif M et TURNEY, Peter D., 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In : Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics. 2010. pp. 26–34.

MOHAMMAD, Saif M. et TURNEY, Peter D., 2013. Nrc emotion lexicon. National Research Council, Canada.

MOHAMMAD, Saif M., 2017. Challenges in sentiment analysis. In : A practical guide to sentiment analysis. Springer. pp. 61–83.

MONCLA, Ludovic, GAIO, Mauro, NOGUERAS-ISO, Javier et MUSTIÈRE, Sébastien, 2016. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. International Journal of Geographical Information Scienc. Vol. 30, n° 6, pp. 1137–1160.

MONDARY, Thibault, NAZARENKO, Adeline, ZARGAYOUNA, Haifa et BARREAUX, Sabine, 2012. The quaero evaluation campaign on term extraction. In : The eighth international conference on Language Resources and Evaluation (LREC). pp. 663–669.

MUZERELLE, Judith, LEFEUVRE, Anaïs, SCHANG, Emmanuel, ANTOINE, Jean-Yves, PELLETIER, Aurore, MAUREL, Denis, ESHKOL, Iris et VILLANEAU, Jeanne, 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures.

NADEAU, David et SEKINE, Satoshi, 2009. A survey of entity recognition and classification. Named entities—recognition, classification and use. Amsterdam/Philadelphia: John Benjamins Publishing Company.

NEISSER, Ulric, 1967. Cognitive psychology appleton-century-crofts. New York. pp. 351.

NISSIM, Malvina et PIETRANDREA, Paola, 2017. MODAL: A multilingual corpus annotated for modality. CLiC-it 2017 11-12 December 2017, Rome. pp. 234.

NOUVEL, Damien, ANTOINE, Jean-Yves, FRIBURGER, Nathalie et SOULET, Arnaud, 2013. Fouille de règles d'annotation pour la reconnaissance d'entités nommées. .

NOUVEL, Damien, EHRMANN, Maud et ROSSET, Sophie, 2015. Les entités nommées pour le traitement automatique des langues. ISTE Group.

NOVAKOVA, Iva et SORBA, Julie, 2014. L'émotion dans le discours. À la recherche du profil discursif de stupeur et de jalousie. Les émotions dans le discours. Emotions in discourse. pp. 161–173.

NUGIER, Armelle, 2009. Histoire et grands courants de recherche sur les émotions. Revue électronique de psychologie sociale. Vol. 4, n° 4, pp. 8–14.

OLMEDO, Élise, 2015. Cartographie sensible. Tracer une géographie du vécu par la recherche-création.

OUERTATANI, Asma, GASMI, Ghada et LATIRI, Chiraz, 2018. Détection d'opinion argumentée à partir de Twitter. In : CORIA.

PAK, Alexander et PAROUBEK, Patrick, 2010. Twitter as a corpus for sentiment analysis and opinion mining. In : LREc. pp. 1320–1326.

PANG, Bo, LEE, Lillian et VAITHYANATHAN, Shivakumar, 2002. Thumbs up?: sentiment classification using machine learning techniques. In : Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics. pp. 79–86.

PANG, Bo et LEE, Lillian, 2008. Using very simple statistics for review search: An exploration. In : Coling 2008: Companion volume: Posters. pp. 75–78.

PAUMIER, Sébastien, NAKAMURA, Takuya et VOYATZI, Stavroula, 2009. UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources. eLEX2009. Vol. 173.

PEDREGOSA, Fabian, VAROQUAUX, Gaël, GRAMFORT, Alexandre, MICHEL, Vincent, THIRION, Bertrand, GRISEL, Olivier, BLONDEL, Mathieu, PRETTENHOFER, Peter, WEISS, Ron, DUBOURG, Vincent et OTHERS, 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research. Vol. 12, n° Oct, pp. 2825–2830.

PHILIPPOT, Pierre, 2013. Émotion et psychothérapie: L'influence des émotions dans la société. Primento.

PIETRANDREA, Paola, 2018. Epistemic constructions at work. A corpus study on spoken Italian dialogues. Journal of Pragmatics. Vol. 128, pp. 171–191.

PLUTCHIK, Robert, 1980. A general psychoevolutionary theory of emotion. In : Theories of emotion. Elsevier. pp. 3–33.

POIBEAU, Thierry, 2014. Le traitement automatique des langues pour les sciences sociales. Réseaux. n° 6, pp. 25–51.

PONCET-JEANNE, Marie, 2007. L'expressivité non verbale des personnes âgées atteintes de démence de type Alzheimer, marqueur de leur affectivité préservée. PhD Thesis. Lyon 2.

PRADINES, Maurice, 1946. Traité de psychologie générale (PUF, 1946), t. II et III.

RIEGEL, Martin, PELLAT, Jean-Christophe et RIOUL, René, 1994. Grammaire méthodique du français. Linguistique nouvelle.

ROSSET, Sophie, GROUIN, Cyril et ZWEIGENBAUM, Pierre, 2011. Entités nommées structurées: guide d'annotation Quaero. LIMSI-Centre national de la recherche scientifique.

SALTON, Gerard et BUCKLEY, Christopher, 1988. Term-weighting approaches in automatic text retrieval. Information processing & management. Vol. 24, n° 5, pp. 513–523.

SCHMID, Helmut, 1999. Improvements in part-of-speech tagging with an application to German. In : Natural language processing using very large corpora. Springer. pp. 13–25.

SÉBILLOT, Pascale, 2015. Natural language processing faced with big and potentially impaired textual data: What difference does it make? In : Big data : nouvelles partitions de l'information. Actes du séminaire IST INRIA, octobre 2014 [en ligne]. De Boeck. pp. 43-60. Information et stratégie. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01056396

SHANNON, Claude Elwood, 1948. A mathematical theory of communication. Bell system technical journal. Vol. 27, n° 3, pp. 379–423.

SINDHU, Chandra S. et VADIVU, G, 2019. Sentiment Analysis and Opinion Summarization of Product Feedback. International Journal of Recent Technology and Engineering (IJRTE). Vol. 8.

STERN, Rosa et SAGOT, Benoît, 2010. Détection et résolution d'entités nommées dans des dépêches d'agence.

SUIGNARD, Philippe, 2017. Wordsurf: un outil pour naviguer dans un espace de «Word Embeddings».

TUAN, Yi-Fu, 1977. Space and place: The perspective of experience. *U of Minnesota Press*.

VANDELOISE, Claude, 1986. *L'espace en français: sémantique des prépositions spatiales*. Seuil.

VAPNIK, Vladimir, 1992. Principles of risk minimization for learning theory. In : Advances in neural information processing systems. pp. 831–838.

VIDAL DE LA BLACHE, Paul, 1913. Des caractères distinctifs de la géographie. Annales de Géographie, t. 22, n°124, 1913. [en ligne]. DOI 10.3406/geo.1913.8245. Disponible à l'adresse : https://www.persee.fr/doc/geo_0003-4010_1913_num_22_124_8245

VILLANEAU, Jeanne, PECORE, Stefania, SAÏD, Farida et MARTEAU, Pierre-François, 2019. Combiner analyse syntaxique de surface et apprentissage supervisé pour la fouille d'opinion ciblée: expérimentations sur des données d'opinion concernant les livres. In : Extraction et Gestion des Connaissances: Actes de la conférence EGC'2019. BoD-Books on Demand.

ZENASNI, Sarah, KERGOSIEN, Eric, ROCHE, Mathieu et TEISSEIRE, Maguelonne, 2016. Extracting new spatial entities and relations from short messages. In : *Proceedings of the 8th International Conference on Management of Digital EcoSystems*. ACM. 2016. pp. 189–196.

LISTE DES FIGURES

LISTE DES TABLEAUX

Hélène FLAMEIN

Etude de la perception d'une ville

Repérage automatique, analyse et visualisation

Résumé :

A l'heure où de plus en plus de corpus et de données sont accessibles, le travail initié s'interroge sur l'exploitation de données linguistiques dans un corpus d'oral à dimension sociolinguistique avec l'objectif d'en extraire automatiquement du contenu subjectif. A partir de l'exploitation du corpus ESLO (Enquête Sociolinguistique à Orléans), l'objectif est de modéliser, détecter et visualiser la perception qu'ont les locuteurs de la ville d'Orléans. Pour cela, une approche pluridisciplinaire associant la linguistique, le Traitement Automatique des Langues (TAL) et la géographie a été suivie.

Une première étape de traitement automatique est la reconnaissance des mentions de lieux en français parlé. Celle-ci s'appuie sur une analyse linguistique de la variation des noms de lieux et prépare l'étude de la perception qui leur est associée. Les techniques de l'apprentissage supervisé ont permis la détection de la subjectivité et de la polarité. Cette détection oriente l'analyse de la perception mais ne suffit pas à en rendre compte. Pour aller plus loin dans la réflexion, la typologie de la perception est approfondie. Afin de confronter les indices subjectifs extraits des transcriptions, des techniques de visualisation de l'information sont employées. Les opérations réalisées aboutissent à une représentation graphique des lieux identifiés associée aux déclarations subjectives des locuteurs, matérialisant la perception de la ville d'Orléans par ses habitants.

Mots clés : Lieu, Perception, ESLO, Traitement Automatique des Langues, Visualisation de l'information

Study of the perception of a city

Automatic identification, analysis and visualization

Abstract :

At a time when more and more corpora and data are accessible, the work initiated in this thesis questions the exploitation of linguistic data in an oral corpus with a sociolinguistic dimension in order to automatically extract subjective content. From the exploitation of the ELSO corpus (Sociolinguistic Survey in Orléans), the objective is to model, detect and visualize the perception the speakers have of the city of Orléans. To this end, a multidisciplinary approach bringing linguistic, Natural Language Processing (NLP) and geography together was followed.

A first necessary step of automatic processing is the automatic recognition of place references in spoken French. It is based on a linguistic analysis of the variation of the place names and prepares the study of the perception associated with them. Supervised machine learning technics have allowed the detection of subjectivity and polarity. This detection guides the analysis of perception but it is not sufficient to report it. To go further, the typology of perception is deepened. In order to compare the subjective clues extracted from the transcripts, techniques of information visualization are used. The operations carried out result in a graphic representation of the identified places associated with the subjective statements of the speakers, giving a tangible view of the perception of the city of Orléans by its inhabitants.

Keywords : Location, Perception, ESLO, Natural Language Processing, Information visualization

Laboratoire Ligérien de Linguistique

Université d'Orléans, UFR LLSH
10 rue de Tours – BP 46527
45065 ORLEANS Cedex 2